JAZYKOVEDNÝ ČASOPIS

JAZYKOVEDNÝ ÚSTAV ĽUDOVÍTA ŠTÚRA

SLOVENSKEJ AKADÉMIE VIED

1

ROČNÍK 76, 2025

sciendo

SAP
SLOVAK ACADEMIC PRESS

JAZYKOVEDNÝ ČASOPIS
VEDECKÝ ČASOPIS PRE OTÁZKY TEÓRIE JAZYKA

JOURNAL OF LINGUISTICS
SCIENTIFIC JOURNAL FOR THE THEORY OF LANGUAGE

Vydáva/Published by: Jazykovedný ústav Ľudovíta Štúra Slovenskej akadémie vied, v. v. i.
- v tlačenej podobe vo vydavateľstve SAP – Slovak Academic Press, s. r. o.
- elektronicky vo vydavateľstve Sciendo – De Gruyter
  https://content.sciendo.com/view/journals/jazcas/jazcas-overview.xml

Adresa redakcie/Editorial address: Jazykovedný ústav Ľ. Štúra SAV, Panská 26, 811 01 Bratislava
Kontakt: jazykovedny.casopis@juls.savba.sk
Elektronická verzia časopisu je dostupná na internetovej adrese/The electronic version of the journal is available at: https://journals.savba.sk/index.php/jc/, https://www.sav.sk/?lang=sk&doc=journal-list&journal_no=26

Vychádza trikrát ročne/Published triannually

Dátum vydania aktuálneho čísla (2025/76/1) – september 2025

Quartile ranking 2024: Q3
CiteScore 2024: 0,6
SCImago Journal Rank (SJR) 2024: 0,150
Source Normalized Impact per Paper (SNIP) 2024: 0,508

JAZYKOVEDNÝ ČASOPIS je evidovaný v databázach/JOURNAL OF LINGUISTICS is covered by the following services: Baidu Scholar; Cabell's Journalytics; CEEAS (Central & Eastern European Academic Source); CEJSH (The Central European Journal of Social Sciences and Humanities); CEEOL (Central and Eastern European Online Library); CNKI Scholar (China National Knowledge Infrastructure); Crossref database; Dimensions; EBSCO; ERIH PLUS (European Reference Index for the Humanities and Social Sciences); ExLibris; Google Scholar; IBR (International Bibliography of Reviews of Scholarly Literature in the Humanities and Social Sciences); IBZ (International Bibliography of Periodical Literature in the Humanities and Social Sciences); International Medieval Bibliography; J-Gate; Jisc - Open Policy Finder; JournalGuide; JournalTOCs; KESLI-NDSL (Korean National Discovery for Science Leaders); Linguistic Bibliography; Linguistics Abstracts Online; MLA International Bibliography; Naver Academic; Naviga (Softweco); OpenAlex; ProQuest; ReadCube; ScienceON/AccessON; SCILIT; SCImago (SJR); Scite_; SCOPUS; Semantic Scholar; TDNet; Ulrich's Periodicals Directory/ulrichsweb; WorldCat (OCLC); X-MOL.

ISSN 0021-5597 (tlačená verzia/print) ISSN 1338-4287 (verzia online) MIČ 49263

JAZYKOVEDNÝ ČASOPIS

**JAZYKOVEDNÝ ÚSTAV ĽUDOVÍTA ŠTÚRA
SLOVENSKEJ AKADÉMIE VIED**

1

ROČNÍK 76, 2025

Natural Language Processing and Corpus Linguistics

SLOVKO 2025

Tematické číslo Jazykovedného časopisu venované počítačovému spracovaniu prirodzeného jazyka a korpusovej lingvistike.

Prizvaní editori:
Kristína Bobeková
Jakob Heinrich Horsch
Adriána Žáková

CONTENTS

CORPUS-BASED AND CORPUS-DRIVEN RESEARCH

CORPUS BUILDING

NATURAL LANGUAGE PROCESSING

CREATION AND USE OF LANGUAGE RESOURCES

DIGITAL HUMANITIES

sciendo

FOREWORD

After a two-year interval, the Journal of Linguistics presents another special issue featuring contributions from the international SLOVKO 2025 conference. This year's contributions cover a wide range of current issues in European corpus linguistics. These issues span the building and annotation of language corpora, natural language processing, creation and use of language resources, and digital humanities. We continue to witness rapid technological development in these areas, which provides a framework and impetus for linguistic research. There is also a fruitful combination of methods – a truly interdisciplinary "crossroads." One of the keywords of this year's conference is creativity and its various forms and contexts. This includes the creative principle in language itself and the creative search to understand it. The contributions in this issue reflect a wide spectrum of such creativity, whether concerning the exploration of colors in poetry or the creative dimension of constructions and other abstract linguistic structures. Creativity is also necessary to expand the possibilities of processing and annotating an ever-widening range of texts and for the intersection of linguistics and artificial intelligence. Creativity is essential for innovatively processing lexicons and developing the humanities in the 21st century. With the increasing availability of digital text data and technological advances in natural language processing, new opportunities arise for analyzing language data and phenomena beyond the scope of traditional descriptive linguistics. As the articles in this year's special issue of the Journal demonstrate, corpus linguistics has long transcended purely linguistic research. Its overlap with areas such as mathematical linguistics, digital humanities, computer science, pedagogy, and media and communication confirms that it is a discipline that brings together diverse research traditions while offering new tools for analyzing linguistic phenomena. Given the growing amount of text data and the development of language technologies – including the rapid development of generative artificial intelligence – new opportunities are emerging to explore the relationship between language's meaning, form, and function in its natural environment.

The papers presented at the **13th annual international SLOVKO 2025 conference** reflect the methodological diversity of corpus research. Topics include the design and construction of specialized corpora, the analysis of corpus data, questions of representativeness and annotation, the automation of corpus processing, and its application in empirical research. Applications include automated text processing, machine learning, language modeling, and the development of language tools. The authors approach language research with an emphasis on accuracy, reproducibility, and the interdisciplinary connection between linguistics and computer science, enabling effective modeling of linguistic phenomena and development of tools for processing large amounts of text data. On the other hand, the authors emphasize the growing trend of incorporating corpus linguistics with language education and translation. The authors cover a wide range of linguistic topics, often with interdisciplinary overlap, thus creating a colorful picture of current corpus research. However, as mentioned in the Foreword from the previous 12th conference issue, this diversity requires a unifying framework that enables mutual understanding and the combination of approaches. Through this special issue of the Journal of Linguistics, we continue to strive to create this space, reflecting the results of the international SLOVKO 2025 conference and building on our long-term efforts to promote scientific dialogue in the field of language technologies.

We thank all authors for their professional contribution, the reviewers for their thorough assessment of the articles, and our colleagues for their help with this special issue of the journal. We believe the articles in this issue will find the reader's interest by offering not only a glimpse of individual research questions but also an outline of the broader connections between corpus linguistics and various scientific fields, and between scientists working in related disciplines. At the same time, we hope the articles inspire discussion about the methodological challenges and perspectives of corpus research in Slovak and international contexts. We hope you find the articles stimulating and inspiring as you reflect on the future of corpus linguistics in the rapidly changing digital age.

Kristína Bobeková and Miroslav Zumrík

PREDHOVOR

Jazykovedný časopis po dvoch rokoch opäť prichádza so špeciálnym číslom príspevkov z medzinárodnej konferencie SLOVKO 2025. Príspevky z tohtoročnej konferencie sú venované širokému spektru aktuálnych otázok európskej korpusovej lingvistiky, budovaniu a anotácii jazykových korpusov a počítačovému spracovaniu prirodzeného jazyka, tvorbe a využitiu jazykových zdrojov a digitálnych humanitných vied. V týchto oblastiach sme aj naďalej svedkami nielen prudkého technologického rozvoja, ktorý poskytuje rámec a impulz pre lingvistický výskum, ale zároveň aj plodného kombinovania metód, skutočnej interdisciplinárnej „križovatky“. Jedným z kľúčových slov tohtoročnej konferencie by mohla byť kreativita a jej rôzne podoby a súvislosti. A to jednak tvorivý princíp v samotnom jazyku, alebo tvorivé hľadanie ciest k jeho poznávaniu. V príspevkoch tohto čísla sa zrkadlí široké spektrum takejto poznávanej i poznávajúcej tvorivosti. Či už ide o skúmanie kreatívneho využívania farieb v poézii či kreatívny rozmer konštrukcií a ďalších abstraktných jazykových štruktúr. Tvorivosť si tiež vyžaduje rozširovanie možností spracovania a anotácie čoraz širšieho spektra textov, a prirodzene aj prekryv jazykovedy a možností umelej inteligencie. Bez tvorivosti sa nezaobíde ani spracovanie lexiky a rozvoj humanitných vied v 21. storočí. V kontexte narastajúcej dostupnosti digitálnych textových dát a technologického pokroku v oblasti spracovania prirodzeného jazyka sa teda otvárajú nové možnosti analýzy jazykových dát a javov, ktoré presahujú rámce tradičnej deskriptívnej lingvistiky. Ako ukazujú príspevky v špeciálnom čísle Jazykovedného časopisu, korpusová lingvistika už dávno prekročila rámec čisto lingvistického skúmania. Jej presah do oblastí ako matematická lingvistika, digitálne humanitné vedy, informatika, pedagogika či mediálne a komunikačné sféry potvrdzuje, že ide o disciplínu, ktorá dokáže prepájať rôznorodé výskumné tradície a zároveň ponúkať nové nástroje na analýzu jazykových javov. V kontexte narastajúceho množstva textových dát a vývoja jazykových technológií, okrem iného v podobe skokového rozvoja generatívnej umelej inteligencie, sa tak otvárajú nové možnosti, ako skúmať prepojenie významu, formy a funkcie jazyka v jeho prirodzenom prostredí.

Publikované príspevky **13. ročníka medzinárodnej konferencie SLOVKO 2025** reflektujú metodologickú rozmanitosť korpusového výskumu – od návrhu a tvorby špecializovaných korpusov, cez analýzu korpusových dát či otázky reprezentatívnosti, anotácie, automatizácie spracovania a využitia korpusov v empirickom výskume, až po aplikácie v oblasti automatizovaného spracovania textu, strojového učenia, jazykového modelovania a vývoja jazykových nástrojov. Autori pristupujú k skúmaniu jazyka na jednej strane s dôrazom na presnosť, reprodukovateľnosť a interdisciplinárne prepojenie lingvistiky s informatikou – čo umožňuje efektívne modelovanie jazykových javov a vývoj nástrojov pre spracovanie textových dát vo veľkom rozsahu. Na druhej strane kladú dôraz na aktuálne stúpajúci trend prepájania korpusovej lingvistiky s jazykovým vzdelávaním a prekladom. Autori sa pohybujú naprieč širokým spektrom lingvistických tém často s interdisciplinárnym presahom, čím vytvárajú pestrý obraz súčasného korpusového výskumu. Táto diverzita si však – tak ako bolo na tomto mieste uvedené v čísle s príspevkami z predošlého 12. ročníka konferencie – vyžaduje aj istý jednotiaci rámec, ktorý umožňuje vzájomné porozumenie a kombinovanie prístupov. Tento priestor sa naďalej snažíme vytvoriť prostredníctvom špeciálneho čísla Jazykovedného časopisu, ktoré zároveň reflektuje výsledky medzinárodnej konferencie SLOVKO 2025 a nadväzuje na dlhodobú snahu podporovať vedecký dialóg v oblasti jazykových technológií.

Ďakujeme všetkým autorom za ich odborný prínos, recenzentom za dôsledné posúdenie príspevkov a spolupracovníkom za ich prínos k tvorbe vydania špeciálneho čísla Jazykovedného časopisu. Veríme, že príspevky v tomto čísle čitateľov zaujmú a neponúknu im len fragmentárny pohľad na jednotlivé výskumné otázky, ale načrtnú aj širšie prepojenia medzi korpusovou lingvistikou a rôznymi vednými odbormi, ako aj medzi vedcami pôsobiacimi v príbuzných disciplínach. Zároveň dúfame, že príspevky podnietia diskusiu o metodologických výzvach a perspektívach korpusového výskumu v slovenskom i medzinárodnom kontexte. Prajeme vám podnetné čítanie a inšpiratívne zamyslenie sa nad smerovaním korpusovej lingvistiky v rýchlo sa meniacej dobe digitálneho veku.

Kristína Bobeková a Miroslav Zumrík

CORPUS-BASED AND CORPUS-DRIVEN RESEARCH

sciendo
DOI 10.2478/jazcas-2025-0001

AS HUNGRY AS A WOLF, AS SILENT AS A LAMB: CORPUS ANALYSIS OF ADJECTIVAL SIMILES WITH AN ANIMAL COMPONENT

KATARÍNA GAJDOŠOVÁ[1] – LUCIA JASINSKÁ[2]
[1]Slovak National Corpus, Ľudovít Štúr Institute of Linguistics, Slovak Academy of Sciences, Bratislava, Slovakia (ORCID: 0009-0005-2995-1146)
[2]Department of Slovak Studies, Slavonic Philologies and Communication, Faculty of Arts, Pavol Jozef Šafárik University, Košice, Slovakia (ORCID: 0000-0003-3078-9930)

GAJDOŠOVÁ, Katarína – JASINSKÁ, Lucia: As Hungry as a Wolf, as Silent as a Lamb: Corpus Analysis of Adjectival Similes With an Animal Component. Journal of Linguistics, 2025, Vol. 76, No 1, pp. 11 – 19.

Abstract: This study focuses on analysis of the Slovak adjectival similes with an animal noun component. Its aim is not only to verify the incidence of individual similes, but also to investigate the qualitative aspects of phrases that follow the structure: adjective + *ako/sťa* (like/as) + noun, e.g. *otravný ako komár* 'as annoying as a mosquito' and *pomalý ako rak* 'as slow as a crayfish'. The analysis is based on data from the Slovak National Corpus and is aimed at describing the semantic properties of the selected phrases.

Keywords: adjectives, corpus, simile, animal component

1 INTRODUCTION

The primary function of an adjective is to enrich the lexical meaning of the noun with which it forms a collocation. This relationship is two-way, i.e. the nouns also influence the lexical meaning of the adjectives and they are involved in defining it. In their semantic structure, the specification and identification semes are only formed in connection with a particular noun (cf. e.g. Horák 1956; Pauliny 1981).

This study assumes that specific adjectives tend to be associated with only a limited set of animals in Slovak similes, reflecting culturally salient or perceptually prominent traits. It aims to explore which adjectives are linked to which animals, and how consistently these associations appear in actual usage. The analysis draws on corpus data to map these patterns. To answer this, the analysis proceeds in two directions: first, by examining which animals are associated with a given adjective in similes (e.g. mlčanlivý ako mulica/ryba 'as silent as a hinny/fish') and second, by investigating which adjectives co-occur with specific animal names (e.g. bocian: dlhý, jednonohý, prostoduchý, vysoký 'stork: long, one-legged, simple-minded, tall').

This dual approach enables us to explore both the fixity and variability of adjectival similes in Slovak, while acknowledging that the findings represent only a partial study of the broader lexical system.

Since the article is primarily focused on verifying the occurrence of characteristics attributed to animals, expressed by adjectives in specific word expressions, at this stage of the research we do not distinguish between random, the so-called textual collocations, and lexicalized expressions (including phrasemes or figurative ones).

2 RESEARCH DATA

There are many collocations that associate a property (of something) with a substance. Narrowing down the selection, this paper focuses on a certain semantic range of noun collocations: phrases combining adjectives and nouns that associate a property or static feature with a certain animal. As part of the synthesis phase, the presented research evaluates which adjectives are most frequently used in these collocations, as well as what properties, features, or states are attributed to specific animals. This research draws its data from the Slovak National Corpus (*prim-10.0-public-sane*; hereinafter referred to as the "SNC"). The corpus was used to search for collocations with the structure adjective + *ako/sťa* (like/as) + noun. Then, collocations where an animal name occurred in the noun position were recorded. The LexiCorp (Benko 2019) tool was used to select materials containing the *zviera* 'animal' lemma from the Dictionary of Contemporary Slovak (Slovník súčasného slovenského jazyka, hereinafter referred to as the "DCS"). The resulting material was then checked manually, forming a list of animals captured in the dictionaries. These were then supplemented with some animal names from a more comprehensive list of animals[1]. The final list included a total of 866 animal names. It contained not only general animal names, but in some cases both their male and female counterparts, e.g. *sliepka – kohút* 'hen – rooster', *mačka – kocúr* 'queen – tomcat', and for some animals it also included diminutive names, e.g. *baranček* 'baby ram', *prasiatko* 'piglet', as well as other derived forms, e.g. *mača* 'kitten', *prasa* 'pig'.

The corpus search made use of CQL: [tag="[AG].*"] [lemma_lc="ako|sťa"] [lemma_lc="*zviera"], and the *zviera (*animal) position used animal names from the final list, e.g. akara|akuči|albatros (acara|acouchi|albatross). In the SNC, 258 lexemes were found that were functionally used in the text in the studied simile. The obtained material was studied in two forms – with a focus on the adjectival feature (different animal names associated with the same adjective, e.g. *mlčanlivý ako mulica, ryba* 'as silent as a hinny, fish') and with a focus on the animal name (studying which adjectives are associated with the name of a specific animal, e.g. *bocian: dlhý, jednonohý, prostoduchý, vysoký* 'stork: long, one-legged, simple-minded, tall'). This paper is the result of a partial analysis and, given its scope, it does not aspire to be exhaustive.

[1] https://www.zones.sk/studentske-prace/biologia/175s-zoznam-zvierat-a-zivocichov-od-a-po-z

3 SIMILE ANALYSIS

This research assumes that each adjective will only be assigned to some animals (possibly only a single one) as bearers of a specific characteristic. The aim of the research is to find out which animals are associated with specific characteristics in the eyes of language users.

3.1 Analysis through adjectival features

The most prominently occurring phrase in the studied material was *hladný ako vlk* ‘as hungry as a wolf’ (503), which is commonly used in communication. Moreover, it is also listed as a fixed expression in the DCS. Other phrases with the conjunctions *ako* and *sťa* that occurred rather frequently included the similes *voľný ako/sťa vták* ‘as free as a bird’ (193), *zdravý ako ryba/rybička/rybka* ‘as healthy as a fish’ (162/123/10), *červený ako rak* ‘as red as a crayfish’ (158), *šťastný ako blcha* ‘as happy as a flea’ (156), *mokrý ako myš* ‘as wet as a mouse’ (147), *slabý ako mucha/muška* ‘as weak as a fly’ (100/5). Once again, these are fixed expressions, some of which are listed in the Short Dictionary of Slovak (Krátky slovník slovenského jazyka 2003; hereinafter referred to as the “SDS”) or the DCS as examples of phrasemes. The collocation *spotený ako myš* ‘as sweat-soaked as a mouse’ is a relatively frequent phrase (89) in which the original passive participle becomes static and acquires the meaning of the adjective *mokrý* ‘wet’. Even the SDS lists the expressive phrases *mokrý/zmoknutý/spotený ako myš* ‘as wet/rain-soaked/sweat-soaked as a mouse’, which implies their status as synonyms. This development is a consequence of the dynamics manifested in the transition of words between different parts of speech, the natural result of which is lexicalization.

Less frequent incidence was recorded for the phrases *krotký ako baránok/baran* ‘as tame as a ram lamb’ (74/4), *čulý ako rybička/rybka* ‘as lively as a fish’ (57/1), *tvrdohlavý ako baran* ‘as stubborn as a ram’ (54), *čerstvý ako rybička* ‘as fresh as a fish’ (53), *opatrný ako had* ‘as careful as a snake’ (50). The first four collocations are commonly used in both standard and spoken Slovak, which is why associations in meaning between the adjective and the animal in question are not surprising. The simile *opatrný ako had* ‘as cautious as a snake’ is not prototypical, as evidenced by its position when phrasemes are listed in the DCS – the examples also show that caution is a less typical characteristic for snakes than deceitfulness, flexibility, insincerity or the ability to make excuses. However, the phrase *opatrní ako hady* ‘as cautious as snakes’ can be found in the Gospel of Matthew (*Behold, I am sending you out as sheep in the midst of wolves, so be cautious as serpents and simple as doves.* [10, 16]), from where the text is paraphrased either in full or in abridged or updated forms (of the listed occurrences in the SNC, as many as 38 are direct quotations).

Some less frequent collocations are also worth mentioning, e.g. *sklesnutý ako ovca* ‘as dejected as a sheep’ (49), *slobodný ako vták*, *čierny ako havran* ‘as free as a bird, as black as a rook’ (48); *lenivý ako voš* ‘as lazy as a louse’ (43), *jednoduchý ako holubica/holub* ‘as simple as a dove'[2] (38/1), as well as those that users of the language are widely familiar with, e.g. *prefíkaný ako líška* 'as sly as a fox' (37), *usilovný ako včelička* 'as industrious as a bee' (33), *pyšný ako páv* 'as proud as a peacock' (27), *hladný ako pes* 'as hungry as a dog' (24).

Many phrases were recorded with minimal occurrences (below 10). Among these, there were often expressions that could be perceived as symptomatic and in some sense expressive, e.g. *bezmocný ako ovca* 'as helpless as a sheep', *múdry ako pes* 'as smart as a dog', *nervózny ako mačka* 'as nervous as a cat', *samotársky ako sob* 'as solitary as a reindeer', *tvrdohlavý ako býk* 'as stubborn as a bull', *škaredý ako opica* 'as ugly as a monkey', *štíhly ako laň* 'as thin as a doe', *žiarlivý ako pes* 'as jealous as a dog'.

3.1.1 The contextual dependence of adjectival similes

In some cases, the characteristic does not apply to the animal in general, but only in a particular situation. As a result, adjectival associations with several seemingly incompatible animal names were recorded. These include, among others, the characteristics *bezbranný ako jarabička* [v hniezde] 'as defenceless as a partridge [in the nest]', *bezmocný ako mucha* [v pavúčom objatí] 'as helpless as a fly [in a spider's embrace]'; *protivný ako ryby* [ktoré neberú] 'as annoying as fish [that don't bite]'; *príťažlivý ako sumec* [zabalený do poťahu na matrac] 'as attractive as a catfish [wrapped in a mattress cover]'.

A specific one is the adjective *nervózny* 'nervous', which requires completion because the meaning of the collocation is not definable without further context. For instance, the phrase *nervózny ako blcha* 'as nervous as a flea', *nervózny ako prasa* 'as nervous as a pig' only gains a clear meaning in a specific textual context (*nervózny ako blcha* [na horúcom tanieri] 'as nervous as a flea [on a hot plate], *prasa* [pred zakáľačkou] 'a pig [before the slaughter]'; similar to other animals in this collocation: *levica* [v klietke] 'a lioness [in a cage]'. Although adjectives reflect a relatively permanent property, their semantic structure may also capture a state that only lasts for a certain time, e.g. an emotion, feeling, sensation, or facial expression. Since the adjective *nervózny* 'nervous' is seemingly semantically unrelated to animal names, its meaning is modified through the author's perspective, which is manifested in creative solutions, e.g. through humorous context.

Another interesting type of simile is based on irony, e.g. *Je asi taká bezbranná ako mačka s ostrými pazúrmi.* 'She's about as defenceless as a cat with sharp claws'[3]; *Ocko*

[2] The phrase is the second part of the biblical quote *opatrní ako hady a jednoduchí ako holubice* 'cautious as snakes and simple as doves', and this meaning prevails in the SNC. In 11 instances, either the whole sentence occurs (including the part about snakes), or only the part referring to the dove in other texts. Surprisingly, an actualizing shift of this quotation to the form *jednoduchý ako holub* 'simple as a pigeon' was found in a text from the religious sphere (a 1996 book by Anton Hlinka titled *Ozvena slova 3. Blahozvesť v horizonte ľudskej skúsenosti* 'Echo of the word 3. Gospel on the Horizon of Human Experience').

[3] Warrenová, T. A. (2013). Krásna pokušiteľka. Bratislava: Slovenský spisovateľ, 271 p. Translation: Diana Ghaniová.

je už doma z roboty, je asi taký milý ako krokodíl, ktorého štuchajú do oka. 'Dad's already home from work, he's about as nice as a crocodile that's being poked in the eye.'[4]. In these examples, a trait is being attributed to an animal that can be characterized by the opposite character trait. It is true, however, that the author appears to be aware of the contradiction within the phrase, thus the simile is preceded by the words *asi taký/-á* 'about as'.

In addition to adjectives, the attributive function of the immediate left-hand position can also be served by participles indicating an implicitly present action. It is worth noting that the classification of the identified lexemes into the group of participles in the SNC is determined by their word-formation structure, which makes them identical to the passive participles of verbs. Although a few isolated occurrences of the above type of collocations were identified in the studied material, there were also phrases with a higher frequency, e.g. *spotený ako myš* 'as sweat-soaked as a mouse' (89), *uťahaný ako kôň* 'as worn out as a horse' (72), *unavený ako pes* 'as tired as a dog' (28), *zbitý ako pe*s 'as beat-up as a dog' (19), *zmoknutý ako myš* 'as rain-soaked as a mouse' (7). Some features are associated with multiple animals, e.g. *opitý ako čajka*, *hus*, *prasa*, *sviňa*, *teľa* 'as drunk as a gull, goose, pig, swine, calf'; *uťahaný ako kôň*, *mača*, *pes*, *somár*, *ťava* 'as tired as a horse, kitten, dog, donkey, camel'; *zviazaný ako jahňa*, *kozľa*, *ovca*, *prasa* 'as tied as a lamb, kid, sheep, pig'. It is also true for the above collocations that attributes with a higher frequency of occurrence can be classified as adjectives in regard to their part-of-speech status (e.g. *uťahaný* 'worn out', *spotený* 'sweat-soaked').

In order to find out whether the most used adjectives in the examined constructions are already lexical phrases, figurative or phraseological, we looked for them in dictionaries. In the SDS and the DCS, the following phrases can be found, likewise those with high frequency in our research, such as *zdravý ako ryba* 'as healthy as a fish', *hladný ako vlk* 'as hungry as a wolf', *voľný ako vták* 'as free as a bird', *šťastný ako blcha* 'as happy as a flea' (a total of 31). In addition to these adjectives our study contains a significantly larger number of similes involving also other adjectives.

3.2 Analysis through animal names

The analysis performed in this research examined what characteristics are attributed to which animals, i.e. which specific animals have a single characteristic attributed in the texts and, conversely, which animals have a rich inventory of characteristics attributed. The reverse approach can be evaluated both quantitatively and qualitatively.

In terms of the number of characteristics associated with animals, the largest numbers of attributes were found in phrases with the nouns *pes* 'dog' (84), *had* 'snake' (65), *mačka/mača/mačiatko* 'cat/kitten' (64/11/10), *kôň* 'horse' (59), *ryba/rybička/rybka* 'fish' (47/13/7), *vták* 'bird' (45), *baran* 'ram' (40), *medveď* 'bear' (38), *mucha* 'fly' (36), *vlk* 'wolf' (32), *opica* 'monkey' (31), *lev* 'lion' (30).

[4] Šulajová, Z. (2012). Džínsový denník 3. Bratislava: Slovenský spisovateľ, 439 p.

The dog as the animal with the greatest number of attributes is perceived in differentiated contexts. In terms of capturing one's outward appearance, it is used as an entity of a particular colour (*biely* 'white', *hnedý* 'brown'[5]), then a characteristic related to physical appearance (*nízky* 'short', *starý* 'old', *škaredý* 'ugly'), a character trait (*bezohľadný* 'reckless', *citlivý* 'sensitive', *lenivý* 'lazy'*, mĺkvy* 'silent', *prísny* 'stern', *sprostý* 'stupid', *šťastný* 'happy'), or an indication of one's current state (*hladný* 'hungry', *smutný* 'sad'). Some adjectives associated with the lexeme *pes* 'dog' through simile have rather vague semantic contours when perceived without context (*podradný* 'inferior', *prirodzený* 'natural', *samotný* 'solitary'). As the examples indicate, dogs are also perceived on the positive-negative axis.

A similar situation arises when comparing characteristics to a cat, often perceived as a pragmatic antonym of the lexeme *pes* 'dog'. In the context of character, cats are described using both neutral and expressive adjectives (*čistotný* 'clean', *falošný* 'deceitful', *opatrný* 'careful', *ostražitý* 'cautious', *prefíkaný* 'cunning', *smilný* 'lustful', *úskočný* 'crafty', *zvodný* 'seductive'). A cat's appearance is associated with a colour that is being likened (*biely* 'white', *čierny* 'black'), an external characteristic (*chudý* 'thin', *krásny* 'beautiful', *mäkký* 'soft'), or an observed state (*čistý* 'clean', *ospanlivý* 'sleepy', *pradúci* 'purring'). Here, too, there were several adjectives where comparisons to a cat may be semantically non-standard (*pružný* 'flexible, elastic', *vláčny* 'smooth'). The word for a baby cat – *mača* 'kitten' – is mostly associated with character traits (*hravý* 'playful', *lenivý* 'lazy', *malý* 'small'). A similar situation occurs in similes with the lexeme *mačiatko* 'kitten', which include adjectives with a positive meaning (*ihravý* 'playful', *krotký* 'tame', *mierumilovný* 'peaceful'), one attribute related to physical aspects (*slabý* 'weak'), and the adjective *veľký* 'big', which, in conjunction with the diminutive, creates a phrase based on irony (*„Tie ostatné, čo boli pred vami, sa v škole dlho neohriali," uškrnul sa, zatiaľ čo do predného skla udierali vločky veľké ako mačiatka.* '„The other ones before you didn't stay in school long," he grinned as snowflakes as big as kittens hit the windshield.'[6]).

Despite the antonymic status of the lexemes *pes* 'dog' and *mačka* 'cat', some characteristics within comparative phrases are perceived as compatible with both animals, these include the attributes *lenivý* 'lazy', *maškrtný* 'sweet-toothed', *múdry* 'smart', *nervózny* 'nervous', *nežný* 'gentle', *prítulný* 'cuddly', *šťastný* 'happy', *zlostný* 'irritable, angry'.

We also examined the pragmatic opposites *dog* and *cat* in the phraseological dictionary of Czech (Čermák et al. 2009) in which several comparative constructions with the lexeme *kočka* 'cat' – listed alphabetically (not by frequency) – are treated as a separate entries. These include expressions such as *falešný jako kočka* 'false as a cat', *mrštný jako kočka* 'agile as a cat' and *utahaný jako kočka* 'exhausted as a cat'.

[5] Given the length of the paper, only a few examples are listed here and elsewhere.

[6] Hospodárske noviny. (2007). Bratislava: Ecopress a.s., Vol. 16 [23/03/2007].

In opposition to the expression *utahaný* 'exhausted', the adjective *čilý* 'lively' appears in the expression *čilý jako rybička* 'lively as a little fish'. The lexeme *pes* 'dog' is part of several word expressions in the dictionaries (e.g. *hubený/vyzáblý/vychrtlý, opuštěný, utahaný, vzteklý jako pes* 'as thin/gaunt/skinny, abandoned, tired, angry as a dog'), while just two of them (*hladový jako pes* 'as hungry as a dog', *věrný jako pes* 'as faithful as a dog') are identical with our data, in which considerably more similes occur. Although this is only an insight into figurative language, several characteristics also appear in our study.

When describing the features associated with the lexeme *had* 'snake', primarily the ones that appear symptomatic will be listed. The reason is that most adjectives denote the negative features usually evoked by this animal (*falošný* 'deceitful', *ľstivý* 'guileful', *podlý* 'wicked') or objectively observed characteristics (*jedovatý* 'venomous', *slizký* 'slimy', *tenký* 'thin'). Therefore, the recorded attributes *krásny* 'beautiful', *múdry* 'wise'[7]*,* and *prítulný* 'cuddly' may appear contradictory in connection with snakes.

A qualitative approach to the compared adjectives reflects the associative views of language users. The texts show that some animals are perceived exclusively negatively (*brav* 'pig, barrow' as *tučný* 'fat', *ťažký* 'heavy'; *aligátor* 'alligator' as *veľký* 'large', *zákerný* 'insidious'), others more positively (*bažant* 'pheasant' as *malý* 'small', *pekný* 'beautiful'; *holúbok* 'dove' as *krotký* 'tame'*, mierny* 'mild', *nevinný* 'innocent'). An example of predominantly positive connotations could be seen in phrases with the lexeme *labuť* 'swan' (*jemný* 'delicate', *krásny* 'beautiful', *pôvabný* 'charming') and exclusively positive perceptions were observed with the diminutive form *včelička/včielka* 'small bee' (*pracovitý* 'hard-working', *usilovný* 'diligent', *šikovný* 'skilful'); on the other hand, the lexeme *včela* 'bee' is can also be associated with negative features (*agresívny* 'aggressive', *nebezpečný* 'dangerous'). Given the above examples, the synthesis of this research can also consider the influence of the word-formation structure of the superordinate noun on attribute selection. For expressive derived nouns, but also those with a diminutive suffix, it seems that language users tend to use an adjective denoting a characteristic with a positive connotation.

Several animal nouns were recorded in unique similes, i.e. with only one characteristic or feature assigned to the bearer (e.g. *bdelý ako ostriež* 'as alert as a perch', *tlstá ako kosatka* 'as fat as a killer whale', *slepý ako krt* 'as blind as a mole'). In most of the cases, these are qualitative adjectives referring to an external characteristic of the animal. In exceptional cases, relational adjectives are used in the comparative structure (*capovitý ako cap* 'as "billygoaty" as a billy goat', *širokoplecí ako býk* 'broad-shouldered like a bull').

[7] The phrase *múdry ako had* 'wise as a snake' can be found in texts as an updated form of the biblical quote *opatrní ako hady* '*cautious as snakes*', and in one of the listed occurrences a metatextual connection is indicated through the use of quotation marks. In another occurrence there is a more significant update in the second part of the original phrase: *múdry ako had a krotký ako holúbok* 'as wise as a snake and as tame as a dove'.

4 CONCLUSION

The conducted research shows that the studied similes are dominated by qualitative adjectives (e.g. *bystrý* ‘clever’, *múdry* ‘smart’, *pyšný* ‘prideful’) that are usually associated with more than one animal. However, the studied phrases also included isolated collocations where an adjective was only used to describe one animal. These included not only qualitative adjectives (*hnusný ako ropucha* ‘as disgusting as a toad’, *prostý ako sova* ‘as simple as an owl’), but also relational ones (*hašterivý ako vrabec* ‘as quarrelsome as a sparrow’, *ochlpený ako potkan* ‘as hairy as a rat’). What was surprising was the absence of certain animals in the similes (e.g. *bzdocha* ‘stink bug’, *svišť* ‘marmot’). From the perspective of evaluativeness, most of the similes are rather neutral (*farebný ako papagáj* ‘as colourful as a parrot’, *ryšavý ako krysa* ‘as ginger as a rat’), some can be viewed along the positive (*chrabrý ako lev* ‘as brave as a lion’, *pilný ako včelička* ‘as diligent as a bee’) – negative (*hlúpy ako hus* ‘as dumb as a goose’, *útočný ako piraňa* ‘as aggressive as a piranha’) axis, while semantically they more often reflect a character trait rather than an external feature. A significant number of similes use the adjective *veľký* ‘large’ (including its comparative forms), which has a wide application (in the SNC: *veľký ako osa, vrabec* ‘as big as a wasp, sparrow’, but also *veľký ako aligátor, medveď* ‘as big as an alligator, bear’). In these cases, the choice of animal is up to the language user, who can also make use of irony.

The semantic divergence of the adjective appears to be useful for the user, as depending on the subjective perception of the speaker one expression can be used for comparisons with differentiated entities, e.g. *bezočivý ako kocúr, opica, ploštica, voš* ‘as impudent as a cat, monkey, bedbug, louse’. However, one animal can also be associated with several, sometimes contrasting characteristics (*baran: hlúpy, krotký, tvrdohlavý, poslušný* ‘ram: stupid, tame, stubborn, obedient’).

The analysis offers suggestions for further investigation of the semantic, syntactic, as well as the word-formation and onomasiological relations of adjectival similes with an animal component. Further plans include observing the semantics of adjectives in comparative constructions containing the names of young animals, as well as constructions with augmentative animal names, investigating transformational nominalization with the consequence of using an identical adjective in both an animal simile and a nominal syntagm in the position of a congruent attribute, as well as analysing the emotions associated with animals.

ACKNOWLEDGEMENTS

The research in this paper was supported by the Slovak National Corpus project supported by the Slovak Academy of Sciences, Ministry of Education, Research, Development and Youth of the Slovak Republic, Ministry of Culture of the Slovak Republic and the Ľudovít Štúr Institute of Linguistics, Slovak Academy of Sciences and by the Slovak Research and Development Agency under the projects APVV-22-0261.

References

Benko, V. (2019). LexiCorp: Corpus Approach to Presentation of Lexicographic Data. In: I. Kosem – T. Zingano Kuhn – M. Correia (eds.): Electronic lexicography in the 21st century: Smart lexicography. Proceedings of the eLex 2019 conference (October 1 – 3, 2019). Brno: Lexical Computing CZ, s.r.o., pp. 957–969.

Čermák, F. (2009). Slovník české frazeologie a idiomatiky 1. Přirovnání. 2., přeprac. a doplň. vyd. Praha: Leda, 512 p.

Horák, G. (1956). K významovému roztriedeniu prídavných mien. In Slovenská reč, 21, pp. 30–34.

Krátky slovník slovenského jazyka. (2003). Red. J. Kačala – M. Pisárčiková – M. Považaj. 4. dopl. a upr. vyd. Bratislava: Veda, 985 p. Accessible at: https://slovniky.juls.savba.sk.

Pauliny, E. (1981). Slovenská gramatika (Opis jazykového systému). Bratislava: SPN, pp. 77–80, 121–129.

Slovenský národný korpus – prim-10.0-public-sane. (2022). Bratislava: Jazykovedný ústav Ľ. Štúra SAV. Accessible at: https://korpus.sk.

Slovník súčasného slovenského jazyka. A – Pn (2006–2021). Bratislava: Veda, vydavateľstvo SAV. Accessible at: https://slovniky.juls.savba.sk.

sciendo
DOI 10.2478/jazcas-2025-0002

CONSTRUCTIONAL CREATIVITY – THE ROLE OF CORPUS DATA

THOMAS HOFFMANN
Department of English Language and Linguistics, Faculty of Languages and
Literatures, Catholic University of Eichstätt-Ingolstadt, Eichstätt, Germany
(ORCID: 0000-0001-6096-3296)

HOFFMANN, Thomas: Constructional Creativity – The Role of Corpus Data. Journal of Linguistics, 2025, Vol. 76, No 1, pp. 20 – 30.

Abstract: One domain of human cognition that has recently received considerable attention in cognitive linguistics is linguistic creativity (e.g. Bergs 2019; Hartmann and Ungerer 2024, 2025; Hoffmann 2024, 2025; Turner 2018). The present paper gives an overview of creativity research from the fields of linguistics and psychology and introduces the 5C model of constructional creativity (constructors, co-constructors, constructs, constructional blending and the constructional network; Hoffmann 2024, 2025a,b). A particular focus of the paper will be on the role of corpus linguistics for the investigation of constructional creativity.

Keywords: constructional network, coverage, creativity, vector space models

1 INTRODUCTION

In their daily interactions, speakers reuse a great number of formulaic, fixed expressions such as *Hi*, *Good morning*, *No worries* or *Love you*. Conklin and Schmitt point out that "studies suggest that formulaic language makes up between one third and one half of discourse" (2012, p. 46). Sinclair (1991) calls this tendency to draw on conventional words and phrases 'the idiom principle'. At the same time, Sinclair notes a competing tendency that he labels the 'open choice principle': People have the ability to combine words into novel utterances that they have never uttered before. For Chomsky (1965, p. 6), this "creative aspect" is "an essential property of language".

In the last couple of years, linguistic creativity has received considerable attention in cognitive linguistics (cf., e.g. Bergs 2018, 2019; Bergs and Kompa 2020; Hartmann and Ungerer 2024, 2025; Herbst 2018; Hoffmann 2018, 2019, 2020, 2022a; Norde and Trousdale 2025; Schneck 2018; Trousdale 2018; Turner 2018, 2020; Uhrig 2018, 2020). The present article first presents a cognitive linguistic model (the 5C model of constructional creativity; Hoffmann 2024, 2025a,b) for the holistic investigation of linguistic creativity (Section 2). Then, it will illustrate how corpus linguistic methods can be used for studies exploring diachronic innovation as well as synchronic creative utterances (Section 3).

2 THE 5C MODEL OF CONSTRUCTIONAL CREATIVITY

While in linguistics creativity is often reduced to productivity (see Sampson 2016 for a detailed discussion), the standard definition in psychology emphasizes that creative products (linguistic or otherwise) have to be **original/novel** as well as **appropriate/useful** (cf., e.g. Simonton 2012; Kaufman 2016). To illustrate this, imagine if I randomly typed away at my keyboard: I might end up with a completely novel sequence of letters that has never been produced before. Yet, while *sdsgd hsd sdaor;rf jejgfb jbwe* might be original it clearly is utterly inappropriate as it is unintelligible. From a linguistic point of view, utterances are therefore creative if they are both novel/original as well as appropriate. While corpus data can help in assessing the degree of novelty of an utterance (see Section 3), appropriateness/usefulness is a subjective criterion that depends on context as well as on the appreciation by listeners or readers (Giora 2003; Veale 2012). In a literary context, such as (1), e.g. a semi-random sequence might become acceptable (or even be considered 'high art'):

(1) The fall (bababadalgharaghtakamminarronnkonnbronntqnnerronntuonnt-hunntrovarrhounawnskawntoohoohoordenenthur-nuk!) of a once wallstrait old-parr […] (James Joyce, *Finegan's Wake*, 1939, 1; cit. in: Bergs 2018, p. 286)

As this brief discussion already illustrates, various perspectives have to be taken into account when assessing the creativity of any phenomenon. Drawing on Glaveanu's 5A model (Glaveanu 2013; Lubart et al. 2021), we can identify the following five major variables:

1. the individual that is creative (the 'actor'),
2. the people who interact with the actor and evaluate the creative product (the 'audience'),
3. the creative product or act (the 'artifact'),
4. the creative process that produces the artifact (the 'action'),
5. the environmental, material and contextual factors that influence the action (the 'affordances').

Hoffmann (2024, 2025a,b) has incorporated the 5As into a cognitive linguistic model of linguistic creativity known as the "5C model of constructional creativity". The theoretical foundation of the model is Construction Grammar (Goldberg 2006, 2019; Hilpert 2019; Herbst and Hoffmann 2024; Hoffmann 2022). The main tenet of Construction Grammar is the claim that constructions, i.e. symbolic pairings of FORM and MEANING are the central units of human language:

Constructions cover the entire lexis-syntax cline and range from words (FORM: /ˈbʊk/ ↔ MEANING: 'concept of a book') over partly schematic patterns

(*Un*-V construction: FORM: /ʌn$_1$-ˈV$_2$/ ↔ MEANING: ‘reversing$_1$ an EVENT$_2$)’ as in *unbind*, *unfreeze*, *undo*) to fully schematic templates (e.g. the Resultative construction: FORM: [SBJ$_1$ [V$_2$ OBJ$_3$ OBL$_4$]VP] ↔ MEANING: ‘Agent$_1$ causes Patient$_3$ to become RESULT-GOAL$_4$ by V$_2$-ing’; adapted from Hoffmann (2022, p. 179), that licenses utterances such as *He wiped the table clean*. or *They cut the man free*.). All of a speakers’ constructions are stored in the long-term memory of their constructional network (Diessel 2019). The production of (creative as well as routine) utterances then involves the activation and combination of various constructions into so-called ‘constructs’ in the working memory. This happens via the domain-general process of Conceptual Blending (Fauconnier and Turner 2002), which is postulated to be the only mental operation required to combine constructions (also known as constructional blending; Herbst and Hoffmann 2024; Hoffmann and Turner fc.).

Fig. 1 provides a visual representation of the full 5C model:

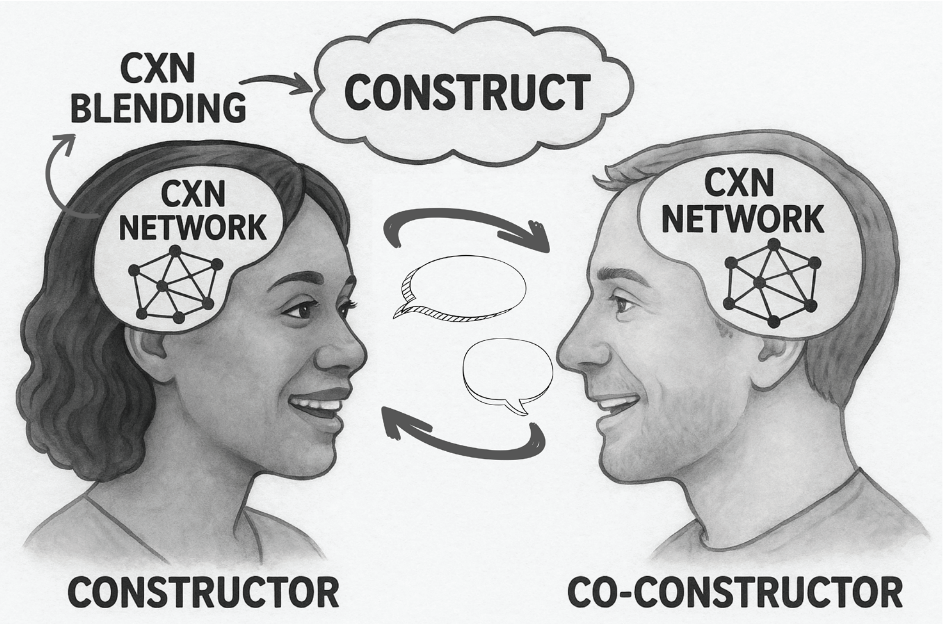


Fig. 1. The 5C Model of Constructional Creativity (available via open access at: https://osf.io/d9hbg)

Glaveanu’s (2013) actor and audience are relabeled as ‘constructor’ and ‘co-constructor’ in the 5C model, and as Hoffmann (2024; 2025a,b) shows it is the dynamic interaction of the two that often leads to creative utterances (the parole elements that realize the mental constructs). The mental process that underlies the production of creative utterances is 'constructional (cxn) blending' (see above). The particular focus of the present paper is on the final part of the 5C model – the 'constructional (cxn) network'.

Usage-based Construction Grammar approaches (such as Goldberg 2006, 2019; Hilpert 2019; Herbst and Hoffmann 2024; Hoffmann 2022) argue that constructions are acquired through language use and that the strength of mental storage (the 'entrenchment' of a construction) depends on frequency effects: If an utterance such as *Good Morning!* is repeatedly encountered without any variation it will become entrenched as a prefab/chunks. When a pattern such as the Resultative construction appears with many different lexicalizations (e.g. *They elected him president*., *He drank himself stupid*., *She danced herself happy.*[1]), it can give rise to the schema above. The specific utterances that underlie the schema are not forgotten immediately: Instead, "partially abstracted (lossy) structured exemplars dynamically cluster within our hyper-dimensional conceptual space" (Goldberg 2019, p. 51). These exemplar clouds, the "coverage" of a construction, play a crucial role when it comes to the acceptability of novel instances:

> Specifically, a potential productive use of an existing construction (a coinage) is acceptable to the degree that the category which would be required to include the previously attested examples and the coinage is well attested within the hyper-dimensional conceptual space in which exemplars cluster (Goldberg 2019, pp. 62–63).

Finally, constructions and their coverage are stored in the long-term memory network of constructions, where they have vertical as well as horizontal links (see, e.g. Diessel 2019; Sommerer and Van de Velde 2025): *Unfair*, *unholy* or *untrue* will be stored as specific taxonomic instances of a more schematic *Un*-Adjective construction. At the same time, a horizontal link to *fair*, *holy* and *true* as well as their schematic positive Adjective construction will encode their antonymic relationship in the network (see Hoffmann 2022, pp. 54–55).

3 CORPUS LINGUISTIC IMPLICATIONS

Most corpora are aggregate data collected from many different individuals. As Schmid (2020) emphasizes, they, therefore, only provide direct evidence for the conventionalization of constructions but not the level of individual mental entrenchment. At the same time, following Schmid's own 'from-corpus-to-cognition principle' (2000, p. 39), Stefanowitsch and Flach (2017, p. 122) argue that corpora at least provide an indirect window onto entrenchment. First of all, corpus data are

[1] *She painted a mess of some pale yellows and dirty greys, got the grown-ups placed, and she danced herself happy.* Source: https://admp.org.uk/wp-content/uploads/E-Motion-Spring-18_Vol-xxviii-No1.pdf [last accessed 08/08/2025].

necessarily the output of individual grammars and as such authentic corpus data can be used to draw “inferences about the mental representations underlying this behavior” (Stefanowitsch and Flach 2017, pp. 102–103; ‘corpus-as-output’ hypothesis). Secondly, according to the ‘corpus-as-input’ view, corpus data can also be seen as a proxy for the input that speakers of a speech community are exposed to and which consequently shape their mental construction networks (Stefanowitsch and Flach 2017, p. 103).

Corpora might, therefore, offer at least indirect evidence for mental constructional networks – but how can they be used in creativity research if creative constructs by definition are supposed to be novel, i.e. should not frequently be found in corpora? One potential use is the evolution of novel constructions in diachronic studies. Hoffmann and Trousdale (2022), e.g. investigated the rise of the construction in (2), which licenses constructs such as (3a,b):

(2) The *Hell*-construction
FORM: [NP$_i$ V$_j$ [the N$_{\text{TABOO}j}$]$_k$ [*out of* [NP$_l$]]
↔
MEANING: [‘SEM$_i$ excessively$_k$ PRED$_j$ SEM$_l$’]
i = Subject/Agent
l = Oblique/Theme

function: [speaker’s heightened emotion]
(adapted from Hoffmann and Trousdale 2022, p. 378)

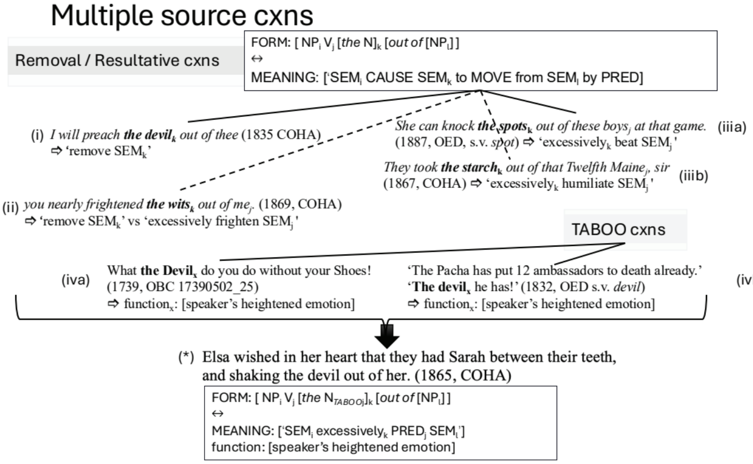
(3) a. I got to beat the Devil out of you, child (2015, COCA)
b. I respect the hell out of those guys (2019, COCA)

The *Hell*-construction in (2) is an extravagant construction (Hartmann and Ungerer 2024, 2025) since it uses a taboo noun (*Devil* (3a), *hell* (3b)) and expresses a speaker’s heightened emotion. Semantically, it expresses the same state of affairs as the Transitive construction (cf. *I got to beat you. I respect those guys*), albeit with the additional meaning that the event is carried out excessively ((3a) means the beating was excessive, (3b) that the level of respect was exceptionally high).

Earlier studies had identified single constructions as the source of the *Hell*-construction. Hoeksema and Napoli (2008), e.g. claimed that constructions that expressed a literal exorcism such as *I will preach the devil out of thee* (1835 COHA[2]) functioned as the source of (2). While the details of the corpus study by Hoffmann and Trousdale (2022) is beyond the scope of the present contribution, Fig. 2 at least

[2] Corpus of Historical American English (COHA). Accessible at: https://www.english-corpora.org/coha/ [last accessed 08/08/2025].

illustrates the various input constructions they identified as potential source constructions in their COHA data:



Multiple source cxns

Removal / Resultative cxns

FORM: [NP$_i$ V$_j$ [*the* N]$_k$ [*out of* [NP$_l$]]
↔
MEANING: [‘SEM$_i$ CAUSE SEM$_k$ to MOVE from SEM$_l$ by PRED]

(i) *I will preach **the devil**$_k$ out of thee* (1835 COHA)
⇨ ‘remove SEM$_k$’

(ii) *you nearly frightened **the wits**$_k$ out of me$_j$.* (1869, COHA)
⇨ ‘remove SEM$_k$’ vs ‘excessively frighten SEM$_j$’

*She can knock **the spots**$_k$ out of these boys$_j$ at that game.* (iiia)
(1887, OED, s.v. *spot*) ⇨ ‘excessively$_k$ beat SEM$_j$’

*They took **the starch**$_k$ out of that Twelfth Maine$_j$, sir* (iiib)
(1867, COHA) ⇨ ‘excessively$_k$ humiliate SEM$_j$’

TABOO cxns

(iva) What **the Devil**$_x$ do you do without your Shoes!
(1739, OBC 17390502_25)
⇨ function$_x$: [speaker’s heightened emotion]

‘The Pacha has put 12 ambassadors to death already.’ (ivb)
‘**The devil**$_x$ he has!’ (1832, OED s.v. *devil*)
⇨ function$_x$: [speaker’s heightened emotion]

(*) Elsa wished in her heart that they had Sarah between their teeth,
and shaking the devil out of her. (1865, COHA)

FORM: [NP$_i$ V$_j$ [*the* N$_{TABOOj}$]$_k$ [*out of* [NP$_l$]]
↔
MEANING: [‘SEM$_i$ excessively$_k$ PRED$_j$ SEM$_l$’]
function: [speaker’s heightened emotion]

Fig. 2. The multiple sources of the modern Hell construction (based on the data of Hoffmann and Trousdale 2022)

Importantly, Hoffmann and Trousdale (2022) note that different speakers might have used different source constructions (e.g. (i) and (iiia) or (ii) and (iva,b) or other routes in Fig. 2) to innovate the novel construction. Similarly, hearers, upon encountering the construction for a first time, could rely on multiple routes in the constructional network to parse and then entrench the new constructions. Careful corpus research, however, is essential to uncover such multiple paths in language change.

Synchronically, corpora are also of great use for creativity studies. Particularly helpful are so-called ‘vector space models’ that use large corpora to produce so-called ‘word embeddings’ (see, e.g. Perek 2016; Surdeanu and Valenzuela-Escárcega 2024). Word embeddings are numerical vectors that encode the context in which a word appears and are an important component of large language models (Surdeanu and Valenzuela-Escárcega 2024, pp. 117–131). Since “words that occur in similar contexts tend to have similar meanings” (Surdeanu and Valenzuela-Escárcega 2024, p. 131), word embeddings can be seen as a proxy for the “semantic representation of words” (Surdeanu and Valenzuela-Escárcega 2024, p. 132). In line with the ‘corpus-as-output’ principle, the word embeddings of a constructional slot, such as the V slot in (2), allows researchers to establish the existing coverage of a construction. This can then be used to assess the creativity of experimental responses.

Hoffmann and Steinhauser (2025), e.g. expanded the Divergent Association Test (DAT; Olson et al. 2021) to constructional contexts. In an online experiment, they asked subjects to “enter 10 verbs that are as different from each other as possible, in all meanings and uses of the words” in constructional contexts such as (4):

(4) Yesterday, the woman _______________ the hell out of the man.

Hoffmann and Steinhauser (2025) randomized the subject and object position of *woman* and *man* (to preclude gender stereotype effects). All in all 55 subjects were recruited via the platform Prolific (age range 18–60, 37 females, 18 males, all L1 speakers of UK English). Again, the specific details of the study are beyond the current paper. Yet, once more, corpus data proved a useful tool to assess the creativity with which subjects responded. Fig. 3 gives the word cloud of all the verbs offered in the study, with the most frequent verbs being larger in size proportional to their frequency:

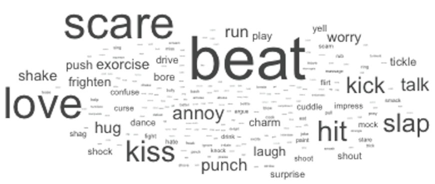


Fig. 3. Word cloud of experimental data of *Hell*-construction (from Hoffmann and Steinhauser 2025)

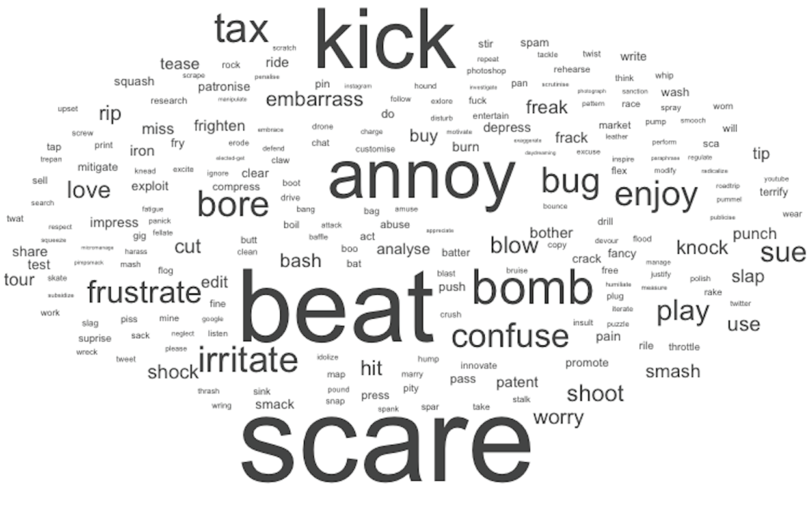


Fig. 4. Word cloud of corpus data of *Hell* construction (based on GB data from GloWbE; Hoffmann 2021)

Compare the results from Fig. 3 to the ones in Fig. 4, which are based on the coverage of the construction in the GB subcorpus of GloWbE (Hoffmann 2021): Since the verbs *beat* and *scare* (as well as their associated semantic frames) dominate the constructional coverage in the corpus, it is not surprising that these were also produced most frequently (and early) by subjects in the experiment.

Even more interesting perhaps is the possibility of using vector space models to plot the range of verbs that were produced in the experiment but did not appear in the corpus (Fig. 5):

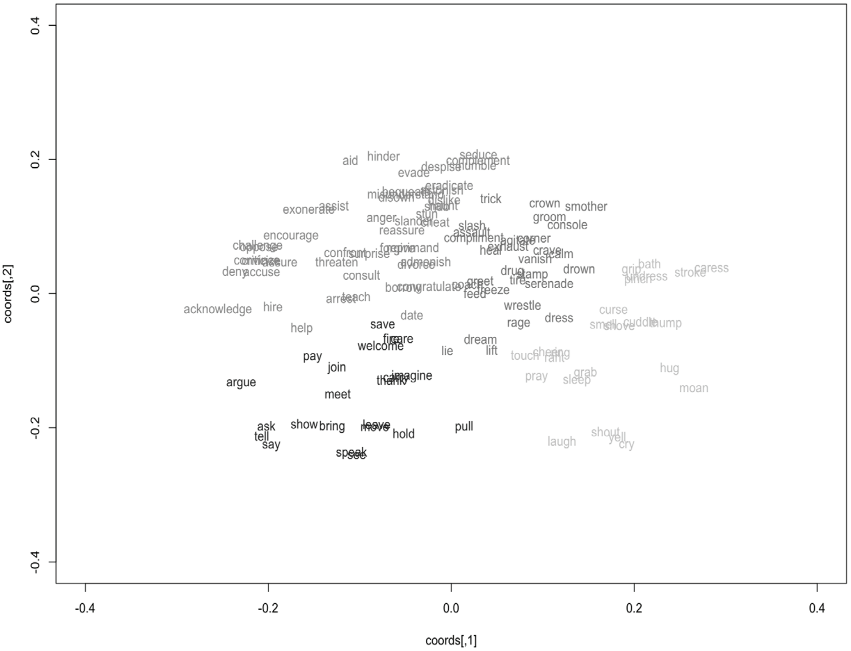


Fig. 5. Vector space model of verbs in *Hell* construction that were produced in the experiment but did not surface in the reference corpus[3] (data from Hoffmann and Steinhauser 2025 and Hoffmann 2021).

Several of the verbs seem to have been activated by the male-female subject-objects (e.g. *caress*, *date*, *seduce*). At the same time, there are many others (e.g. *acknowledge serande*, *vanish*) that warrant closer future attention. It is only through corpus-based vector space models that such relevant instances can be identified.

[3] The vector space that was used for reference was “verbs.coca.w5.skip.1000d.txt” by F. Perek. Accessible at: https://osf.io/3gynu/files/osfstorage/630e4092171132006f5a8dba [last accessed 08/08/2025].

4 CONCLUSION

The present paper has outlined that cognitive linguistic research on creativity must take into account the 5Cs (constructor, co-constructor, constructional network, construction blending, construct). A particular focus was placed on the constructional network and how corpus linguistic studies can fruitfully be used to investigate constructional creativity.

ACKNOWLEDGEMENTS

The research by Hoffmann and Steinhauser (2025) is part of the project Predictive Construction Grammar – Developing a neuro-cognitive model of linguistic creativity funded by the VolkswagenStiftung.

References

Bergs, A. (2018). 'Learn the rules like a pro, so you can break them like an artist' (Picasso): Linguistic aberrancy from a constructional perspective. Zeitschrift für Anglistik und Amerikanistik, 6(3), pp. 277–293.

Bergs, A. (2019). What, if anything, is linguistic creativity? Gestalt Theory, 41(2), pp. 173–183.

Bergs, A. and Kompa, N. (2020). Creativity within and outside the linguistic system. Cognitive Semiotics, 13(1).

Conklin, K., and Schmitt, N. (2012). The processing of formulaic language. Annual Review of Applied Linguistics 32, pp. 45–61.

Giora, R. (2003) On our mind: Salience, context, and figurative language. Oxford: Oxford University Press, 259 p.

Glaveanu, V. (2013). Rewriting the language of creativity: The five A's framework. Review of General Psychology, 17(1), pp. 69–81.

Hartmann, S., and Ungerer, T. (2024). Attack of the snowclones: A corpus-based analysis of extravagant formulaic patterns. Journal of Linguistics, 60(3), pp. 599–634.

Hartmann, S., and Ungerer, T. (2025). Chaos theory, shmaos theory: Creativity and routine in English *shm*-reduplication. In: S. Arndt-Lappe – N. Filatkina (eds.): Dynamics at the lexicon-syntax Interface: Creativity and routine in word-formation and multi-word expressions. De Gruyter, pp. 295–322.

Fauconnier, G., and Turner, M. (2002). The Way We Think: Conceptual Blending and the Mind's Hidden Complexities. New York: Basic Books, 464 p.

Goldberg, A. E. (2006). Constructions at Work: The Nature of Generalization in Language. Oxford: Oxford University Press, 286 p.

Goldberg, A. E. (2019). Explain Me This: Creativity, Competition and the Partial Productivity of Constructions. Princeton: Princeton University Press, 195 p.

Herbst, Th. (2018). Collo-Creativity and blending: Recognizing creativity requires lexical storage in constructional slots. Zeitschrift für Anglistik und Amerikanistik, 66(3), pp. 309–328.

Herbst, Th., and Hoffmann, Th. (2024). A Construction Grammar of English: A Constructionist Approach to Syntactic Analysis (CASA). Amsterdam: John Benjamins, 315 p.

Hilpert, M. (2019). Construction Grammar and its Application to English. 2nd ed. Edinburgh: Edinburgh University Press, 296 p.

Hoeksema, J., and Napoli, D. J. (2008). Just for the hell of it: A comparison of two taboo-term constructions. Journal of Linguistics, 44(2), pp. 347–378.

Hoffmann, Th. (2018). Creativity and construction grammar: Cognitive and psychological issues. Zeitschrift für Anglistik und Amerikanistik, 66(3), pp. 259–276.

Hoffmann, Th. (2019). Language and creativity: A construction grammar approach to linguistic creativity. Linguistics Vanguard, 5(1).

Hoffmann, Th. (2020). Construction grammar and creativity: Evolution, psychology and cognitive science. Cognitive Semiotics, 13(1).

Hoffmann, Th. (2021). The Cognitive Foundation Of Post-colonial Englishes: Construction Grammar as the Cognitive Theory for the Dynamic Model. (Cambridge Elements in World Englishes). Cambridge: Cambridge University Press.

Hoffmann, Th. (2022). Construction Grammar: The Structure of English. (Cambridge Textbooks in Linguistics). Cambridge: Cambridge University Press, 315 p.

Hoffmann, Th. (2024). The 5C model of linguistic creativity: Construction Grammar as a cognitive theory of verbal creativity. Journal of Foreign Languages and Cultures, 8(1), pp. 139–154.

Hoffmann, Th. (2025a). Creativity. In Reference Module in Social Sciences, Elsevier. Accessible at: https://doi.org/10.1016/B978-0-323-95504-1.00588-3.

Hoffmann, Th. (2025b). Cognitive approaches to linguistic creativity. In: X. Wen – Ch. Sinha (eds.): The Cambridge Encyclopedia of Cognitive Linguistics. Cambridge: Cambridge University Press.

Hoffmann, Th., and Steinhauser, M. (2025). Constructional Divergent Association Task (CXN-DAT): Using constructional contexts to measure linguistic creativity. Poster presented at the 2025 Society for the Neuroscience of Creativity (SfNC) conference, Paris Brain Institute (May 22 – 23, 2025).

Hoffmann, Th., and Trousdale, G. (2022). On multiple paths and change in the language network. Zeitschrift für Anglistik und Amerikanistik, 70(3), pp. 359–382.

Hoffmann, Th., and Turner, M. (fc.). Creative Construction Grammar. (Cambridge Elements in Cognitive Linguistics). Cambridge: Cambridge University Press.

Kaufman, J. C. (2016). Creativity 101. 2nd ed. New York: Springer Publishing Company, 368 p.

Lubart, T., Glăveanu, V., de Vries, He., Camargo, A., and Storme, M. (2021). Cultural perspectives on creativity. In: J. C. Kaufman – Robert J. Sternberg, (eds.): Creativity: An introduction. Cambridge: Cambridge University Press, pp. 128–151.

Norde, M., and Trousdale, G. (2025). Creativity, paradigms and morphological constructions: evidence from Dutch pseudoparticiples. Linguistics, 63(4), pp. 1029–1063.

Olson, J. A., Nahas, J., Chmoulevitch, D., Cropper, S. J., and Webb, M. E. (2021). Naming unrelated words predicts creativity. Proceedings of the National Academy of Sciences of the United States of America, 118(25), e2022340118. Accessible at: https://doi.org/10.1073/pnas.2022340118.

Perek, F. (2016).Using distributional Semantics to study syntactic productivity in diachrony: A case study. Linguistics, 54(1), pp. 149–188.

Sampson, G. (2016). Two ideas of creativity. In: M. Hinton (ed.): Evidence, Experiment and Argument in Linguistics and Philosophy of Language. Bern: Peter Lang, pp. 15–26.

Schmid, H.-J. (2000). English Abstract Nouns as Conceptual Shells: From Corpus to Cognition. Berlin: Mouton de Gruyter, 468 p.

Schneck, P. (2018). Creative grammarians: Cognition, language and literature: An exploratory response. Zeitschrift für Anglistik und Amerikanistik, 66(3), pp. 381–391.

Simonton, D. K. (2012). Creative productivity and aging. In: S. Krauss Whitbourne – M. J. Sliwinski (eds.): The Wiley-Blackwell Handbook of Adulthood and Aging. Malden, MA: Wiley-Blackwell, pp. 477–496.

Sommerer, L., and Van de Velde, F. (2025). Constructional networks. In: M. Fried – K. Nikiforidou (eds.): The Cambridge Handbook of Construction Grammar. Cambridge: Cambridge University Press, pp. 220–246.

Stefanowitsch, A., and Flach, S. (2017). The corpus-based perspective on entrenchment. In: H.-J. Schmid (ed.): Entrenchment and the Psychology of Language Learning: How We Reorganize and Adapt Linguistic Knowledge. Berlin: de Gruyter, pp. 101–127.

Surdeanu, M., and Valenzuela-Escárcega, M. A. (2024). Deep Learning for Natural Language Processing. A Gentle Introduction. Cambridge: Cambridge University Press, 344 p.

Trousdale, G. (2018). Creativity parallels between language and music. Zeitschrift für Anglistik und Amerikanistik, 66(3), pp. 371–380.

Trousdale, G. (2020). Creativity, reuse, and regularity in music and language. Cognitive Semiotics, 13(1).

Turner, M. (2018). The role of creativity in multimodal construction grammar. Zeitschrift für Anglistik und Amerikanistik, 66(3), pp. 357–370.

Turner, M. (2020). Constructions and creativity. Cognitive Semiotics, 13(1).

Uhrig, P. (2018). I don’t want to go all Yoko Ono on you. Zeitschrift für Anglistik und Amerikanistik, 66(3), pp. 295–308.

Uhrig, P. (2020). Creative intentions – The fine line between ‘creative’ and ‘wrong’. Cognitive Semiotics, 13(1).

Veale, T. (2012). Exploding the Creativity Myth: The Computational Foundations of Linguistic Creativity. London et al.: Bloomsbury, 184 p.

sciendo
DOI 10.2478/jazcas-2025-0003

A CORPUS-BASED APPROACH TO THE STUDY OF PERMISSIVE CONSTRUCTIONS IN ENGLISH AND UKRAINIAN

IRYNA KARAMYSHEVA
Department of Applied Linguistics, Institute of Computer Science and Information Technologies, Lviv Polytechnic National University, Lviv, Ukraine
(ORCID: 0000-0001-8683-2040)

KARAMYSHEVA, Iryna: A Corpus-based Approach to the Study of Permissive Constructions in English and Ukrainian. Journal of Linguistics, 2025, Vol. 76, No 1, pp. 31 – 40.

Abstract: This paper presents a corpus based contrastive study of permissive constructions in English and Ukrainian. Sentences with studied permissive constructions are viewed as sentences with secondary predication constructions expressed by a non-finite complement after the complement taking a predicate with the meaning of permission. In English the secondary predicate within the non-finite complement is expressed by the Infinitive, Participle I, Participle II and Nominals (adjective/noun). In Ukrainian the only type of permissive constructions available are with the Infinitive as secondary predicate. Permissive constructions present a network of constructions with partly schematic, partly substantive 'meso-constructions' playing an important role in the entrenchment process. Identification of 'meso-constructions' is helpful for shaping the correct search query. The research material is gathered from two corpora, namely COCA and General Regionally Annotated Corpus of Ukrainian (GRAC). The frequency analysis of data helps to draw conclusions within the Usage-based Construction Grammar approach.

Keywords: permissive construction, meso-construction, type frequency, corpus-based study, English/Ukrainian

1 INTRODUCTION

Corpus Linguistics has truly revolutionized the world of language study, spreading into many applied spheres such as language teaching and second language acquisition. Recent publications (Hunston 2022) reflect not only technological advances but also focus on methodological progress and the social impact of Corpus Linguistics, highlighting an extensive use of corpus data for research within linguistic frameworks.

The present paper is meant as a contribution to corpus-based contrastive studies, which can be considered a successful merger of Contrastive Linguistics with achievements within the field of Corpus Linguistics (Hasselgård 2020, p. 185). To be more precise, the research is a corpus-based contrastive grammar study (English-Ukrainian language pair), embedded in the framework of Usage-based Construction Grammar.

The English sentence with the permissive construction, as in (1), is a sentence with secondary predication that is expressed by the non-finite complement '*students to choose*', licensed by the complement taking predicate '*allow*', which is a primary predication predicate with the meaning of permission.

(1) *Most schools* ***allow students to choose*** *from a list of books*. (NEWS: Christian Science Monitor, 1997)

A secondary predication construction itself consists of a secondary subject, expressed by a pronoun in the objective case or a noun in the common case, and a secondary predicate, expressed by a non-finite. Consider example (1) with the syntactic roles described:

Most schools (S1) allow (P1) [students (S2) to choose (P2) from a list of books (adjunct)][object].

Ukrainian learners of English typically have difficulties with such sentences. Practical grammars of English for Ukrainian learners prevailingly advise rendering secondary predication constructions with the help of subordinate sentences with the tensed finite forms of the verb. However, our experience of working with different types of secondary predication constructions in English has shown that we do have equivalent non-finite complement constructions in Ukrainian. Here the merits of contrastive analysis come into play, since it helps pay more attention to some specific phenomena in a language that may otherwise go unnoticed.

This paper has the following structure: Section 1 introduces the research object and states the main research purpose and Section 2 gives a relevant theoretical background to the study. Section 3 outlines the material and methods used. The subsections of Section 3 present the analysis of data gathered from two corpora, describing the studied constructions in English and Ukrainian. Section 4 presents a discussion of the main findings, with some general concluding remarks.

2 THEORETICAL BACKGROUND

2.1 Permissive constructions with non-finite complements

Permissive constructions are viewed in this research as constructions that contain a matrix verb encoding the act of permission and non-finite complementation that makes a secondary predication in addition to that of the main verb. The understanding of 'construction' is accepted in this paper as defined by Croft (2022, p. 17): "any pairing of form and function in a language [...] used to express a particular combination of semantic content and information packaging". The packaging of the semantic content can be organized as predication. Events prototypically function as predications. A complement clause construction is defined, according to Croft, in terms of encoding one event as the argument of a second event. Only certain predicates allow events as arguments; these predicates are called complement-taking predicates or CTPs (Noonan 2007, p. 53; Croft 2022, pp. 551–558). Noonan (2007, pp. 52–150) distinguishes among others such types of CTP events that can have a second event encoded by a complement construction: perception events, desiderative events, and manipulative events. Manipulatives include the closely related causative and permissive predicates, both involving an element of causation (Noonan 2007, p. 136). Manipulative predicates express a relation between an agent or a situation which functions as a cause, an affectee, and a resulting situation. The affectee must be a participant in the resulting situation. Moreover, manipulative predicates may in addition encode information about the manner of causation (compare, for example, causative '*force*' and permissive '*let*').

Permissive constructions, also called enablement constructions of the different-subject construction type (e.g. Egan 2008, pp. 13, 23), have been the focus of attention less often in comparison to their causative 'kins'. Therefore, this case study is devoted to revealing the range of permissive constructions and CTPs they are used with with the help of corpus data.

2.2 Permissive constructions as a network of constructions

Following the constructionist approach, we describe permissive constructions as constructs with their form and meaning/function with the specific information packaging as structures with secondary predication embedded into the primary predication structure in the form of the non-finite complement, performing the function of object after the CTP (P1) with the specific meaning of "permission for performing some action":

FORM: [X permits/allows Y Vnon-finite]
MEANING: X represents a manipulative force or agent, while Y represents a patient/an affectee who is permitted to perform an act (Vnon-finite): 'an agent permits a patient to perform some action'.

Permissive constructions form a certain subnetwork within the network of non-finite complement constructions. Constructions as mental representations also vary according to the degree of their schematicity. Traugott and Trousdale (2013, p. 16) propose the following minimal set of constructional levels: schemas, subschemas, and micro-constructions. In the same vein Hoffmann et al. (2019, pp. 6, 26), also Horsch (2023a, p. 705–707) speak about 'micro-constructions' (specific, substantive instances of a construction), 'macro-constructions' (abstract schematic constructional templates) and 'meso-constructions' (semi-productive, partly substantive, partly schematic intermediate entities).

According to Diessel (2023, p. 29) we can speak not only about taxonomic relations of grammatical patterns in the constructicon as an inheritance network: “Every (schematic) construction includes at least one slot that is associated with a class of lexical and/or phrasal fillers”. Data extracted from corpora (Section 3) reveal that we have a larger range of permissive constructions in English in comparison to Ukrainian with more lexical fillers (in our case CTPs). Moreover, we have filler types: Infinitive, Participle I and II used as non-finites, as well as Nominal (adjective/noun), as a result of ‘to be’ deletion, treated as a separate type. In Ukrainian only the use of Infinitive is possible.

One of the basic assumptions of the usage-based approach is that constructions are entrenched as a consequence of input frequency. Evidence of the entrenchment of constructions can be found by employing two fundamental concepts of usage-based language study – ‘token’ and ‘type’ frequency. Whereas ‘token frequency’ is evidence for specific and substantive constructions, ‘type frequency’ “[…] plays a crucial role in identifying types, or meso-constructions” (Horsch 2023b, p. 291). Reflecting upon the advances in statistical analysis in recent decades, Gries (2023, p. 562) affirms that token frequency per corpus is supposed to be causally related to entrenchment whereas type frequency, by contrast, productivity, acquisition, and grammaticalization. Therefore, the study of grammatical constructions with the help of corpora is inevitably connected with analyzing their token and type frequencies.

3 CORPORA, DATA AND METHODS APPLIED

3.1 Methods

The present study is meant as a contribution to corpus-based contrastive grammar studies. Consequently, the main methods applied are contrastive analysis and frequency analysis.

The novelty of this research is that English permissive constructions are compared with Ukrainian ones, following claims that “[…] the notion of construction provides us with a valuable and useful concept for cross-linguistic comparison and analysis” (Boas 2010, p. 16) and that constructions, as the basic unit at all levels of analysis, can be found in Slavic as well (Fried 2017, pp. 243–244). In line with Boas’ suggestion that English should serve as “basis” (2010, p. 14) for contrastive CxG-based investigations, Horsch applied the methodology from a study on the English Comparative Correlative constructions in Slovak and in Spanish (Hoffmann et al. 2019; Horsch 2023a; Horsch 2024).

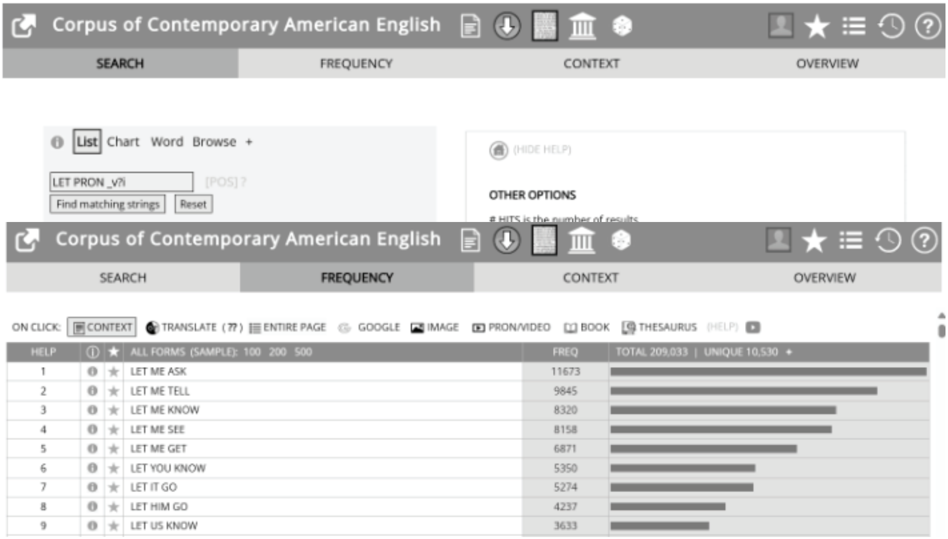
The collection of data from corpora in order to study the token and type frequency of permissive constructions in English and Ukrainian presupposes the application of frequency analysis in this research.

3.2 Corpora

Data extraction procedure from COCA. The Corpus of Contemporary American English (COCA) was used to extract English permissive constructions.

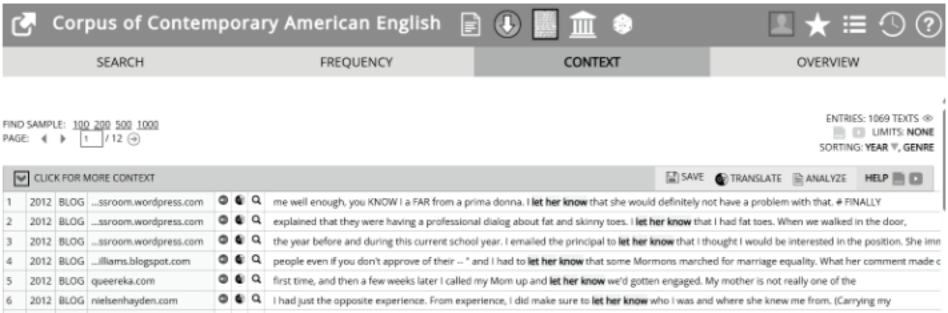
Permissive constructions with non-finite complements follow the pattern: N1 V N2 non-f V (N3). This pattern can be regarded as a maximally abstract 'constructional template', reflecting the sequence of parts of speech used to express the primary predication as the main clause and the secondary predication as an embedded non-finite complement construction. To build a proper search query it was necessary to take into account: 1) the verb (P1) that serves as a permissive CTP, taking the non-finite complement; 2) the expression of N2/S2 (secondary subject or the semantic subject of the non-finite complementation), which is most often expressed by a pronoun in the objective case or less often a noun/noun phrase in the common case; 3) the expression of the P2 (secondary predicate) which is expressed by such non-finite forms as Infinitive (with and without 'to'), Participle I and II as well as Nominal (adjective/noun/noun phrase), which has to be reflected in the choice of correct POS tags.

The set of tokens within the first 100 hits yielded the following most frequent verbs serving as CTPs: 1) VERB PRON _V?I (bare infinitive): *let, make, help, hear, see*; 2) VERB PRON TO _V?I (infinitive with 'to'): *want, would like, expect, lead, find, need, ask, tell, allow, help, believe, invite*. This list remains practically unchanged up to 1000 hits and is topped by the CTP '*let*' used with permissive constructions. Other CTPs are used with causative, desiderative, evaluative and perception subtypes of constructions with the non-finite complement in English. Therefore, the second stage of the search procedure was to use more specialized queries with specific verbs as CTPs. Consider Fig.1 (example with the verb '*let*' as a CTP):



LET PRON _v?i

HELP	ⓘ	★	ALL FORMS (SAMPLE): 100 200 500	FREQ	TOTAL 209,033 | UNIQUE 10,530 +
1	ⓘ	★	LET ME ASK	11673	
2	ⓘ	★	LET ME TELL	9845	
3	ⓘ	★	LET ME KNOW	8320	
4	ⓘ	★	LET ME SEE	8158	
5	ⓘ	★	LET ME GET	6871	
6	ⓘ	★	LET YOU KNOW	5350	
7	ⓘ	★	LET IT GO	5274	
8	ⓘ	★	LET HIM GO	4237	
9	ⓘ	★	LET US KNOW	3633	



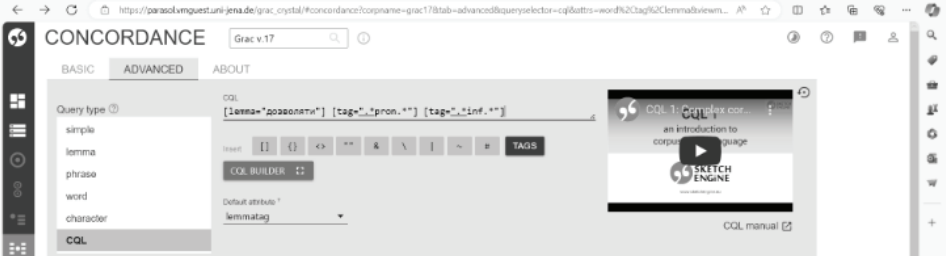
ENTRIES: 1069 TEXTS
LIMITS: NONE
SORTING: YEAR ▼, GENRE

1	2012	BLOG	...ssroom.wordpress.com	me well enough, you KNOW I a FAR from a prima donna. I **let her know** that she would definitely not have a problem with that. # FINALLY
2	2012	BLOG	...ssroom.wordpress.com	explained that they were having a professional dialog about fat and skinny toes. I **let her know** that I had fat toes. When we walked in the door,
3	2012	BLOG	...ssroom.wordpress.com	the year before and during this current school year. I emailed the principal to **let her know** that I thought I would be interested in the position. She imn
4	2012	BLOG	...illiams.blogspot.com	people even if you don't approve of their -- " and I had to **let her know** that some Mormons marched for marriage equality. What her comment made c
5	2012	BLOG	queereka.com	first time, and then a few weeks later I called my Mom up and **let her know** we'd gotten engaged. My mother is not really one of the
6	2012	BLOG	nielsenhayden.com	I had just the opposite experience. From experience, I did make sure to **let her know** who I was and where she knew me from. (Carrying my

Fig. 1. A combination of screenshots, exemplifying a specialized query with the examples of sentences

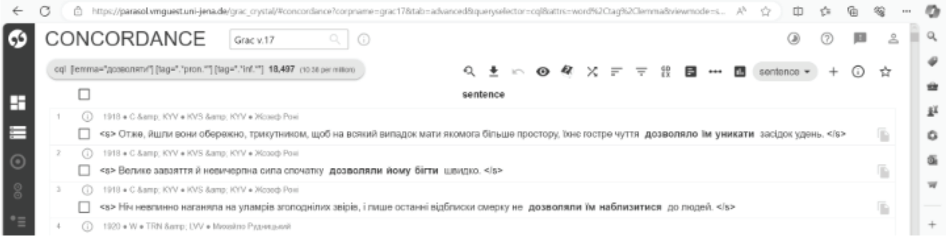
Consequently, more specialized queries reflecting meso-constructions appeared to be more efficient.

Data extraction procedure from GRAC. The General Regionally Annotated Corpus of Ukrainian (GRAC; uacorpus.org) is a general-purpose reference corpus of Ukrainian and is the largest and the most representative corpus of Ukrainian by far. The corpus counts 1.781 billion tokens. One of the newest versions, Grac.v.17, was used. The search procedure with GRAC is more complicated since it requires a specialized CQL expression for producing a correct query (Fig. 2). The search with a maximally abstract 'constructional template', similar to the procedure with COCA, was not successful with GRAC. Only specialized queries, with CTPs included, yielded the necessary results. Similarly, separate queries should have been produced to search for constructions where S2 was expressed by a pronoun, or by a noun/noun phrase, e.g.: `[lemma="дозволяти"] [tag=".*pron.*"] [tag=".*inf.*"]`, `[lemma="дозволяти"] [tag=".*noun.*"] [tag=".*inf.*"]`. The query with the noun tag contained samples of sentences intermixed with pronouns, which called for the manual sorting out of such cases. Therefore, a more specialized query for the search of permissive constructions with S2 expressed by a noun/noun phrase was applied: `[lemma="дозволяти"] [tag=".*noun.*" & tag!=".*pron.*"] [tag=".*inf.*"]`


https://parasol.vmguest.uni-jena.de/grac_crystal/#concordance?corpname=grac17&tab=advanced&queryselector=cql&attrs=word%2Ctag%2Clemma&viewm...
CONCORDANCE
Grac v.17
BASIC
ADVANCED
ABOUT
Query type
simple
lemma
phrase
word
character
CQL
Insert
[]
{}
<>
""
&
\
|
~
#
TAGS
CQL BUILDER
CQL 1: Complex cor...
CQL 1
an introduction to corpus ... language
SKETCH ENGINE
CQL manual

CQL
[lemma="дозволяти"] [tag=".*pron.*"] [tag=".*inf.*"]

Default attribute
lemmatag



https://parasol.vmguest.uni-jena.de/grac_crystal/#concordance?corpname=grac17&tab=advanced&queryselector=cql&attrs=word%2Ctag%2Clemma&viewmode=s...

CONCORDANCE

Grac v.17

cql [lemma="дозволяти"] [tag=".*pron.*"] [tag=".*inf.*"] 18,497 (10.38 per million)

sentence

1 1918 • C &amp; KYV • KVS &amp; KYV • Жозеф Роні

<s> Отже, йшли вони обережно, трикутником, щоб на всякий випадок мати якомога більше простору, їхнє гостре чуття **дозволяло їм уникати** засідок удень. </s>

2 1918 • C &amp; KYV • KVS &amp; KYV • Жозеф Роні

<s> Велике завзяття й невичерпна сила спочатку **дозволяли йому бігти** швидко. </s>

3 1918 • C &amp; KYV • KVS &amp; KYV • Жозеф Роні

<s> Ніч невпинно наганяла на уламрів зголоднілих звірів, і лише останні відблиски смерку не **дозволяли їм наблизитися** до людей. </s>

4 1920 • W • TRN &amp; LVV • Михайло Рудницький

Fig. 2. A combination of screenshots, exemplifying a specialized query in GRAC with examples sentences

The appropriateness of the extracted examples was checked by looking through combinations of strings in order to 'weed out' the so called 'false positives', especially in COCA. There was a careful check of examples within 1000 hits. This procedure revealed a certain tendency: the more frequent the construction in the corpus is, the fewer incorrect matches are yielded. Subtypes of constructions, comparatively small in number, were checked in a full amount. Therefore, we believe that the arrangement of constructions with a certain CTP in order of descending frequency as well as the percentage correlation of data can be considered correct.

3.3 DATA

Types of English permissive constructions. The data obtained from COCA corroborates the availability of four subtypes of permissive constructions according to the filler type: Infinitive, Participle I and II, and Nominal (adjective/noun). The list of CTPs, triggering the subtype with the Infinitive, includes 6 verbs listed in order of their frequency: *let, help, allow to, enable to, permit to, leave to*. These CTPs were used to build two types of queries – with pronoun and noun as S2 correspondingly, e.g.: LET PRON _V?I and LET NOUN _V?I Consider example (2).

(2) *I don*'*t* ***let him have*** *a Facebook account [...].* (BLOG: momfaze.com, 2012)

The list of CTPs triggering the subtype with Participle I includes 3 verbs listed in order of their frequency: *leave, let, allow.* Consider the query sample LEAVE PRON _V?G and example (3):

(3) *The river chill had* ***left him feeling*** *feverish and brittle*. (FIC: Dennis Mahoney. Bell weather: a novel, 2016)

The subtype with Participle II is triggered only by one CTP: *leave.* Consider the query sample LEAVE PRON _V?N and example (4):

(4) *They* ***left him tied*** *to the fence*. (NEWS: USA Today, 1998)

The list of CTPs triggering the subtype with Nominal includes 2 verbs: *leave, let.* Consider the query sample LET PRON ADJ and example (5):

(5) *He just wouldn*'*t* ***let me alone***. (FIC: O'Shaughnessy, Perri. Show no fear. New York: Pocket Books, 2008)

The type frequency of English permissive constructions is the following: with the Infinitive (413,488 tokens – 94.38%), with Participle I (5,269 tokens – 1.2%), with Participle II (739 tokens – 0.17%), with Nominal (18,648 tokens – 4.25%). It is obvious that the most prototypical representative of permissive constructions in English is the subtype with the infinitive, triggered by the largest number of CTPs (6) and containing the CTP '*let*' with the highest number of tokens (221,627 tokens – 50.58% out of the total number of permissive constructions 438,144 tokens (100%)). This proves that the construction with '*let*' has the highest degree of entrenchment and the subtype of permissive constructions with the Infinitive is a highly productive one.

Types of Ukrainian permissive constructions. Data obtained from GRAC corroborates the availability of permissive constructions but only with one filler type – the Infinitive. The number of CTPs, triggering this subtype, contains only three verbs given in order of their frequency: '*дозволяти*' *(let, allow to*), '*допомагати*' *(help)*, '*залишати*' (*leave to*). These CTPs were used in two queries – with a pronoun or a noun as S2 correspondingly. Consider example (6).

(6) *Такі зустрічі […]* ***дозволяють нам обмінятися*** *досвідом*
Such meetings allow PRES PL us Pronoun DAT PL exchange NFINITIVE experience
(Онлайн-ЗМІ "Чернігівщина: події і коментарі", 2013)
'*Such meetings [...]* ***allow us to exchange*** *experiences*' (Online media "Chernihiv Region: Events and Comments", 2013)

The peculiar feature of Ukrainian permissive constructions is that S2 expressed by the personal pronoun takes the dative case with CTPs '*дозволяти*' (*let, allow to*), '*допомагати*' (*help*), and the accusative case with the CTP '*залишати*' (*leave to*). Altogether, Ukrainian permissive constructions are less frequent, if we compare the subtype with the Infinitive, with the total number of tokens 54,757.

4 RESULTS AND DISCUSSION

The present study relied on insights from Usage-based Construction Grammar. Corpus data play a crucial role in Usage-based CxG, based on the assumption that grammar is shaped by the frequency of use. The data harvested from two corpora, COCA and GRAC, allows to state the availability of the following types of permissive constructions within the contrasted English-Ukrainian language pair (Tab. 1):

Permissive construction with	English (quantity of lexical fillers/CTP's)	Relative frequency per million	Ukrainian (quantity of lexical fillers/CTP's)	Relative frequency per million
Infinitive	6	413.488	3	30.72
Participle I	3	5.269	-	
Participle II	1	0.739	-	
Nominal (adjective/noun)	2	18.648	-	

Tab. 1. Subtypes of permissive constructions in English and Ukrainian with the CTPs used and their quantity given as relative frequency per million

The notion of 'construction' itself served as *tertium comparationis* to carry out the contrastive analysis of English and Ukrainian permissive constructions with non-finite complements as secondary predication constructions. The correct choice of *tertium comparationis* proved that 'construction' indeed is a viable instrument, serving as a comparative concept. It has to be remarked that meso-constructions (partly schematic, partly substantive constructions) play an important role in the taxonomic constructional network, being intermediate between micro-constructions (attested tokens) and macro-constructions (maximally abstract templates). Meso-constructions are also useful in building correct specialized queries for extracting the necessary data from corpora.

The analysis helped reveal that English permissive constructions can be truly regarded as a network of constructions within English constructions containing non-finite complements with four attested types of fillers as a secondary predicate: the Infinitive, Participle I and II, Nominal. The permissive construction with the Infinitive, triggered by the CTP '*let*', has the highest degree of entrenchment and, therefore, the subtype with the Infinitive is highly productive in modern English. This cannot be said about the Ukrainian permissive constructions, which are used only with one filler type – the Infinitive. Nevertheless, the corpus-based contrastive analysis helped reveal a specific feature of Ukrainian constructions: the secondary subject expressed by a personal pronoun/noun is used not only in the accusative case but as well in the dative. This is a fact worth paying attention to since in English traditional grammars, non-finite complement constructions can be found under the terms 'Accusativus cum Infinitivo/Partizipio' and the English objective case of personal pronouns is considered to be equivalent to the accusative case. Therefore, the analysis can be useful for Ukrainian learners of English grammar in many respects. Consequently, the presented study makes an important contribution to the disciplines of corpus-based contrastive studies, in particular Contrastive Grammar of English and Ukrainian Languages, as well as to Usage-based Construction Grammar.

References

Boas, H. C. (2010). Comparing constructions across languages. In: H. C. Boas (ed.): Contrastive Studies in Construction Grammar (Constructional Approaches to Language 10). Amsterdam: John Benjamins, pp. 1–20.

COCA. Accessible at: https://www.english-corpora.org/coca/.

Croft, W. (2022). Morphosyntax: Constructions of the World's Languages. Cambridge University Press, 688 p.

Diessel, H. (2023). The Constructicon: Taxonomies and Networks. Cambridge University Press, 75 p.

Egan, Th. (2008). Non-finite complementation. A usage-based study of infinitive and -ing clauses in English. Amsterdam – New York, NY: Rodopi, 432 p.

Fried, M. (2017). Construction Grammar in the Service of Slavic Linguistics, and Vice Versa. Journal of Slavic Linguistics, 25(2), pp. 241–276.

Gries, S. Th. (2023). New Technologies and Advances in Statistical Analysis in Recent Decades. In: M. Díaz-Campos – S. Balasch (eds.): The Handbook of Usage-Based Linguistics. New Jersey: Wiley-Blackwell, pp. 561–579.

Hasselgård, H. (2020). Corpus-based contrastive studies: Beginnings, developments and directions. In Languages in Contrast: International Journal for Contrastive Linguistics, 20(2), pp. 184–208.

Hoffmann, Th. et al. (2019). The More Data, The Better: A Usage-based Account of the English Comparative Correlative Construction. Cognitive Linguistics, 30(1), pp. 1–36.

Horsch, J. (2024). English as a basis for contrastive constructional studies: A case study. In Book of Abstracts ICCG 13. The 13th International Conference on Construction Grammar. Göteborg (August 26 – 28, 2024), pp. 61–64.

Horsch, J. (2023a). From corpus data to constructional networks: Analyzing language with the Usage-based Construction Grammar framework. Journal of Linguistics, 74(3), pp. 701–740.

Horsch, J. (2023b). The comparative correlative construction in World Englishes: a usage-based construction grammar approach. Eichstätt (Germany): Catholic University of Eichstätt-Ingolstadt. PhD thesis, 346 p.

Hunston, S. (2022). Corpora in Applied Linguistics. 2nd edition. Cambridge, Cambridge University Press, 341 p.

Noonan, M. (2007). Complementation. In: T. Shopen (ed.): Language typology and Syntactic Description, Vol. II: Complex constructions, 2nd edition, pp. 52–150.

Shvedova M., von Waldenfels R., Yarygin S., Rysin A., Starko V., Nikolajenko T. et al. (2017–2025): GRAC: General Regionally Annotated Corpus of Ukrainian. Electronic resource: Kyiv, Lviv, Jena. Accessible at: uacorpus.org.

Traugott, E., and G. Trousdale. (2013). Constructionalization and Constructional Changes (Oxford Studies in Diachronic & Historical Linguistics 6). Oxford: Oxford UP, 304 p.

sciendo
DOI 10.2478/jazcas-2025-0004

COMMA DISTRIBUTION IN CZECH TEXTS: VARIATION BY GENRE AND AUTHOR, AND ERROR ANALYSIS

JAKUB MACHURA[1] – HANA ŽIŽKOVÁ[2] – VOJTĚCH KOVÁŘ[3]
[1]Department of Czech Language, Faculty of Arts, Masaryk University, Brno, Czech Republic (ORCID: 0000-0002-6623-3064)
[2]Department of Czech Language, Faculty of Arts, Masaryk University, Brno, Czech Republic (ORCID: 0000-0002-6483-6603)
[3]Department of Czech Language, Faculty of Arts, Masaryk University, Brno, Czech Republic (ORCID: 0009-0005-0307-9046)

MACHURA, Jakub – ŽIŽKOVÁ, Hana – KOVÁŘ, Vojtěch: Comma Distribution in Czech Texts: Variation by Genre and Author, and Error Analysis, 2025, Vol. 76, No 1, pp. 41 – 51.

Abstract: This article investigates the distribution and typology of commas in Czech texts, combining genre-differentiated samples with an annotated error corpus to offer a comprehensive view of punctuation usage and misuse. Building on previous work, we expand the analysis from a small newspaper sample to a broader set of texts, encompassing fiction, blogs, translations, and school dictations. Using a consistent typology of comma usage, we classify 1,000 manually selected instances and identify trends in different textual genres. Furthermore, we examine over 1,000 missing comma errors and more than 200 redundant ones from the self-built error corpus. The results reveal genre-dependent tendencies in comma types, especially in the use of commas preceding connectives and within asyndetic structures. The study offers insights for improving automatic comma insertion systems and deepens our understanding of punctuation norms and deviations in Czech.

Keywords: Comma typology, Punctuation errors, Czech language, Automatic comma insertion

1 INTRODUCTION

Punctuation plays a critical role in written communication, structuring sentences and guiding interpretation. Among punctuation marks, the comma is both frequent and frequently misused, making it a prime subject of computational linguistic research and grammatical annotation. In Czech, comma placement follows a complex set of syntactic rules and conventions, which are not always intuitively understood by writers—particularly in informal contexts or in translation.

A detailed typology of comma usage in Czech was proposed by Machura et al. (2022), laying the groundwork for further empirical analysis. The primary aim of the proposed typology was to systematically classify the positions of commas within Czech sentence structures, with particular attention to syntactic, semantic, and lexical factors. Such classification enables the identification of comma types that can be reliably defined through explicit linguistic rules, making them suitable for implementation in rule-based systems for automatic comma correction. Conversely, the typology also highlights those cases where comma placement is more ambiguous or context-dependent and thus requires statistical modeling or more advanced morphological, syntactic, and semantic analysis. While their initial study provided a valuable classification framework and a rough frequency estimate based on newspaper articles, the limited size and genre scope of the sample called for a broader, more representative dataset. At the same time, the need for improved evaluation methods for automatic comma insertion models has grown alongside the development of proofreader tools.

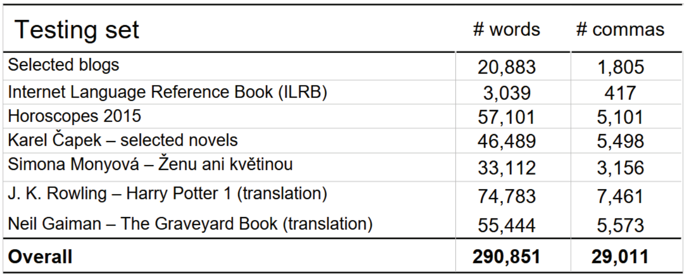
This article presents two complementary studies: first, a distributional analysis of 1,000 commas drawn from a variety of text genres; second, an error analysis using the self-built annotated corpus. Together, these perspectives illuminate both how commas are typically used in Czech writing and where writers most often go wrong.

2 COMMAS DISTRIBUTION IN CZECH TEXTS

In (Machura et al. 2022) the typology of comma insertion place was comprehensively described. This allows 1) to specify the place (boundary) in the sentence structure where a comma is inserted, 2) to analyze the type of commas that users of the language omit or overuse, or 3) to evaluate the results of language models that are pre-trained namely for the task of inserting commas into text, and then subsequently improve these models. Based on a relatively small sample of newspaper articles, which consisted of 183 sentence commas, a very rough frequency distribution of commas by type was outlined in that paper. Therefore, it was decided to analyze a larger sample that would more accurately determine the comma type distribution while also being representative, as it would consist of texts of different kinds, not just newspaper articles.

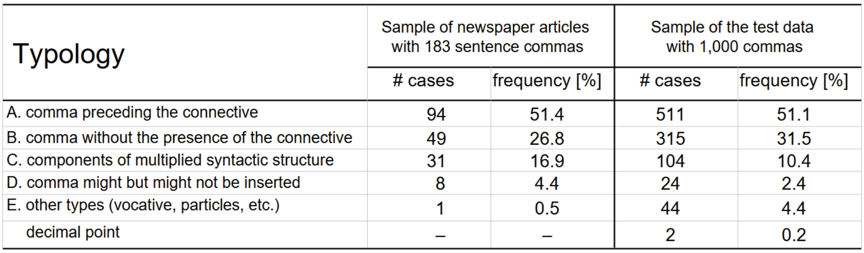
The new larger sample consisting of 1,000 commas was created from the same data presented in (Kovář et al. 2016), which are used specifically for the evaluation and comparison of methods for automatic comma insertion into Czech text. Since the data are exactly the same, it is also possible to compare the current results with testing done in the past (Machura et al. 2022; Machura et al. 2023). In total, seven texts of different natures and styles are used as testing data, see Tab. 1.

From each of the 7 texts, a sample containing 125 commas was selected. To add to the total of 1,000 commas, a sample from school dictations was included, which also contained 125 commas. All 1,000 commas were classified according to the selected typology and compared with a previous smaller sample (183 commas) from newspaper articles, see Tab. 2. The largest group, *A. a comma preceding the connective*, again reached slightly more than half of all commas (51.1%). Type *B. comma without the (near) presence of the connective* reached less than one-third (31.5%), while type *C. comma separating components of multiplied syntactic structure* decreased to only about one-tenth of all commas (10.4%). It turns out that type *D. cases where a comma is not obligatory or can change the meaning of the utterance* is even less frequent (2.4%), whereas type *E. commas around vocative phrases or particles and interjections* (standing outside the structure and syntactically independent) is more frequent (4.4%). This increase can also be explained by the selection of texts, as there were four fiction texts in the sample where vocative phrases may appear more often. There were also two commas in the sample which are used as a decimal point in numeral notation (in English, the symbol of the period is used as a decimal point whereas the comma also works as a thousand separator comma, and therefore both the period and the comma are ambiguous punctuation marks).



Testing set	# words	# commas
Selected blogs	20,883	1,805
Internet Language Reference Book (ILRB)	3,039	417
Horoscopes 2015	57,101	5,101
Karel Čapek – selected novels	46,489	5,498
Simona Monyová – Ženu ani květinou	33,112	3,156
J. K. Rowling – Harry Potter 1 (translation)	74,783	7,461
Neil Gaiman – The Graveyard Book (translation)	55,444	5,573
Overall	**290,851**	**29,011**

Tab. 1. Statistics of the test data for automatic comma insertion



Typology	Sample of newspaper articles with 183 sentence commas		Sample of the test data with 1,000 commas	
	# cases	frequency [%]	# cases	frequency [%]
A. comma preceding the connective	94	51.4	511	51.1
B. comma without the presence of the connective	49	26.8	315	31.5
C. components of multiplied syntactic structure	31	16.9	104	10.4
D. comma might but might not be inserted	8	4.4	24	2.4
E. other types (vocative, particles, etc.)	1	0.5	44	4.4
decimal point	–	–	2	0.2

Tab. 2. Comparison of the distribution of commas on a small (genre-specific) and a larger (genre-diverse) sample

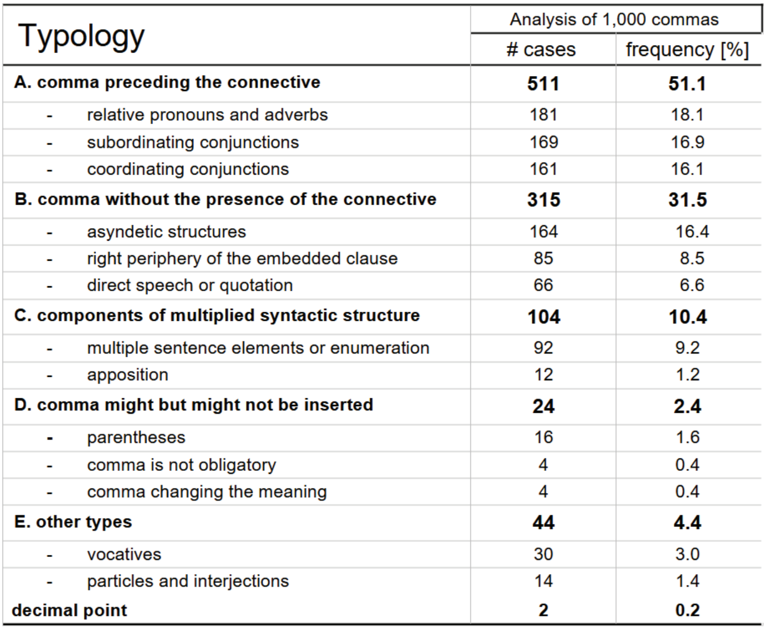
The table below presents a typological classification of 1,000 commas according to their syntactic function and context. Although this sample is not genre-balanced, the distribution confirms earlier findings about the predominance of type A commas, those preceding a connective, which account for 51.1% of all cases. Within this category, relative pronouns and adverbs (18.1%), subordinating conjunctions (16.9%), and coordinating conjunctions (16.1%) are represented in a relatively balanced manner, showing that various clause-linking strategies are equally comma-dependent in Czech syntax.

Type B commas, which appear without an explicit connective, form the second largest group (31.5%). Most notable within this type are asyndetic structures (16.4%), where elements are listed or juxtaposed without a linking word. Additionally, the right periphery of embedded clauses (8.5%) and direct speech or quotation (6.6%) reflect cases where comma placement relies more on syntactic and pragmatic cues than explicit connectives.

Type C commas, used in multiplied syntactic structures (e.g. enumerations and appositions), account for 10.4% of the sample. This relatively moderate proportion underscores the syntactic regularity of comma use in such constructions, with enumerations (9.2%) being more frequent than apposition (1.2%).

Optional commas (Type D) make up a small share (2.4%), typically found in parenthetical structures (1.6%), typically clauses with *"prosím" 'please'* or cases where punctuation can subtly alter the meaning or is simply not obligatory (0.4% each). This highlights the comparatively rare—but linguistically interesting—cases of stylistic or interpretative punctuation.

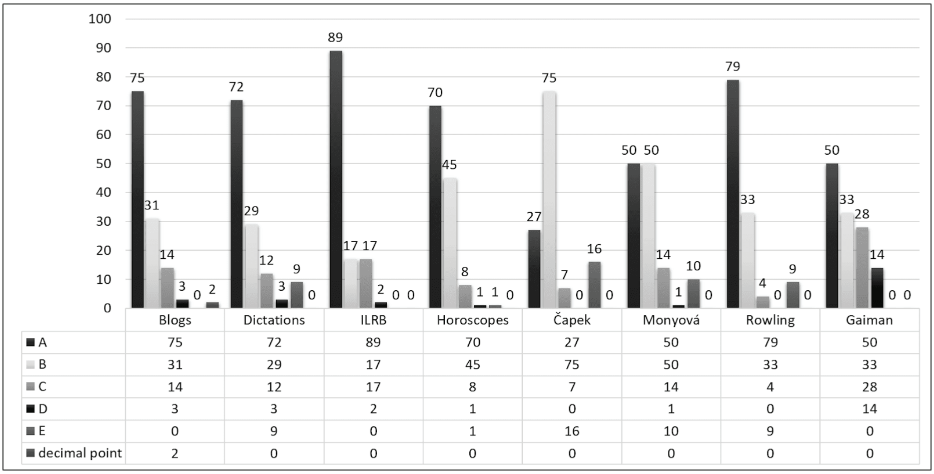
Other types (Type E) include vocatives (3.0%) and particles or interjections (1.4%), together forming 4.4%. These categories often fall outside the core syntactic structure and rely on discourse-level functions. Tab. 3 also includes decimal points (0.2%), which, while not true syntactic commas, are relevant for punctuation processing in computational contexts.



Typology	Analysis of 1,000 commas	
	# cases	frequency [%]
A. comma preceding the connective	**511**	**51.1**
- relative pronouns and adverbs	181	18.1
- subordinating conjunctions	169	16.9
- coordinating conjunctions	161	16.1
B. comma without the presence of the connective	**315**	**31.5**
- asyndetic structures	164	16.4
- right periphery of the embedded clause	85	8.5
- direct speech or quotation	66	6.6
C. components of multiplied syntactic structure	**104**	**10.4**
- multiple sentence elements or enumeration	92	9.2
- apposition	12	1.2
D. comma might but might not be inserted	**24**	**2.4**
- parentheses	16	1.6
- comma is not obligatory	4	0.4
- comma changing the meaning	4	0.4
E. other types	**44**	**4.4**
- vocatives	30	3.0
- particles and interjections	14	1.4
decimal point	**2**	**0.2**

Tab. 3. Observed distribution of 1,000 commas in detail

Different trends in comma distribution can be seen for each sample (see Tab. 4). Type *A. a comma preceding the connective* is prevalent for most texts, except for Čapek (21.6%, 27 commas) and Monyová (40%, 50 commas). Type B *comma* *without the (near) presence of the connective* is most frequent in Čapek (60%, 75 commas), Monyová (40%, 50 commas), in horoscopes this type is more common than in general (36%, 45 commas). However, sentences from the Internet Language Reference Book (2025) contain type B, which is far below average (13.6%, 17 commas). Type C is below average in horoscopes (6.4%, 8 commas), Čapek (5.6%, 7 commas) and Rowling (3.%, 4 commas). Gaiman, on the other hand, contains twice as many type C (22.4% with 28 commas) and more than 4 times as many type D (11.2%, 14 commas) as the average. Surprisingly, besides Gaiman, all samples of fiction texts contain type E, and the dictations contain an over-average of this type (7.2%, 9 commas; it can be assumed that this type was included in the dictations for didactic purposes).



	Blogs	Dictations	ILRB	Horoscopes	Čapek	Monyová	Rowling	Gaiman
■ A	75	72	89	70	27	50	79	50
■ B	31	29	17	45	75	50	33	33
■ C	14	12	17	8	7	14	4	28
■ D	3	3	2	1	0	1	0	14
■ E	0	9	0	1	16	10	9	0
■ decimal point	2	0	0	0	0	0	0	0

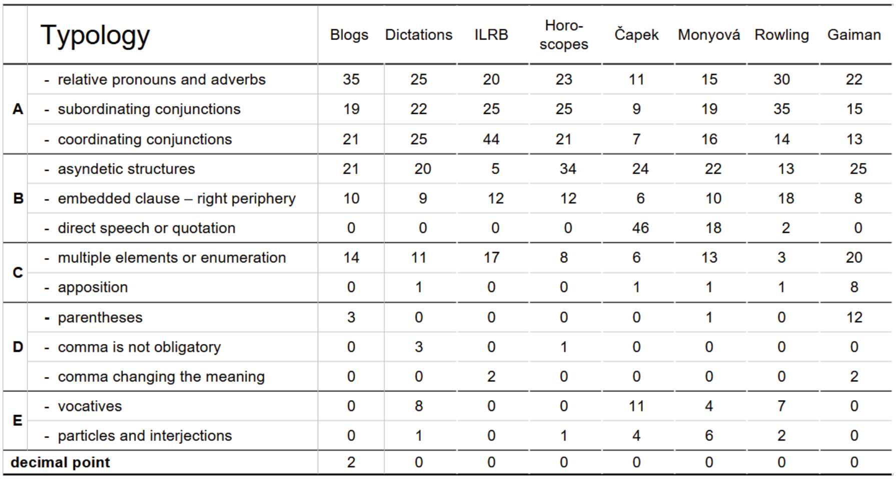
Tab. 4. Comparison of comma distribution in texts of different genres

Comparing the application of each sub-category across the samples provides a distinct perspective, see Tab. 5. The highest number of commas preceding relative clauses appears in blogs (35), while subordinating conjunctions are most frequently used in Rowling (35). Additionally, the ILRB example sentences contain the highest number of coordinating conjunctions (44). Hypotactic comma + connective is prevalent in all texts except ILRB, where the ratio of hypotactic to paratactic comma + connective is balanced (20 + 25 : 44). Čapek and Monyová, in particular, exhibit a significantly lower overall comma frequency throughout Section A.

The highest frequency of asyndetic structures, characterized by the absence of connectives, is found in horoscopes. Rowling's sample contains the greatest number of embedded sentences requiring separation at the right periphery (18). Additionally, more than two-thirds of all commas used around direct speech occur in Čapek (46). Notably, despite being a work of fiction, the Gaiman sample contains no instances of direct speech requiring the use of commas.

A more detailed analysis of type C. *components of multiplied syntactic structure* revealed that all samples contained instances of multiple elements or enumeration, while apposition appeared only marginally. Notably, it was entirely absent in blogs, ILRB, and horoscopes, whereas the Gaiman sample contained eight occurrences, which is relatively high given the small sample size. Similarly, nearly all instances of parentheses were found in Gaiman's text (12), with minimal representation in the other samples.

Commas marking vocatives were present in all fiction texts except for the Gaiman sample, with more than one-third occurring in Čapek's text (11). Additionally, eight instances of vocative commas were identified in dictation texts, where they likely were included deliberately for didactic purposes. The majority of commas surrounding particles and interjections also appeared predominantly in fiction, particularly in Monyová's text (6 instances).



	Typology	Blogs	Dictations	ILRB	Horo-scopes	Čapek	Monyová	Rowling	Gaiman
A	- relative pronouns and adverbs	35	25	20	23	11	15	30	22
	- subordinating conjunctions	19	22	25	25	9	19	35	15
	- coordinating conjunctions	21	25	44	21	7	16	14	13
B	- asyndetic structures	21	20	5	34	24	22	13	25
	- embedded clause – right periphery	10	9	12	12	6	10	18	8
	- direct speech or quotation	0	0	0	0	46	18	2	0
C	- multiple elements or enumeration	14	11	17	8	6	13	3	20
	- apposition	0	1	0	0	1	1	1	8
D	- parentheses	3	0	0	0	0	1	0	12
	- comma is not obligatory	0	3	0	1	0	0	0	0
	- comma changing the meaning	0	0	2	0	0	0	0	2
E	- vocatives	0	8	0	0	11	4	7	0
	- particles and interjections	0	1	0	1	4	6	2	0
decimal point		2	0	0	0	0	0	0	0

Tab. 5. Distribution of 1,000 commas in each subcategory

The sample sizes from individual authors are insufficient to provide conclusive insights into either the author or the overall syntactic structure of the text. However, they offer an indication of prevailing trends in individual texts. Furthermore, variations in the use of different types of commas can influence the effectiveness of automatic comma insertion, among other factors.

3 THE ERROR CORPUS

During the development of the online proofreader Opravidlo.cz, an error corpus was created. It is published in Sketch Engine (Killgariff et al. 2014) and can be searched by CQL queries. We used authentic texts from these domains: Autorevue.cz; Babinet.cz; Doktorka.cz; Hyperinzerce.cz; Seznamka.cz; Super.cz and Zpovednice.cz. Nine sets of hand-corrected sentences were compiled, each containing up to 2,000 sentences. Three annotators annotated each set, i.e. there are three versions of the corrections for each sentence. For a sentence to be included in the corpus, at least 2 of these 3 annotators had to agree on the correction, and the agreement had to be accurate to the letter (or 2 of the 3 had to say that the sentence was correct). With the agreement counted this way, we selected 13,829 sentences on which at least 2/3 agreed. This represents 82.5% (there were 2,939 disagreements). Of these, 4,136 sentences contain at least one tagged error, and the total set contains 6,411 tagged errors.

It should be added that annotators often marked errors in sentences by marking the wrong section in red. This method proved to be problematic in the subsequent evaluation of the sentences since the annotators used different editors (MS Excel, Libre Office, Google Docs). For this reason, these markings were ignored, and the error locations were calculated by comparing the error and the correction. For this reason, the section with an error is always the section between two spaces that contains the error, e.g. for punctuation errors, both the comma and the word before it are marked:

<s> Strašně mě << **vyděsilo | vyděsilo,** >> co se tu někde píše. </s>
'I was horrified by what is written somewhere here.'

On the one hand, this is a rather primitive practice, and it might be worth marking it more precisely, at least in some cases; on the other hand, it is somewhat consistent with the idea that the proofreader tool being developed should instead underline some larger section of text so that the warning is visible.

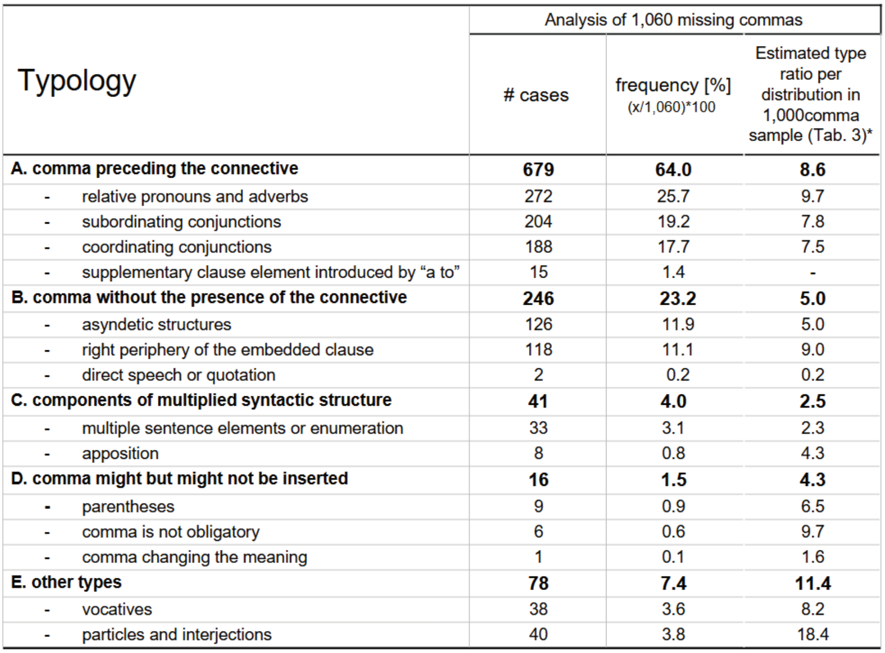
When analyzing the annotated sentences, some systematic problems became apparent, for example, roughly one-tenth of all marked errors are corrections of hyphenation to hyphen, which may be due to the normalization of the texts. The situation is similar to other typographical errors such as quotation marks or other characters.

3.1 Punctuation errors

The *Error corpus* contains a total of **15,516 commas**, including 19 decimal points, leaving **15,497 valid sentence commas**. In total, there are 1,066 corrections where a comma was added after a word (found using CQL: `(<err/> !containing ",")(<corr/> containing ","))`. Of these, **1,060 instances** were analyzed as **missing comma errors**. This means that the writers of these blog texts achieved a **recall (R) of 93.2%** for correctly written sentence commas (R = 14,437 / (14,437 + 1,060)).

Conversely, using the query `(<err/> containing ",")(<corr/> !containing ",")`, there are 247 corrections where a comma was removed. In **218** of these cases, the comma was identified as **redundant**. If we assume that the writers produced 14,437 commas correctly and 218 commas redundantly, their **precision (P)** would be **98.5%** (P = 14,437 / (14,437 + 218)). The relatively high recall and precision can be related to the simpler sentence structure of the blogs where type *A. a comma preceding connective* usually prevails and writers do not usually have problems writing a comma before a connective.

In a more detailed analysis of the missing commas (see the following Tab. 6), two-fifths of the missing commas of type A represent a comma before relative pronouns and adverbs (25.7% of all missing commas, 272 commas), which in Czech usually separate relative clauses from the rest of the sentence. In 204 cases (19.2% of all missing commas), writers omitted a comma before subordinating conjunctions that separate the subordinate clause from the main clause. To a slightly lesser extent, the writers failed to insert a comma before the connective that separates sentences that are formally coordinated (17.7%, 188 commas). If we consider the distribution of commas according to Tab. 3, then writers had slightly more difficulty placing commas before relative pronouns and adverbs (they omitted 9.7% of commas that should be placed before relative pronouns and adverbs). The last group, which cannot be fully classified under the previous three, consists of supplementary clause elements introduced by *"a to" 'and this'* (1.4%, 15 commas).



Typology	Analysis of 1,060 missing commas		
	# cases	frequency [%] (x/1,060)*100	Estimated type ratio per distribution in 1,000comma sample (Tab. 3)*
A. comma preceding the connective	**679**	**64.0**	**8.6**
- relative pronouns and adverbs	272	25.7	9.7
- subordinating conjunctions	204	19.2	7.8
- coordinating conjunctions	188	17.7	7.5
- supplementary clause element introduced by "a to"	15	1.4	-
B. comma without the presence of the connective	**246**	**23.2**	**5.0**
- asyndetic structures	126	11.9	5.0
- right periphery of the embedded clause	118	11.1	9.0
- direct speech or quotation	2	0.2	0.2
C. components of multiplied syntactic structure	**41**	**4.0**	**2.5**
- multiple sentence elements or enumeration	33	3.1	2.3
- apposition	8	0.8	4.3
D. comma might but might not be inserted	**16**	**1.5**	**4.3**
- parentheses	9	0.9	6.5
- comma is not obligatory	6	0.6	9.7
- comma changing the meaning	1	0.1	1.6
E. other types	**78**	**7.4**	**11.4**
- vocatives	38	3.6	8.2
- particles and interjections	40	3.8	18.4

Tab. 6. Observed distribution of the missing 1,060 commas in the Error corpus

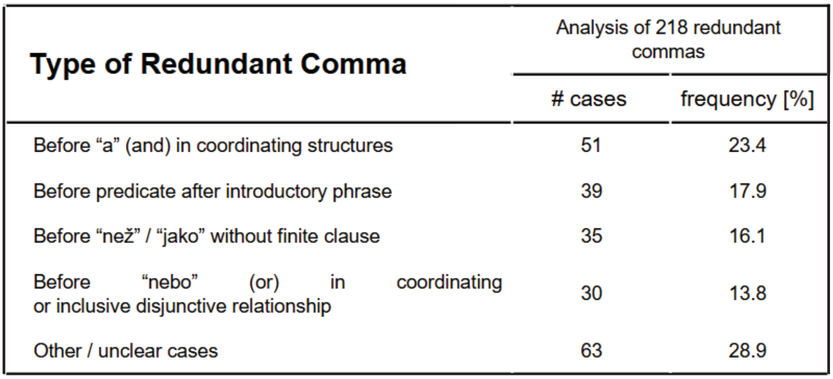
*E.g., the Error corpus is expected to contain approximately 15,500 correctly placed sentence commas. Of these, an estimated 51.1% are of type A (see Tab. 3), which corresponds to about

7,920 commas. Out of this subset, 679 commas — or 8.6% of type A — were omitted by the writers.

More than half of the missing commas of type B are asyndetic structures with no presence of a connective (11.9%, 126 commas). Embedded clauses usually contain a connective on the left side, but are separated asyndetically from the right. In these cases, the comma was missing 118 times (11.1%). This type appears to be more problematic, as we estimate that the writers did not close 9% of the embedded clauses from the right periphery.

The lowest number of errors was recorded for types C (4.0%, 41 commas) and D (1.5%, 16 commas). Almost uniformly, commas were missing around vocative phrases (3.6%, 38 commas) and particles and interjections (3.8%, 40 commas). Writing commas around particles and interjections appears to be the most difficult for writers (they forgot to insert a comma in 18.4% of cases), whereas they only forgot commas around vocatives in 8.2% of cases.

A closer look at the 218 cases of redundant commas (see Tab. 7) reveals several recurring patterns. The most frequent error type (23.4%, 51 instances) was the insertion of a **comma before the conjunction *a* (‘and’) in coordinating structures**, where no comma is required. Surprisingly, the second most common type (17.9%, 39 instances) involved a **comma erroneously placed between the initial phrase and the predicate** — e.g. *“Dům s praktickou dispozicí, nabízí příjemné bydlení”* ‘*A house with a handy layout provides a comfortable living experience’* or *“Z hlediska praktičnosti využití jeho patentů*, má ohromný náskok před Edisonem”* ‘*In terms of the practical use of his patents, he has a significant advantage over Edison’*. These commas may reflect either a prosodic pause (as in spoken language) or influence from English syntactic patterns (e.g. introductory adverbs or phrases).



Type of Redundant Comma	Analysis of 218 redundant commas	
	# cases	frequency [%]
Before “a” (and) in coordinating structures	51	23.4
Before predicate after introductory phrase	39	17.9
Before “než” / “jako” without finite clause	35	16.1
Before “nebo” (or) in coordinating or inclusive disjunctive relationship	30	13.8
Other / unclear cases	63	28.9

Tab. 7. The most common types of redundant commas

Another common issue, observed in 35 instances (16.1%), was a **redundant comma before the conjunctions *než* ‘*than*’ and *jako* ‘*as*’**. In Czech, these conjunctions only require a comma if they are followed by a finite verb clause. Omitting this distinction often leads to unnecessary punctuation. The last more frequent group (13.8%, 30 commas) was the **redundant comma before the conjunction *nebo* *'or'* in a coordinating or inclusive disjunctive relationship** (in Czech, the comma before *nebo* is written when using any of correlative conjunctions such as *ať–nebo* 'whether–or'*, buď–nebo* 'either–or' or in exclusive disjunction). The remaining redundant comma cases were less frequent and often lacked a clear syntactic or prosodic motivation.

4 CONCLUSION

This study offers a comprehensive analysis of comma usage in Czech texts, integrating typological classification with distributional and error analyses. The expanded sample of 1,000 classified commas corroborates previous findings regarding the predominance of commas preceding connectives while also emphasizing genre-specific variation—particularly in fiction, where commas not accompanied by a connective, as well as those marking vocatives and syntactically independent expressions, occur more frequently. In the next phase of research, it would be useful to compare comma distribution with other genres (primarily non-fiction).

The analysis of the Error corpus further reveals systematic patterns in punctuation errors. Writers most commonly omit commas before relative pronouns and subordinating conjunctions or within asyndetic structures, while redundant commas often appear in positions influenced by prosody or interference from English syntax. Despite these challenges, the high overall precision and recall of comma usage in informal web texts suggests a strong intuitive grasp of fundamental rules among Czech writers. These findings not only enhance our understanding of punctuation norms in Czech but also provide valuable feedback for the development of automated comma insertion tools.

ACKNOWLEDGEMENTS

The authors acknowledge that this work was supported by the OSCARS project, funded by the European Commission's Horizon Europe Research and Innovation programme (grant agreement No. 101129751), led by the five Science Clusters: ENVRI, ESCAPE, LS RI, PaNOSC, and SSHOC.

References

Hlaváčková D. et al. (2022). Opravidlo.

Internet Language Reference Book. (2025). Praha: Ústav pro jazyk český AV ČR.

Kilgariff, A. et al. (2014). The Sketch Engine: ten years on. Lexicography. Springer Berlin Heidelberg, 1(1) pp. 7–36. Accessible at: https://dx.doi.org/10.1007/s40607-014-0009-9.

Machura, J. et al. (2022). Automatic Grammar Correction of Commas in Czech Written Texts: Comparative Study. Online. In: P. Sojka – A. Horák – I. Kopeček – K. Pala (eds).: Text, Speech, and Dialogue: 25th International Conference, TSD 2022, Brno, Czech Republic (September 6 – 9, 2022) Proceedings. Cham (CH): Springer, pp. 113–124. Accessible at: https://dx.doi.org/10.1007/978-3-031-16270-1_10.

Machura, J. et al. (2023). Is it Possible to Re-educate RoBERTa? Expert-driven Machine Learning for Punctuation Correction. Jazykovedný časopis, (74)1, pp. 357–368. Accessible at: https://dx.doi.org/10.2478/jazcas-2023-0052.

sciendo
DOI 10.2478/jazcas-2025-0005

THE WORDS ARE FALLING: THE SYNTAGMATIC COMBINATION OF THE LEXEMES *SLOVO* ‘WORD’ AND *PADNOUT/PADAT* ‘TO (BE) FALL(ING)’ AS A BUILDING BLOCK OF OTHER STEREOTYPED EXPRESSIONS AND PHRASEMES

KAMILA MRÁZKOVÁ
Institute of Translations Studies, Faculty of Arts, Charles University, Prague, Czech Republic & Department of Stylistics and Sociolinguistics, Czech Language Institute, Czech Academy of Sciences, Prague, Czech Republic (ORCID: 0000-0002-9382-5305)

MRÁZKOVÁ, Kamila: *The words are falling*: The Syntagmatic Combination of the Lexemes *slovo* ‘word’ and *padnout/padat* ‘to (be) fall(ing)’ as a Building Block of Other Stereotyped Expressions and Phrasemes. Journal of Linguistics, 2025, Vol. 76, No 1, pp. 52 – 62.

Abstract: This study examines the syntagmtic combination of the lexemes *slovo* ‘word’ and *padnout/padat* ‘to (be) fall(ing)’ in contemporary Czech, focusing on how this structure functions as a building block for various multiword expressions and phrasemes. Based on data from the Czech National Corpus (Syn v13), the research analyses the syntactic and lexical development of this syntagm and its semantic functions in different contexts. The findings reveal that this combination appears in several distinct multiword expressions with varying degrees of stability and idiomaticity, rather than representing modifications of a single phraseme. While some expressions serve as verba dicendi that can be freely developed syntactically and lexically, others form the nucleus of established phrasemes, most notably the biblical expression *slova padla na úrodnou půdu* ‘words fell on fertile soil’. The study demonstrates significant differences in the usage patterns of perfective versus imperfective verb forms, and singular versus plural forms of the noun, which correlate with specific phraseological meanings and contexts.

Keywords: multiword expression, phraseme, journalistic style, language corpora

1 INTRODUCTION

The language or style of the mass media, or in the classical concept of functional stylistics, journalistic style, is characterized by a high frequency of automatism, i.e. using established stereotyped expressions or clichés. At the same time, media texts are also the domain of the opposite phenomenon, namely the de-automatization of expression (Havránek 1932, pp. 52–60), which is achieved, among other things, by variations of stereotyped expressions and phrasemes,[1] or by using phrasemes that are unexpected in the given context. In the previous study (Mrázková 2025), I examined

[1] Čermák (2007, p. 82) distinguishes between phrasemes and idioms on the one hand and stereotypical expressions on the other. He considers the distinctive feature of phraseology to be an anomaly, consisting in the impossibility of paradigmatic substitution of one component by another.

the use of the stereotyped multiword expression *ostrá slova padají* ‘sharp words are falling’ and its near-synonymous (in media contexts) lexical variants with adjectives *silný* ‘strong’ and *tvrdý* ‘hard’, comparing traditional journalism and new online media. The research on these multiword expressions revealed, among others, that the syntagm formed by the lexeme *slovo* ‘word’ as subject and the verb *padat* (to be falling, imperfective) as predicate is part of other multiword expressions. Some of these can be understood as amplifications or modifications of the multiword expression *ostrá slova padají* ‘harp words are falling’, but some cannot; their meaning and use are different. The aim of the present study, based on a corpus of mostly media texts, is to find out how the syntagm containing the subject *slovo* ‘word’ and the predicate *padnout* (to fall, perfective) or *padat* (to be falling, imperfective) is complemented and developed both syntactically and lexically, and what is the meaning and use of the multiword expressions formed in this way.

2 ANALYSIS

2.1 Data

The analysis is based on the corpus Syn version 13 (hereafter Syn v13), which, with its 6.4 billion positions, is currently the largest corpus of the Czech National Corpus. Syn v13 is a referential corpus, not a representative one, as it consists of about 80% journalism. The remaining 20% is specialist literature and fiction. Although I am primarily interested in the use of stereotypical multiword expressions in media texts, I have not created a sub-corpus of only media texts to exclude other genres. This is because fiction can provide instances of usage that are atypical (e.g. creative variations of a phraseme), and thus represent a certain contrast to the prevailing media usage.

I have searched in Syn v13 corpus for the co-occurrence of the lemmas *slovo* ‘word’ and *padnout* ‘to fall’, respectively *slovo* and *padat* ‘to be falling’ in the context of two positions. That is, evidence was sought for the use of the verbs in the present tense, the preterite, the future and the infinitive, in different word order, the noun *slovo* could have been in singular or plural and in either case. This search was supplemented with queries combining co-occurrence of lemmas with the grammatical category of number to show the frequency of the conjunction of the searched verbs with the noun in singular or plural. The concordances found that did not match the syntagm in which *slovo* is the subject and *padnout* or *padat* is its predicate were removed manually.

2.2 The syntagm *slovo* ‘word’ + *padnout/padat* ‘fall/be falling’ as verbum dicendi

In utterances like *jelikož byl dotyčný kuřák, okamžitě padlo slovo „rakovina”* ‘since the person in question was a smoker, the word “cancer” immediately has fallen’ or *padají už i slova o tom, kdo koho vlastně živí* ‘even words about who actually feeds whom are falling', the syntagm formed by the noun *slovo* as subject and the verb *padnout/padat* as predicate works as a verbum dicendi. However, it would probably be more accurate to speak of a metalinguistic or metacommunicative expression, as it does not meet the definition of verbum dicendi as formulated by Daneš (1999). According to this definition, verbum dicendi should be "an action verb whose agent participant (the one who in a sentence with an active verb finite will be in the position of the subject) is the agent of the action that consists in the use of language" (Daneš 1999, p. 105), which is not the case here. To avoid unnecessary complications, however, I will continue to refer to it as verbum dicendi.

As multiword expression, "*slovo* + *padnout/padat*" is a lexicalized metaphor whose metaphorical nature is no longer perceived in language. It is one form of ontological metaphor in which "message […] (i.e. what is said)" is conceived as a "thing" (Vaňková 2007, p. 147) and can be treated as such.

Phrases like "the word has fallen" or "words were falling" hide the speaker, conceptualizing speech in a depersonalized way, as a process separate from the producer. But instances such as *padala slova, pochodovaly emoce* 'words were falling, emotions were marching' or *slovo padlo a musí za ním stát* 'the word has fallen and he must stand behind it' are not very common. Unlike the common verba dicendi, "*slovo* + *padnout/padat*" is not much used on its own, without further syntactic and lexical development. The sentence is usually completed by identifying the speaker, the content of the speech (by quotation or paraphrase), or the "words" are characterized by attributes that indicate the content of the speech. In the case of 'words were falling, emotions were marching', the nature of the speech and, in part, the content are inferred from the context, in the latter case, 'the word has fallen' means 'a promise has been made', so it is a different verbum dicendi than above.

Not all uses of the syntagm are verba dicendi, among others because the verb *padnout* has also other meanings than 'to fall', e.g. "to fit" as in *Každé slovo musí padnout, aby se neporušil rytmus* 'Every word has to fit so as not to break the rhythm'. Individual multiword expressions containing the syntagm "*slovo* + *padnout/padat*" also differ in whether they prefer the perfective or imperfective verb and whether the noun is more often in the singular or plural form, as shown by the results of the corpus search presented below.

2.3 Perfective vs. imperfective, singular vs. plural

The absolute frequency of the syntagm "*slovo* + *padnout*" (with perfective verb) in the Syn v13 corpus in context of two positions is 7,932, its relative frequency being 1.24 i.p.m. By contrast, the syntagm "*slovo* + *padat*" (with imperfective verb), also within a context of two positions, has an absolute frequency of 3,888, which equates to a relative frequency of 0.61 i.p.m. This more or less corresponds to the proportional representation of each verb: 95 i.p.m. for the perfective *padnout* 'to fall' and 38.74 i.p.m. for the imperfective *padat* 'to be falling'. As a predicate of the subject *slovo*, the imperfective *padat* is slightly more frequent (33% of total usage) compared to its overall frequency (28% of total usage).

Fig. 1 shows the frequency of combinations of the perfective and imperfective verbs with the singular and plural, respectively. Of these combinations, the imperfective verb with the singular noun has the lowest frequency, 0.01 i.p.m., and accounts for 0.8% of all uses of the syntagm "*slovo* + *padnout/padat*". In contrast, the perfective verb *padnout* with the singular of *slovo* is the most frequent, with a relative frequency of 0.95 i.p.m. and a 54.6% share of the usage of all the variants of the syntagm. The combination of the imperfective verb *padat* and the plural noun *slova* has a frequency of 0.51 i.p.m. in Syn v13 and accounts for 31% of the usage of all variants, while the frequency of the perfective verb and its subject in plural is about half as low, exactly 0.27 i.p.m. and accounts for 15.4% of the usage of all variants of the given syntagm.

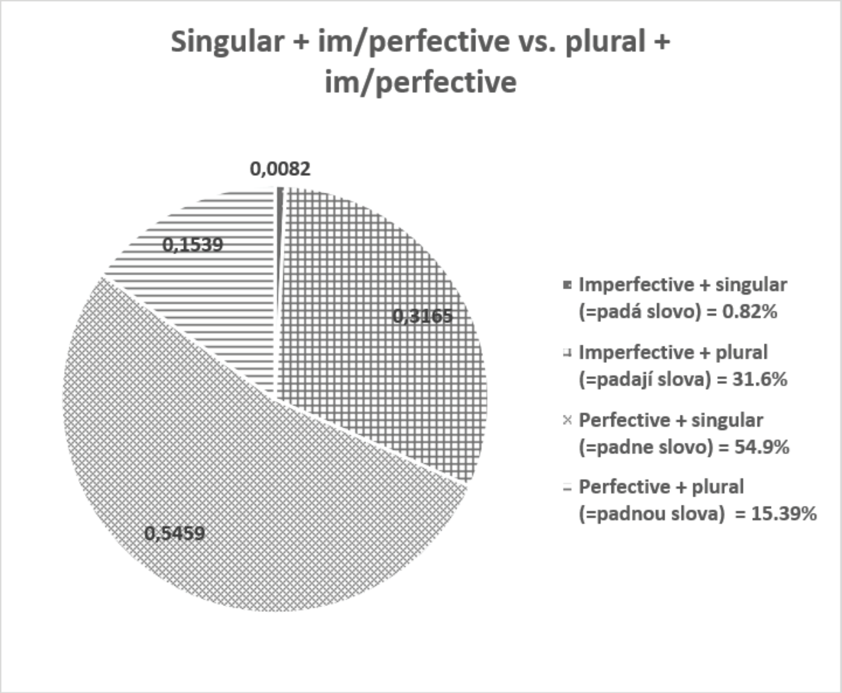


Fig. 1. Frequency of singular + im/perfective vs. plural + im/perfective

The reasons for the above-mentioned differences in the frequency of individual combinations lie in the different meanings of perfective and imperfective verbs, and thus in their different involvement in multiword expressions and phrasemes. Below I will try to show the most significant cases of these differences in usage.

2.4 Reported speech

The results of the corpus search confirm that the syntagm formed by the subject *slovo* 'word' and the predicate *padnout/padat* 'to (be) fall(ing)' in both the perfective and imperfective variants is overwhelmingly used as a verbum dicendi, i.e. to quote or report someone's speech. The perfective and imperfective verbs differ, however, in the ways in which a speech is usually reported. The perfective verb *padnout* is most often used to report on a speech that did not take place, serving as a variant of *o tom nepadlo ani/jediné slovo* 'not a single/no one word was spoken about'. This usage (Ex. 1) accounts for almost half of all occurrences of the syntagm *slovo* + *padnout*, i.e. with the perfective verb (see Fig. 2). In contrast, analogous evidence of negative constructions with the imperfective such as in Ex. 2 is a bare minimum; in the Syn v13 there are 15 such instances in total which, given the size of the corpus, means a relative frequency of zero.

(1) *O presumpci neviny nepadne ani* slovo.
'No one word falls of the presumption of innocence'.

(2) *O tom, že by mohly mít nějaké problémy také banky, nepadá ani slovo.*
'Not a word is falling about the fact that banks might have some problems too'.

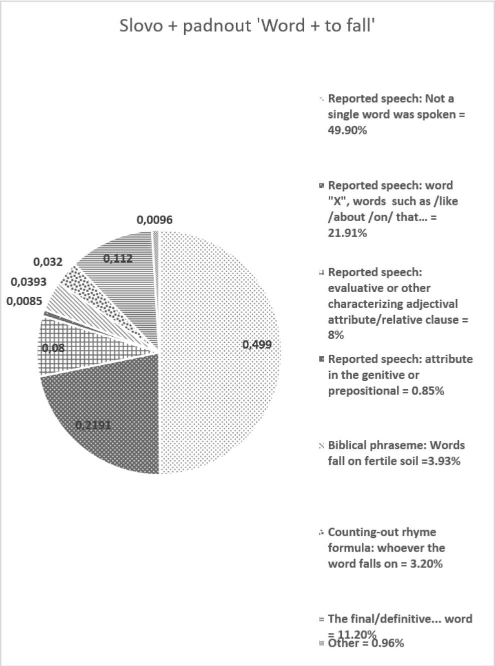


Fig. 2. Frequency of individual types of use of "*slovo* + *padnout*" 'word(s) + to fall'

Positive constructions with a singular subject and an imperfective verb, expressing the repeated use of a single word such as in Ex. 3, are slightly more frequent; the overall frequency of both positive and negative constructions of the imperfective with its subject in singular is 0.01 i.p.m. Quotation of a specific word, words or sentences, sometimes in quotation marks (Ex. 4) or using the conjunction *jako* 'as', and indirect quotations, i.e. in a dependent clause with the conjunctions *že* 'that' (Ex. 5), *aby* 'to', *jestli/zda* 'if/whether', etc. belong to the most frequent uses of the syntagm *slovo* + *padat* as a verbum dicendi. The frequent thematizations of speech content using the preposition *o* 'about', such as in the expression 'word(s) fall(s) about' (Ex. 6), are also included in this category. This group accounts for almost 22% of *slovo* + *padnout* (perfective verb) usage and 38% of *slovo* + *padat* (imperfective verb) usage. Thus, imperfective verb is relatively almost twice as frequent for this type of speech reproduction, although the absolute number of combinations with the perfective is slightly more prevalent (1 737 perfectives vs. 1 491 imperfectives).

(3) *Z úst jedněch i druhých padá nejčastěji slovo "voliči".*
'The word 'voters' is falling most often from the mouths of both.'

(4) *Přesně padala ta slova* "*Proč podporovat soukromou akci?*"
'The exact words were falling "Why support a private event?"'

(5) *Přitom loni po velkém zátahu padala slova, že drogová scéna v kraji dostala velkou ránu.*
'Yet last year, after the big bust, the words were falling that the drug scene in the region had taken a big hit.'

(6) *Mezi funkcionáři občas padnou slova o koupení zápasu* […]
'Words about buying the game occasionally fall amongst the officials […]'

Likewise, when the subject *slovo* is complemented by evaluative or other specific attributes—either adjectives in grammatical agreement with the subject or substantives in the genitive or prepositional case—it can also be regarded as a form of reported speech or a report on speech. Attributes such as *ostrý* 'sharp', *sprostý* 'vulgar', *velký* 'big', etc., or genitive constructions *slova díků* 'words of thanks' or *slova chvály i opovržení* 'words of praise and words of scorn', give a more definite or vague information about what was said. The same is true of relative clauses complementing the subject *slovo.* In total, this group presents 53% of uses with the imperfective (see Fig. 3).

A comparison of the frequencies of collocations with the adjectival attribute *ostrý* 'sharp' shows the clear predominance of the imperfective, even in terms of absolute numbers. While the combination of the syntagm *ostrý* + *slovo* 'sharp + word' with the imperfective has 742 occurrences in Syn v13 (Ex. 8), which is 19% of all uses of the imperfective in syntagm *slovo* + *padat*, there are only 164 combinations with the perfective (Ex. 7), which corresponds to 2% of the total use of this syntagm with the perfective verb *padnout.*

(7) *Když se zkouší, tak samozřejmě padnou ostřejší slova, jsme v časovém presu* [...]
‘When you’re rehearsing, of course the sharper words fall, we are in a time crunch […]’

(8) *Na adresu představitelů města padala ostrá slova.*
‘Sharp words were falling down on the town officials.’

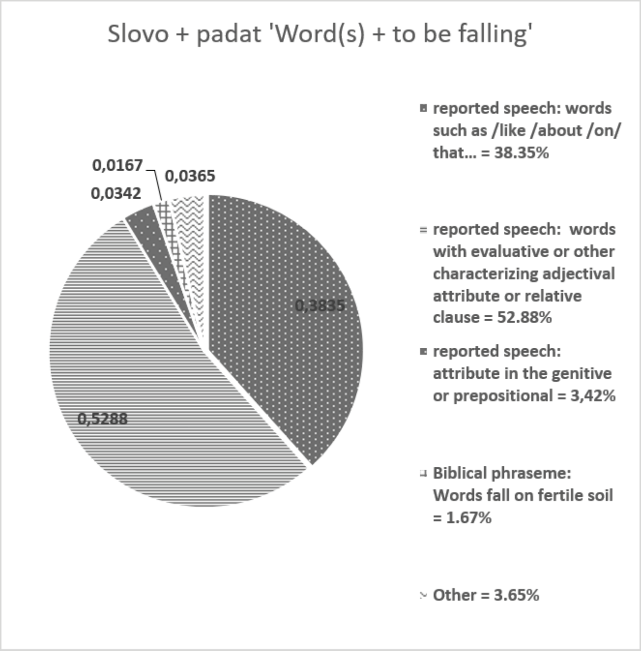


Fig. 3. Frequency of individual types of use of “*slovo*+ *padat*” ‘word + to be falling’

2.5 “Words fell on fertile soil”

The syntagm *slovo* + *padat* is also part of larger established multiword expressions, which refer to speech or talk as well, but it is not a mere development of the above verba dicendi but a separate phraseme. It is especially the biblical phraseme (*nějaká/něčí*) *slova padla na úrodnou půdu* ‘(some/someone’s) words fell on fertile soil’ and its various modifications and variants. The phraseme is related to the parable of the sower in the Gospel of St. Matthew, in which the working of God’s word is compared to seeds falling on different places, on a path, on a rock or also on fertile soil (Matthew 13). The subject *slovo* is nearly always in the plural and the perfective verb prevails over the imperfective, both in absolute number of occurrences (312 vs. 65) and relatively, given the ratio of the variant in relation to the overall use of the syntagm *slovo + padnout/padat* with perfective (3.9%, Ex. 9, 11, 12) and imperfective (1.7%, Ex. 10, 13), respectively.

(9) *Jeho slova padla na úrodnou půdu a po přestávce vyběhlo na hřiště úplně jiné mužstvo*.
‘His words fell on fertile soil, and after the break a completely different team ran onto the field.’

(10) *Jak se však zdá, půda, na kterou dobře míněná slova padala, byla stejně kamenitá jako Kalifornie sama.*
‘But apparently the soil on which the well-meant words were falling was as rocky as California itself.’

(11) *Na tuto půdu padla slova misionářů, jejich odhodlanost a obětavost.*
‘On this soil fell the words of the missionaries, their determination and dedication.’

(12) *Na úrodnou půdu ale jeho slova nepadla.*
‘But his words did not fall on fertile soil.’

(13) *Vltavský dokument* […] *je otevřenou výpovědí moudré ženy, jejíž slova padají do půdy církve těhotné změnou* […]
‘Vltava’s documentary […] is the open testimony of a wise woman whose words fall on the soil of the Church pregnant with change […]’

The nucleus (Jelínek, Kopřivová, Petkevič and Skoumalová 2018) of the biblical phraseme consists of the lexemes *slovo* ‘word’, *padnout/padat* ‘to (be) fall(ing)’, *na* ‘on’, *půda* ‘soil’, and overwhelmingly, *úrodný* ‘fertile’. *Slovo*, nearly always in the plural, is the subject of the verb *padat/padnout*, *půda* is in accusative case with the preposition *na*, and *úrodný* is the adjectival attribute of the noun *půda*. The lexical variations allow the substitution of *půda* ‘soil’ for *zem* ‘earth’ and the substitution of *úrodný* ‘fertile’: here the lexical variations are more varied: *živný* ‘nurturing’, *nakypřený* ‘ploughed’, *vděčný* ‘grateful’ or *těhotný změnou* ‘pregnant with change’. The phraseme is of course also used as a negative statement, and this can be achieved not only by negating the verb in the predicate (Ex. 12), but by complementing the noun ‘soil’ with adjectives such as *hluchý* ‘deaf’ (Ex. 14), *neúrodný* ‘barren’ or *tvrdý* ‘hard’ as well. A similar meaning can be expressed using an attribute in the genitive case, as in Ex. 15.

(14) *Tentokráte však slova padla na hluchou půdu.*
'This time, however, the words fell on deaf soil.'

(15) *Začasté padají má slova na půdu totálního nepochopení a nezájmu.*
'Often my words fall on the soil of total misunderstanding and disinterest.'

Syntactically, this phraseme is further developed by adding to the subject *slova* some attribute referring to the author of the "words" or specifying the "words" in some way. Most often this attribute is a possessive pronoun (Ex. 9, 12, 15) or a possessive adjective; the author of the words can also be referred to by attributes in the genitive case (Ex. 11). Among the adjectival attributes, lexemes such as *ostrý* 'sharp' appear in the evidence from media texts; however, they are not nearly as common as in the metalinguistic variants of 'sharp/strong/hard etc. words are falling'. If the subject *slovo* is not syntactically complemented by any attribute, it is because it has already been specified in the previous context.

Examples 10, 13 or 15 show that this biblical phraseme is indeed used in a variety of ways in texts and its individual elements often appear in the syntactic structure in distant positions, as in Ex. 10. On the other hand, most of the instances are rather stereotypical and come mainly from sports journalism: it is the words of football or hockey coaches that most often fall or fail to fall on fertile soil (Ex. 9).

2.6 "The final word has fallen"

Another notable use of the syntagm *slovo* + *fall*, which is close to verbum dicendi but has a somewhat shifted meaning, is *konečné/poslední/definitivní etc. slovo padlo* 'the final/last/definitive etc. word has fallen'. Its meaning is not just to report on the speech but means 'a final decision has been made'. Similarly to *nepadlo ani slovo* 'not a word has fallen', this established multiword expression is exclusively associated with the perfective verb *padnout* and accounts for 11.2% of the total usage of the syntagm *slovo* + *padnout*. Ex. 16 is a typical example.

(16) *Konečné slovo padne, až bude uzavřeno i poslední čtvrtletí minulého roku.*
'The final word will be given (= fall) when the last quarter of the previous year is closed.'

2.7 "Whoever the word falls on..."

The last specific multiword expression containing the syntagm *slovo* + *padnout* that will be mentioned here is the penultimate verse of the children's counting-out rhyme *na koho to slovo padne* 'whoever the word falls on' / *ten musí jít z kola ven* 'has to go out of the round'. In the corpus evidence, the last verse, "has to go out of the round," is usually just inferred and the usual meaning of this multiword expression here is "who will be chosen" (Ex. 17). The form of this phraseme is fixed, the subject is always in the singular and the verb is always perfective; the use of this phraseme accounts for almost 4% of the total use of the syntagm "*slovo* + *padnout*", i.e. the variant with perfective (see Fig. 2).[2]

(17) *Do dalších parlamentních voleb nebude jasné, na koho to slovo padne, a o to bude politický diskurz bouřlivější.*
'Until the next parliamentary elections, it will not be clear who the word will fall on, and the political discourse will be all the stormier.'

3 CONCLUSION

The analysis confirmed that syntagm *slovo* + *padnout/padat* 'word + to (be) fall(ing)' is a segment of different types of multiword expressions with different degrees of stability and idiomaticity; it cannot be said that these are variations or modifications of one phraseme. On the one hand, there are formulations in which the syntagm serves as a verbum dicendi and can be complemented and developed both lexically and syntactically according to what the speaker is referring to, without being limited by the idiomaticity of the expression. The figurative meaning of these multiword expressions is not perceived, they are very stereotyped and are typical of journalism; they are an example of automatism that allows easy production and reception of the text. On the other hand, the syntagm *slovo* + *padat/padnout* is the nucleus of multiword expressions that are clearly phrasemes, the most significant of them being biblical phraseme *slova padla na úrodnou půdu* 'the words fell on fertile soil'. Its idiomaticity and fixedness is manifested by the preference for the plural noun and the perfective verb, while its integration into the context is often quite creative.

ACKNOWLEDGEMENTS

The research has been supported by the Czech Science Foundation, project *Czech phraseology and changes in its use in contemporary and genre contexts*, no. 24-10254S.

References

Čermák, F. (2007). Frazeologie a idiomatika česká a obecná. Czech and general phraseology. Praha: Karolinum, 718 p.

Daneš, F. (1999). Verba dicendi a výpovědní funkce. In Jazyk a text II. Výbor z lingvistického díla Františka Daneše. Praha, pp. 105–114.

[2] The corpus evidence is quite heavily influenced by citations of the titles of two plays so named, one by Alena Vostrá and the other by a Hungarian playwright Gábor Görgey.

Havránek, B. (1932). Úkoly spisovného jazyka a jeho kultura. In: B. Havránek – M. Weingart (eds.): Spisovná čeština a jazyková kultura. Praha: Melantrich, pp. 32–84.

Jelínek, T., Kopřivová, M., Petkevič, V., and Skoumalová, H. (2018). Variabilita českých frazémů v úzu. Časopis pro moderní filologii 100(2), pp. 151–175.

Křen, M., Cvrček, V., Čapka, T., Hnátková, M., Jelínek, T., Kocek, J., Kovářiková, D., Křivan, J., Milička, J., Petkevič, V., Skoumalová, H., Šindlerová, J., and Škrabal, M. (2024). Corpus SYN, version 13 from 27/12/2024. Ústav Českého národního korpusu FF UK, Praha 2024. Accessible at: https://www.korpus.cz.

Mrázková, K. (2025). „Ostrá slova padají z obou stran." Užívání metakomunikačních frazémů v současných česky psaných online médiích. Časopis pro moderní filologii 107(2), pp. 141–155.

Vaňková, I. (2007). Nádoba plná řeči. Praha: Karolinum, 312 p.

sciendo
DOI 10.2478/jazcas-2025-0006

ADVERBS AND PARTICLES: PART-OF-SPEECH HOMONYMY IN CORPUS DATA AND MEDIA DISCOURSE

KRISTÍNA PIATKOVÁ[1] – MÁRIA STANKOVÁ[2]
[1]Department of Language Cultivation and Terminology, Ľudovít Štúr Institute of Linguistics, Slovak Academy of Sciences, Bratislava, Slovakia
(ORCID: 0009-0000-3128-2141)
[2]Department of Journalism, Faculty of Arts, Comenius University, Bratislava, Slovakia (ORCID: 0000-0003-4450-9370)

PIATKOVÁ, Kristína – STANKOVÁ, Mária: Adverbs and Particles: Part-of-speech Homonymy in Corpus Data and Media Discourse. Journal of Linguistics, 2025, Vol. 76, No 1, pp. 63 – 74.

Abstract: The study addresses the part-of-speech homonymy between adverbs and particles in Slovak, with a focus on linguistic data from corpora and media discourse. Four lexemes were analysed: *absolútne*, *konečne*, *očividne* and *prirodzene*, examining differences in their frequency, collocations, and contextual functions. The results revealed that lexical and pragmatic factors are crucial for distinguishing adverbs from particles, and that the meaning of the context plays a significant role in their interpretation. The study contributes to the debate on the exact criteria for distinguishing parts of speech and highlights the importance of a comprehensive approach in corpus and media linguistics.

Keywords: adverbs, particles, part-of-speech homonymy, corpus linguistics, media

1 INTRODUCTION

The study explores the adverb-particle relationship using the Slovak National Corpus[1] and media texts on words that are classified as both adverbs and particles: *absolútne* 'absolutely', *konečne* 'finally', *očividne* 'obviously', and *prirodzene* 'naturally'. It aims to examine the differences and similarities between four words in the corpus and media discourse, focusing on part-of-speech homonymy, adverb collocations, and the role of context. It also sought to identify inconsistencies in distinguishing adverbs from particles and propose criteria for their differentiation across the corpus, dictionaries, and media discourse.

In monographs on Slovak morphology, the distinction between adverbs and particles is described in functional, syntactic and semantic terms (Dvonč 1966, pp. 27–32; Oravec, Bajzíková and Furdík 1984, pp. 13–17; Závodný 2016, pp. 61, 142). Adverbs are one of major word classes with both lexical and grammatical meaning

[1] Omnia Slovaca IV Maior Beta corpus (23.01). Status as of 07/03/2025.

and can function as sentence constituents, while particles lack these properties, therefore, are classified as grammatical words. According to M. Šimková, this classification is problematic, and she considers particles semantic and textually-communicative words (2003, p. 234). J. Šindlerová and B. Štěpánková confirm a similar state in Czech linguistics, arguing that particles, compared to adverbs, have a partially weakened or modal meaning (2021, p. 444). The weakening of meaning and the subsequent reduction of syntagmatic combinability of adverbs lead to their secondary transformation into particles (Oravec, Bajzíková and Furdík 1984, p. 176). M. Ološtiak refers to the transition of adverbs into particles as *deadverbial particulization* (2017, p. 70).

The specific relationship between adverbs and particles, as well as the problematic identification of word classes, is mentioned in academic morphology (MSJ; Dvonč 1966, pp. 804–805) and in the textbook *SSJ: Morfológia* (Oravec, Bajzíková and Furdík 1984, pp. 201–202).

In the study of part-of-speech homonymy, J. Kačala (1984) examined the word *ťažko* 'hardly' and its adverb/particle classification based on substitution and its relationship to the verb. M. Šimková (2002) analysed the category of state, while J. Šindlerová and B. Štěpánková focused on "intensifiers" (2021).

According to academic literature, we have outlined potential criteria for differentiating adverbs from particles:

Area	**Property/Function**	**Adv.**	**Part.**
Morphosyntax	**Sentence constituent**	●	
	Predicate	●	
	Grammatical meaning	●	●
	Core of the utterance	●	
	Gradability	●	
	Positional flexibility		●
	Formal distinction		●
	Relation to the verb	●	
Pragmatics	**Context**		●
	Intention/Attitude		●
Semantics	**Full meaning**	●	
	Communicative potential	●	●

Tab. 1. Criteria for part-of-speech classification

2 METHODOLOGY

Our study focuses on comparing corpus data with examples from media to characterize the adverb-particle part-of-speech homonymy. From the corpus database, we randomly selected 100 instances of the selected words: *absolútne*, *konečne*, *očividne* and *prirodzene*, separately classified as adverbs and particles, including occurrences in media discourse. Subsequently, this sample was manually verified in the context of information from the dictionaries: *Krátky slovník slovenského jazyka* (KSSJ 2020), *Ortograficko-gramatický slovník slovenčiny* (OGSS 2022), *Pravidlá slovenského jazyka* (PSP 2013), *Slovník slovenského jazyka* (SSJ 1959–1968), *Slovník súčasného slovenského jazyka* (SSSJ 2006–2021), as well as MSJ.

Then, we examined the collocational profiles of the adverbs, focusing on their most frequent combinations with verbs, adjectives and other adverbs. The third phase of the research was aimed at comparing corpus data with occurrences in media texts[2], from which we collected a total of 40 instances. The examples from journalistic texts also serve to demonstrate the importance of context in differentiating between adverbs and particles.

3 FINDINGS

3.1 Corpus data

Based on the corpus search results, we present the relative frequency of the analysed expressions as adverbs and particles. The results indicate that while the words *absolútne* and *očividne* occur notably more frequently as adverbs, the words *konečne* and *prirodzene* show a more balanced distribution between the two word classes.

	Adverb	**Particle**
absolútne (‘absolutely’)	285,389	2,210
konečne (‘finally’)	370,476	357,843
očividne (‘obviously’)	94,260	200
prirodzene (‘naturally’)	117,415	124,769

Tab. 2. Relative frequency

[2] The analysed data are extracted from commentaries published on the websites of *Denník N* and *SME*. We selected five examples of each examined word from the published commentaries in both media (as of 10/03/2025).

The deadjectival adverbs *absolútne*, *konečne*, *očividne*, and *prirodzene* were analysed, primarily using lexicographical data.

Absolútne

The word *absolútne* can be found in both KSSJ and OGSS as an adverb. In SSSJ (A–G, 2006), the adverb has three meanings: "with unlimited power, as an autocrat, self-ruler"; "independently of any conditions, unconditionally" and as a colloquial word meaning "completely, fully, entirely". It is a qualitative adverb expressing degree (MSJ 1966, p. 603). According to Šikra (1991), it expresses the maximum possible degree. In the media, we encounter the following usage:

Školstvo je ***a.*** *kľúčové*. 'Education is absolutely crucial.'
Krajina je ***a.*** *závislá od ruského plynu a ropy*. 'The country is absolutely dependent on Russian gas and oil.'
Žiadny zákon však nie je ***a.*** *dokonalý*. 'No law is absolutely perfect.'

The lexical meaning of degree is found in all the listed examples. The lexeme *absolútne* is synonymous with terms like *úplne* 'completely' and *celkom* 'entirely', but it is considered a colloquial lexical unit, which may be related to the gradual determinologization of the words *absolútno* 'the absolute' and *absolútny* 'absolute'. In journalism, the adverb *absolútne* often appears in clichés with adjectives, as shown by corpus data:

A. *kľúčovou zložkou treného cesta je teplota*. 'An absolutely key component of choux pastry is temperature.'
Ale na podobu týchto pravidiel už nemá mať ***a.*** *nijaký vplyv*. 'But it should have absolutely no influence on the form of these rules anymore.'

The word *absolútne* also functions as an emphatic focusing particle, which "emphasizes the extreme degree of the following specification, meaning 'at all' or 'by no means'" (SSSJ A–G 2006), although in KSSJ and PSP, it is classified solely as an adverb. Focusing particles resemble adverbs more than introductory particles, making classification harder for users. They modify statements, add expressiveness, emphasise key parts and allow the author to express a subjective attitude or emotion:

Nepredviedli sme ***a.*** *nič*. 'We didn't show absolutely anything.'
Spartak Trnava získal titul ***a.*** *čestne a korektne*. 'Spartak Trnava won the title absolutely fairly and correctly.'
„A zrazu zisťujeme, že z toho, čo povedal pán Uhliarik, neplatí ***a.*** *nič," konštatoval dnes Fico*. '"And suddenly, we realize that nothing Mr. Uhliarik said is absolutely true," Fico stated today.'

Očividne

According to the interpretation in MSJ (1966, p. 603), the qualitative adverb *očividne* primarily conveys the meaning of manner and expresses degree indirectly, often in an expressive or hyperbolic way. According to SSSJ (2021), it indicates degree and means "in a visible, distinct and clear manner". The particle *očividne* "expresses conviction about a certain assumption" (SSSJ) and is synonymous with the words *zjavne* 'manifestly' and *evidentne* 'evidently'). In KSSJ (2020, p. 419), the adverb is illustrated with the example *o. chradol* 'o. waste away', while the particle is explained using the equivalents *navidomoči* 'visibly' and *zjavne*. The interpretation solely through synonyms may be confusing for an ordinary language user.[3]

Based on *corpus* data, the word is used in the media as:

a) adverb:

*Verejnosť tomu **o.** rozumie a prejavilo sa to aj na eurovoľbách*. 'The public obviously understands this, and it was reflected in the European elections.'
*Nemci zo severu sú **o.** milovníci architektúry so symbolikou*. 'Germans from the north are obviously lovers of architecture with symbolism.'
*Podľa novinárov bol **o.** znechutený*. 'According to journalists, he was obviously disgusted.'

b) particle:

*Dôvod na radosť mali **o.** obe finálové súperky*. 'Both finalists obviously had reason to be happy.'
*Niektorým stuhol **o.** úsmev na tvári, ale nikto sa nesťažoval*. 'Some obviously had their smiles frozen on their faces, but no one complained.'
*Medzi jeho voličov patrili **o.** i organizátori petície na podporu rodiny*. 'His voters obviously included the organizers of the petition in support of the family.'

The adverb *očividne* characterizes the predicate in terms of its meaning (*Verejnosť tomu **o.** rozumie*. – 'The public obviously understands this.'), it only modifies the meaning of the verb and forms a syntagm with the superior sentence element. The presence of the adverb can also be verified by asking *How*? However, the situation changes if we modify the sentence: *O., verejnosť tomu rozumie.* 'Obviously, the public understands this.' Here, *očividne* is not syntactically related to any sentence element. Instead, it functions as an independent utterance referring to a broader context, from a word-class perspective, it is classified as a particle.[4]

[3] This statement is based on questions addressed to the language advisory service of Ľ. Štúr Institute of Linguistics of the Slovak Academy of Sciences.

[4] For this type of word, which expresses an attitude toward the entire statement, the term sentence adverb or (semi)particle is used (compare: Šikra, 1991; Uhlířová, 1979).

Prirodzene

The qualitative adverb *prirodzene* is processed in KSSJ (2020, p. 583) in the form of examples. Secondarily, it holds the function of a delimitative evaluative particle expressing certainty or assurance. According to MSJ (1966, p. 788), it is synonymous with words such as *samozrejme* 'of course', *isteže* 'certainly' or *pochopiteľne* 'understandably'. In SSJ (1963), the adverb *prirodzene* means "in a natural, unforced, unconstrained way", while the particle is characterised by synonyms such as *pravda* 'indeed', *pravdaže* 'of course' and *zaiste* 'certainly', with the qualifier "colloquial". As a particle, it is typically separated by commas in a statement, which emphasizes its subjectivising potential and contextuality.

The position of an adverb in a sentence can serve as a clue for distinguishing the part of speech, since its position next to the verb is neutral. In contrast, the position next to an adjective or another adverb is syntactically indicated by the adverb's close placement to the superior sentence element. Let us examine the following two examples:

V nasledujúcich rokoch ceny ***p.*** *rástli*. 'In the following years, prices naturally increased.'
Súčasťou servisu je pre nás ***p.*** *aj pomoc klientom s likvidáciou poistných udalostí*. 'As part of our service, we naturally also assist clients with the settlement of insurance claims.'

In the first sentence, the adverb is linked to the verb *rásť* 'increase'. In contrast, in the second sentence, there is no such connection between *prirodzene* and the verb. Another criterion is whether the adverb can be further syntactically expanded, e.g. *veľmi* ***p.*** *rástli* 'they grew very naturally'. A similar expansion of the adverb can also be found in corpus examples:

Geostratégom však vzhľadom na politiku i ekonomiku celkom ***p.*** *prichádza na myseľ rok 1989*. 'To geostrategists, however, the year 1989 quite naturally comes to mind in the context of politics and economics.'
Ak politici dostanú právomoc regulovať médiá, budú úplne ***p.*** *v pokušení zneužiť svoje právomoci*. 'If politicians are given the power to regulate the media, they will be completely naturally tempted to abuse their authority.'

Konečne

The lexeme *konečne* is defined as an adverb meaning "after a certain period of time" (SSSJ 2011), with the addition of "often with a sense of satisfaction from the speaker", thus approaching the meaning of a particle. The particle *konečne* is also characterized in KSSJ and SSSJ as a statement of a certain fact, a summary of a certain finding, a synonym for the particles *napokon* 'eventually' and *koniec* *koncov* ‘altogether’, and as an expression of satisfaction from the completed action. As a delimitative explanatory particle, it is synonymous with expressions like *napokon* and *naostatok* ‘ultimately’ (MSJ, p. 775).

*Podnikatelia a živnostníci by mali **k.** dostať servis na úrovni*. ‘Entrepreneurs and business owners should finally receive proper service.’
*Rokovania o budúcnosti Le Monde prichádzajú v čase, keď sa noviny **k.** dostali do zisku.* ‘Negotiations about the future of Le Monde come at a time when the newspaper has finally become profitable.’
*Grécka finančná kríza stále spôsobuje ľuďom bolesť a marí ich nádeje a sny, aj keď krajina vlani v auguste, po ôsmich rokoch, **k.** vystúpila zo záchranného programu*. ‘The Greek financial crisis still causes pain for people and shatters their hopes and dreams, even though the country finally emerged from the bailout program last August after eight years.’

In the first case, both parts of speech can be considered, but in the other two, the word is better treated as an adverb due to its reference to an adverbial time expression or temporal subordinate clause, which *konečne* often relates to. Corpus examples of the particle confirm its explanatory character:

*Možno nastal **k.** čas na verejnú diskusiu aj na tému, či chceme, aby spoločnosť z pozadia riadili oligarchovia*. ‘Perhaps the time has finally come for a public discussion on whether we want oligarchs to run society from behind the scenes.’
*Siahol po jazyku, ktorému jeho volič **k.** rozumie*. ‘He adopted a language that his voters finally understand.’
*Ministerka školstva a vláda **k.** priznali, že pre situáciu na ministerstve školstva vedci prichádzajú o ďalšie milióny eur*. ‘The Minister of Education and the government have finally acknowledged that due to the situation at the Ministry of Education, scientists are losing millions of euros.’

3.2 Collocations

In the next step, we identified the most frequent collocations.[5]

Collocations	Number of occurrences	Collocations	Number of occurrences
a. súhlasiť (‘a. agree’)	1,168	*a. nesúhlasiť* (‘a. disagree’)	789
a. chápať (‘a. understand’)	200ň	*a. nechápať* (‘a. not understand’)	848

[5] Given the scope of the work, we select only some of the examples that were relevant in journalistic texts.

a. rozumieť (‘a. comprehend’)	134	*a. nerozumieť* (‘a. not comprehend’)	927
a. zaujímať (sa) (‘be a. interested’)	18	*a. nezaujímať (sa)* (‘be a. disinterested’)	1,353
a. zaujímavý (‘a. interesting’)	49	*a. nezaujímavý* (‘a. uninteresting’)	410

Tab. 3. Collocations with the adv. *absolútne*

The adverb *absolútne* is often used in journalistic texts with adjectives and verbs carrying a negative meaning, formally expressed with the prefix ne-. Therefore, we also examined the usage of their “non-negative” neutral forms, and it turned out that the combination with the verb *súhlasiť* ‘agree’ (1,168 occurrences) predominates over the negative form *nesúhlasiť* (789 occurrences).

We again observe the tendency for redundant evaluative or intensifying expressions in media discourse. As corpus data cannot distinguish between news and journalism, this will be revisited in section 4.3.

Regarding the adverb *očividne*, the following word combinations were significantly represented:

Collocations	Number of occurrences	Collocations	Number of occurrences
o. spokojný (‘o. satisfied’)	584	*o. nespokojný* (‘o. dissatisfied’)	72
o. páčiť (sa) (‘o. like’)	338	*o. nepáčiť (sa)* (‘o. dislike’)	175
o. prekážať (‘o. bother’)	57	*o. neprekážať* (‘o. not bother’)	218
o. nervózny (‘o. nervous’)	193	*o. zaskočený* (‘o. startled’)	130
o. vadiť (‘o. be a problem’)	30	*o. nevadiť* (‘o. not be a problem’)	164

Tab. 4. Collocations with the adv. *očividne*

Although, in the case of the adverb *očividne,* collocations with words with the negative prefix do not predominate, expressions with a negative meaning are found among the most frequent collocations. Based on presuppositional semantics, it can be assumed that the context in which the word combinations ***o.*** *neprekážať* ‘o. not bother’ and **o.** *nevadiť* ‘o. not be a problem’ are used refers to negative facts. This is confirmed by examples:

Nemeckému ministrovi ***o.*** *neprekážalo, že zhromaždenie na Majdane bolo v rozpore s platnými ukrajinskými zákonmi*. 'The German minister obviously didn't mind that the gathering at Maidan was in violation of the applicable Ukrainian laws.'
To, že zaberajú miesto ďalším vodičom, im ***o.*** *neprekáža*. 'The fact that they occupy space for other drivers obviously doesn't bother them.'
Hoci Ye trvá krátko a celý album si vypočujete cestou do práce, fanúšikom to ***o.*** *nevadilo*. 'Although Ye is short and the entire album can be listened to on the way to work, it was obviously not a problem for the fans.'

The adverb *prirodzene* is often accompanied in texts by additional synonymous adverbs, which express an extreme degree. Also, it is frequently combined with verbs that express sensory and emotional impressions, e.g. *pôsobiť* 'to appear'. Copular verbs expressing existence or the pretense of existence (see MSJ, pp. 374–376). For example:

Umožňuje mu to podľa nej atakovať exguvernéra a vyzerať ***p***. 'It allows him, according to her, to attack the ex-governor and appear naturally.'

Collocations	Number of occurrences
celkom p. ('quite n.')	5,773
úplne p. ('entirely n.')	3,353
p. vyskytovať sa ('n. occur')	2,071
p. vyzerať ('look n.')	1,980
p. pôsobiť ('appear n.')	1,676
p. nachádzať sa ('be n. found')	1,118

Tab. 5. Collocations with the adv. *prirodzene*

The first two adverbs mainly pair with words of evaluation or approach, while *prirodzene* primarily collocates with expressions of existence and form. Its four most frequent verbs form two synonymous pairs, and it often combines with intensifying adverbs.

The adverb *konečne* is commonly associated with verbs that have a temporal meaning (e.g. *začať (sa)* – 'to begin', *dočkať s*a – 'to wait for'), as well as verbs that refer to the fundamental meaning of this adverb related to the passage of time.

Collocations	Number of occurrences
k. začať (sa) ('f. begin')	14,019
k. dostať ('f. get')	11,608
k. nájsť ('f. find')	8,138
k. dočkať sa ('f. wait for')	6,459
k. prísť ('f. come')	7,915
k. podariť sa ('f. succeed')	7,822

Tab. 6. Collocations with the adv. *konečne*

3.3 Media discourse

There are three reasons for supplementing our corpus data analysis with data from commentaries: we aimed to determine (1) whether the genre is reflected in the predominance of particles over adverbs; (2) whether the results of the corpus analysis would differ from those in journalistic texts; and (3) what role context plays in the analysis of part-of-speech homonymy.

Despite the genre's evaluative nature, particles did not notably dominate; 26 of 40 examples were particles, 14 were adverbs. Both media sources showed a tendency to separate particles with commas, sentence positioning, or isolation as sentence adverbs, particularly with *očividne* and *prirodzene*.

Regarding collocations, the corpus data did not overlap significantly with our "media" research sample, which is understandable given its limited size. Still, similar principles emerged – adverb *konečne* was more frequently paired with dynamic verbs, while *očividne* tended to co-occur with stative verbs. Also, the adverb *absolútne* was often used with words carrying a negative meaning, most frequently adjectives (e.g. *a. nevhodná* 'absolutely inappropriate'). In terms of collocations, the use of modal verbs is prominent in the sample (e.g. *a. nevedeli odpovedať* 'they absolutely could not answer', *a. musíme* 'we absolutely must', *musíme k. precitnúť* 'we must finally wake up' or *o. chcú* 'they obviously want').

The sample showed a tendency to use adverbs like *absolútne* in questions, answers, and interview references (e.g. *a. nevhodná otázka* 'an absolutely inappropriate question', *a. nevedeli odpovedať* 'they absolutely could not answer', *a. kritické otázky* 'absolutely critical questions'). All four analysed words occasionally acted as contextual connectors, linking extralinguistic reality and enhancing text coherence.

4 CONCLUSION

Determining part-of-speech category requires considering multiple aspects, as the “classic” factors presented in morphologies are not always sufficient. The difficulty of distinguishing between individual adverbs and particles is also stressed by the ambiguous description of the lexical meaning in dictionaries with examples.

Part-of-speech homonymy is not a marginal phenomenon in the grammars of inflected languages, and thus, it deserves both academic and didactic consideration. The aim of this text was to highlight possible, established, as well as relatively innovative principles for determining part-of-speech category in the case of homonymy between adverbs and particles. It has been shown that although this phenomenon may initially seem to belong solely to the domain of grammar, its analysis and study must be conducted at the semantic level (including lexical and presuppositional semantics), always taking context into account. Particles primarily refer to what is outside the utterance, and thus modifying its meaning.

From a practical point of view, our study has shown that the formal separation of particles by commas within an utterance may not always be justified. Hence, we consider such exercises inappropriate for pedagogical practice in secondary or higher education. In media discourse, the consistent separation of particles by commas may rather reflect the preferences of a particular language proofreader or the editorial style of the given media.

Based on the analysed examples, we believe that efforts to simplify and generalise the rules for distinguishing between adverbs and particles may have the opposite effect. This issue is even more pronounced in mechanical or automated distinctions made during the annotation of texts into linguistic corpora.

ACKNOWLEDGEMENTS

The study is part of the VEGA project 2/0121/24 *Tvorba kolokačného slovníka prísloviek a s tým spojený výskum spájateľnosti prísloviek na báze Slovenského národného korpusu* (‘Creation of a Collocational Dictionary of Adverbs and research on the linking of adverbs based on the Slovak National Corpus’).

References

Dvonč, L. et al. (1966). Morfológia slovenského jazyka. Bratislava: Veda, 896 p.

Kačala, J. (1984). Slovo ťažko ako príslovka a ako častica. Kultúra slova 18(6), pp. 193–197.

Krátky slovník slovenského jazyka. (2020). 5. dopl. a upr. vyd. Martin: Matica slovenská. 960 p.

Ološtiak, M. (2017). Slovotvorba, slovnodruhové prechody, preberanie a skracovanie lexém. Prešov: Prešovská univerzita v Prešove, 120 p.

Oravec, J., Bajzíková, E., and Furdík, J. (1984). Súčasný slovenský jazyka: Morfológia. Bratislava: Veda, 227 p.

Ortograficko-gramatický slovník slovenčiny. A – Ž. Red. M. Sokolová – A. Jarošová.

Pravidlá slovenského jazyka. (2013). 4. nezmenené vyd. Bratislava: Veda, 592 p.

Slovenský národný korpus – prim-9.0.-juls-all. – Bratislava: Jazykovedný ústav Ľ. Štúra SAV 2023. Accessible at: http://korpus.juls.savba.sk [cit. 21/03/2025].

Slovník súčasného slovenského jazyka. A – G. (2006). Hl. red. K. Buzássyová – A. Jarošová. Bratislava: Veda, 1134 p.

Slovník súčasného slovenského jazyka. H – L. (2011). Ved. red. A. Jarošová – K. Buzássyová. Bratislava: Veda, 1087 p.

Slovník súčasného slovenského jazyka. O – Pn. (2021). Ved. red. A. Jarošová. Bratislava: Veda, 1128 p.

Slovník slovenského jazyka. 6 zv. (1959–1968). Red. Š. Peciar. Bratislava: Vydavateľstvo SAV.

Šikra, J. (1991). Sémantika slovenských prísloviek. Bratislava: Veda, 212 p.

Šimková, M. (2002). Slovnodruhová príslušnosť vetných prísloviek. Slovenská reč 67(4–5), pp. 193–210.

Šimková, M. (2003). Tzv. gramatické slová v morfológii a syntaxi. In: M. Šimková (ed.): Tradícia a perspektívy gramatického výskumu na Slovensku. Bratislava: Veda, pp. 233–239.

Šindlerová, J., and Štěpánková, B. (2021). Between adverbs and particles: A corpus study of selected intensifiers. Journal of Linguistics, 72(2), pp. 444–453.

Uhlířová, L. (1979). K postavení tzv. větných příslovcí v aktuálním členění. Slovo a slovesnost, 40(2), pp. 143–148.

Závodný, A. (2001). Prednášky a praktiká z morfológie slovenského jazyka I. Trnava: Trnavská univerzita, 208 p.

sciendo
DOI 10.2478/jazcas-2025-0007

NO DOUBT, BUT… ON CONNECTIVE FUNCTIONS OF EPISTEMIC MARKERS

LUCIE POLÁKOVÁ[1] – JANA ŠINDLEROVÁ[2] – BARBORA ŠTĚPÁNKOVÁ[3]
[1]Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics, Charles University, Prague, Czech Republic (ORCID: 0000-0002-4879-5530)
[2]Department of Linguistics, Faculty of Arts, Charles University, Prague, Czech Republic (ORCID: 0000-0002-9610-4618)
[3]Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics, Charles University, Prague, Czech Republic (ORCID: 0000-0001-9498-7165)

POLÁKOVÁ, Lucie – ŠINDLEROVÁ, Jana – ŠTĚPÁNKOVÁ, Barbora: No doubt, but... on Connective Functions of Epistemic Markers. Journal of Linguistics, 2025, Vol. 76, No 1, pp. 75 – 84.

Abstract: Epistemic markers are usually treated on the basis of their primary function to express the level of certainty of a speaker about a given proposition. They are often described as items operating on higher levels than syntax. In this paper, we focus on cases in which epistemic markers actually contribute to the syntactic organization of text, either by developing a text-organizing function or a discourse connective function. Specifically, we address three patterns which appeared in the corpus data: a confrontation of modalities, a function of a topic orientation marker, and a contrastive pattern with concessive features.

Keywords: epistemic markers, connectives, concessive constructions, Czech

1 INTRODUCTION

Epistemic markers (EMs)[1] are function words which express the level of certainty of a speaker towards a proposition. Their predominant function is pragmatic and as such, they belong to the group of pragmatic markers.[2] The certainty levels expressed range from full certainty to probability, possibility, down to uncertainty or doubt. Typical examples for Czech are expressions like *jistě* ‘certainly’, *asi* ‘perhaps’, *možná* ‘maybe’ or *stěží* ‘hardly’, i.e. words the meaning of which is defined by their subjective, attitudinal and modifying character.

[1] This study is part of a project investigating expressions that are considered to be typical epistemic markers (EMs). According to Komárek et al. (1986), we also include evidential (*očividně* ‘obviously’) and confirmatory (*samozřejmě* ‘of course’) expressions in this group. For the purposes of this study, we use the term ‘EM’ (instead of ‘examined expression’, for example) even when the epistemic modality of a marker is completely or significantly reduced.

[2] In Czech linguistics, epistemic markers belong to particles, a word class defined through their attitudinal and syn-syntactic character (cf. e.g. Komárek et al. 1986).

EMs often etymologically derive from adverbials (*určitě*, *jasně*, *zřejmě*), so they retain word class ambiguity. The same expression can be used either as an autosemantic adverbial (1a), or as a synsemantic pragmatic marker (1b).

(1a) *Její vraždu viděl zcela* ***jasně****.*
‘He saw her murder quite **clearly**.’

(1b) ***Jasně*** *že ti zavolám.*
‘I will call you **for sure**.’

The shift from adverbial to pragmatic meaning seems to be quite a systematic language change across the respective group of words (cf. Šindlerová et al. 2023; Traugot 1989), the original adverbial meaning may even disappear completely with time.

Secondarily, epistemic pragmatic markers can also have other, derived pragmatic functions, e.g. they can gain an independent syntactic status and become affirmative markers, see (2).

(2) *Zavoláš mi? –* ***Jistě****!*
‘Would you call me? – **Sure**!’

In this study, we approach epistemic markers **from a syntactic perspective**, i.e. from the point of view of **their possible connective functions**. During our analysis of these expressions, we came across a certain group of syntactic patterns, or contexts,[3] where the original epistemic marker occurs in a position typical of a connective device or a certain part of it, while also exhibiting similar semantics.

The connective function of EMs has not been consistently described using Czech data yet. It is not mentioned either in standard grammars, or dictionaries, including the most recent ones. Nevertheless, individual studies of Czech particles suggest that connective function may be in fact quite common for various pragmatic markers (see e.g. Kolářová 1998 on the connective functions of *teda* (roughly) ‘so’, or Mladová 2008 on the functions of the prototypical focusing particle *také* ‘also’).

We will try to answer the following research questions:

1. What are the types of constructions and contexts for the EMs in connective function?
2. Does the different epistemic strength of individual EMs have an impact on the interpretation of the propositions?
3. What other aspects do the EMs bring into the interpretation of an utterance compared to neutral connective devices?

[3] We prefer to speak about contexts, where the setting of the studied marker(s) is inter-sentential, goes beyond the sentence boundary.

We will focus on the description of such patterns with epistemic markers that appear to be the most distinctive demonstration of text linking functions in our data.

2 DATA AND METHOD

The research is inspired by the works on the SEEM-CZ project, within which a lexicon of Czech epistemic and evidential markers SEEMLex is being composed (cf. Štěpánková et al. 2024) via the annotation of the data of the InterCorp corpus, specifically the core part of its Czech and English sections (Rosen et al. 2022). A list of EMs was compiled manually after analysis of the Czech grammars, dictionaries, and corpora. The forty most frequent expressions in the InterCorp corpus were then selected from this list and annotated with 100 examples each. The annotation scheme (see Štěpánková et al. 2024) includes information such as the use of the EM and the presence of contrast in the sentence. During the annotation, the connective function of certain EMs was identified based on these clues. The patterns found in the InterCorp data were then confronted with their occurrences in the large Czech National Corpus (Křen et al. 2024), in the annotated data of the newest version of the Prague Discourse Treebank 4.0 (Synková et al. 2024) and in CzeDLex, the lexicon of Czech discourse connectives (Mírovský et al. 2021).

3 ANALYSIS

In this section, we describe the patterns of the connective use of EMs we found in the corpus data. We proceed from cases where the presence of an EM supports, emphasizes, or builds upon a contrastive syntactic relation, to cases that signal a topic diversion, up to those with a clear connective function.

3.1 Confrontation of modalities

The first pattern contrasts two EMs expressing differing – sometimes even opposing – degrees of certainty, by juxtaposing them in a comparative construction: maybe A, but definitely B, see (3). Typically, in such constructions, the confrontational meaning arises from the opposition between the semantic features of the lexical items themselves. This semantic tension remains strong even in the absence of an explicit contrastive connective. Furthermore, one part of the construction may be negated (yielding opposite polarity, as in (3a), though this is not always the case (3b).

(3a) *V lepším případě **možná** i šest – ale **rozhodně** ne víc.* (InterCorp)
‘**Maybe** six at best – but **certainly** no more.’

(3b) *Kniha Viktimologie pro forenzní praxi nepatří podle Ludmily Čírtkové do ruky těm, o kterých v publikaci najdeme následující vtip – **možná** cynický, **ale rozhodně** výstižný pro naši dobu [...].* (SYNv13)

‘According to Ludmila Čírtková, the book Victimology for Forensic Practice does not belong in the hands of those about whom the following joke can be found in the publication – **perhaps** cynical, **but certainly** appropriate for our times [...].’

Since the conjoining “ability” of the EMs in this case lies only in the cooperation within a semantic (lexical) contrast, we do not consider this a case of proper connective use. Nevertheless, there is a clear text-orienting function present in the construction.

3.2 Topic orientation

A different pattern is typical of the word *každopádně* ‘anyway’ in Czech when it appears in the left periphery of a sentence. There are (at least) two subtypes of this pattern, though the boundaries between them are often unclear. The first subtype typically involves a list of possibilities stated by the speaker, followed by a summarizing conclusion introduced by *každopádně* (4). The conclusion presents an idea that the speaker considers valid regardless of whether the previous context is relevant. In this respect, *každopádně* works as a discourse marker[4], more specifically, as a marker of topic orientation (cf. Fraser 2009).

(4) *Prosila ji, aby přijela, protože se cítí slabá, nebo je něco v nepořádku s domem,* ***každopádně*** *protože potřebuje pomoc.* (SYNv13)
‘She begged her to come because she felt weak or something was wrong with the house, **either way** because she needed help.’

This function of *každopádně* is even more perceivable in an inter-sentential setting (5).

(5) *Možná loď bloudila vesmírem velmi dlouho, nebo její posádka změnila v zoufalé snaze uniknout nebezpečí kurz.* ***Každopádně*** *se nenašlo nic, podle čeho by šlo zjistit, odkud pocházela.* (SYNv13)
‘Perhaps the ship had been wandering in space for a very long time, or its crew had changed course in a desperate attempt to escape danger. **Either way**, there’s nothing to indicate where it came from.’

In other cases, *každopádně* signals a more distinct shift in topic, or a transition from unimportant (irrelevant) information to an important (relevant) one (6). In such instances, the expression’s relation to the preceding context is not semantic; the clauses adjacent to the marker are only loosely related in terms of content, or may even be entirely unrelated.

[4] In accordance with Fraser (2009), we understand “discourse markers” to be a class of pragmatic markers that signal an aspect of the organization of ongoing discourse.

(6) *Než jsem vyrazil, postříkal jsem vnitřek vozu deodorantem s vůní borovic, ale moc rozdílu v tom nebylo.* ***Každopádně*** *jsem myslel pouze na jediné – aby se Georgina neprobudila, dřív než dorazím do Raytonu.* (SYNv13)
'I sprayed the inside of the car with pine scented deodorant before I left, but it didn't make much difference. **Anyway**, my only thought was to make sure Georgina didn't wake up before I got to Rayton.'

In both cases, the epistemicity of *každopádně* is weakened. In the former case, the speaker summarizes the previous utterances, which cannot be verified, with a general statement they regard as certain. In the latter case, the speaker dismisses the prior statements as irrelevant in contrast to the following one – they shift their attention from one situation (or an aspect of a situation) to another.

3.3 Concessive pattern

One of the most compelling examples of a connective function of EMs occurs when an epistemic expression stands as a direct **part of the connecting expression** in a correlative contrastive pattern (7).

(7) *Císař* ***možná*** *křesťany podporoval,* ***ale*** *u dvora bylo plno lidí, kteří se na tu novou sektu dívali v lepším případě s pobavením, v horším případě s podezřením či dokonce nepokrytě nepřátelsky*. (SYNv13)
'The Emperor **may** have supported the Christians, **but** the court was full of people who looked at the new sect with amusement at best, suspicion at worst, or even outright hostility.'

Here, the EM stands in the left part of the clause (sentence), taking the position and function of the prototypical Czech connective *sice* in the correlative multiword connective *sice – ale*, cf. (8).

(8) *Císař* **sice** *křesťany podporoval,* **ale** *u dvora bylo plno lidí, kteří se na tu novou sektu dívali v lepším případě s pobavením, v horším případě s podezřením či dokonce nepokrytě nepřátelsky*.

Sice prototypically signals the speaker's admission of the validity of the content of the proposition in which it appears, a partial assent. It is non-autonomous in that it presupposes some contradiction in the second proposition. Sentences with *sice – ale* are typically formally analyzed as adversative in Czech grammars (Grepl and Karlík 1998, p. 341), nevertheless their status as a paratactic formulation of a concessive relation is sometimes mentioned (Karlík 1995, p. 112). To our knowle.g. English grammars do not recognize a multiword connective of this type. Rather, translations reveal that other linguistic devices are usually employed, including *may* (9), *while,* *although*, etc. (cf. Vavřín and Rosen 2015). Compare also example (10), where, in the English translation, the acknowledgment of the content of the first proposition is conveyed through the intensifying use of *did.*

(9) *Já vždycky říkám, že kachna je* ***sice*** *dost tuhá,* ***ale*** *má svou zvláštní chuť.*
‘I always say that duck **may** be tough, **but** it has its own special taste.’ (InterCorp)

(10) *Tato strategie úspor* ***sice*** *vyřešila problémy s bilancí, měla* ***však*** *za následek nízký růst a obrovskou nezaměstnanost.*
‘The savings strategy **did** sort out the imbalances, **but**, in turn, resulted in low growth and increasingly high unemployment.’ (InterCorp)

The concessive interpretation of the constructions with modals is discussed e.g. by Palmer (2001, p. 31), or Leclerq (2024), who offers an account of a concessive *may/might* construction as a newly developed strong pattern in English syntax. Leclerq speaks primarily about the secondary grammaticalization of a modal verb (a potential epistemic marker) as a connective device, thus we believe that other epistemic markers also do have a potential to constitute contrastive and concessive relations.

Other expressions in English, which are, more rarely, used in the same position/setting, are also related to veridicality: apart from *may/might*, literature focused on English primarily mentions the use of: *true*, *fact*, *well*, *indeed*, *granted…* in these contexts (König 1988, pp. 154–155). In the more conversational or argumentative settings, expressions like *no doubt* or *of course* were identified. Also the German *zwar*[5] etymologically comes from *es ist wahr* ‘it is true’.

The modal expressions occurring in concessive connective contexts express varying degrees of certainty – from high certainty (*bezpochyby*, *určitě*, *rozhodně,* etc.) to low certainty (*možná*). We believe that these markers can indicate different communicative functions and speaker strategies. For example, in (11), the use of the high-certainty expression *bezpochyby* ‘no doubt’ does not primarily indicate the speaker’s strong epistemic conviction regarding the proposition. Rather, it reflects a highly polite attitude toward the interlocutor’s claim, preceding the speaker’s own contradictory statement. Similarly, in (12), *možná* is used to express a directive assumption[6] about the interlocutor’s state of mind. In both cases, the initial proposition is weakened in favour of the following, contradictory claim.[7]

[5] The German construction *zwar – aber* is a direct parallel to Czech *sice – ale*.

[6] By directive assumption we mean a situation in which the speaker firmly assumes a particular (agreeing or disagreeing) stance of their communication partner, but the acknowledgement or denial from the partner is not expected at all.

[7] Cf. also Rossari (2018) or Ivanová (2019) for the analysis of non-epistemic use of *may be* in constructions of the type *I may be a woman, but I can change a tire*.

(11) ***Bezpochyby*** *máš pravdu, že je někdo na palubě, ale to nemusí nutně znamenat, že to má něco společného s námi…* (SYNv13)
‘You’re **no doubt** right that someone’s on board, but that doesn’t necessarily mean it’s anything to do with us…’

(12) ***Možná*** *že jsem vás nepřesvědčila, ale já věřím, že hlavní slovo v té záležitosti měl pan Dixon.* (InterCorp)
‘I **may** not have convinced you **perhaps**, but I am perfectly convinced myself that Mr. Dixon is a principal in the business.’

Analogous structures can as well be found in an inter-sentential setting, see (13). This is particularly interesting, since the original Czech construction *sice – ale*, as well as the German *zwar – aber*, is syntactically strongly constrained, it occurs within one complex sentence unit, otherwise it is considered ungrammatical. It seems that similar constructions with epistemic markers operate independently of sentence boundaries in this respect.

(13) [Context: *Když je někdo opravdu dobrý, tak si poradí.*] ***Samozřejmě*** *i dnes máme fenomenální osobnosti. Na jednu takovou připadne* ***však*** *devět jiných, které nevhodné školní vzdělávání srazí do průměru.* (PDiT 4.0)
‘[Context: If someone is really good, they will find a way.] **Of course,** we still have phenomenal individuals today. For every one of them, **however**, there are nine others whom inadequate schooling pushes down to mediocrity.’

Moreover, a sentence-initial EM can become syntactically independent, separated by a comma, and its function may shift to that of an affirmative particle (14).

(14) ***Jistě****, skepticismus je někdy na místě.* ***Ale*** *že by dětem někdo podsouval vzpomínky na minulé životy pořád dokola?* (SYNv13)
‘**Sure**, scepticism is sometimes appropriate. **But** that someone would keep implanting memories of past lives into children over and over?’

The abovementioned patterns, be it the intra-sentential, inter-sentential, or the independent affirmative markers, share typical characteristics and conversational structure of a concessive relation. Among the typical features shared with concessive structure, the following are mentioned: adversativity and causal relation between the two parts of the sentence, resulting in an unmet expectation, triadic logical structure and polyphonic character (underlying dialogic interaction, the sentence presents opinions of at least two people, there are two “voices” present) (Schwenter 2000; Drobník 2024). Couper-Kuhlen and Thompson (2000, p. 382) work with the following conversational structure of the conceding act:

1st move A: States something or makes some point
2nd move B: Acknowledges the validity of this statement or point (the conceding move)
3rd move B: Goes on to claim the validity of a potentially contrasting statement or point

In our data, the first move is either present in the preceding context (actual dialogue between two subjects) (15), implicitly assumed on the basis of the speaker's experience with the partner or a third person (16), or it is represented as a generally accepted fact (17). The concessive construction then involves the explicitly contrasted 2nd and 3rd move, respectively.

(15) *„Musím ji jít hledat." „**Jasně**. **Ale** zatím se posaď tuhle na lavici."* (SYNv13)
"'I have to go look for her." "**Sure**. **But** in the meantime, take a seat on the bench here."'

(16) *Strážci jazyka se mnou **možná** nebudou souhlasit, **ale** to je v pořádku, nemusejí.* (SYNv13)
'The guardians of the language **may** disagree with me, **but** that's okay, they don't have to.'

(17) *Z těch dvou možností by **nepochybně** byla bezpečnější puška, **ale** čím víc Frank myslel na horkou krev, stříkající z proříznutého hrdla George T. Nelsona a skrápějící mu ruce, tím víc se mu zamlouvala i druhá možnost.* (SYNv13)
'Of the two options, the rifle would **undoubtedly** have been the safer, **but** the more Frank thought about the hot blood spurting from George T. Nelson's slit throat and scraping his hands, the more he liked the second option.'

The EM then serves to emphasize the acknowledgement of the validity of the 1st move. In other words, "in the epistemic domain concessive connection expresses the idea that the speaker, in spite of being convinced of the content of the concessive clause, still reaches the opposite conclusion contained in the main clause" (Crevels 2000, p. 318). The same idea is expressed by Karlík (1995), who suggests that concessive sentences should be viewed pragmatically as a way to prevent misunderstanding in communication, to prevent conflict (cf. also Barth 2000, or Čermáková et al. 2019).

The epistemic strength of the EM then indicates the actual amount of credit we give to the 1st move statement, or the level of politeness we wish to grant to its author, cf. (18a) and (18b).

(18a) ***Bezpochyby** je pravda, že může nastolovat nějaká témata, může je debatovat s vládou, **ale** jeho pravomoci jsou někde jinde.* (SYNv13)
'It is **undoubtedly** true that he can raise issues, he can debate them with the government, **but** his powers are elsewhere.'

(18b) ***Možná** máš pravdu, **ale** to ho nijak neomlouvá.* (SYNv13)
'You **may** be right, **but** that's no excuse.'

4 CONCLUSION

In this paper, we have examined several patterns of the connective use of Czech epistemic markers (EMs). The first pattern involves two EMs expressing differing degrees of certainty, which are juxtaposed to highlight a contrast in modality. We describe the role of EMs in this context as *text-structuring*. The second pattern concerns *topic orientation* – here, the EM functions as a topic shifter, marking the boundary between irrelevant and relevant utterances, or summarizing preceding, less relevant content with a more general and more relevant idea.

The third, and most prominent, pattern is the *concessive* use, in which the EM takes on the role of the Czech connective *sice* within the multiword connective expression *sice – ale*. Unlike the neutral *sice*, however, the EM also conveys politeness – expressing a greater or lesser degree of respectful acknowledgment of the interlocutor's opinion before presenting one's own, contrasting view.

While observing the connective functions of the EMs in corpora, we can see that the (secondary) pragmaticalization is an ongoing process. Nevertheless, it is important to try to capture the full range of their functions, in order to offer credible and relevant accounts of the epistemic markers in Czech.

ACKNOWLEDGEMENTS

The research has been supported by the Czech Science Foundation under the project GA23-05240S.

References

Barth, D. (2000). "that's true, although not really, but still": Expressing concession in spoken English. In Cause – Condition – Concession – Contrast: Cognitive and Discourse Perspectives. Berlin: De Gruyter Mouton, pp. 411–438.

Čermáková, A. et al. (2019). Diskurzní markery. In: J. Hoffmannová et al. (eds.): Syntax mluvené češtiny, pp. 244–351.

Couper-Kuhlen, E., and Thompson, S. (2000). Concessive patterns in conversation. In Cause – Condition – Concession – Contrast: Cognitive and Discourse Perspectives. Berlin. Mouton de Gruyter, pp. 381–410.

Crevels, M. (2000). Concessives on different semantic levels: A typological perspective. In Cause – Condition – Concession – Contrast: Cognitive and Discourse Perspectives. Berlin: Mouton de Gruyter, pp. 313–340.

Drobník, O. (2024). Přípustka v mluvnicích češtiny. Bohemistyka, 24(3), pp. 323–348.

Grepl, M., and Karlík, P. (1998). Skladba češtiny. Olomouc: Votobia, 603 p.

Fraser, B. (2009). An Account of Discourse Markers. International Review of Pragmatics, 1(2), pp. 293–320.

Ivanová, M. (2019). Epistemicita ako koncept deiktického odkazovania na propozíciu. In: J. Kesselová (ed.): Personálna a sociálna deixa v slovenčine. Prešov: Filozofická faktulta Prešovskej univerzity, pp. 153–210.

Karlík, P. (1995). Studie o českém souvětí, Brno: MU.

Kolářová, I. (1998). Významy slova tedy (teda) v souvislých textech. Naše řeč, 81(2–3), pp. 118–123.

Komárek, M., Kořenský, J., and Petr, J. (1986). Mluvnice češtiny 2. Praha: Academia.

König, E. (1988). Concessive connectives and concessive sentences: cross-linguistic regularities and pragmatic principles. In: J. Hawkins (ed.): Explaining Language Universals. Oxford: Blackwell, pp. 145–166.

Křen, M. et al. (2024). Corpus SYN, version 13 from 27/12/2024. ÚČNK FFUK, Praha. Accessible at: https://www.korpus.cz.

Leclercq, B. (2024). The post-modal grammaticalisation of concessive may and might. Constructions and Frames, 16, pp. 130–161.

Mírovský, J., Synková, P., Poláková, L., Kloudová, V., and Rysová, M. (2021). CzeDLex 1.0. Data/software, UK, Prague. Accessible at: http://hdl.handle.net/11234/1-4595.

Mladová, L. (2008). K problematice vztahu rematizátorů a textových konektorů. (On the Relation between Rhematizers and Discourse Connectives). In Čeština doma a ve světě, 3 and 4, pp. 126–133.

Palmer, F. R. (2001). Mood and Modality. Cambridge University Press.

Rosen, A., Vavřín, M., and Zasina, A. J. (2022). InterCorp, Release 15 of 11 November 2022. ÚČNK UK. Accessible at: http://www.korpus.cz.

Rossari, C. (2018). The representation of modal meaning of French sentence adverbs in a qualitative and quantitative approach. In Linguistik online 92(5), Special Issue, pp. 235–255.

Schwenter, S. (2000). Viewpoints and polysemy: Linking adversative and causal meanings of discourse markers. In Cause – Condition – Concession – Contrast: Cognitive and Discourse Perspectives. Berlin: Mouton de Gruyter, pp. 257–281.

Synková, P., Mírovský, J., Paclíková, M., Poláková, L., Rysová, M., Scheller, V., Zdeňková, J., Zikánová, Š., and Hajičová, E. (2024). Prague Discourse Treebank 4.0. Data/software, LINDAT/CLARIAH-CZ. Accessible at: https://ufal.mff.cuni.cz/pdit4.0.

Šindlerová, J., Štěpánková, B., and Andrén, I. L. (2023). Epistemická částice zřejmě pohledem paralelního korpusu. Korpus – gramatika – axiologie, 27, pp. 37–52.

Štěpánková, B., Poláková, L., Šindlerová, J., and Novák, M. (2024). What Can Dictionaries Tell Us About Pragmatic Markers – Building the Lexicon of Epistemic and Evidential Markers in Czech. In Lexicography and Semantics. Proceedings of the XXI EURALEX International Congress, Zagreb: Institut za hrvatski jezik, pp. 728–741.

Traugott, E. C. (1989). On the rise of epistemic meanings in English: An example of subjectification in semantic change. Language 65(1), pp. 31–55.

Vavřín, M., and Rosen, A. (2015). Treq. FFUK. Praha. Accessible at: http://treq.korpus.cz.

sciendo
DOI 10.2478/jazcas-2025-0008

A MOUNTAIN OF EVIDENCE? A CORPUS STUDY OF THE ARGUMENT STRUCTURE OF TRANSFERRED SENSES OF NOUNS IN ENGLISH

JAKUB SLÁMA
Department of Contemporary Lexicology and Lexicography, Czech Language Institute, Czech Academy of Sciences, Prague, Czech Republic & Department of Grammar, Czech Language Institute, Czech Academy of Sciences, Prague, Czech Republic (ORCID: 0000-0002-6555-0471)

SLÁMA, Jakub: A Mountain of Evidence? A Corpus Study of the Argument Structure of Transferred Senses of Nouns in English. Journal of Linguistics, 2025, Vol. 76, No 1, pp. 85 – 94.

Abstract: Based on an earlier observation, the study poses the question of whether the presence or absence of a valency complement of a noun relates to whether the noun is used literally or non-literally, in a transferred sense (typically based on metaphor). Two case studies are presented, one concerning 13 body part nouns (such as *foot*) and their transferred uses, and the other concerning two landscape nouns, *mountain* and *flood*. Both studies show that non-literal uses of nouns are much more likely to take an overt complement. This might relate in part to the type shift of sortal nouns into relational nouns and in part to the low degree of lexicalization of some transferred senses, which renders them more reliant on contextual cues, such as the use of a complement, for adequate interpretation.

Keywords: argument structure, complement, Part-*of*-Whole construction, metaphor, pseudo-partitive construction, quantifier, valency

1 INTRODUCTION

While noun valency is still somehow "in the shadow of the valency of verbs" (Spevak 2014, p. ix), there is a relatively large body of literature on this subject. Generativist linguists – largely in continuation of the tradition started by such influential works as Chomsky (1970) and Grimshaw (1990) – have focused almost exclusively on deverbal nominalizations, often overindulging in theorizing without much regard for empirical data (cf. Newmeyer 2009; Lieber 2016). On the other hand, functional linguists have studied various types of complex noun phrases in English empirically – but often without explicitly and systematically addressing issues of argument structure (e.g. Keizer 2007; ten Wolde 2023). In this somewhat scattered landscape of literature on English noun valency, it appears that various phenomena have gone unexplored. One of them is the interaction between metaphor (or transferred, non-literal senses more generally) and noun valency.

In a previous corpus study of noun valency (Sláma 2020, p. 445), it was suggested that nouns that appear not to be valent (i.e., not to require arguments) might be used in the part-*of*-whole pattern both literally and metaphorically, as illustrated by examples (1) and (2) below, taken from the British National Corpus (BNC). What is interesting here is the fact that the *of*-phrase in the second, metaphoric example cannot be omitted (without this resulting in an essentially nonsensical sentence).

(1) *The underlying cause for this decision was the awful damage caused by* ***the savage winter of 1709****.* (BNC)

(2) *When a man reaches* ***the winter of his life****, there's nothin' he can look forward to but death.* (BNC)

Even though the metaphoric reading of the noun might be somewhat responsible for the fact that the *of*-phrase is essentially obligatory (and thus perhaps somehow closer to being a valency complement rather than a modifier), to my knowle.g. the interaction of metaphor and argument structure has not been studied. This paper is thus an attempt to contribute towards bridging this gap.

Section 2 introduces the distinction between sortal and relational uses of nouns, which I believe to be relevant here, as what we see in example (2) appears to be the reinterpretation of an inherently sortal noun (*winter*) as a relational noun. Section 3 provides a little background on the role of metaphor in language and in grammatical constructions more specifically, and presents two corpus studies, one focusing on transferred senses and complementation of polysemous body part terms (such as *foot*) in a subcorpus of the corpus InterCorp (Section 3.1), and the other examining two landscape nouns, *mountain* and *flood*, in the BNC (Section 3.2). Section 4 proposes some explanations for the observations reported on in this paper.

2 SORTAL VS. RELATIONAL USES OF NOUNS

Behaghel (1932, p. 22) was perhaps one of the first scholars to distinguish between absolute concepts (*absolute Begriffe*) and relative concepts (*relative Begriffe*), a distinction commonly interpreted today as one between **sortal nouns** and **relational nouns** (e.g. Mackenzie 1997). For instance, *cat* is a sortal noun; when hearing the word *cat*, one knows what is meant, and upon seeing a cat, one can (generally) identify it as a cat without any further information. On the other hand, when seeing a woman, one cannot identify her beyond any doubt as a mother or a non-mother, as a woman is a mother only in relation to some other person (hence the frequent relational use of the noun, as in, for instance, *my mother*, or *the mother of my friend*); *mother* is thus a relational noun. This distinction has been repeatedly implicated as relevant for valency: while sortal nouns are avalent, relational nouns have valency properties (e.g. Löbner 1985, p. 292; Plag 2003, p. 148).

Löbner (2011, 2015) discusses the distinction in more detail and further distinguishes sortal nouns into (unique) **individual nouns** (e.g. *Paula*, *pope*, and *weather*) and (non-unique) **sortal nouns** proper (e.g. *cat*, *table*, and *water*), and relational nouns into (unique) **functional nouns** (e.g. *father*, *mouth*, and *surface*) and (non--unique) **relational nouns** proper (e.g. *brother*, *part*, and *eye*). These four concept types differ with respect to their use with markers of definiteness, number, and possession.

Löbner differentiates between congruent and incongruent uses of nouns; congruent uses are those in which the use of the noun corresponds to its inherent semantics. For instance, *father* is a functional noun (i.e., an inherently unique and relational noun), and thus its use in *The father of Peter is tall* with the definite article is a congruent one; on the other hand, its use in *A father has called* is an incongruent one, leading to a concept shift, whereby *father* is interpreted as a sortal rather than a relational noun (Brenner et al. 2014, pp. 22–23). Instead of viewing sortal and relational nouns as two separate classes of nouns, it thus appears more adequate to think of nouns in terms of their relational or sortal uses, with many nouns commonly crossing the boundary. Having conducted two corpus studies and a psycholinguistic experiment, Brenner et al. (2014) conclude that their results "support the hypothesis that nouns are lexically specified with respect to the conceptual features uniqueness and relationality but that a relatively high proportion of their actual uses is incongruent with their lexical specification."

3 THE ROLE OF METAPHORS: TWO CASE STUDIES

Especially since the advent of Cognitive Linguistics as a new theoretical framework (cf. Croft and Cruse 2004), a renewed interest in metaphor has flourished. It has been recognized that the metaphor is much more than an ornate device used in literature, and the pervasiveness of conceptual metaphors in language has been documented. For instance, the conceptual metaphor TIME IS MONEY is reflected in everyday English expressions such as *You're wasting my time*, *That flat tire cost me an hour*, or *This gadget will save you hours* (Lakoff and Johnson 1980, pp. 7–8), which parallel the way we talk about money.

Some attention has been paid to the fact that conceptual metaphors also affect the way some grammatical constructions are used (cf. Dancygier and Sweetser 2014, pp. 127–161; Sullivan 2025). For instance, Sláma (2022, p. 259; 2024, p. 171) suggests that Czech perfective verbs with the prefix *pro-* that require an obligatory direct object referring either to an amount of money or an amount of time are instances of a construction (in the sense of Construction Grammar) with two senses, also based on the TIME IS MONEY conceptual metaphor: 'to spend money by doing something' (as in (3) below) and 'to spend time by doing something' (as in (4)).

(3) *Jinde lidé vždy vice peněz* ***projedí*** *než „****probydlí****“.* (Sláma 2022, p. 258)
lit. ‘Elsewhere people always **eat away** more money than they **live away**.’
‘People elsewhere always spend more money on food than on housing.’

(4) *Celá devadesátá léta* ***jsme provečírkovali****.* (Sláma 2022, p. 259)
lit. ‘The whole nineties we partied away.’
‘In the nineties we spent/wasted all the time partying.’

Within a project concerned with noun valency in English, I created a database of complex nominals with potential complements. The nominals were identified manually in a corpus of the seven *Harry Potter* novels by J. K. Rowling and three accompanying books by the same author. The database contains about 22,000 complex nominals with further annotation; the details are not of importance here. What is relevant, however, is that in the database it was also annotated whether the head noun of a nominal is used in its literal sense (e.g. *the* ***feet*** *of a man with hair and beard so overgrown Harry could see neither eyes nor mouth*) or in its transferred (e.g. metaphorical) sense (e.g. *the* ***foot*** *of the stairs/page/bed*). Tab. 1 shows the ten nouns with potential complements that are found most frequently in the database in their transferred senses:

Noun	**Transferred sense uses**
foot	78
head	61
cloud	25
stream	19
shower	18
trace	15
heart	15
arm	15
wave	15
sea	14

Tab. 1. Ten head nouns in the database with the highest number of uses in transferred senses

It is apparent from Tab. 1 that two semantic groups of nouns are represented most often: body part nouns (*foot*, *head*, *heart*, *arm*) and nouns referring to parts of the landscape and related natural phenomena (*cloud*, *stream*, *shower*, *wave*, *sea*). This is not surprising, as both body part terms and landscape terms have been shown to often involve polysemy and metaphor (e.g. Lewandowska-Tomaszczyk 2020; Wierzbicka 2007; Burenhult and Levinson 2008; Bromhead 2013). The following two corpus-based case studies thus focus on these two groups of nouns.

3.1 A corpus study of body part nouns

From the above-mentioned *Harry Potter* database, I filtered out all body part nouns that are attested with an *of*-phrase relating the (body) part meronymically to a whole, which could be seen as a complement (e.g. Müller 2000, p. 75). I selected only nouns attested in the database non-marginally in both a literal and a transferred sense. This resulted in a list of 13 English nouns, listed here with an example of each in a transferred sense: *arm* (*each arm of Harry's chair*); *back* (*the back of his seat*); *face* (*the face of the white moon*); *foot* (*the foot of the stairs*); *hand* (*the luminous hands of his clock*); *head* (*the head of the stairs/broom*); *heart* (*the heart of the Forest/maze*); *knee* (*the knees of his jeans*); *leg* (*the legs of the chair*); *mouth* (*the mouth of the tent/alleyway/cave*); *neck* (*the neck of the dressing gown*); *spine* (*the leather spines of books*); *tail* (*the tails of his frock-coat*).

Given the necessity to annotate the uses of the nouns manually both for their sense (literal vs. non-literal) and the presence or absence of the complement (as not every *of* following a body part noun is relevant) and given the high frequency of the body part nouns, I had to work with a rather small subcorpus. Given the fact that the seven main *Harry Potter* novels are included in the corpus InterCorp v16 – English, I created a subcorpus containing only these seven novels, identified in it all instances of the 13 lemmas, excluded all irrelevant instances (e.g. *back* used as a verb), and annotated the rest for their sense (literal vs. non-literal) and the presence or absence of a complement.

The results are provided in Tab. 2.

Noun	**Literal interpretation**		**Non-literal interpretation**	
	Complement	**No complement**	**Complement**	**No complement**
arm	3 (0.49%)	612 (99.51%)	15 (93.75%)	1 (6.25%)
back	6 (2.36%)	248 (97.64%)	195 (89.45%)	23 (10.55%)
face	23 (1.47%)	1,545 (98.53%)	7 (50.00%)	7 (50.00%)
foot	2 (0.30%)	674 (99.70%)	81 (31.15%)	179 (68.85%)
hand	5 (0.31%)	1,604 (99.69%)	4 (20.00%)	16 (80.00%)
head	31 (2.34%)	1,291 (97.66%)	17 (77.27%)	5 (22.73%)
heart	1 (0.39%)	258 (99.61%)	18 (100.00%)	0 (0.00%)
knee	1 (0.64%)	156 (99.36%)	4 (100.00%)	0 (0.00%)
leg	6 (1.84%)	320 (98.16%)	3 (25.00%)	9 (75.00%)
mouth	1 (0.22%)	451 (99.78%)	6 (85.71%)	1 (14.29%)
neck	5 (2.55%)	191 (97.45%)	12 (70.59%)	5 (29.41%)
spine	3 (42.86%)	4 (57.14%)	1 (25.00%)	3 (75.00%)
tail	8 (9.64%)	75 (90.36%)	4 (50.00%)	4 (50.00%)
Total	95 (1.26%)	7,429 (98.74%)	367 (59.19%)	253 (40.81%)

Tab. 2. (Non-)literal senses of 13 body part nouns and the presence/absence of a complement

In their literal uses, body part nouns are used with a complement only marginally (1.26% of instances); note that this in no way contradicts my assumption that body part nouns are inherently relational, as very common uses with, for instance, possessives (*his hand*) also showcase the relational behavior of these nouns while not featuring an *of*-complement.

In the transferred senses of body part nouns, however, the proportion of uses with a complement rises to 59.19%. If we exclude the noun *foot*, as it skews the data (in its frequent sense of a unit of measure, in which it never appears with an *of*-phrase functioning as a complement), the percentage rises even higher – to 79.44%. Mostly (in cases different from that of *foot* in the sense of a unit), when there is no complement with a non-literal use of the noun, the underlying argument is expressed as a premodifier or an adnominal determiner:

(5) *"The tree was placed at* ***the tunnel mouth*** *to stop anyone coming across me while I was dangerous."* (InterCorp v16 – English)

(6) *Harry opened his eyes and stared through his fingers at* ***the wardrobe's clawed feet****, remembering what Fred had said* […]. (InterCorp v16 – English)

3.2 A corpus study of landscape nouns

Landscape nouns (e.g. *mountain* and *sea*) and similar, typically weather-related nouns (e.g. *shower* and *cloud*) are also often prone to polysemy based on metaphor and have been identified in the *Harry Potter* database as a significant group illustrating the association between metaphor and argument structure. For a case study of such nouns, I originally chose to work with the British National Corpus (BNC). However, in the corpus, nouns such as *mountain* and *sea* are highly frequent and it would be impossible to annotate manually all of their occurrences if multiple high-frequency nouns were chosen. Since this study is intended as a first step towards investigating the interactions of metaphor and valency, I decided to annotate all occurrences of only two nouns: *mountain*, a noun presumably quite representative of this group, and the less frequent *flood*. Both can be used as quantifiers in pseudopartitive constructions (e.g. Koptjevskaja-Tamm 2001), i.e., with presumable complements, as in *mountains of debt* or *the flood of complaints*.

For both lemmas, all instances in the BNC were identified, compounds (e.g. *mountain biking* and *flood gates*) were excluded, and so were other irrelevant examples (e.g. the verb in *I can flood them with data* or the proper noun in *James Flood*). The remaining instances were annotated for whether their use is literal or transferred, and for the presence or absence of the complement. Instances with *of*-phrases that are more plausibly seen as modifiers rather than complements (e.g. *the destructive floods of autumn 1981*) and instances of specific constructions (as in *a veritable mountain of a man*; cf. ten Wolde 2023) were not included as complements. The results are summarized in Tab. 3.

Noun	**Literal interpretation**		**Non-literal interpretation**	
	Complement	**No complement**	**Complement**	**No complement**
mountain	156 (4.28%)	3,485 (95.72%)	209 (71.33%)	84 (28.67%)
flood	1 (0.13%)	785 (99.87%)	360 (89.55%)	42 (10.45%)
Total	157 (3.55%)	4,270 (96.45%)	569 (81.87%)	126 (18.13%)

Tab. 3. Literal vs. non-literal senses of *mountain* and *flood* and the presence vs. absence of a complement in the BNC

In their literal uses, the two nouns are used with an *of*-complement in 3.55% of cases only (the only instance of this with *flood* is found in the context *a flood of water gushed from the McMonnies Lake*, which can be understood both as a literal flood and an expression of great quantity); in their non-literal, usually quantifying uses, the two nouns occur with an *of*-complement in 81.87% of cases. This is clearly a significant difference. The cases in which a non-literal reading co-occurs with the absence of a complement are generally accounted for by idiomatic expressions (most often, *to be in full flood*, as in *discussion was already in full flood*), instances where the underlying argument is expressed as a premodifier, as in examples (7) and (8), and instances in which the argument is inferable from the context, as in examples (9) and (10):

(7) *It is* ***a huge and rapidly growing rubbish mountain****, the largest, per citizen, in the world.* (BNC)

(8) […] *could lead to* ***an immigration flood*** *exceeding "the worst fears" of many of his backbench colleagues.* (BNC)

(9) *Michael was happy enough with his "batburger", but preferred the chips on Karen's plate to* ***the mountain*** *on his own.* (BNC)

(10) *The flow of Albanian escapees across the southern border into Greece accelerated dramatically in December, and turned into* ***a flood*** *in early January.* (BNC)

4 CONCLUSION

The first case study of 13 body part nouns and their transferred senses (Section 3.1) and the second case study of two landscape nouns, *mountain* and *flood* (Section 3.2), both show beyond any doubt that at least with some nouns, transferred senses are significantly more likely to take an overt *of*-complement than literal senses.

In part, especially with the landscape nouns, this arguably relates to what was discussed in Section 2. While *mountain* and *flood* are by default sortal nouns in the narrow sense discussed by Löbner (i.e., they refer to non-unique and non-relational concepts), they might be used incongruently with their inherent concept type and be shifted into relational nouns when used as quantifiers.

In part, especially since body part nouns are already inherently relational and do not need to undergo a type shift to be used relationally, this can be accounted for if we presume that transferred senses that are not very strongly lexicalized (unlike the fully lexicalized *foot* in the sense of a unit of measurement) might need some sort of contextual support: if we are talking about the face of a clock, for instance, we need to somehow specify this (as in *the face of a clock*, *the clock's face*, or *its face*) – unless it is evident from the context that we are talking about a clock (rather than a person's face, which is presumably the default expectation when the word *face* is used), as in:

(11) *Harry liked <u>the clock</u>. It was completely useless if you wanted to know the time, but otherwise very informative. It had nine golden hands, and each of them was engraved with one of the Weasley family's names. There were no numerals around* ***the face****, but descriptions of where each family member might be.* (InterCorp v16 – English)

On a general level, I hope to have illustrated that metaphor and other phenomena giving rise to transferred senses, such as metonymy, are relevant for the study of valency and grammatical constructions more generally. Hopefully, further studies of similar phenomena might illuminate more clearly some of the reasons why the interaction between literal vs. non-literal senses and the presence vs. absence of complements might be so significant.

ACKNOWLEDGEMENTS

The preparation of this paper was financed within the long-term development program of the Czech Language Institute of the Czech Academy of Sciences (RVO No. 68378092).

References

BNC: The British National Corpus, version 2 (BNC World). Distributed by Oxford University Computing Services on behalf of the BNC Consortium. Praha: Ústav Českého národního korpusu FF UK. Accessible at: http://www.korpus.cz/.

InterCorp v16 – English: Klégr, A., Kubánek, M., Malá, M., Rohrauer, L., Šaldová, P., Šebestová, D., Vavřín, M., and Zasina, A. J. (2023). InterCorp v16 – English. Praha: Ústav

Českého národního korpusu FF UK. Accessible at: http://www.korpus.cz/.

Behaghel, O. (1932). Deutsche Syntax: Die Wortklassen und Wortformen. Heidelberg: Winter.

Brenner, D., Indefrey, P., Horn, C., and Kimm, N. (2014). Evidence for four basic noun types from a corpus-linguistic and a psycholinguistic perspective. In: D. Gerland – C. Horn – A. Latrouite – A. Ortmann (eds.): Meaning and Grammar of Nouns and Verbs. Düsseldorf: Düsseldorf University Press, pp. 21–48.

Bromhead, H. (2013). Mountains, Rivers, Billabongs: Ethnogeographical Categorization in Cross-linguistic Perspective. Ph.D. thesis. Australian National University.

Burenhult, N., and Levinson, S. C. (2008). Language and landscape: a cross-linguistic perspective. Language Sciences, 30(2–3), pp. 135–150.

Chomsky, N. (1970). Remarks on nominalization. In: L. Jacobs – P. Rosenbaum (eds.): Readings in English Transformational Grammar. Waltham: Ginn & Co., pp. 184–221.

Croft, W., and Cruse, D. A. (2004). Cognitive Linguistics. Cambridge: Cambridge University Press.

Dancygier, B., and Sweetser, E. (2014). Figurative Language. Cambridge: Cambridge University Press.

Grimshaw, J. (1990). Argument Structure. Cambridge: The MIT Press.

Keizer, E. (2007). The English Noun Phrase: The Nature of Linguistic Categorization. Cambridge: Cambridge University Press.

Koptjevskaja-Tamm, M. (2001). “A piece of the cake” and “a cup of tea”: Partitive and pseudo-partitive nominal constructions in the Circum-Baltic languages. In: Ö. Dahl – M. Koptjevskaja-Tamm (eds.): Circum-Baltic Languages. Volume 2: Grammar and Typology. Amsterdam: John Benjamins Publishing Company, pp. 523–568.

Lakoff, G., and Johnson, M. (1980). Metaphors We Live By. Chicago – London: The University of Chicago Press.

Lewandowska-Tomaszczyk, B. (2020). Polysemic chains, body parts and embodiment. In: I. Kraska-Szlenk (ed.): Body Part Terms in Conceptualization and Language Use. Amsterdam – Philadelphia: John Benjamins Publishing Company, pp. 31–52.

Lieber, R. (2016). English Nouns. The Ecology of Nominalization. Cambridge: Cambridge University Press.

Löbner, S. (1985). Definites. Journal of Semantics, 4(4), pp. 279–326.

Löbner, S. (2011). Concept Types and Determination. Journal of Semantics, 28(3), pp. 279–333.

Löbner, S. (2015). Functional Concepts and Frames. In: T. Gamerschlag – D. Gerland – R. Osswald – W. Petersen (eds.): Meaning, Frames, and Conceptual Representation. Düsseldorf: Düsseldorf University Press, pp. 15–42.

Mackenzie, J. L. (1997). Nouns are avalent – and nominalizations too. In: K. van Durme (ed.): The Valency of Nouns. Odense: Odense University Press, pp. 89–118.

Müller, H. H. (2000). Noun phrases in specialized communication. The cognitive processing of the Danish s-genitive construction. In: L. Lundqvist – R. J. Jarvella (eds.): Language, Text, and Knowledge. Mental Models of Expert Communication. Berlin – New York: Mouton de Gruyter, pp. 49–81.

Newmeyer, F. (2009). Current challenges to the lexicalist hypothesis: An overview and a critique. In: W. Lewis – S. Karimi – H. Harley (eds.): Time and Again: Theoretical Perspectives on Formal Linguistics. Amsterdam: John Benjamins Publishing Company, pp. 91–117.

Plag, I. (2003). Word-Formation in English. Cambridge – New York: Cambridge University Press.

Sláma, J. (2020). The Skewed Frequency Hypothesis and the identification of valent nouns in English. Prace filologiczne, 75(1), pp. 437–452.

Sláma, J. (2022). Konstrukční gramatika a popis české slovotvorby [Construction Grammar and the description of Czech word-formation]. Naše řeč, 105(5), pp. 247–266.

Sláma, J. (2024). Prefixálně-valenční konstrukce a jejich význam pro teorii slovotvorby [Argument structure constructions with prefixes and their relevance to the theory of word--formation]. In: I. Bozděchová – B. Niševa – K. Skwarska (eds.): Česká slovotvorná koncepce v kontextu slovanské jazykovědy. Praha: Academia, pp. 169–178.

Spevak, O. (2014). Editor's foreword. In: O. Spevak (ed.): Noun Valency. Amsterdam – Philadelphia: John Benjamins Publishing Company, pp. ix–xiii.

Sullivan, K. (2025). Metaphors and Constructions. In: M. Fried – K. Nikiforidou (eds.): The Cambridge Handbook of Construction Grammar. Cambridge: Cambridge University Press, pp. 129–145.

Wierzbicka, A. (2007). Bodies and their parts: An NSM approach to semantic typology. Language Sciences, 29(1), pp. 14–65.

ten Wolde, E. (2023). The English Binominal Noun Phrase: A Cognitive-Functional Approach. Cambridge: Cambridge University Press.

sciendo
DOI 10.2478/jazcas-2025-0009

MAPPING RECURRENT LEXICO-GRAMMATICAL PATTERNS IN ENGLISH THROUGH SUBTREE FRAGMENTS

ALEKSANDAR TRKLJA
Institute for Translation Studies, University of Innsbruck, Innsbruck, Austria
(ORCID: 0000-0002-7287-5338)

TRKLJA, Aleksandar: Mapping Recurrent Lexico-grammatical Patterns in English Through Subtree Fragments. Journal of Linguistics, 2025, Vol. 76, No 1, pp. 95 – 112.

Abstract: This paper examines subtree fragments (StF) as a corpus-informed method for identifying recurrent lexico-grammatical structures and compares them to two established approaches: collocational frameworks (Sinclair and Renouf 1988) and pattern grammar (Hunston and Francis 2000). StFs differ from these approaches in two major respects. First, they are grounded in a theoretical linguistic assumption that lexical heads project syntactic structures, incorporating part-of-speech categories, phrase structures, and thematic role assignment. Second, StFs are identified semi-automatically from parsed corpora by exploring patterns of grammatical words and syntactic categories, in contrast to the predominantly manual, concordance-based methods of the other two approaches. The findings suggest that StFs provide a productive interface between theory-driven syntactic analysis and data-driven corpus linguistics, allowing for fine-grained mapping between form, meaning, and use while retaining compatibility with probabilistic and statistical perspectives.

Keywords: subtree fragments, collocational frameworks, grammar patterns, thematic roles, argument structure, vector representation

1 INTRODUCTION

This paper proposes an approach to identifying subtree fragments (StFs) in corpora. StFs share similarities with both collocational frameworks (Renouf and Sinclair 1991) and grammar patterns (Hunston and Francis 2000). Developed during the peak of the Cobuild project, collocational frameworks refer to discontinuous sequences of two high-frequency grammatical words with a lexical word in between, such as *a + ? + of* or *too + ? + to*. Warren and Leung (2016) later proposed a broader definition of frameworks not limited to two grammatical words. Renouf and Sinclair regarded frameworks as genuine components of language rather than mere analytical tools although they provided no empirical references to support this claim. The key insight from their study is that grammatical words combine with each other to form regular 'scaffolds' into which certain lexical items fit. These combinations are not random but systematic, frequent, and selective. This systematicity makes frameworks valuable for investigating statistical tendencies, such as the distribution of lexical words within specific grammatical environments, as well as for identifying potential semantic classes. The classification of lexico-grammatical sequences into semantic categories is explored in greater depth in the pattern grammar approach proposed by Hunston and Francis (2000). Grammar patterns defined as "a phraseology frequently associated with (a sense of) a word, particularly in terms of the prepositions, groups, and clauses that follow the word" (Hunston and Francis 2000, p. 3). It is assumed that a pattern, together with all its lexical items, constitutes an extended unit of meaning (building on Sinclair 1996).

StFs proposed in the present paper are akin to both collocational frameworks and grammar patterns in that they concern the association of lexical items with sequences of grammatical words. However, as will be explained in the next section, StFs differ from these two notions in specific ways. I will then demonstrate how StFs can be identified semi-automatically in corpora, how their distribution can be explored, and how they can be classified using word embeddings and the information about thematic structures.

2 SUBTREE FRAGMENTS

2.1 Subtree fragments and their identification in corpora

Two major characteristics of both collocational frameworks and grammar patterns are that

i. they are explored without regard to syntactic structures as they are conventionally defined in theoretical linguistics, and
ii. they are identified through the manual exploration of concordance lines. The former follows from the general scepticism in Sinclairian corpus linguistics towards the notions from theoretical linguistics and from the idea that only minimal assumptions should be made when approaching language (Sinclair 1994; Mahlberg 2005). As Sinclair (1994, p. 25) puts it: "we should trust the text. We should be open to what it may tell us. We should not impose our ideas on it, except perhaps to get started. We should only apply loose and flexible frameworks until we see what the preliminary results are in order to accommodate the new information that will come from the text."

This view is understandable given the fact that it stems from lexicographic research, which attempts to provide item-specific descriptive information for practical uses. However, aside from ignoring decades of theoretical and empirical research in syntax, the problem with this view is that it risks treating structural generalisations as irrelevant or even obstructive. By focusing exclusively on surface co-occurrence patterns, such an approach loses explanatory depth since it does not account for why certain combinations are possible or impossible in terms of underlying grammatical relations. It also has limited generalisability, as observations remain tied to attested forms and do not easily extend to potential but unattested structures. Finally, it may lead to misclassification, grouping together formally similar sequences that are structurally distinct (for more details see Trklja, forthcoming).

As for the second common feature, at the time when this research was conducted, the main analytical tool in corpus linguistics was the concordance line, supported by tools for displaying collocations. Since then, both corpus resources and computational tools have developed considerably, making it possible today to automate to a much greater extent the exploration of patterning in corpora.

Subtree fragments (StFs) differ from collocational frameworks and grammar patterns in relation to the features discussed above. First, they are based on the generally accepted assumption in theoretical linguistics that lexical items project syntactic structure – an idea central to the Projection Principle in generative grammar (Jackendoff 1977; Chomsky 1981). In other words, the lexical properties of a head determine the syntactic configuration in which it can appear. As their name suggests, StFs are derived from syntactic trees (see below for more details). These syntactic structures are associated with semantic interpretation and contribute to the construction of thematic structures (theta-grids or argument structures) that encode the roles of participants in events (Williams 1994). Although there is no consensus on whether part-of-speech categories are universal – with Baker (2003) arguing in favour of universality and Croft (2001) arguing against – I will assume here that such categories do exist.

Second, in the present study StFs are identified semi-automatically by analysing the patterning of grammatical words and syntactic categories in corpora, rather than through the manual investigation of concordance lines. This involves using parsed corpora and computational tools capable of extracting and classifying structural configurations according to specified grammatical and lexical criteria. While some manual checking may still be required to ensure accuracy, the reliance on syntactic annotation and automated search distinguishes this approach from the purely concordance-based, manual methods used in the early studies of collocational frameworks and grammar patterns. Crucially, because StFs are grounded in syntactic theory, their identification and interpretation are linked to an explicit model of grammar, rather than to surface-level co-occurrence patterns alone.

What kinds of structures are StFs? The notion of subtrees used here is adopted from Bod (1995) and the following three generalizations define subtrees:

> “A subtree of a tree T is a subgraph t of T such that
> (1) t consists of more than one node
> (2) t is connected
> (3) except for the frontier nodes of t, each node in t has the same daughter nodes as the corresponding node in T” (Bod 1995, p. 36).

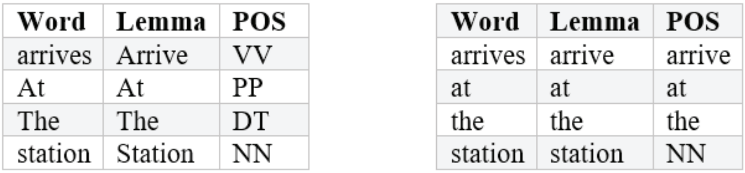
Unlike some other approaches that rely on the notion of subtrees or similar concepts (Aravind et al. 1975; Marcus 2001), Bod (1998) explicitly states that subtrees are elements of the speaker's linguistic experience. Bod (1998) argues that grammatical knowledge consists of a "statistical ensemble of language experiences" (Bod 1998). In this view, the corpus is regarded as a representation of the speaker's past language experience, and statistical learning is implicitly assumed as the mechanism through which this experience is encoded. The frequency with which utterances have previously been used influences the probability with which speakers will produce expressions and sentences in the future[1]. In particular,

> "this means that new utterances are constructed by combining fragments that occur in the corpus, while the frequencies of the fragments are used to determine the most probable utterance for a given meaning" (Bod 1998).

This does not mean that speakers are unable to produce novel sentences or expressions, but the proposal emphasises that previous experience contributes to the production of such units.

The level of detail in the representation, in terms of sub-trees derived from corpora, depends on the availability of annotation sets and will therefore vary from language to language. Grammatical information is encoded in corpora using parts-of-speech (POS) tag sets and/or syntactic parsers. For the purposes of this study, I will assume a sparse representation of functional categories using the TreeTagger PoS tagset (Marcus et al. 1993). The focus of the study will be on English verbs for illustrative purposes, making use of the English TreeTagger PoS tagset and the British National Corpus (BNC) (Leech 1992). This tagset contains the grammatical categories which are annotated with basic features. For example, verbs are annotated with information about tense. In the present study only the general grammatical information is included (e.g. V, N, A) with the lexical categories being represented without any grammatical features. Pronouns are included in the category N. No claim is made here that the data are representative of the English language as a whole or that register- and genre-specific differences are irrelevant. The present approach enables the identification of sequences of POS categories with function words (such as *V the N of the N*), as well as the combination of particular lexical words with POS categories and function words, (such as *find the A N)*. I will explore both types of StFs below. The tabular representation of PoS tags from TreeTagger for the expression *arrives at the station* is as follows:

[1] From a statistical learning perspective, this proposal aligns with findings in cognitive science and psycholinguistics showing that speakers are sensitive to distributional regularities in their linguistic input (e.g. Ellis 1996, 2002; Armstrong et al. 2017). Thus, high-frequency substructures become entrenched in memory and are more readily retrieved and recombined, whereas low-frequency or novel combinations are less predictable and may require greater processing effort.



Word	Lemma	POS
arrives	Arrive	VV
At	At	PP
The	The	DT
station	Station	NN

Word	Lemma	POS
arrives	arrive	arrive
at	at	at
the	the	the
station	station	NN

Tab. 1. Tabular representation of POS categories in a tagged corpus

I wrote a Perl script to identify StFs by detecting sequences of POS categories and function words. The script can also replace a POS category with the lemma form of a lexical item enabling StFs associated with a lexical word to be identified in a manner similar to the representation in the pattern grammar. In the next step, an n-gram function was used to compile combinations of the actual word within a defined window size. I explored n-grams of three, four and five words. To give an example, one StF associated with for the verb *arrive* is *arrive at the NN*, which occurs 715 times in the BNC. But, the resulting n-grams are not always grammatically complete sequences. Thus, the structure *find the N of* which is generated from the corpus is excluded because it contains a syntactically incomplete prepositional phrase. On the other hand, the structures such as *find the N of the N*, *find the N of a N* or *find the N of the A N* are regarded as StF because they constitute complete VP. At the final stage, all sequences were manually inspected, and only those forming a grammatically complete verb phrase (VP) were included for further analysis. Fig. 1 shows a tree representation of StFs for the verb *find*, with the four types of StFs identified in the BNC.

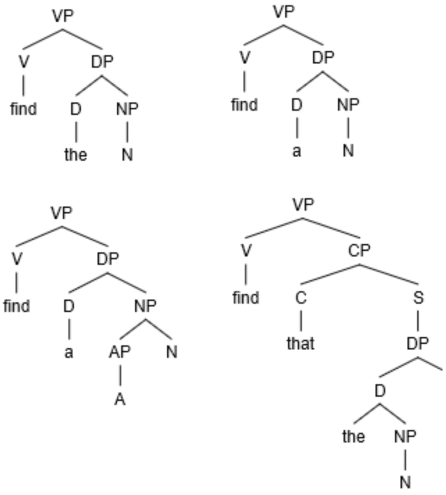
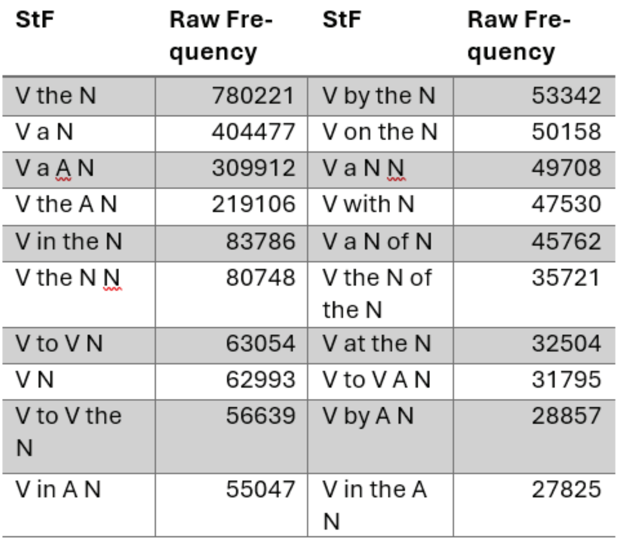


Fig. 1. Four StFs associated with the verb *find* identified in the BNC

2.2 Distribution of general StFs in the BNC

In this subsection I will explain how the distribution of StFs consisting of sequences of POS categories and function words was explored. For illustrative purposes, the present data focuses only on the most representative StFs defined as those that occur at least 1,000 times. In total, 84 StFs that build a VP were identified in this manner in the BNC. The top 20 StFs are presented in Tab. 2 and the comprehensive list can be found in Appendix A.



StF	Raw Frequency	StF	Raw Frequency
V the N	780221	V by the N	53342
V a N	404477	V on the N	50158
V a A N	309912	V a N N	49708
V the A N	219106	V with N	47530
V in the N	83786	V a N of N	45762
V the N N	80748	V the N of the N	35721
V to V N	63054	V at the N	32504
V N	62993	V to V A N	31795
V to V the N	56639	V by A N	28857
V in A N	55047	V in the A N	27825

Tab. 2. The 20 most frequent VP StFs identified in the BNC

Unlike grammar patterns, but like collocational frameworks, StFs include not only the obligatory elements of an argument structure but also modifiers. This has both disadvantages and advantages. The disadvantage is that it overlooks the fact that these instances still belong to the same verb phrase. The advantage is that it provides detailed information about the specific kinds of modifiers typically used. Both types of information can be explored further. In the present study, however, I focus on a more general classification. All subtrees were grouped into broader structural types. For example, the sequences *V the N*, *V a N*, *V the A N*, *V a A N*, *V the N of the N* are all classified into the same category: transitive verbs serving as the head of the verb phrase and selecting a determiner phrase (DP) as their complement. The final classification comprises 23 distinct classes (see Appendix B), encompassing a total of 84 individual StFs.

The initial descriptive statistics reveal a clear tendency in the distribution of VP across types. The most frequent structures (Type 1), such as *V the N* or *V a N*, involve direct NP complements typically associated with core arguments (e.g. Theme, Patient). In contrast, more complex or marked structures, such as those involving directional PPs (*V into the N*), resultative phrases (*V the A N to N*), or role-identifying as-phrases (*V as a N*), are markedly less frequent. The data indicate that Type 1 overwhelmingly dominates usage, accounting for approximately 66.2% of all VP subtree occurrences and 17% of all sequence types (Fig. 2). Other types are much less frequent, each contributing between 0.2% and 8.4%. As the second pie chart (Fig. 3) indicates structural diversity of VP is more evenly distributed across types, with many contributing around 2–6% of the total.

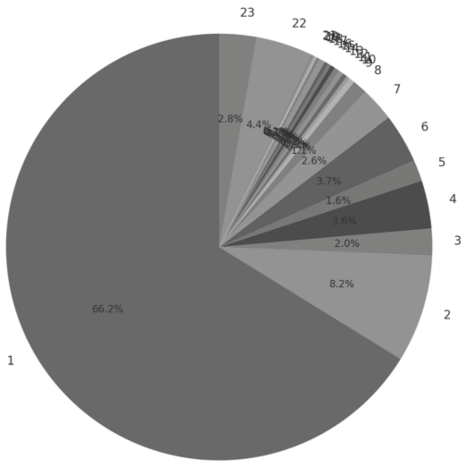


Fig. 2. Total frequency share per VP type

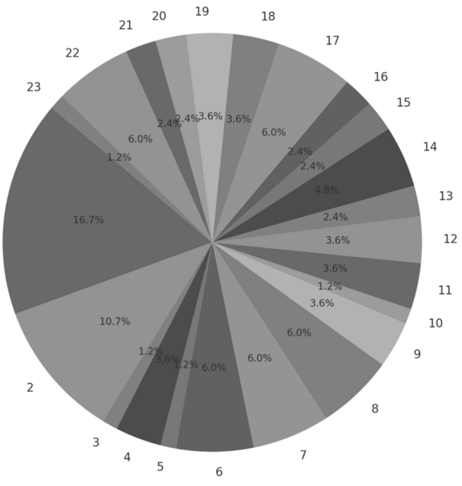


Fig. 3. Distribution of the number of StFs per VP type

To further investigate the data I explored syntactic variety and usage frequency associated with the present set of StFs. I define syntactic variety as the number of distinct StFs grouped under a given VP type (e.g. *V the N*, *V a N*, *V a N of N*), which reflects the degree of formal diversity or grammatical flexibility permitted by a given type. Usage frequency refers to the total number of occurrences of all StFs of a particular VP type.

One may assume that the two dimensions are positively correlated. In other words, constructions that are structurally more productive – that is, capable of supporting a greater number of grammatical variants – are also expected to occur more frequently in actual language use. In order to test this assumption empirically, I formulated the following hypothesis:

– H_0 (Null Hypothesis): There is no relationship between the syntactic variety of a VP type and its usage frequency in the corpus.
– H_1 (Alternative Hypothesis): There is a positive relationship between syntactic variety and usage frequency in the corpus.

To test this hypothesis, I conducted correlation and regression analyses using VP types classified into 23 categories. Descriptive statistics suggests that VP types with more subtree variants tend to show higher overall frequencies. For instance, Type 1 includes some of the most common VP patterns (e.g. *V the N*, *V a N*, *V the A N*), with 14 distinct subtree structures. This type accounts for 66% of the total frequency across all types (of 2,356,436 occurrences). However, this dominance was not matched by other types with comparable structural diversity. Type 6, which comprises different *with N* structures (e.g. *V with N*, *V with the N*) has a low frequency (of just 116,710 occurrences) Similarly, Type 7, which includes five variants appears 79,800 times in total. To determine whether the observed trend is statistically significant and generalisable, I applied Pearson and Spearman correlation tests. The Pearson test yields a strong linear correlation ($r = 0.830$, $p < .001$), and the Spearman rank correlation also shows a significant monotonic association ($\rho = 0.583$, $p = 0.0047$). These results allow us to reject the null hypothesis, suggesting that VP types with greater syntactic variety do tend to occur more frequently in corpus data. However, further analysis complicates this finding. A linear regression using raw frequency values indicates that syntactic variety explains 69.4% of the variance in frequency ($R^2 = 0.694$). Yet, this model was highly influenced by Type 1, a clear outlier with both high variety and extraordinarily high frequency. A second model using log-transformed frequency values reduces this distortion and explains 57.5% of the variance ($R^2 = 0.575$), showing that the association remains significant, but not uniformly strong across all types. The diminishing returns observed in the log-scale model further suggest that the relationship is not strictly proportional: each additional subtree type adds progressively less to the overall frequency. Taken together, these findings support a partial rejection of the null hypothesis. There is indeed a statistically significant relationship between syntactic variety and usage frequency but the relationship is not linear and is heavily skewed by a small number of functionally entrenched constructions. Type 1, as we saw, is not only syntactically diverse but also highly conventional and semantically general, which likely enhances its functional entrenchment which is a property that cannot be reduced to structural variety alone.

2.3 Investigation of specific StFs in the BNC

At the next stage, it is possible to investigate specific StFs and fine-grained semantic distinctions within a syntactically uniform pattern. I selected for illustrative purposes the StF *V the N* which belongs to Type 1. This subtree instantiates numerous semantically diverse expressions (e.g. *accept the offer*, *cut the cost*, *feel the pain*), making it a good candidate for further analysis. 200 of the most frequent *V the N* expressions was collected in the BNC and passed through the pretrained BERT-based model all-MiniLM-L6-v2 from the sentence-transformers library. This model produces high-dimensional vector representations (384 dimensions) for short texts, encoding rich semantic information learned from large corpora. These vectors were then subjected to KMeans clustering with k = 10 to discover semantically coherent groups. To visualize the structure of these clusters, a t-SNE projection was used to reduce the high-dimensional embeddings to two dimensions. Fig. 4 illustrates the spatial distribution of the clusters where each cluster is related to distinct semantic types. For example, one cluster groups expressions such as *accept the offer*, *assess the situation*, and *address the issue*, which all share a judgmental or evaluative function, with the noun denoting an abstract Theme or Proposition. Another cluster includes verbs like *buy the house*, *cut the cost*, and *cover the expense*, associated with economic transactions or resource manipulation, where the noun represents a Patient or Affected Object.

At the final stage, one may select a specific verb and analyse its distribution across StF-types. For the present purposes, I selected the verb *find* and explored its distribution across 2,000 concordance lines from the BNC. The results indicate that it occurs in the following StF: *find N*, *find the N*, *find a N*, *find the A N* (Type 1) and *find that S* (Type 23). I have excluded fragments containing the verb *find out* as this is a distinct lexical item. Unlike pattern grammar, this analysis does not indicate semantic interpretation where verbs and patterns are classified into semantic classes. The classification used in pattern grammar is based on intuition and ignores the higher argument structure representation. An alternative approach that I propose here is to explore the thematic structures of the fragments. Let us consider *find* as an example. Thematic roles assigned by find to its complement are typically Theme referring to something located or discovered as in *find the book* or *find a job*. Occasionally, however, the complement receives the role of Patient if it is affected by the action. This occurs in secondary predication constructions (Rothstein 1983, 2004), where *find* takes a DP complement together with an additional predicate that describes a state or property of that DP such as in *find the defendant guilty*, *find the room in a mess* or *find the door locked*. In these complex transitive uses the DP is both the object of *find* and the subject of the secondary predicate and is understood as undergoing or being in the state described and hence its interpretation as a Patient rather than a Theme. This suggests that the syntactically complete fragments in the latter case includes an additional element which can be realised either as a past participle, prepositional phrase or a predicative adjective. In *find that S* constructions, the complement expresses a proposition or a piece of information or a cognitive result: what is found to be true (e.g. *find that it was closed*). This indicates a cognitive or evaluative use of find (semantic overlap with realize or discover).

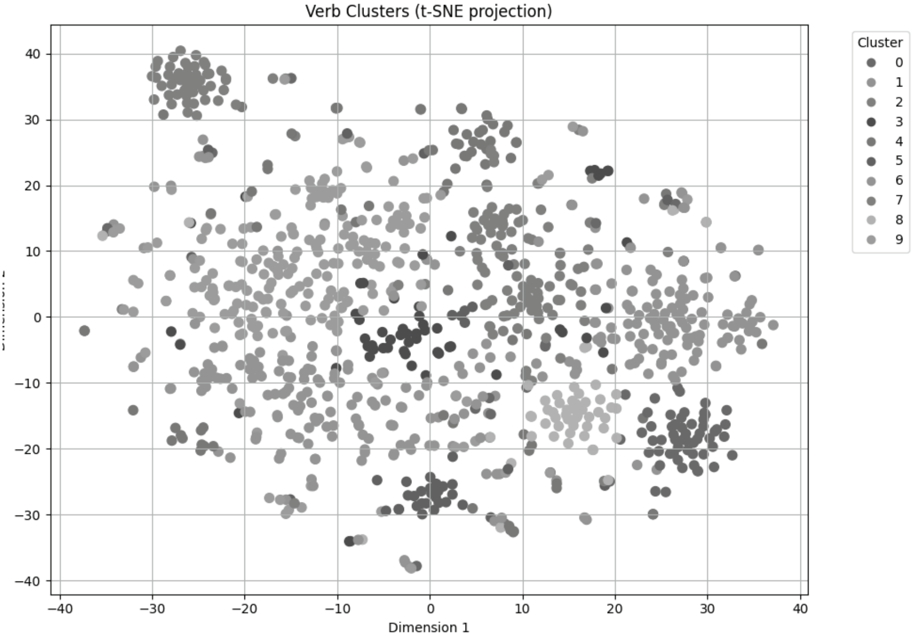


Fig. 4. Clusters of *V the N*-expressions from the BNC

In the current BNC sample, the DP that occur in subject position with the *find*-fragments predominantly fulfils the Experiencer or Cogniser role. But, in addition to its canonical argument structure (Experiencer finds Theme/Proposition), the verb *find* also supports extended argument realizations that introduce Source (*She found the message from John* and *He found her a job*), Beneficiary (*He found a gift for her*), and Means (*They found the solution with a tool*) roles via prepositional phrases or double object constructions.

Formally, this can be represented as:

FIND(x, y, [s], [z], [p])
where
- x = Experiencer (subject NP)
- y = Theme / Patient (Theme if s absent; Patient if s present)
- s = Secondary Predicate (optional; AdjP, Ved, V-ing, PP; makes y = Patient)
- z = Beneficiary (optional; NP or PP)
- p = Source / Asset / Instrument (optional; PP).

3 CONCLUSION

This study proposed the use of StFs as an analytical tool to explore lexical and syntactic patterns in corpora. The aim was to clarify the theoretical assumptions, methodological procedures and potential advantages of the StF approach in relation to existing corpus linguistic approaches, while situating it within the broader corpus linguistic and syntactic theoretical landscape. Unlike collocational frameworks and pattern grammar, which do not commit to syntactic categories beyond those minimally required for corpus annotation, StFs draw directly on syntactic structure and its semantic interpretation. This includes the assignment of thematic roles and the representation of argument structure. Secondly, StFs are identified using a semi-automatic method involving parsed corpora and the computational extraction of patterns defined over syntactic and lexical categories. This method relies less on manual inspection of concordance lines than the other two approaches. Overall, the StF method should offer a bridge between theory-driven and data-driven approaches. This combination enables a more precise mapping of form, meaning, and use than is possible with purely surface-based methods while accommodating probabilistic and statistical insights from corpus linguistics. The findings suggest that incorporating syntactic structure into corpus pattern analysis can enrich theoretical and applied descriptions of language, particularly in contexts where thematic role distinctions and variation in argument structure are important.

4 APPENDIX A: DISTRIBUTION OF THE MOST FREQUENT STFS IN THE BNC

Subtree fragments	Frequency of StFs in the BNC
V the N	780221
V a N	404477
V a A N	309912

V the A N	219106
V in the N	83786
V the N N	80748
V to V N	63054
V N	62993
V to V the N	56639
V in A N	55047
V by the N	53342
V on the N	50158
V a N N	49708
V with N	47530
V a N of N	45762
V the N of the N	35721
V at the N	32504
V to V A N	31795
V by A N	28857
V in the A N	27825
V with the N	26183
V with A N	25867
V to V a N	24230
V by N N	24191
V N of N	22543
V for the N	22028
V into N	21027
V in a N	20181
V into the N	18009
V by the A N	16696
V as a N	15752
V in a A N	13750
V N of the N	12218
V by a N	12199
V for a N	11465
V the A N of the A N	10385

V the A N of the N	10370
V with a N	9873
V through the N	9296
V with the A N	8335
V at the A N	8002
V over the N	7846
V for the A N	7787
V about the N	7768
V as a A N	7612
V into A N	7557
V out of the N	7013
V a N of the N	6261
V with a A N	6257
V as a A N	6104
V for a A N	6016
V into a N	5320
V off the N	5316
V N from N	4305
V under the N	4250
V N for the N	4228
V the N in the N	4152
V against the N	3953
V among the N	3953
V N to N	3712
V across the N	3584
V the N to N	3288
V N from the N	3223
V the N to the N	3060
V as the A N	2665
V after the N	2650
V N as a N	2553
V through the A N	2355
V N to the N	1983

V a N to N	1972
V about the A N	1895
V between N and N	1895
V over the A N	1800
V the N from the N	1771
V through a N	1729
V N as a N	1434
V over a N	1266
V about a N	1220
V N into N	1203
V the N in the A N	1184
V N as N	1135
V a N from the N	1042
V N into the N	1022
V after a N	1016
V the A N to N	1003

5 APPENDIX B: DISTRIBUTION OF THE STF-TYPES

Subtree fragments	**Frequency of StFs in the BNC**	**Type**
V the N	780221	1
V a N	404477	1
V a A N	309912	1
V the A N	219106	1
V the N N	80748	1
V N	62993	1
V a N N	49708	1
V a N of N	45762	1
V the N of the N	35721	1
V N of N	22543	1
V N of the N	12218	1
V the A N of the A N	10385	1
V the A N of the N	10370	1

V a N of the N	6261	1
V in the N	83786	2
V in A N	55047	2
V in the A N	27825	2
V into N	21027	2
V in a N	20181	2
V into the N	18009	2
V in a A N	13750	2
V into A N	7557	2
V into a N	5320	2
V to V N	63054	3
V to V the N	56639	4
V to V A N	31795	4
V to V a N	24230	4
V on the N	50158	5
V with N	47530	6
V with the N	26183	6
V with A N	25867	6
V with a N	9873	6
V with a A N	6257	6
V at the N	32504	7
V for the N	22028	7
V for a N	11465	7
V for the A N	7787	7
V for a A N	6016	7
V as a N	15752	8
V as a A N	7612	8
V as a A N	6104	8
V as the A N	2665	8
V N as a N	2553	8
V through the N	9296	9
V through the A N	2355	9
V through a N	1729	9

V at the A N	8002	10
V over the N	7846	11
V over the A N	1800	11
V over a N	1266	11
V about the N	7768	12
V about the A N	1895	12
V about a N	1220	12
V out of the N	7013	13
V off the N	5316	13
V N from N	4305	14
V N from the N	3223	14
V the N from the N	1771	14
V a N from the N	1042	14
V under the N	4250	15
V against the N	3953	15
V N for the N	4228	16
V among the N	3953	16
V N to N	3712	17
V the N to N	3288	17
V the N to the N	3060	17
V N to the N	1983	17
V a N to N	1972	17
V across the N	3584	18
V after the N	2650	18
V after a N	1016	18
V between N and N	1895	19
V N as a N	1434	19
V N as N	1135	19
V the N in the N	4152	20
V the N in the A N	1184	20
V N into N	1203	21
V the A N to N	1003	21
V by the N	53342	22

V by A N	28857	22
V by N N	24191	22
V by the A N	16696	22
V by a N	12199	22
V that S	87245	23

References

Armstrong, B. C., Frost, R., and Christiansen, M. H. (2017). ‘The long road of statistical learning research: past, present and future.’ Philos. Trans. R. Soc. B Biol. Sci. 372(1711).

Baker, M. (2003). Lexical Categories: Verbs, Nouns and Adjectives. Cambridge: Cambridge University Press.

Bod, R. (1998). Beyond Grammar: An experience-based theory of language. Stanford, CA: Center for the Study of Language and Information.

Chomsky, N. (1965). Aspects of the theory of syntax. Cambridge, MA: MITPress.

Croft, W. (2001). Radical construction grammar: Syntactic theory in typological perspective. Oxford: Oxford University Press.

Ellis, N. C. (1996). Sequencing in SLA: phonological memory, chunking and points of order. Studies in Second Language Acquisition, 18, pp. 91–126.

Ellis, N. C. (2002). Frequency effects in language processing. Studies in Second Language Acquisition, 24(2), pp. 143–188.

Jackendoff, R. (1977). X-bar Syntax: A Study of Phrase Structure. Cambridge, MA: MIT Press.

Hunston, S., and Francis, G. (1999). Pattern grammar. A corpus-driven approach to the lexical grammar of English. Amsterdam and Philadelphia: John Benjamins.

Leech, G. (1992). 100 million words of English: The British National Corpus (BNC). Language Research, 28(1), pp. 1–13.

Mahlberg, M. (2005). English General Nouns: A Corpus Theoretical Approach. Amsterdam/Philadelphia: John Benjamins.

Marcus, M., Santorini, B., and Marcinkiewicz, M. A. (1993). Building a large annotated corpus of English: The Penn Treebank. Computational Linguistics, 19(2) pp. 313–330.

Marcus, G. F. (2001). The Algebraic Mind: Integrating Connectionism and Cognitive Science. Cambridge: MIT Press.

Renouf, A., and Sinclair, J. McH. (1991). Collocational frameworks in English. In: K. Aijmer – B. Altenberg (eds.): English corpus linguistics: Studies in the honour of Jan Svartvik, pp. 128–143. London: Longman.

Rothstein, S. (1983). The syntactic forms of predication. Cambridge, MA: MIT.

Rothstein, S. (2004). Structuring events: A study in the semantics of lexical aspect. Malden, MA: Blackwel.

Sinclair, J. M. (1994). Trust The Text. In: M. Coulthard (ed.), pp. 12–25.

Sinclair, J. (1996). The search for units of meaning. Textus, 9(1), pp. 75–106.

Trklja, A. (forthcoming). ‘Distributional properties of near synonyms in lexical domains: A formal and metric-based approach.’ Corpora.

Warren, M., and Leung, M. (2016). Do Collocational Frameworks have Local Grammars?, International. Journal of Corpus Linguistics, 21(1), pp. 1–27.

Williams, E. (1994). Thematic structure in syntax. Linguistic inquiry monographs, Vol. 23. Cambridge, MA: MIT press.

sciendo
DOI 10.2478/jazcas-2025-0010

IDIOMS IN DISGUISE: HOW GRAMMATICAL PROFILING REVEALS PHRASEOLOGICAL PATTERNS

ANNA VYSLOUŽILOVÁ[1] – DOMINIKA KOVÁŘÍKOVÁ[2]
[1]Institute of Linguistics, Faculty of Arts, Charles University, Prague, Czech Republic
(ORCID: 0009-0001-0670-7993)
[2]Institute of Linguistics, Faculty of Arts, Charles University, Prague, Czech Republic
(ORCID: 0000-0002-4419-6901)

VYSLOUŽILOVÁ, Anna – KOVÁŘÍKOVÁ, Dominika: Idioms in Disguise: How Grammatical Profiling Reveals Phraseological Patterns. Journal of Linguistics, 2025, Vol. 76, No 1, pp. 113 – 122.

Abstract: This article examines how morphological anomalies – specifically, the unusually high frequencies of certain singular noun forms – can reveal idiomatic usage in Czech. Using data from the GramatiKat tool, 1,102 noun lemmas were analyzed, of which 28% participated in idiomatic expressions. The study identifies clear distributional patterns across grammatical cases, with idioms most frequent in the accusative, genitive, locative, and instrumental singular. Monocollocational idioms are distinguished, as they are associated with specific structural patterns. The results show that idiomatic expressions can influence morphological distributions and leave measurable traces in corpus data. The approach is further applicable to other parts of speech, such as verbs and adjectives, suggesting a broader role for grammatical profiling in the identification of idiomatic and phraseological patterns.

Keywords: grammatical profiling, morphological anomalies, idiomatic expressions

1 INTRODUCTION

In morphologically rich languages like Czech, lexemes rarely exhibit uniform frequency distributions across their paradigm forms (Janda and Tyers 2021). Instead, certain grammatical forms often occur with markedly higher frequency than others, creating distinctive grammatical profiles. This paper investigates whether such morphological anomalies – particularly nouns with unusually high frequencies in specific case forms – can serve as reliable indicators of idiomatic expressions in Czech.

Research into the frequency distribution of grammatical forms has a solid tradition in Czech linguistics (Jelínek, Bečka and Těšitelová 1961; Bartoň et al. 2009; Cvrček et al. 2010). These studies have established that grammatical anomalies often correlate with specific lexical combinations found in idiomatic expressions. As Kodýtek (in Cvrček et al. 2010) observes, morphological distributions are influenced by semantic factors: nouns denoting animate entities typically show higher frequencies in nominative forms, while inanimate nouns often display increased occurrences in genitive and accusative cases.

The analysis builds upon previous research by extending Kováříková's (in press) study of morphological anomalies in the dative singular and incorporating Vysloužilová's (Dittrichová 2024) findings on the relationship between morphological anomalies and multi-word units.

By analyzing 1,102 anomalous noun lemmas from the corpus tool GramatiKat (Kováříková and Kovářík 2021), this study addresses three questions: (1) Can paradigmatic imbalance indicate idiomatic expressions? (2) Which idiom types most frequently underlie such anomalies? and (3) How does the relationship between morphological anomalies and idiomaticity vary across grammatical cases? The findings reveal that over a quarter of lemmas displaying distributional outliers participate in idioms, with proportions exceeding 80% in certain cases, suggesting that morphological anomalies can serve as pathways for identifying phraseological patterns.

2 THEORETICAL FRAMEWORK

This study positions itself at the intersection of corpus linguistics, Construction Grammar, and phraseology. From a corpus linguistics perspective, our approach follows Sinclair's (1991) emphasis on examining actual usage patterns rather than linguistic intuitions, while employing frequency-based criteria for identifying phraseological units as outlined by Gries (2008). By using distributional outliers as the entry point for analysis, we employ a corpus-driven rather than corpus-based approach (Tognini-Bonelli 2001), allowing patterns to emerge from frequency data rather than testing predefined hypotheses.

Within Construction Grammar (Goldberg 1995; Croft 2001), linguistic patterns are understood as form-meaning pairings with varying degrees of fixedness and conventionality. When certain grammatical forms appear with unusual frequency in particular contexts, this often signals their entrenchment in linguistic usage.

The phraseological dimension builds on Čermák's (2007) conception of idioms as involving both formal and semantic anomaly. We developed a modified classification system to accommodate the specific patterns revealed in our corpus analysis. This approach investigates whether paradigmatic imbalance serves as an effective entry point for idiom identification across different grammatical cases.

3 DATA SOURCE AND SAMPLE SELECTION

This study utilizes the corpus tool GramatiKat (Kováříková and Kovářík 2021), which analyzes data from the SYN2015 corpus to provide detailed information on the distribution of word forms across word classes and specific lemmas. It compares these distributions to class-wide patterns through interactive tables that help identify lemmas with anomalous behavior.

The tool distinguishes between two anomaly types: **upper outliers** (lemmas with unusually high frequency of specific forms) and **lower outliers** (forms with very low or zero frequency). Upper outliers are defined as lemmas whose frequency in a given form exceeds 1.5 times the interquartile range above the 75th percentile; lower outliers typically lack any corpus attestation (Kováříková 2021).

This study focuses exclusively on upper outliers – noun lemmas with disproportionately high frequency in particular singular forms. Using GramatiKat's "Anomalous lemmas" function, we selected the "Noun" category and examined each singular case separately. To ensure comparability and manage sample size, we selected the top 20% of anomalous lemmas for each case based on frequency deviation scores. Lemmas with tied values at the cutoff point were also included.

The selection used GramatiKat version 1. For each case form, we exported, sorted, and thresholded the lemmas. The final sample also contained mistagged or duplicate lemmas, which were kept for transparency but excluded from idiom analysis. In total, 1,102 lemmas were analyzed out of 5,120 anomalous entries (see Tab. 1).

Case	**Total**	**Sample**
Nominative sg.	769	162
Genitive sg.	361	75
Dative sg.	1,169	252
Accusative sg.	321	66
Vocative sg.	839	201
Locative sg.	809	169
Instrumental sg.	852	177
Sum	**5,120**	**1,102**

Tab. 1. Number of anomalous lemmas per case and sample size (top 20%)

4 METHODOLOGY[1]

The analysis focused on identifying and classifying idiomatic constructions associated with the anomalous noun lemmas. All lemmas included in the study were drawn from the *GramatiKat* tool as described above. For each lemma, the specific anomalous form – typically a case form with unusually high frequency – served as the starting point for corpus exploration.

[1] A more detailed description of the methodology can be found in Vysloužilová's thesis (Dittrichová 2024).

The corpus analysis was conducted using the KonText application, drawing on two corpora, SYN2015 and SYNv11 (synchronic corpora of written Czech). Each lemma was searched in the form in which it exhibited the anomaly, using a corresponding CQL query adapted to the grammatical case.

Idiom identification employed two complementary approaches:

1. **Automatic annotation** using the FRANTA annotation tool, which tags multi-word units based on a predefined list of approximately 40,000 phraseological units, mostly from Čermák's *Slovník české frazeologie a idiomatiky* (Čermák 2009; Čermák and Hronek 2009a–c). We queried both the SYN2015 subcorpus within SYNv11 and the full SYNv11 corpus.
2. **Collocational analysis** using KonText's Collocations function with the following parameters: collocation window span of –3 to +3, minimum collocate frequency of 3, and sorting by the logDice association measure.

For idiom classification, we developed a custom typology with nine categories that combined structural and functional criteria, as the traditional tripartite typology of verbal, non-verbal, and propositional idioms (Čermák and Hronek 2009c) was found to be too coarse, while Čermák's detailed typology based on structural components (Čermák 2007) proved too fine-grained for the purposes of this study.

Six primary categories:

- Grammatical idioms (such as multi-word prepositions, e.g. *z hlediska* 'from the perspective of')
- Monocollocational idioms (containing a component with extremely limited collocability, e.g. *být k mání* 'to be available')
- Binomials (characterized by repetition of two formaly similar components, e.g. *alfa a omega* 'the alpha and omega')
- Similes (e.g. *žít si jako v bavlnce* 'to live in cotton wool')
- Contact idioms (e.g. *pozdrav pánbůh* 'God bless you')
- Foreign-language units (e.g. *alma mater*).

Three broader types for remaining idioms:

- Nominal idioms (e.g. *od malička* 'since childhood')
- Verbal idioms (e.g. *hodit zpátečku* 'to shift into reverse')
- Propositional idioms (e.g. *andělíčku, můj strážníčku, opatruj mi mou dušičku* 'little angel, my guardian, protect my little soul').

5 IDIOMATICITY ACROSS GRAMMATICAL CASES

Of the 1,102 anomalous noun lemmas analyzed, 28% (306 lemmas) participated in one or more idiomatic expressions. The distribution of idiomaticity varied markedly across grammatical cases, revealing significant asymmetries in how cases participate in phraseological patterns (Tab. 2).

Case	**Number of lemmas**	**Idiomatic lemmas**	**Percentage**
Nominative sg.	162	15	9%
Genitive sg.	75	39	52%
Dative sg.	252	47	19%
Accusative sg.	66	58	88%
Vocative sg.	201	10	5%
Locative sg.	169	71	42%
Instrumental sg.	177	66	37%
Total	**1,102**	**306**	**28%**

Tab. 2. Proportion of idiom-participating lemmas by case

The accusative singular exhibited the strongest correlation with idiomatic usage (88% of analyzed lemmas). These idioms frequently involved the preposition *na* and included numerous monocollocational idioms (expressions in which one component appears almost exclusively in that specific phrase) such as *dávat si bacha* ('to watch out') and *brát v potaz* ('to take into account'). Verbal idioms in this case often featured substantivized adjectives, as in *být na pováženou* ('to be questionable') or *dát někomu čas na rozmyšlenou* ('to give someone time to think it over'). Many of these expressions combined with the verb *dát/dávat* ('to give'), including *dát někomu na srozuměnou* ('to make something clear to someone') or *dát někomu něco na požádání* ('to provide something upon request'). Several idioms also referenced cultural or temporal contexts, such as *na Zelený čtvrtek* ('on Green Thursday') or *na doživotí* ('for life').

The genitive singular showed idiomaticity in 52% of cases and was associated with binomials more than other cases: *ani vidu, ani slechu* ('not a trace') and *bez ladu a skladu* ('without order or structure'). The genitive also appeared in numerous prepositional idioms with *bez, do*, and *od*, as in *bez prodlení* ('without delay'), *dostat se do ráže* ('to get fired up') or *od malička* ('since early childhood').

The locative singular displayed idiomatic usage in 42% of analyzed lemmas and was particularly rich in grammatical idioms, especially multi-word prepositional constructions with *v* or *na*: *v rámci* ('within the framework of') and *na základě* ('on the basis of'). The locative also featured monocollocational idioms like *být ve střehu* ('to be on alert') and *v mžiku* ('in an instant').

In contrast, nominative (9%) and vocative (5%) forms rarely participated in idioms. When they did, they typically appeared in contact idioms (ty vole – 'dude'), exclamatory formulas (*pane bože* – 'oh my God'), or foreign expressions (*alma mater, Ave Maria*).

Idiom type	**Nom. sg.**	**Gen. sg.**	**Dat. sg.**	**Acc. sg.**	**Voc. sg.**	**Loc. sg.**	**Instr. sg.**	**Total**
Grammatical idioms	0	1	0	0	0	11	3	**15**
Monocollocational idioms	1	7	3	13	0	10	8	**42**
Binomials	1	3	0	1	0	1	0	**6**
Similes	3	0	0	0	0	1	2	**6**
Contact idioms	3	0	0	4	8	0	2	**17**
Foreign-language units	3	2	0	0	0	0	0	**5**
Nominal idioms	2	17	4	13	0	19	19	**74**
Verbal idioms	2	9	40	27	1	28	31	**138**
Propositional idioms	0	0	0	0	1	1	1	**3**

Tab. 3. Idiom types by case (number of lemmas)

The identified idioms were classified into nine types based on structural and functional properties (see section 4). Verbal idioms emerged as the most prevalent, accounting for 138 lemmas and showing particular concentration in the dative, instrumental, locative, and accusative cases. Though less common, nominal idioms (74 lemmas) clustered notably in the locative and instrumental cases, with significant presence in the genitive and accusative as well. Monocollocational idioms, comprising 42 lemmas, revealed a widespread distribution pattern across multiple cases, particularly favouring the accusative and locative (more about this type in section 6). The analysis uncovered clear case preferences among certain idiom types – grammatical idioms appeared almost exclusively in the locative case, while contact idioms gravitated toward vocative. The remaining categories – binomials, similes, and foreign-language units – appeared infrequently in the corpus, with just 5–6 lemmas each distributed sparsely across different cases. This uneven distribution pattern confirms that idiom types do not spread randomly across grammatical cases but rather reflect underlying structural constraints and functional contexts of language use.

6 MONOCOLLOCATIONAL IDIOMS: STRUCTURE AND DISTRIBUTION

Among the nine idiom types identified, monocollocational idioms represent a particularly distinctive category characterized by containing components that rarely appear outside the specific idiomatic construction, creating strong lexical restrictions that contribute to morphological anomalies. They often contain the verb *být* (‘to be’) or a light verb (e.g. *dát*, ‘to give’, *mít* ‘to have’) combined with a fixed noun phrase, often introduced by a preposition.

The 42 monocollocational idioms identified in our sample exhibited clear distributional patterns across grammatical cases, with the accusative (13 lemmas), locative (10), instrumental (8), and genitive (7) showing the highest frequencies. This distribution indicates that certain cases offer especially favourable conditions for these fixed expressions, while others – notably the vocative, with no occurrences – do not support this idiom type.

6.1 Case-based distribution of monocollocational idioms

The accusative singular is especially productive for monocollocational idioms, typically following a verb + *na* + noun pattern. These often incorporate substantivized adjectives such as *srozuměnou* or *rozmyšlenou*:

- *dát na srozuměnou* (‘to make clear’)
- *dát na rozmyšlenou* (‘to give time to think’)
- *vystavovat něco na odiv* (‘to flaunt something’)
- *brát v potaz* (‘to take into account’).

The locative singular is also common, typically with *v*:

- *být ve střehu* (‘to be on alert’)
- *zmizet v propadlišti dějin* (‘to disappear into the abyss of history’)
- *v mžiku* (‘to be in an instant’)
- *v hloubi duše* (‘deep down’).

Instrumental singular idioms occur more often without prepositions:

- *zářit novotou* (‘to shine with novelty’)
- *končit fiaskem* (‘to end in a fiasco’)
- *nehnout ani brvou* (‘not even blink’)
- *mít něco za lubem* (‘to have something up one’s sleeve’).

Genitive singular idioms often involve *do*:

- *vyšumět do ztracena* (‘to fade away into nothing’)
- *nemít potuchy* (‘to have no idea’)
- *do třetice všeho dobrého* (‘third time’s the charm’)
- *dostat něco do vínku* (‘to be endowed with something at birth’).

6.2 Productive constructions beyond morphological anomaly

The monocollocational idioms identified in our study revealed several productive patterns, with *být/nebýt k* + noun in the dative singular standing out as particularly notable. Monocollocational expressions such as *být k mání* (‘to be available’), *být k nesnesení* (‘to be unbearable’), *být k popukání* (‘to be hilarious’), and *být k snědku* (‘ready to be eaten’) exemplify this pattern. While these constructions were initially identified as part of our search for morphological anomalies, their recurring formal structure suggested a more systematic phenomenon deserving deeper investigation.

Further analysis, as documented in Kováříková (in print), showed that the *být/nebýt k* + dative construction is far more productive than initially expected. This pattern encompasses dozens of items, many of which do not display the stark morphological anomalies that first drew our attention or do not qualify as strictly monocollocational. The identified construction *být/nebýt k* + noun is not merely a random collection of idioms but rather a partially schematic template that combines fixed grammatical elements (the verb *být* and the preposition *k*) with a variable nominal component.

This expanded perspective also revealed the existence of additional constructional types with similar syntactic foundations but different preposition-case combinations. For example, constructions with the accusative case and preposition *na* typically express states approaching a limit or breakdown: *být na spadnutí* ('to be about to collapse'), *být na vyhození* ('to be fit for disposal'), *být na vymření* ('to be on the verge of extinction'). In parallel, *být v* + locative constructions like *být v pokušení* ('to be tempted'), *být v ohrožení* ('to be in danger'), or *být v napětí* ('to be tense') typically denote internal or situational states that involve an element of danger, pressure, or tension.

These patterns demonstrate that the *být* + preposition + noun in a certain case frame represents a broader system of idiomatic expressions in Czech, within which the *k* + dative variant stands out for its productivity and formal coherence. This finding illustrates how initial observations about monocollocational idioms can lead to the discovery of more extensive constructional patterns that blur the boundary between grammar and lexicon.

7 CONCLUSION

This study has shown that grammatical anomalies – defined as unusually high frequencies of particular singular noun forms – can serve as useful indicators of idiomatic expressions. In many cases, what first appears to be a morphological irregularity turns out to reflect the influence of fixed multi-word combinations. When a noun occurs disproportionately in one case form, it is often because it regularly appears in a specific idiomatic construction. This tendency is especially clear in the accusative (88%), genitive (52%), locative (42%), and instrumental (37%) singular, whereas the nominative (9%) and vocative (5%) show minimal idiomatic usage. Of the 1,102 anomalous lemmas analyzed, 28% participated in idioms. Verbal idioms were the most frequent (138 lemmas), followed by nominal idioms (74) and monocollocational idioms (42). These findings demonstrate the potential of corpus-based methods to uncover idiomatic patterns that may remain unnoticed in dictionary-based or introspective approaches.

The findings suggest that paradigmatic imbalance can reflect syntagmatic regularity. Idioms and other multi-word constructions appear to influence the frequency of specific forms within a paradigm, shaping usage patterns in observable ways. This distributional signature is measurable through corpus analysis, suggesting that idiomaticity functions not just semantically but also as a morphological phenomenon with quantifiable effects.

While this study focused on nouns, the profiling method used in GramatiKat may also be useful for exploring distributional patterns in other word classes. Preliminary observations point to promising directions: in verbs, certain lexical items appear disproportionately in feminine or masculine forms. For instance, *háčkovat* (‘to crochet’), *zachichotat se* (‘to giggle’), or *proplakat* (‘to cry through’) are more frequent in feminine past tense forms, while *narukovat* (‘to enlist’), *vloupat se* (‘to break in’), or *habilitovat se* (‘to obtain habilitation’) occur more often in masculine animate. In adjectives, anomalies often arise from multi-word terms, where the gender of the adjective is determined by the head noun of complex noun phrases – for example, *akciová společnost* (‘joint-stock company’), *ministerská vyhláška* (‘ministerial decree’), or *vysoká škola* (‘university’). These regularities may be phraseological rather than idiomatic, but they still show how lexical, syntactic, and discursive conventions shape morphological distributions.

By combining computational anomaly detection with careful qualitative analysis, this research contributes to data-driven approaches to phraseology and grammatical profiling. The methodology presented here demonstrates how corpus evidence can complement traditional idiom identification methods, potentially uncovering patterns that might otherwise remain undetected using conventional approaches.

ACKNOWLEDGEMENTS

This work was supported by the European Regional Development Fund project “Beyond Security: Role of Conflict in Resilience-Building” (reg. No. CZ.02.01.01/00/22_008/0004595).

References

Bartoň, T., Cvrček, V., Čermák, F., Jelínek, T., and Petkevič, V. (2009). Statistiky češtiny. Nakladatelství Lidové noviny.

Croft, W. (2001). Radical Construction Grammar: Syntactic Theory in Typological Perspective. Oxford University Press.

Cvrček, V. et al. (2010). Mluvnice současné češtiny. 1, Jak se píše a jak se mluví. Karolinum.

Čermák, F. (2007). Frazeologie a idiomatika: česká a obecná = Czech and General Phraseology. Karolinum.

Čermák, F. (2009). Slovník české frazeologie a idiomatiky. 4, Výrazy větné. Leda.

Čermák, F., and Hronek, J. et al. (2009a). Slovník české frazeologie a idiomatiky. 2, Výrazy neslovesné. Leda.

Čermák, F., and Hronek, J. et al. (2009b). Slovník české frazeologie a idiomatiky. 3, Výrazy slovesné. Leda.

Čermák, F., and Hronek, J. et al. (2009c). Slovník české frazeologie a idiomatiky. 1, Přirovnání. Leda.

Dittrichová, A. (2024). Vyhledávání frazémů na základě anomálie v distribuci tvarů. Diploma thesis, supervisor Kováříková, D. Ústav českého jazyka a teorie komunikace, FF UK, Praha. Accessible at: http://hdl.handle.net/20.500.11956/188429.

Goldberg, A. E. (1995). Constructions: A Construction Grammar Approach to Argument Structure. Chicago: University of Chicago Press.

Gries, S. T. (2008). Phraseology and linguistics theory: A brief survey. In: S. Granger – F. Meunier (eds.): Phraseology: An interdisciplinary perspective, pp. 3–25. John Benjamins.

Janda, L., and Tyers, M. (2021). Less is more: why all paradigms are defective, and why that is a good thing. Corpus Linguistics and Linguistic Theory, 17(1), pp. 109–141.

Jelínek, J., Bečka, J. V., and Těšitelová, M. (1961). Frekvence slov, slovních druhů a tvarů v českém jazyce. Státní pedagogické nakladatelství.

Kováříková, D. (in press). Morfologické anomálie jako klíč k idiomatickým konstrukcím. Studie z korpusové lingvistiky, svazek 30. Nakladatelství Lidové noviny.

Kováříková, D. (2021). Sharing data through specialized corpus-based tools: the case of GramatiKat. Jazykovedný časopis, 72(2), pp. 531–544.

Kováříková, D., and Kovářík, O. (2021). GramatiKat. Nástroj pro výzkum gramatických kategorií a gramatických profilů. FF UK. Accessible at: http://www.korpus.cz/gramatikat.

Křen, M. et al. (2015). SYN2015: reprezentativní korpus psané češtiny. FF UK. Accessible at: https://www.korpus.cz.

Křen, M. et al. (2022). Korpus SYN, verze 11 ze 14/12/2022. FF UK. Accessible at: https://www.korpus.cz.

Machálek, T. (2014). KonText – rozhraní pro vyhledávání v korpusech. FF UK. Accessible at: http://kontext.korpus.cz/.

Sinclair, J. (1991). Corpus, Concordance, Collocation. Oxford University Press.

Tognini-Bonelli, E. (2001). Corpus Linguistics at Work. John Benjamins.

sciendo
DOI 10.2478/jazcas-2025-0011

THE SONORITY SEQUENCING PRINCIPLE IN HISTORICAL CZECH: A CORPUS-BASED STUDY

MARKÉTA ZIKOVÁ[1] – RADEK ČECH[2] – MARTIN BŘEZINA[3] – PAVEL KOSEK[4]
[1]Department of Czech Language, Faculty of Arts, Masaryk University, Brno, Czech Republic (ORCID: 0000-0002-0635-8893)
[2]Department of Czech Language, Faculty of Arts, Masaryk University, Brno, Czech Republic (ORCID: 0000-0002-4412-4588)
[3]Department of Czech Language, Faculty of Arts, Masaryk University, Brno, Czech Republic (ORCID: 0000-0002-6986-9754)
[4]Department of Czech Language, Faculty of Arts, Masaryk University, Brno, Czech Republic (ORCID: 0000-0001-6678-9989)

ZIKOVÁ, Markéta – ČECH, Radek – BŘEZINA, Martin – KOSEK, Pavel: The Sonority Sequencing Principle in Historical Czech: A Corpus-based Study. Journal of Linguistics, 2025, Vol. 76, No 1, pp. 123 – 131.

Abstract: This paper investigates the application of the Sonority Sequencing Principle (SSP) in historical Czech through a corpus-based approach. Drawing on texts from the 14th and 17th centuries, we examine the structure of word-initial and word-final consonant clusters with respect to both the strict and mild versions of the SSP. The results reveal two frequent types of violations: those involving liquids—specific to the diachronic development of Czech—and those involving sibilants, which are common cross-linguistically. Our findings provide new empirical evidence for the study of historical phonotactics in Slavic languages.

Keywords: Sonority Sequencing Principle, syllable structure, phonotactics, consonant clusters, corpus linguistics, historical Czech

1 INTRODUCTION

The Sonority Sequencing Principle (SSP) accounts for cross-linguistic phonotactic patterns in syllable structure. It states that syllables follow a universal sonority contour, with sonority peaking at the syllable nucleus and decreasing toward the margins, i.e. the onset and coda (Clements 1990; Zec 1995). This contour is determined by sonority—a scalar property reflecting relative loudness and acoustic energy—which follows a universal hierarchy: vowels > glides > liquids > nasals > fricatives > plosives (Parker 2011).

Although the SSP is considered a *phonological* principle, its application is strongly influenced by *morphological* structure. Specifically, the principle applies more strictly within words than at word edges, where SSP violations are more common. This asymmetry between word-internal and word-peripheral positions is illustrated by two Czech words: *rvala* ‘she tore’ and *larva* ‘larva’. Both are bisyllabic and contain the consonant cluster *rv*, in which a more sonorous liquid is followed by a less sonorous fricative. Whether this cluster is tautosyllabic or heterosyllabic depends on its position within the word.

In *rvala*, the cluster occurs word-initially and thus forms the onset of the first syllable (*rva.la*). In contrast, in *larva*, where the cluster is word-internal, the analogous syllabification is ungrammatical: **la.rva*. Instead, the cluster is heterosyllabic (*lar.va*), meaning that only the fricative forms the onset (of the second syllable), while the liquid serves as the coda (of the first syllable). In sum, the word-internal cluster *rv* conforms to the SSP, whereas the word-initial cluster in *rva.la* violates it. In the syllable #*rva*, sonority does not decrease away from the nucleus; instead, it rises within the onset, contrary to the expected sonority contour.

Such SSP-violating peripheral clusters are typical of Slavic languages, including Czech, and originate in their historical development (Bethin 1998). They evolved from Proto-Slavic forms containing *jer* vowels (ь/ъ), which were lost in weak positions. The SSP-violating form *rva.la* thus evolved from the SSP-conforming *rъ.va.la*.

In this paper, we examine peripheral consonant clusters in Czech. While the phonotactic patterns of contemporary Czech are relatively well documented—albeit typically without reference to the SSP (cf. Bičan 2013; Lukeš and Šturm 2017)—our analysis focuses on historical Czech, which remains understudied from this perspective. This research provides new empirical evidence that may serve as a foundation for future comparative work on the development of syllable structure in Czech.

The aim of this paper is to determine the extent to which peripheral clusters in historical Czech conform to the SSP, and the extent to which they violate it. In the case of violating clusters, we focus on distinguishing between accidental violations—those arising from the diachronic development of Proto-Slavic—and systematic ones, which are attested cross-linguistically and not limited to Slavic languages, as discussed, for example, in Yin et al. (2023).

The paper is organized as follows. Section 2 introduces the theoretical background of the SSP. Section 3 presents the corpus of historical Czech and outlines the methodology used for data extraction. Sections 4 provides the results of our corpus-based analysis and their interpretation, respectively. Finally, Section 5 concludes the paper.

2 THE SONORITY SEQUENCING PRINCIPLE

Following Parker (2011), we adopt a seven-level sonority scale, with vowels at the highest levels and obstruents at the lowest levels, as illustrated in Tab. 1. The table shows the correspondences between sonority classes, IPA segments, and graphemes.

sonority level	sonority class	segments	graphemes
7	non-high vowels	/a aː e eː o oː/	*a á e ě é o ó*
6	high vowels	/i iː u uː/	*i y í ý u ú ů*
5	glides	/j/	*j*
4	liquids	/r l/	*r, ŕ, l, ľ, ĺ, ł*
3	nasals	/m n ɲ/	*m n ň*
2	fricatives	/f v s z ʃ ʒ r̝ x ɦ/	*v, f, s, ś, z, ź, š, ž, ř, ch, h*
1	stops and affricates	/p b t d t͡s t͡ʃ c ɟ k g/	*p, b, t, d, c, ć, č, ť, ď, k, g*

Tab. 1. The sonority scale and the segmental inventory of historical Czech

Based on this scale, sonority profiles of words can be constructed, as illustrated below. Fig. 1 displays the sonority profiles of the words *trám* 'beam' and *nárt* 'instep', which contain consonant clusters in onset and coda positions, respectively. In both cases, the sonority profile features a single peak—formed by a vowel—which corresponds to the syllable nucleus (N). From this nucleus, sonority decreases toward both syllable margins, in accordance with the predictions of the SSP.

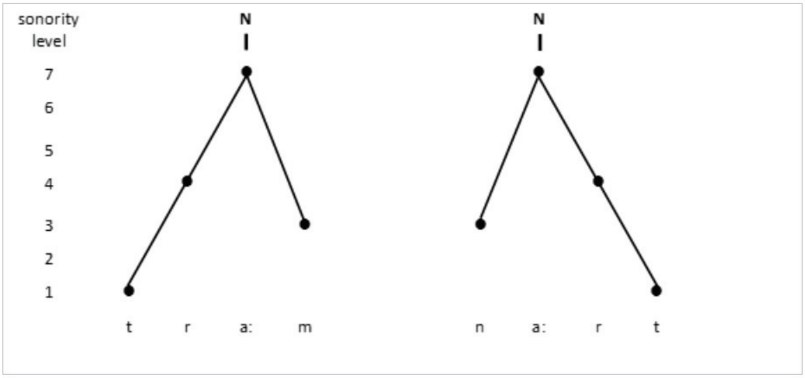


Fig. 1. Sonority profiles of *trám* and *nárt*

Fig. 2 displays two further examples attested in Czech, where additional sonority peaks are formed by consonants at word margins. In the word *lotr* 'rascal', the peak-forming consonant (a liquid *r*) appears word-finally. It functions as a syllable nucleus, and the resulting bisyllabic structure *lo.tr* is thus consistent with the SSP.

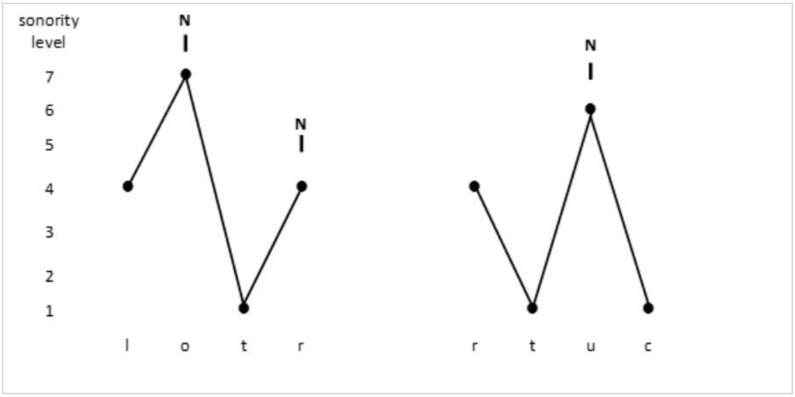


Fig. 2. Sonority profiles of *lotr* and *rtuť*

In the monosyllabic word *rtuť* 'mercury', the peak-forming liquid appears word-initially. Unlike in the previous example, however, it does not function as a syllable nucleus, and the onset cluster #*rt* therefore violates the SSP.

To distinguish between two types of peak-forming consonants attested in Czech, we adopt the terms *syllabic* consonants for those that function as syllable nuclei, and *trapped* consonants for those that violate the SSP (cf. Scheer 2009). In Czech, these two types differ in two main respects. First, they exhibit an asymmetrical distribution at word margins: syllabic consonants occur only at the right margin (as in *lo.tr* above), while trapped consonants appear at both the left margin (e.g. *r* in *rtuť*) and the right margin (e.g. *s* in *koks* 'coke'). Second, they differ in their segmental properties: syllabic consonants are limited to liquids (and sometimes nasals). Syllabic and trapped liquids are discussed further in Section 4.1.

Summing up, liquids that form sonority peaks in word-final position are syllabic consonants. Other consonants in word-final peak position are trapped and violate the SSP. Likewise, all peak-forming consonants in word-initial position are trapped.

Trapped consonants that form sonority peaks represent one SSP-violating type. A second type involves word-peripheral consonants such as the plosive *p* in *pták* 'bird' and *t* in *fakt* 'fact'. Like trapped consonants, these plosives violate the SSP because sonority does not decrease toward the onset margin in #*pt* or toward the coda margin in *kt*#. However, unlike trapped consonants, these plosives do not form sonority peaks, as they are adjacent to another consonant of the same sonority level.

The contrast between trapped consonants such as *r* in #*rt* and non-trapped SSP violations such as *p* in #*pt* is sometimes captured by distinguishing two versions of the SSP: a strict version and a mild version. According to the strict version, sonority must decrease continuously from the syllable nucleus toward both syllable margins. In contrast, the mild version requires only that sonority must not rise toward the onset or the coda. Under the strict SSP, both *#rt* and *#pt* violate the principle, as sonority does not decrease in either cluster. Under the mild version, however, *#pt* is a well-formed onset: it constitutes a sonority plateau, where sonority remains constant but does not rise, unlike in the trapped cluster *#rt*.

In what follows, we test both versions of the SSP on historical Czech data. Our aim is to provide a typology of SSP violations attested in a corpus of Czech texts written between the 14th and 17th centuries, and to interpret them from a broader cross-linguistic perspective.

3 CORPUS DATA EXTRACTION

The corpus of historical Czech used in our analysis is based on 26 texts of varying token counts, written between the 14th and 17th centuries.[1] It comprises 113,159 tokens, understood as graphical words, i.e. sequences of graphemes delimited by spaces on both sides. To obtain relevant data for testing the validity of both the strict and mild versions of the SSP, tokens were processed according to the following algorithm.

In the first step, non-syllabic prepositions such as *k* 'to', *s* 'with', *v* 'in', and *z* 'from', were joined with the tokens that followed them. This means that originally separate tokens in prepositional phrases—such as *z toho* 'from it'—were treated as a single unit for analysis, yielding onsets like *#zt*. The rationale behind this step is that non-syllabic prepositions do not form independent phonological units; they always procliticize to the following word. Similarly, the non-syllabic second-person auxiliary *s* forms a single unit with the preceding host, as in the verbal token *byls* 'you were'. In this case, however, the *s*-encliticization is already reflected in the orthography.

In the next step, each token was annotated according to the sonority scale presented above in Tab. 1. The annotation was carried out automatically using a sonority parser developed by the authors, as described in Ziková et al. (2023).

From these sonority annotated data, we extracted word-initial and word-final demisyllables. The demisyllable is defined as a unit consisting of an onset or coda together with its adjacent vowel nucleus. For example, a word *křtalt* 'character' contains the word-onset demisyllable *#křta* and the word-coda demisyllable *alt#*.

[1] Rather than working with original manuscripts, we relied on published editions to facilitate automatic processing. We used the following critical editions. Cejnar, J. (1964) Nejstarší české veršované legendy; Daňhelka, J. (1952) *Husitské skladby Budyšínského rukopisu*; Flajšhans, V. (1882) *Staročeská píseň o božím těle ze XIII. století*; Hrabák, J. – Vážný, V. (1959) *Dvě legendy z doby Karlovy*; Janošík-Bielski, M. (2008) *Modlitba Kunhutina*; Kolár, J. (1959) *Frantové a grobiáni: z mravokárných satir 16. věku v Čechách*; Lomnický z Budče, Š. (1572) *Instrukcí aneb Krátké naučení*; Patera, A./Černá, A. M. (1881/2008) *Hradecký rukopis*; Vážný, V. (1963) *Alexandreida*; Žampachová, K. (2021) *Umučení rajhradské – edice a jazykový rozbor*.

In the next step, we extracted the consonantal parts of these demisyllables, yielding 702 word-onset types (e.g. *#křt*) and 132 word-coda types (e.g. *lt#*). The observed asymmetrical distribution of onsets and codas confirms the tendency toward open syllables, that is syllables lacking codas. This finding aligns with previous observations for contemporary Czech, as reported by Lukeš and Šturm (2017).

The collected word onsets and codas were then analyzed according to both versions of the SSP, as presented in the following section.

4 TWO TYPES OF SSP VIOLATIONS

A closer inspection of the data reveals two frequent types of SSP violations, involving clusters with liquids and sibilants, respectively. While the first type is specific to the historical development of Czech (and Slavic languages more generally), the second type is well attested beyond Slavic as well.

4.1 Liquids

Liquids are default components of well-formed complex onsets (as in *trick*) or codas (as in *culp*). Moreover, liquids are often syllabic. Although consonant strings containing syllabic liquids such as *pl#* in *people* may appear to violate the SSP, this is not the case, as they are not true clusters: the syllabic liquid forms the syllable's nuclear peak, just like a vowel.

In Section 2, we discussed word-final syllabic liquids (such as *r* in *lo.tr*). In addition to these, word-internal syllabic liquids are also attested in Czech. For example, the words *držet* 'to hold' and *pokrm* 'food' may at first appear to contain an onset cluster *#drž* or a coda cluster *krm#*, respectively. However, these are not true clusters, as the liquid is syllabic. As a result, bisyllabic forms *dr.žet* and *po.krm* are fully consistent with both versions of the SSP.[2]

In addition to clear-cut cases of syllabic liquids, our corpus also includes ambiguous forms such as *slza* 'tear' or *řekl* 'he said'. These words may be either monosyllabic or bisyllabic, and we currently lack sufficient evidence to definitively support either interpretation; see Ziková et al. (2025) for details on the method used to distinguish between the liquid types.

This ambiguity stems from diachronic development. Originally, these words were monosyllabic, containing trapped liquids that violated the SSP. Over time, however, these trapped liquids gradually shifted to SSP-conforming syllabic consonants between the 14th and 16th centuries, resulting in bisyllabic forms *sl.za* and *ře.kl*, as attested in contemporary Czech.

[2] Recall that our analysis is restricted to complex onsets and codas. This criterion also applies to tokens with syllabic liquids. Accordingly, words like *čtvr.tek* 'Thursday' (with the complex onset *#čtv*) and *prst* 'finger' (with the complex coda *st#*) were included in the analyzed sample, whereas words like *dr.žet* and *po.krm*, which contain both a simple onset (*#d*, *#p*) and a simple coda (*t#*, *m#*), were excluded.

As shown in Tab. 2, these ambiguous forms constitute a relatively large share of all recorded SSP violations in word codas—approximately one third under the mild version and one quarter under the strict version of the SSP.

	mild SSP		strict SSP	
	word onset	word coda	word onset	word coda
all violating types	418	43	574	57
violating types with liquids	66	15	67	15
proportion of liquid types	15.8%	34.8%	11.6%	26.3%

Tab. 2. The proportion of SSP-violating types with ambiguous liquids

4.2 Sibilants

The second type of SSP-violating forms attested in our corpus involves sibilants. Unlike ambiguous liquids, which are specific to historical Slavic, sibilants violate the SSP cross-linguistically. For example, Harris (1994) shows that nearly every two-segment word onset in English that conforms to the strict SSP has a corresponding variant expanded by a sibilant that violates both versions of the SSP; cf. pairs such as /pr/*ize* – /spr/*ead*, /tr/*ick* – /str/*ike*, or /pl/*ain* – /spl/*it*, where the first member conforms to the SSP in onset position, while the second violates it due to the initial sibilant.

The reason why sibilants behave in this specific way remains a matter of debate in the literature (Goad 2011). Nevertheless, it is clear that sibilants—including fricatives /s z ʃ ʒ/ and affricates /t͡s t͡ʃ/—contribute significantly to SSP violations at word edges in historical Czech as well, as illustrated in Tab. 3. The table shows that both word-onset clusters with sibilants (e.g. *#spr* in *spravuje* '(s)he manages') and word-coda clusters (e.g. *dž#* in *poněvadž* 'whereas') account for more than 50% of all violating types across all examined parameters.

	mild SSP		strict SSP	
	word onset	word coda	word onset	word coda
all violating types	418	43	574	57
violating types with sibilants	299	23	400	33
proportion of sibilant types	71.5%	53.5%	69.7%	57.9%

Tab. 3. The proportion of SSP-violating types with sibilants

5 CONCLUSION

This paper has examined the extent to which the Sonority Sequencing Principle is reflected in the syllable structure of historical Czech. Using a corpus-based approach, we evaluated both strict and mild versions of the SSP and identified substantial violation rates in both word-initial and word-final positions.

The most frequent sources of these violations are sibilants and, to a lesser extent, liquids that alternate between trapped and syllabic status. The latter type of violation is specific to historical Czech and is connected to the evolution of Proto-Slavic *jer* vowels. Since all trapped liquids were gradually eliminated in word-internal and word-final positions, we expect SSP conformity to differ significantly between historical and contemporary Czech in this regard.

In contrast, sibilants are well-known SSP violators cross-linguistically. In this respect, historical Czech follows a general pattern, and we expect the same for contemporary Czech. Moreover, sibilants are often involved in morphologically complex clusters. For example, English features three productive affixes consisting of the sibilant /s/ that mark possessive, nominal plural, and verbal agreement. The concatenation of these sibilant markers produces complex codas, as seen in *hy*/mn's/, *atte*/mpt-s/, and *he tru*/st-s/, which are not found in monomorphemic words.

A similar situation is observed in Czech. As mentioned above, sibilants appear both in proclitic prepositions and in the enclitic verbal auxiliary. As a result, they give rise to phonotactically specific clusters, such as the onset #*zž* in the prepositional phrase *z života* 'from (the) life', which, once again, have no counterparts in morphologically simplex words. Since the set of proclitic prepositions also includes non-sibilant forms like *v* 'in' and *k* 'to', these too contribute to the formation of word onsets that violate the strict version of the SSP, as #*vb* and #*kpr* in prepositional phrases *v bázni* 'in fear' and *k práci* 'to work', respectively.

In general, morphological complexity undoubtedly influences violations of the SSP. The extent to which this applies to both historical and contemporary Czech—where proclitic prepositions also serve as productive verbal prefixes—remains an open question for future research.

ACKNOWLEDGEMENTS

We express our gratitute to Tomáš Vlček for his assistance with the processing of historical data. This research was supported by the Czech Science Foundation under the project *Development of Syllabic Sonorants in Czech* (GA23-04719S), awarded to Markéta Ziková.

References

Bethin, Ch. (1998). Slavic prosody. Language change and phonological theory. Cambridge: Cambridge University Press.

Bičan, A. (2013). Phonotactics of Czech. Frankfurt am Main: Peter Lang.

Clements, G. N. (1990). The role of the sonority cycle in core syllabification. In: J. Kingston – M. E. Beckman (eds.): Papers in laboratory phonology I: Between the grammar and physics of speech. Cambridge: Cambridge University Press, pp. 283–333.

Goad, H. (2011). The representation of *s*C clusters. In: M. van Oostendorp et al. (eds.): The Blackwell companion to phonology. Oxford: Wiley-Blackwell, pp. 898–923.

Harris, J. (1994). English sound structure. Oxford: Blackwell.

Parker, S. (2011). Sonority. In: M. van Oostendorp et al. (eds.): The Blackwell companion to phonology. Oxford: Wiley-Blackwell, pp. 1160–1184.

Scheer, T. (2009). Syllabic and trapped consonants in the light of branching onsets and licensing scales. In: G. Zybatow et al. (eds.): Studies in formal Slavic phonology, morphology, syntax, semantics, and information structure. Frankfurt am Main: Peter Lang, pp. 411–426.

Šturm, P., and Lukeš, D. (2017). Fonotaktická analýza obsahu slabik na okrajích českých slov v mluvené a psané řeči. Slovo a slovesnost, 78(2), pp. 99–118.

Yin, H., van de Weijer, J., and Round, E. (2023). Frequent violation of the sonority sequencing principle in hundreds of languages: How often and by which sequences? Linguistic Typology, 27(1), pp. 131–175.

Zec, D. (1995). Sonority constraints on syllable structure. Phonology, 12, pp. 85–129.

Ziková, M., Březina, M., Čech, R., and Kosek, P. (2023). Syllabic consonants in historical Czech and how to identify them. Jazykovedný časopis, 74(1), pp. 391–400.

Ziková, M., Březina, M., Čech, R., and Kosek, P. (2025). The shift away from the marked: Syllabic consonants in historical Czech. Glossa: a journal of general linguistics, 10(1), pp. 1–24.

sciendo
DOI 10.2478/jazcas-2025-0012

A QUANTITATIVE ANALYSIS OF SLOVAK AND CZECH SUPREME COURTS DECISIONS

MIROSLAV ZUMRÍK
Slovak National Corpus, Ľudovít Štúr Institute of Linguistics, Slovak Academy of Sciences, Bratislava, Slovakia (ORCID: 0000-0001-5749-082X)

ZUMRÍK, Miroslav: A Quantitative Analysis of Slovak and Czech Supreme Courts Decisions. Journal of Linguistics, 2025, Vol. 76, No 1, pp. 132 – 141.

Abstract: The paper focuses on the comparison of the quantitative characteristics of the decisions of the Slovak and Czech supreme courts, using publicly available databases of their decisions, which can be analyzed from the point of view of quantitative linguistics and stylometry using the Czech application QuitaUp and further investigated mainly by means of the non-parametric Mann-Whitney U test. The aim of the paper is to identify possible statistically significant differences in the representation of selected quantitative measures in the samples of decisions of both courts and to consider the possibilities of how to interpret these differences in terms of a closer comparative stylistic research aimed at Slovak and Czech legal texts.

Keywords: comparative legal linguistics, quantitative linguistics, legal style, Slovak and Czech law, institutional communication

1 INTRODUCTION

Law and language are deeply intertwined. This relationship is already evident from the fact that communication between the law-making subject and the addressees, the recipients of law, takes place exclusively through language (Knapp 2024, p. 179). Equally intertwined are the cognition of law and the cognition of language, since one of the prerequisites for the interpretation of authoritative texts is the cognition of the peculiarities of the language in which these texts are formulated (Holländer 2012, p. 284). The latter statement can be seen as the background for the emergence of the now well-established domains of legal linguistics (e.g. Cvrček 2016), law and corpus linguistics (Mouritsen 2010), corpus linguistics in legal discourse (Goźdź-Roszkowski 2021) or legal corpus linguistics (Bernstein 2021). This paper follows the line of corpus-based and quantitative investigation of legal texts, namely judicial decisions of the Supreme Court of the Slovak Republic and the Supreme Court of the Czech Republic.

In doing so, it draws on extensive databases of decisions of these courts[1], as well as the Slovak Ministry of Justice[2]. The anonymized decisions in these databases

[1] https://www.nsud.sk/rozhodnutia/; https://sbirka.nsoud.cz/

[2] https://www.justice.gov.sk/sudy-a-rozhodnutia/sudy/rozhodnutia/a štatistiky

are available as PDF files, which are explored in the paper using online calculator *QuitaUp*[3]*. The tool allows the analysis of texts in both Slovak and Czech, as well as other languages* using 16 quantitative measures, including h-point, hapax legomena frequency, entropy or average token length. From these, the following seven were selected: 1. token frequency, 2. type frequency, 3. verb distance, 4. activity, 5. descriptivity, 6. secondary thematic concentration, 7. moving average type-token ratio. The values for the above measures for in samples of Slovak and Czech Supreme Court decisions were then analyzed using freely available statistical calculators from the *Statistics Kingdom* portal[4].

Using these sources and tools, the paper examines the following questions:

a) How can the relationship of the two samples be characterized in terms of these quantitative measures and the statistical significance of any differences in their values?

b) How can these findings be interpreted in terms of stylistic differences between Slovak and Czech Supreme Court decisions?

2 THEORY

Legal texts of various genres (laws, decisions, etc.) are produced in large volumes and are often freely accessible to general public, which makes them an available and potentially interesting object for linguistic research (Smejkalová 2021; Wilfling 2013; Bobek 2010). For example, the open database of the Supreme Court of the Slovak Republic contains a total of 102,072 decisions, with the first ones dating back to 1970. The database of the Supreme Court of the Czech Republic does not indicate the total number of decisions, but the first ones date back to 1956. The database of published anonymised Slovak judicial decisions of courts of all types, published at the portal of the Slovak Ministry of Justice contains a total of 4,447,053 decisions that have been published since 2000[5].

The linguistic investigation of general, quantitative properties may also prove interesting in the field of comparing global and more local, national legal systems of different legal cultures (Fábry, Kasinec and Turčan 2019, pp. 152–158). It can be suggested that the institutional and conceptual differences in the various systems and subsystems should also be reflected in the legal communication and style of legal language within these systems. In terms of the stylistic difference between Civil and Common Law, then, we compare (a) the ratio of the predominance of deductive and analytical approaches in decision-making, (b) the ratio of formalism and value-oriented arguments (Smejkalová 2021, p. 231). While a civil law judge decides with

[3] https://korpus.cz/quitaup/

[4] https://www.statskingdom.com

[5] https://www.justice.gov.sk/sudy-a-rozhodnutia/sudy/rozhodnutia/

regard to a particular situation and retrospectively, a common law judge also prospectively makes future law, as his decision can serve as a precedent. In terms of style, this difference translates, for example, into the style of reasoning, which in common law should be more comprehensive and also such that a general rule can be extracted from it in the future (ibid.).

For an analysis of the Czech style in this regard, we can refer to the work of Matczak, Bencze and Kühn (2010), who compared different types of reasoning in decisions of administrative courts in Poland, Hungary and the Czech Republic in 1999 – 2004. They found that judges in Poland and Hungary tend to apply legal rules in a formal way, and value-based or principled reasoning associated with teleological interpretation remains the domain of mainly higher court instances. Only Czech courts have features of a certain "deformalizing", i.e. not purely formal, reasoning, which, according to the authors of the survey, hints at the stimulating "educational" role of the Czech Constitutional Court (Smejkalová 2021, pp. 232–233). The decisions of the Civil Law and the Common Law can, according to Terezie Smejkalová can be perceived as mutual opposites, while the style of Czech court decisions seems to balance "somewhere in the middle" (ibid., p. 233). Referring to Zdeněk Kühn, a "hybrid" model of "complex sophisticated subsumption" can be applied to Czech judicial decision-making, which is characterized by the fact that the Czech judge tries to support each conclusion with multiple arguments (ibid.).

Here one may ask whether the style of Slovak judicial decisions is similar to the Czech one. On the one hand, given the long and long-studied common history and the period of common statehood, one might expect similarities between Czech and Slovak law, and thus also similarities in the styles of legal texts. On the other hand, differences in the culture and law of the two countries exist and are reflected, whether through scholarly journals or at the institutional level in the form of, for example, regular bilateral contacts between the constitutional courts of the two countries.

The motivation for the analysis of the decisions of, in particular, the Supreme Court of the Slovak Republic[6] and the Supreme Court of the Czech Republic[7] was twofold. On the one hand and given the complexity of law and legal institutions, it was necessary to limit oneself to one type of decision-making body. On the other hand, it can be assumed that the supreme courts, as the highest instances of judicial decision-making in both countries, deal with rather complex cases of law application. This legal complexity may manifest itself even in the complexity of the textual form of the decision, which would make the texts an even more potentially interesting object of linguistic analysis compared, for example, to the more concise and formalized criminal orders of the first instance courts, in cases dealing with, e. g., driving under influence. Furthermore, given the complexity of the various legal

[6] https://www.nsud.sk/postavenie-a-posobnost/

[7] https://www.nsoud.cz/o-nejvyssim-soudu/obecne-informace

domains, it was also necessary to focus on one particular domain. In this case, the fairly prominent domain of criminal law was chosen. According to the number of published decisions in the database of the Ministry of Justice of the Slovak Republic, criminal cases are not among the most numerous. 384,330 of them have been published (as of April 2025), compared to decision count in civil law (1,909,279), family law (514,221) and commercial law (407,388).[8]

3 METHODOLOGY

3.1 Selection of data

The basic sources were the databases of decisions of the Supreme Court of the Slovak Republic (furthermore referred to as SK) and the Supreme Court of the Czech Republic (CZ). Only in the case of the latter court it is possible to filter decisions according to their genre (judgment, resolution etc.). Since each text was to be tested in terms of seven quantitative measures, a smaller size was chosen, n=20 for Slovak and n=20 for Czech decisions. The individual texts to be analyzed using QuitaUp were selected by generating 20 random numbers from the range 1 – 200. The range of randomly generated numbers for the Slovak and Czech decisions then served as a key to identify the decisions in either sample as they are added to the decision database over time (starting with the most recently published decisions). This process resulted in the selecting 19 resolutions and 1 judgment for the Slovak court, and 19 resolution and 1 declaratory judgment for the Czech one. The lengths of texts are specified in this table, while it was found that texts of one genre (at least in the sample) are not necessarily longer than texts of the other genre:

	SK	CZ
Token range	780–13418	1710–14050
Standard deviation	3549	3051
Average token count	4483	7555
Median token count	3426	6575

Tab. 1. Length characteristics of the decisions

3.2 Selection and description of measures

When selecting measures, preference was given to those that are less dependent on text length (with the obvious exception of token frequency) or have other advantages over alternative metrics. Measures dependent on text length (such as type frequency) were interpreted according to this dependency. With reference to their definitions on the application page, the selected measures can be described as follows:

[8] https://www.justice.gov.sk/sudy-a-rozhodnutia/sudy/rozhodnutia/

a) token frequency (abbreviated as N) expresses the length of the text;
b) type frequency (V) tells about the number of different words in the text;
c) verb distance (VD), calculated as the arithmetic mean of the number of tokens between two consecutive verbs in the text (excluding auxiliaries), expresses a certain "density" of verbs in the text;
d) activity (Q), calculated as the ratio of the number of verbs to the sum of verbs + adjectives in the text, expresses the degree of how much activeness there is in the text;
e) descriptivity (D), expresses the degree of descriptiveness of the text. It is thus the inverse of the activity value: D = 1 – Q;
f) secondary thematic concentration (STC) is a modification of thematic concentration (TC), which expresses "the degree to which a text is focused on a central theme or themes" (the central theme is detected using thematic words); STC was chosen since it can be calculated even for shorter texts where TC could not be calculated;
g) moving average type-token ration (MATTR) is one of the measures for analyzing lexical diversity; MATTR is based on the segmentation of the text into the so-called "windows" that overlap each other, where for each window (in this case of size 100 tokens) a type-token ratio is computed; the MATTR is calculated from these windows as their arithmetic mean; the advantage of this measure is its independence from the length of the text, as opposed to the measure of entropy (H).

The obtained samples of 20+20 texts are analyzed by QuitaUp for all 7 quantitative measures and the values were recorded in a summary table (total of 40 texts x 7 measures = 280 values) in 7 columns for each measure and their values in the respective Slovak and Czech sample texts. This table was the basis for statistical testing.

3.3 Selection of tests

The values in all columns are first analyzed using Shapiro-Wilk test for normality. If the normality assumption required for parametric tests was not met, the values for these measures were further tested using the non-parametric Mann-Whitney U test, also due to the small *n* of samples. Because of the directionality of the H_1 hypothesis (*SK has smaller values than CZ*), a left-tailed version of the test was chosen. Where normality has been confirmed in any of the data columns for individual measures, the samples have been further tested using Welch's t-test, which has some advantages over Student's t-test (no assumption of equal variances, more reliable with unequal variances, recommended for small sample sizes). However, outliers were identified in values for each normally distributed measure. Since outliers can distort the mean and inflate variance, making the Welch's t-test less reliable, the normally distributed samples concerned were tested using the Mann-Whitney U test instead, which is relatively robust to the presence of outliers. When interpreting the results of Mann-Whitney, the shapes of the distributions (skewness, kurtosis) and spread (checked by comparing standard deviations) were also taken into account. If the shapes and spread were similar, it was possible to interpret a significant Mann-Whitney U test result as indicating a difference in medians. If the spread values differed, a significant result could reflect differences in distribution shape, spread[9], or central tendency, and not just the medians. The difference in the shapes of the distribution and spread was then reflected in the different interpretation of the test results. For each test are reported the values of p, U, test statistics Z and standardized effect size Z/√(n1+n2)[10].

4 ANALYSIS

4.1 Shapiro-Wilk normality test

Significance level (α): 0.05
Normality assumption violated in at least one data column (SK or CZ): N, V, STC, MATTR
Normality assumed for both SK and CZ columns: VD, Q, D

4.2 Mann-Whitney U test (left-tailed) for individual measures

Significance level (α): 0.05
H_0: SK ≥ CZ
H_1: SK < CZ

4.2.1 Token frequency (N)

p = 0.001605; U = 93; Z = -2.9468; Z/√(n1+n2) = medium (0.46);
p-value < α, **H_0 rejected**.

The randomly selected value in SK sample is considered to be less than the randomly selected value in CZ sample. The distributions differed in skewness shape (asymmetrical for SK; potentially symmetrical for CZ), which means there is a difference in overall rank distribution, rather than a direct comparison of medians.

4.2.2 Type frequency (V)

p = 0.0008834; U = 87; Z = -3.1269; Z/√(n1+n2) = medium (0.48);
p-value < α, **H_0 rejected**.

Test indicated that there was a statistically significant difference in median type frequency between SK (median = 868) and CZ (median = 1795). The distributions

[9] The spread captures the scale differences in data.

[10] This measure indicates that the magnitude of the difference between groups.

were similarly shaped as to skewness (potentially symmetrical) and kurtosis (potentially mesocurtical), as well as they had similar spread (695 vs. 630), so it should be safe interpreting differences as being about central tendency. Since V depends on the length of the text, caution should be exercised when interpreting the strength of this finding.

4.2.3 Verb distance (VD)

p = 0.007149; U = 110; Z = -2.4497; Z/√(n1+n2) = medium (0.38);

p-value < α, **H_0 rejected**.

The randomly selected value in SK sample is considered to be less than the randomly selected value of CZ sample. The distributions differed in at least skewness shape (potentially symmetrical for SK; asymmetrical for CZ).

4.2.4 Activity (Q)

p = 0.9988; U = 311.5; Z = 3.0303; Z/√(n1+n2) = medium (0.48);

p-value > α, **H_0 not rejected**.

The randomly selected value in SK sample is considered to be greater than or equal to the randomly selected value in CZ sample. The distributions were similarly shaped as to skewness (potentially symmetrical) and kurtosis (potentially mesocurtical), but differed significantly as to spread of standard deviations (0.050 vs. 0.026). In this case, the distributions are essentially scaled versions of each other – one is just "stretched" more, but the overall form (shape) is the same. Mann-Whitney U test is still valid, it will test for difference in central tendency (usually median). But it could still be influenced by the fact that one distribution is more variable.

4.2.5 Descriptivity (D)

p = 0.001336; U = 88.5; Z = -3.0033; Z/√(n1+n2) = medium (0.47);

p-value < α, **H_0 rejected**.

The randomly selected value in SK sample is considered to be less than the randomly selected value of CZ sample. As was the case with activity, the distributions were similarly shaped as to skewness (potentially symmetrical) and kurtosis (potentially mesocurtical), but differed significantly as to spread of standard deviations (0.050 vs. 0.026).

4.2.6 Secondary thematic concentration (STC)

p = 0.9291; U = 254; Z = 1.4695; Z/√(n1+n2) = small (0.23);

p-value > α, **H_0 not rejected**.

The randomly selected value of SK sample is considered to be greater than or equal to the randomly selected value of CZ sample. The test indicated that there was a statistically significant difference in STC between SK (mdn = 0.13905) and CZ (mdn = 0.11755). The distributions were similarly shaped as to skewness (asymmetrical) and kurtosis (leptokurtic); the difference in spread (standard deviations) is noticeable, but not extreme (0.069 vs 0.056), the test should still give a valid result and can be interpreted as comparing medians without too much distortion.

4.2.7 Moving average type-token ration (MATTR)

p = 0.00001144; U = 43; Z = -4.2347; Z/√(n1+n2) = large (0.67);
p-value < α, **H_0 rejected**.

The randomly selected value of SK sample is considered to be less than the randomly selected value of CZ sample. The test indicated that there was a statistically significant difference in MATTR between SK (mdn = 0.7605) and CZ (mdn = 0.8075). Because the distributions differed at least in skewness shape (potentially symmetrical for SK; asymmetrical for CZ), this result should be interpreted as a difference in overall rank distribution, rather than a direct comparison of medians.

4.3 Summary

Measure	H_0 rejected	Shape diff.	Spread diff.	Overall rank diff.	Median diff.	Effect size
N	x	x		x		medium
V	x				x	medium
VD	x	x		x		medium
Q			x		x/?	medium
D	x		x		x/?	medium
STC			x/?		x	small
MATTR	x	x		x		large

Tab. 2. Summary of test results (x = confirmed, x/? = confirmed with some reservations)

5 CONCLUSION

The null hypothesis (SK ≥ CZ) was rejected for measures N, V, VD, D (effect size being medium) and MATTR (effect size being large). Only for measure V is there a genuine difference in medians (although V is dependent of the text length), but the Mann-Whitney test is likely to be valid for measure D as well. In the case of the measures N, VD and MATTR the test is potentially weakened by differences in the shape of the distribution. Thus, we could say that the decisions in the Slovak Supreme Court sample have smaller V and STC values compared to the Czech ones, and to some extent also are shorter (with smaller N), with smaller MATTR, VD and D. The null hypothesis was not rejected for the Q and STC measures (medium and small effect size, respectively). Here we can say that Slovak decisions have larger median values for STC and with some reservations (and, compared to STC, a bigger effect size) for Q.

On the basis of these quantitative findings, it is possible to interpret with a certain amount of simplification that Czech Supreme Court decisions seem to be generally longer, richer in types, with greater MATTR (effect size being large) and thus a more diversified vocabulary, a greater distance between verbs and a greater descriptivity. This might be seen as consistent with the initial "hybrid" model of Czech judicial decisions of "complex sophisticated subsumption", where value-oriented arguments, requiring more space (greater length), and greater type and lexical richness, have a place. The decisions of the Supreme Court of the Slovak Republic differ from this model in most of the respects used, which on the one hand can be seen as a possible shift from more complex sophisticated subsumption to – speculatively speaking – greater formality. On the other hand, the Slovak decisions show a greater degree of activity and secondary thematic concentration, which relativizes the shift towards greater formalization and – following on the assertion of T. Smejkalová (2021, p. 235) suggests the need for a closer analysis of individual texts as the next step.

ACKNOWLEDGEMENTS

The paper has been written by the Slovak National Corpus project supported by the Ministry of Education, Research, Development and Youth of the Slovak Republic, Ministry of Culture of the Slovak Republic and the Ľudovít Štúr Institute of Linguistics, Slovak Academy of Sciences.

References

Bernstein, A. (2021). Legal linguistics and the half-empirical attitude. Cornell Law Review, Vol. 106, pp. 1398–1456. Accessible at: https://www.cornelllawreview.org/wp-content/uploads/2021/11/Bernstein-final.pdf.

Bobek, M. (2010). O odůvodňování soudních rozhodnutí. Právní rozhledy (6), pp. 204–211.

Cvrček, F. (2016). Právní informatika a lingvistika. Jurisprudence 25(6), pp. 49–53.

Cvrček, V., Čech, R., and Kubát, M. (2020). QuitaUp – nástroj pro kvantitativní stylometrickou analýzu. Czech National Corpus and University of Ostrava. Accessible at: https://korpus.cz/quitaup/.

Fábry, B., Kasinec, R., and Turčan, M. (2019). Teória práva. Bratislava: Wolters Kluwer SR. 324 p.

Goźdź-Roszkowski, S. (2021). Corpus Linguistics in Legal Discourse. International Journal for the Semiotics of Law – Revue internationale de Sémiotique juridique, 34(3), pp. 1515–1540. Accessible at: https://link.springer.com/article/10.1007/s11196-021-09860-8.

Holländer, P. (2012). Filosofie práva. Plzeň: Aleš Čeněk, s. r. o., 424 p.

Knapp, V. (2024). Teorie práva. Plzeň: Aleš Čeněk, s. r. o., 357 p.

Matczak, M., Bencze, M., and Kühn, Z. (2010). Constitutions, EU Law and Judicial Strategies in the Czech Republic, Hungary and Poland. Journal of Public Policy, 30(1), Performing to Type? Institutional Performance in New EU Member States, pp. 81–99.

Mouritsen, S. C. (2010). The Dictionary Is Not a Fortress: Definitional Fallacies and a Corpus-Based Approach to Plain Meaning. BYU Law Review, (5). Accessible at: https://digitalcommons.law.byu.edu/lawreview/vol2010/iss5/10.

Smejkalová, T. (2021). Soudní rozhodnutí jako autoportrét českého soudnictví. Brno: Nakladatelství Masarykovy univerzity, 281 p. Accessible at: https://munispace.muni.cz/library/catalog/book/2120.

Wilfling, P. (2013): Kvalitatívne požiadavky na odôvodnenie súdneho rozhodnutia. Vybrané otázky. Pezinok: Via Iuris, 78 p. Accessible at: https://viaiuris.sk/pravny-stat/kvalitativne-poziadavky-na-odovodnenie-sudneho-rozhodnutia/.

CORPUS BUILDING

DOI 10.2478/jazcas-2025-0013

ANNOTATION OF RHETORICAL ROLES AND SYLLOGISTIC RELATIONS IN CZECH ARGUMENTATIVE LEGAL AND ADMINISTRATIVE TEXTS

SILVIE CINKOVÁ[1] – JANA ŠAMÁNKOVÁ[2] – BARBORA KUBÍKOVÁ[3] – TEREZA NOVOTNÁ[4] – VÍTEK EICHLER[5]

[1]Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics, Charles University, Prague, Czech Republic (ORCID: 0000-0003-4526-3915)

[2]Department of Legal Skills, Faculty of Law, Charles University, Prague, Czech Republic (ORCID: 0009-0008-6251-8223)

[3]Office of the Public Defender of Rights and Defender of Children's Rights, Brno, Czech Republic (ORCID: 0009-0002-7226-389X)

[4]Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics, Charles University, Prague, Czech Republic (ORCID: 0000-0002-1426-4547)

[5]Faculty of Law, Charles University, Prague, Czech Republic (ORCID: 0009-0008-3070-8997)

CINKOVÁ, Silvie – ŠAMÁNKOVÁ, Jana – KUBÍKOVÁ, Barbora – NOVOTNÁ, Tereza – EICHLER, Vítek: Annotation of Rhetorical Roles and Syllogistic Relations in Czech Argumentative Legal and Administrative Texts. Journal of Linguistics, 2025, Vol. 76, No 1, pp. 145 – 156.

Abstract: We present KUKY 1.0, a publicly available corpus annotated by lawyers for rhetorical structure and relevance. We explain the concepts as well as the annotation procedure. Eventually we discuss the challenge of inter-annotator agreement.

Keywords: rhetorical structure, argumentation, annotation, corpus, legal, law, syllogism, relevance

1 INTRODUCTION

1.1 Readability of legal and administrative documents

This study describes a newly released annotated corpus of Czech legal and administrative documents with readability assessment and a multi-layer annotation of relevance, rhetorical roles, and, partly, of rhetorical structure. The purpose of the corpus is prototyping readable Czech legal/administrative documents with large language models.

There are enough grammatical and lexical strategies to increase readability, such as shorter sentences, avoiding passives, or replacing deverbal names with verbs (DuBay 2004) – strategies that work across languages. Some of the classic readability formulas have even been adapted to Czech (Bendová and Cinková 2021). So it seems all what legal writers should do is to internalize these rules and check results with a formula. However, recent research suggests that the effect of grammatical and lexical features on readability is rather modest (Cinková 2024).

This may not surprise lawyers, who have anyway traditionally concentrated on rhetorical structure as a component of logical coherence rather than on stylistics *per se*. Therefore, we have shifted focus in this very direction, and we involved lawyers to shape the annotation scheme. The resulting annotation scheme faithfully implements the steps plain legal writing experts take when perusing a document to optimize it for readability.

1.2 Related work

In the last decades, annotated legal corpora were produced to aid automatic *summarization* and *argumentation mining*. Their annotation schemes either capture the conventional macrostructure, such as preambles and decisions sentences (de Vargas Feijó and Moreira 2018; Šavelka and Ashley 2018), or they concentrate on the pragmatics of sentences or clauses.

For instance, Grover et al. (2003) analyzed judgments of the British House of Lords, using three labels: *Background* (references to law and precedents), *Case* (events and lower court decisions), and *Own* (speaker's judgments and interpretations of *Background*). Bhattacharya et al. (2023), as well as Malik et al. (2022) arrived at a more fine-grained set of *Facts*, *Ruling by Lower Court*, *Argument* (of the present Court), *Statute* (laws references by the present court), *Precedent*, *Ratio of the Decision*, and *Ruling by Present Court*. The best-known legal corpus with argumentation mining annotation is the ECHR judgments corpus (Teruel et al. 2018), along with its recent extensions (Habernal et al. 2024). Yamada et al. (2019) built a corpus of legal argumentation of Japanese civil law judgments. Unlike in most corpora, the annotation segments are not strictly defined by single sentences or clauses but are allowed to span across. This annotation scheme provides both labels of rhetorical roles and relations between them.

KUKY 1.0 resembles the Japanese corpus by allowing for a free segmentation of annotation spans. The labels draw on Grover et al. (2003) and seek to link the corresponding spans into syllogistic triples of premises and conclusions. Apart from the rhetorical roles, the corpus provides an annotation of relevance. This annotation highlights incomprehensible or superfluous text.

2 RELEVANCE AND RHETORICAL ROLES

2.1 Relevance

Irrelevant information increases the cognitive burden of the reader in two ways: for the first, a longer message takes a longer time to read; for the second, the reader will waste their cognitive capacity on integrating disparate inputs to make sense as a whole. Empirical research (Tyler 1990; Song and Schwarz 2010; Wagner and Walker 2019) has proven that the easier a message is to perceive, the more persuasive or authoritative it appears to the reader. Hence, irrelevant information hampers real-world processes in administration and justice by obscuring the actual messages.

When annotating relevance in KUKY 1.0, the annotators mimic the first step in redesigning documents, where the editor deliberates which original content to preserve in the new version. They mark spans as *Relevant*, *Superfluous/Irrelevant*, and *Incomprehensible/Confusing beyond repair*.

2.2 Syllogism in argumentative writing

Argumentative texts judge whether or not a fact contradicts the law. That requires setting out the relevant law and project it on a fact in such a way that the applicability of the given law to the given fact becomes indisputable. The stronger the link between the law and the fact, the more persuasive is the resulting judgment.

According to J. Gardner (1993), the rhetorical centerpiece of legal persuasive reasoning is *syllogism*[1]. Gardner argues that "all legal argument should be in the form of syllogisms" because "syllogistic argument provides the requisite appearance of certainty. It makes the outcome of a case seem as certain and as mechanical as the output of a mechanical equation, and achieves this effect not by actual mathematical operations, but, paradoxically, by exploiting human intuition" (Gardner 1993 §1.1.). Syllogism consists of three components:

1. The *major premise* (a broad statement of general applicability),
2. The *minor premise* (a narrower statement able to serve as an instance of the major premise),
3. The *conclusion* (a statement that evidently holds for the major premise, and, hence, it must also hold for the minor premise).

In the legal domain, the major premise is populated by the law, the minor premise by the fact, and the conclusion by the judgment. Hence, it is not by chance that the law, fact, and conclusion are the central labels in the annotation of rhetorical roles in legal corpora, and KUKY 1.0 is no exception.

2.3 Rhetorical roles in KUKY 1.0

KUKY 1.0 distinguishes between argumentative and normative documents. Tab. 1 presents the annotation scheme of argumentative documents.

Narrative	Minor premise. Facts, testimonies, and past decisions by authorities.
Law	Major premise. Law references, quotes, interpretations, and summaries.
Conclusion	Conclusion. Ruling, finding, judgment.
Advice	Optional information to aid the recipient.

[1] The concept of syllogism is attributed to Aristotle (Aristotle 2004).

Command	Recipient's obligations resulting from the document.
Legal Issue	Summary of the matter of dispute in legal terms, typically formulated as a yes-no question.
Metatext	Processing matters.

Tab. 1. Rhetorical roles in argumentative documents

2.4 Syllogism in KUKY 1.0

Among the argumentative documents in KUKY 1.0, the presence of syllogistic structures distinguishes the top-readable documents from the ordinary ones. We will illustrate syllogism on two authentic examples in Tab. 2 and Tab. 3.

Tab. 2 presents numbered and labeled segments of a court order. Court orders always start with the court ruling. A good court ruling is the summary of a Conclusion in the reasoning part. A good reasoning part contains a Conclusion, which can be divided into several partial ones, but each Conclusion must be backed up by at least one Law and one Narrative, which would ideally match each other in the syllogistic way.

The court uses syllogistic argumentation in this order. It rules that a certain Pavel Boháč can legally represent his elderly mother. The conclusions (1 and 5) are in accordance and Sentence 1 does not add anything new to 5. Conclusion 5 is supported by one Law (2) and one Narrative (3). Hence Conclusion 1 is supported by the same Law and Narrative. The Law and the Narrative match point by point: the familial relationships, the mother's health conditions, and her deliberate approval.

ID	Text span	Label
1	***Order*** *The District Court [...], has decided in the legal matter concerning Jitka Boháčová's approval for representation by a household member as follows:* ***The court approves the representation of Jitka Boháčová by her household member, Mr. Pavel Boháč.***	Conclusion
2	***Legal Framework*** *If a mental disorder prevents an adult from legally acting on their own behalf, they can be represented by a household member [...]. The representative must inform the represented [...] and clearly explain [...]. If the person to be represented refuses, the representation does not arise [...]* *Court approval is required [...]. Before issuing a decision, the court must make the necessary efforts to ascertain the opinion of the represented person [...]*	Law
3	***Assessment of the petition*** *The court verified that there is indeed a familial relationship [...]* *The court visited the subject [...]. The court verified that the subject's health condition [...]. The court confirmed that the subject understands the nature and consequences of the representation, agrees with it, and agrees that the petitioner, her son, will represent her.*	Narrative

4	*The court thus found that the conditions for approving the petitioner as the subject's representative were met [...]. The court also confirmed that representation by a household member is sufficient to protect the rights and interests of the subject.*	Conclusion
5	***Some Rights and Duties of the Household Member Representative***	Advice

Tab. 2. A court order about legal representation by a household member

Another example (Tab. 3) shows the use of syllogism in a last-warning letter for a neighbour to confine her overgrown trees to her lot. Note how two rhetorical roles can appear within one sentence: the first sentence starts with a Conclusion (1) and continues with a Narrative (2). In the Czech original, the narrative is structured as a subordinate content clause (*tím, že...*). The supporting Law appears in 3 and 5. In addition, two Commands (4, 6) are each backed up by a Law (3, 5) as well as by the Narrative, and act very much as Conclusions and actually form two other syllogisms. So, virtually all statements in this documents are components of a syllogism, and this is what makes the text particularly succinct and the train of thought so easy to follow.

ID	**Text span**	**Label**
1	*I am informing you that you are violating my property rights.*	Conclusion
2	*by having planted and grown trees in close proximity of the border between our lots as well as continuously planting new trees without maintaining them, so that their branches and roots are reaching over to my lot.*	Narrative
3	*According to §1016 of the Civil Code you are obliged to maintain all hanging and underground parts of your trees that trespass the border on my lot.*	Law
4	*Therefore I urge you to cut the branches that reach over on my lot and to remove the undergrowing roots on my lot, all of this within 30 days from the delivery of this letter.*	Command
5	*According to §1017 of the Civil Code you can only plant tree species that usually grow above 3 meters at least 3 meters from the border of your lot, lower growing trees then at least 1,5 meters from the border of your lot.*	Law
6	*Therefore I urge you to stop planting new trees at our common lot border in a way that contradicts the law.*	Command
7	*I firmly believe that you are going to stand up to your obligations. Otherwise I will take you to court.*	Advice

Tab. 3. Syllogistic structure in a final demand

3 DATA

3.1 Statistics

KUKY 1.0 is a curated selection of 224 Czech administrative and legal documents (totalling of 374,251 tokens) for readability research, formatted in plain text with or without markdown. The document length lies between 159 and 6,239 tokens, with the median at 1,250 and the mean at 1,671 tokens.

Document contributors were legal experts dedicated to plain legal writing, who sought to select a range of examples from high-quality documents (to serve as blueprints for the given genre), over somewhat accessible documents, to the standard production, which is generally hard to comprehend. The collection is somewhat biased towards the best and good documents, since they require a more careful selection than the standard production, which can be acquired bulk-wise from other sources.

3.2 Document sources

The main sources of documents in KUKY 1.0 are the publicly available databases of the Office of the Czech Public Defender of Rights, the Supreme Administrative Court, and a free legal advice database of a legal company (Frank Bold). Besides, the corpus contains various contributions from individual legal experts: communications between clients and authorities, public local administration announcements, or legal memos. Such documents were thoroughly pseudonymized, including local names, dates, and all other numeric strings, to preclude tracing of the parties involved.

On the top level, the documents are grouped into *argumentative* (174 documents) and *normative* (50 documents). Argumentative documents are always case-related. They map a past event or its result on existing legal norms to judge it. Typical argumentative documents are findings and decisions by authorities, such as courts and supervisory bodies, or personalized client advice by legal experts. Normative documents, on the other hand, set norms (laws) or guide a generic reader through an administrative procedure (e.g. how to register a society). They can even model life situations (e.g. how to deal with a noisy neighbor), but they never address a concrete case.

3.3 Metadata and structure

The main document distinction is the *argumentative* vs. *normative,* but a few more criteria were used to classify the documents and captured in the metadata by single judgments of document contributors (Tab. 4).

The argumentative documents and the normative documents come in two JSON files. Both files consist of three JSON arrays: *documents*, *labels*, and *annotations*. The *documents* array contains objects that represent the individual documents. Each object in the *documents* array contains the document's text, along with metadata, as object properties. The *labels* array lists the labels defined by the annotation scheme. The annotation schemes of the argumentative and the normative documents slightly differ, which is why they are stored in separate files. The *annotations* array lists individual annotations: texts spans marked with labels. Each object within the *annotations* array contains an annotation label and maps on the source text with a reference to the document's ID and with offsets.

Each document has two annotation layers: the rhetorical roles and relevance. These two layers were annotated independently, so each segments the text differently. Their definitions in the *labels* JSON object are merged, but their instances in the *annotations* objects are distinguished by a property called *task_type*.

3.4 Access

The entire KUKY 1.0 corpus along with the documentation is stored in the LINDAT/CLARIAH-CZ repository under the persistent ID http://hdl.handle.net/11234/1-5812 and a CC BY-NC-SA 4.0 license. The documentation is also available at the non-persistent URL https://ufal.mff.cuni.cz/grants/ponk/kuky.

Metadata property	**Values**	**Description**
doc_id, doc_name		Unique document ID, name
Readability	*Low,* *Medium,* *High*	Expert assessment, relative to other documents in the corpus.
Anonymized	*Anonymized by source,* *On-site anonymization,* *No*	Is the document anonymized/ pseudonymized?
SyllogismBased	*True,* *False*	Does this document systematically use syllogism?
DocumentVersion	*Original,* *Partial Redesign,* *Redesign*	Default: Original. Some documents come in an original version and revision(s).
ParentDocumentID		Redesigned documents contain a reference to the *doc_id* of their corresponding Original.
LegalActType	*Individual,* *Normative*	The key distinction between documents in this corpus.
Objectivity	*Quasiobjective,* *Persuasive*	Judgments are quasiobjective. Lawsuits etc. are persuasive.
Bindingness	*True, False*	Is the document legally binding?
AuthorType	*Authority,* *Individual*	Does the author write in the capacity of an authority?
RecipientType	*Natural person,* *Legal person,* *Combined*	Natural persons are not likely to hire an expert to interpret the document for them, while legal persons (e.g. companies) often employ lawyers.
RecipientIndividuation	*Individual,* *Bulk,* *Public*	How familiar are the recipients with the matter?

Tab. 4. Metadata properties in documents

4 ANNOTATION

4.1 Procedure

The annotation proceeded in two separate steps: relevance and rhetorical roles including the syllogistic relations. Both steps were carried out in a cloud installation of Gloss (Poudyal et al. 2020), by courtesy of its developer Jaromír Šavelka. The texts were selected, assessed, edited, and subsequently annotated by lawyers, mostly ones with an extensive experience with practicing as well as teaching plain legal writing.

Deliberately segmented data pose a challenge for measuring the inter-annotator agreement (IAA). Differences in segmenting should not be penalized, as long as the words were identically labeled. Therefore we considered each token one annotator judgment. We report IAA for ten documents and two annotators.

4.2 Inter-annotator agreement on relevance

IAA on Relevance reached accuracy 0.78. Cohen's Kappa over all labels was only 0.47, which is not too bad considering that the *Relevant* label very strongly prevailed, and hence each disagreement was heavily penalized. The prevalence of the *Relevant* label is evident from the confusion matrix in Tab. 5.

	Incomprehensible	Superfluous	Relevant
Incomprehensible	183	495	196
Superfluous	0	999	2071
Relevant	128	437	10330

Tab. 5. Confusion matrix of Relevance annotation (numbers stand for count of tokens with the given combination of labels)

4.3 Inter-annotator agreement on rhetorical roles

IAA on rhetorical roles varied very strongly across documents. Fig. 1 illustrates the IAA as Fleiss' Kappa on individual labels within individual documents. The dashed line represents the average Fleiss Kappa across all labels within the document. The solid line is placed at 0.6, a rule-of-thumb threshold for semantic tasks.

4.4 Inter-annotator agreement on syllogistic relations.

To compute IAA on syllogisms, we modeled the relations between segments as relations between individual tokens. There were possible relations per document (each token with each token). Actual relations between segments were modeled on each word of one segment to each word of the second segment. We neglected their rhetorical role labels. The average accuracy was 0.95 (standard deviation 0.05), precision 0.53 (standard deviation 0.29), and recall 0.3 (standard deviation 0.31).

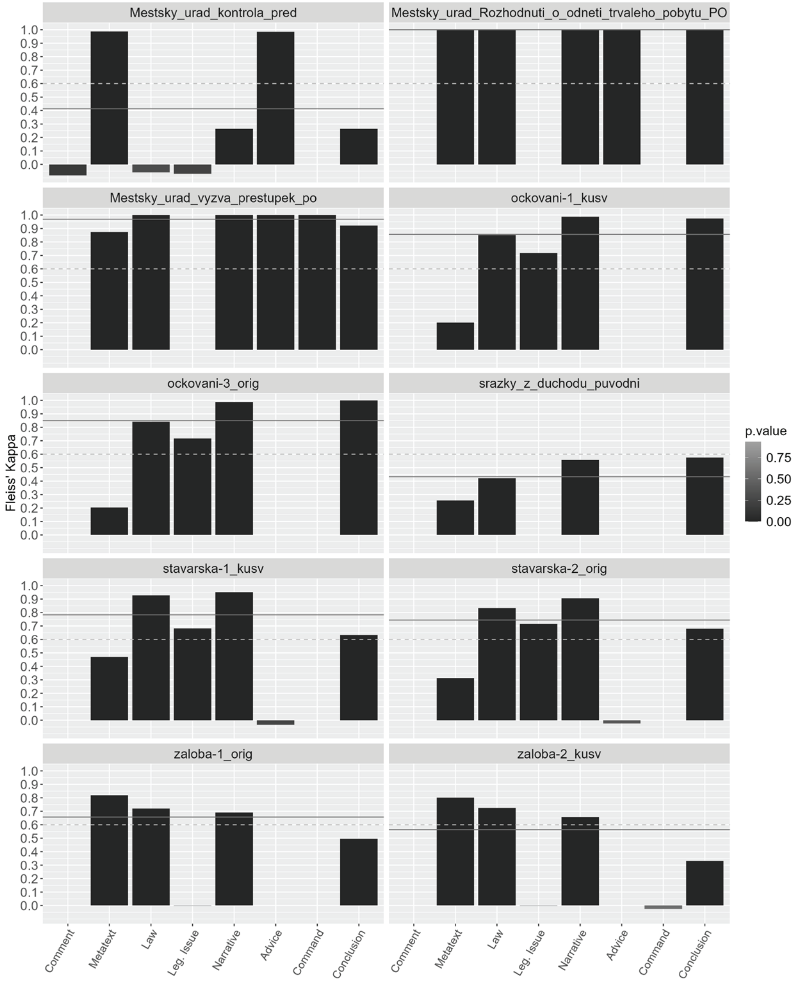


Fig. 1. Inter-annotator agreement on rhetorical roles, document-wise and label-wise

5 DISCUSSION

The tagsets mimic the deliberation phase of a human editor, pursued manually with crayons before drafting the redesigned version—from scratch, with occasional copy-pasting. Even though the two editors follow the same principles, they might pursue them differently, just as the resulting redesigns would never be identical across authors, although both could be equally good.

The IAA is not impressive, but this could be expected with a task that is closer to translation rather than classification. The Relevance annotation has a low IAA because the distribution of labels is very uneven. In practice, annotators recognize most of document content as relevant, no matter how clumsy the style: hence the corresponding accuracy of almost 80%.

The average IAA on Rhetorical Roles does not say much because of the wide dispersion among documents. In fact, only three documents of ten do not reach the 0.6 average Fleiss' Kappa. None of these three had been classified as highly readable in the metadata (before the annotation). Even a most shallow disagreement analysis reveals that rhetorical roles are blurred in unreadable documents, suggesting that the speaker does not care to organize their utterance into purposeful units.

In the document with the worst IAA (Mestsky_urad_kontrola_pred), Narrative, Law, and Conclusion mingle with Metatext even within one sentence, such as in this example:

V průběhu kontroly bylo zjištěno podezření z porušení ustanovení § 21 odst. 4 zákona o ochraně veřejného zdraví, kterého se kontrolovaná osoba dopustila tím, že v průběhu kontroly nebyl v kontrolované provozovně vyvěšen provozní řád schválený orgánem ochrany veřejného zdraví, a to v souladu s výše uvedeným ustanovením, přestože je v dotčené provozovně vykonávána činnost „Pedikúra, manikúra", která je zákonem o ochraně veřejného zdraví považovaná za činnost epidemiologicky závažnou.
'During the inspection, a suspicion of violation of instruction § 21 Par 4 of the Public Health Protection Law was detected, that the inspected person committed by the fact that during the inspection at the inspected shop the operation rules approved by the Public Health Protection officer were not on display, that in accordance with the aforementioned instruction, although in the aforementioned shop was carried out the activity "Pedicure, Manicure", which is considered an epidemiologically relevant activity by the Public Health Protection Law.'

It goes without saying that IAA is hard to maintain when untangling such a scramble into discrete communicative intents. So, for instance, a poorly referenced law might be recognized as such by one annotator, while the other would "downgrade" it to Metatext. The same could easily happen to a sloppily formulated Legal Issue or a nebulous Conclusion.

The syllogistic annotation heavily depends on the Rhetorical Roles annotations. When a text contains numerous ambiguous segments or other forms of incongruity, annotators are often reluctant to scour it for potential partial syllogisms.

Qualitative observations suggest that comprehensible documents are easier to agree on, reminding us of the proverbial Anna Karenina principle saying that all happy families are alike, while each unhappy family is unhappy in its own way. We speculate that, in a machine-learning setup, readability assessment it is not going to be aided as much by the automatic classification of the rhetorical roles themselves as by the confidence levels of the predictions, and the same would apply to the detection of syllogistic structures.

ACKNOWLEDGEMENTS

The research has been supported by the Technology Agency of the Czech Republic under the project TQ01000526 (9/2023-12/2025).

References

Aristotle (2004). Rhetoric. New York: Dover Publications.

Bendová, K., and Cinková, S. (2021). Adaptation of Classic Readability Metrics to Czech. In 24th International Conference on Text, Speech and Dialogue. Cham, Switzerland: Springer, pp. 159–171.

Bhattacharya, P. et al. (2023). DeepRhole: Deep Learning for Rhetorical Role Labeling of Sentences in Legal Case Documents. Artificial Intelligence and Law, 31(1), pp. 53–90. Accessible at: https://doi.org/10.1007/s10506-021-09304-5.

Cinková, S. (2024). Linguistic Factors in the Readability of Czech Administrative and Legal Texts. In: Z. Bohušová – M. Dove (eds.): To Understand Is to Be Free. Interdisciplinary Aspects of Comprehensibility and Understanding. Vienna, Austria: Praesens Verlag, pp. 303–325.

DuBay, W. H. (2004). The Principles of Readability. Costa Mesa, California: Impact Information. Accessible at: https://www.researchgate.net/publication/228965813_The_Principles_of_Readability.

Gardner, J. A. (1993). Legal Argument: The Structure and Language of Effective Advocacy. LexisNexis. Accessible at: https://store.lexisnexis.com/en-us/legal-argument-the-structure-and-language-of-effective-advocacy-sku-us-ebook-03082-epub.html.

Grover, C., Hachey, B., and Korycinski, C. (2003). Summarising Legal Texts: Sentential Tense and Argumentative Roles. In Proceedings of the HLT-NAACL 03 Text Summarization Workshop, pp. 33–40. Accessible at: https://aclanthology.org/W03-0505/.

Habernal, I. et al. (2024). Mining Legal Arguments in Court Decisions. Artificial Intelligence and Law, 32(3), pp. 1–38. Accessible at: https://doi.org/10.1007/s10506-023-09361-y.

Malik, V. et al. (2022). Semantic Segmentation of Legal Documents via Rhetorical Roles. In: N. Aletras et al. (eds.): Proceedings of the Natural Legal Language Processing Workshop 2022. Abu Dhabi, United Arab Emirates (Hybrid): Association for Computational Linguistics, pp. 153–171. Accessible at: https://doi.org/10.18653/v1/2022.nllp-1.13.

Poudyal, P. et al. (2020). ECHR: Legal Corpus for Argument Mining. In: E. Cabrio – S. Villata (eds.): Proceedings of the 7th Workshop on Argument Mining. Online: Association for Computational Linguistics, pp. 67–75. Accessible at: https://aclanthology.org/2020.argmining-1.8/.

Šavelka, J., and Ashley, K. D. (2018). Segmenting U.S. Court Decisions into Functional and Issue Specific Parts. In Frontiers in Artificial Intelligence and Applications. IOS Press. Accessible at: https://doi.org/10.3233/978-1-61499-935-5-111.

Song, H., and Schwarz, N. (2010). If It's Easy to Read, It's Easy to Do, Pretty, Good, and True. Bulletin of the British Psychological Society, 23(2), pp. 108–111.

Teruel, M. et al. (2018). Increasing Argument Annotation Reproducibility by Using Inter-Annotator Agreement to Improve Guidelines. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018).

Tyler, T. R. (1990). Why People Obey the Law. New Haven and London: Yale University Press.

de Vargas Feijó, D., and Moreira, V. P. (2018). RulingBR: A Summarization Dataset for Legal Texts. In: A. Villavicencio et al. (eds.): Computational Processing of the Portuguese Language. Cham: Springer International Publishing, pp. 255–264.

Wagner, W., and Walker, W. (2019). Incomprehensible!: A Study of How Our Legal System Encourages Incomprehensibility, Why It Matters, and What We Can Do About It. Cambridge Core. Cambridge: Cambridge University Press. Accessible at: https://doi.org/10.1017/9781139051774.

Yamada, H., Teufel, S., and Tokunaga, T. (2019). Building a Corpus of Legal Argumentation in Japanese Judgement Documents: Towards Structure-Based Summarisation. Artificial Intelligence and Law, 27(2), pp. 141–170. Accessible at: https://doi.org/10.1007/s10506-019-09242-3.

sciendo
DOI 10.2478/jazcas-2025-0014

WHEN DATA MEET TOOLS: USING THE MONITOR CORPUS FOR THE ANALYSIS OF LANGUAGE DEVELOPMENT

VÁCLAV CVRČEK[1] – MARTIN STLUKA[2] – KLÁRA PIVOŇKOVÁ[3]
[1]Department of Linguistics, Faculty of Arts, Charles University, Prague, Czech Republic (ORCID: 0000-0003-3977-2393)
[2]Department of Linguistics, Faculty of Arts, Charles University, Prague, Czech Republic (ORCID: 0000-0003-3294-3583)
[3]Department of Linguistics, Faculty of Arts, Charles University, Prague, Czech Republic & Department of Philosophy and History of Science, Faculty of Science, Charles University, Prague, Czech Republic (ORCID: 0009-0000-0990-7424)

CVRČEK, Václav – STLUKA, Martin – PIVOŇKOVÁ, Klára: When Data Meet Tools: Using the Monitor Corpus for the Analysis of Language Development. Journal of Linguistics, 2025, Vol. 76, No 1, pp. 157 – 166.

Abstract: The aim of this paper is to introduce an infrastructure developed within the HiČKoK project to enable full-fledged corpus-based diachronic research of Czech. The individual sections of the paper present the components of this infrastructure, which links well-balanced, representative and annotated data with tailor-made tools for diachronic research. The forthcoming monitor corpus, covering the entire period of written Czech, along with its composition and annotation strategies, is briefly introduced. In the following sections, the potential of the application and its four modules—simple query, comparison, time-based associations, and diachronic collocations—are demonstrated through mini case studies. Combining large-scale data (as representative as possible) with a tool that enhances standard corpus functionalities, enriches them with a diachronic perspective, and enables result visualization makes diachronic research on language change more accessible and comprehensive.

Keywords: diachronic research, corpus querying, annotation, language change, monitor corpus, frequency, Czech

1 INTRODUCTION

The use of corpus linguistic methods for research on language change is accompanied by a number of specific challenges and issues. What has been addressed several times are the issues related to corpus compilation and annotation (Davidse and de Smet 2020). Compiling a diachronic corpus often requires sophisticated solutions for taming the variability of spelling (or even standardization of different writing systems); annotation systems have to be painstakingly adjusted to deal with the changes in morphology, syntax and lexicon in order to provide output that is both usable and adequate. A problem of its own, which in many cases cannot be solved satisfactorily, is the composition of the corpus and its representativeness, where we do not find the same spectrum of text types at different stages of language development.

The other group of issues is related to the tools and methods we use for exploring diachronic data. This has been addressed significantly less in the literature despite the fact that diachronic description calls for specific approaches and often requires specific measures for analyzing data.

In this paper, we would like to argue that for full-fledged corpus-based diachronic research, we need to address issues stemming from both the data and the tools for exploring them. This is because, in our view, a truly full-fledged infrastructure for working with diachronic data requires both well-constructed and annotated data and custom-tailored tools for examining them. This type of infrastructure for Czech is being developed within the HiČKoK project (Historie češtiny v korpusovém kontinuu 'History of Czech in the Corpus Continuum', see https://korpus.cz/hickok) which is supported by the Technology Agency of the Czech Republic within the Programme for Support of Applied Research and Innovation SIGMA. Its duration is from September 2023 to November 2026.

First, we will describe the Monitor Corpus that is being developed within the HiČKoK project (and its current version), briefly introduce the plan for its annotation, and in the second part we will focus more on the tools that we develop for effective work with this diachronic data.

2 THE MONITOR CORPUS OF CZECH

The aim of the HiČKoK (History of Czech in the Corpus Continuum) project is to create data, software and knowledge resources for the study of Czech throughout its history (13th–21st century).

One of the main objectives of the project is to create a Monitor Corpus of Czech that covers all eight centuries of the development of Czech in three main text types (fiction, non-fiction and journalism) where possible. It consists of all available diachronic data of the participating institutions (Czech National Corpus, Charles University and Institute of the Czech Language, Czech Academy of Sciences); part of the data that was not available in corpus format (covering the period of 1900–1990) was obtained in cooperation with the National Library of the Czech Republic. At the beginning of 2025, a working internal version of the corpus was created from all linguistically non-annotated texts by harmonizing data from each of the periods being processed (13th–15th centuries, 16th–18th centuries, 19th century and 20th century up to the present). The size of the corpus in its periods is summarized in Tab. 1.

Period	Tokens	Documents
13th–15th century	6 524 459	271
16th–18th century	2 809 406	197
19th century	18 233 175	999
20th–21st century	66 033 366	4 029
Total	93 600 406	5496

Tab. 1. Number of tokens and documents in the main periods of development of Czech and the total numbers for the whole corpus

The main goal, with respect to data compilation, is that the Monitor corpus of Czech should be able to represent the entire development of Czech in a single corpus, uniformly tokenized and annotated according to the latest standards. This is unique not only in the context of Czech, but also worldwide. Similar projects such as COHA (covering the period of 1920–2010), the Helsinki Corpus of English Texts (850–1710) or EEBO (containing over 25,000 books of various genres printed between 1475 and 1700) usually cover shorter periods of time, do not include contemporary language or are not designed as genre-balanced (fiction, non-fiction and journalism).

2.1 Corpus annotation

The methodology chosen for the project enables the corpus processing and annotation of Czech texts produced over eight centuries. The chosen processing approach takes into account the needs of the contemporary user while reflecting the linguistic evolution of Czech.

In attempting to cover eight centuries, care must be taken for comparability and consistency in annotation. For these reasons (and for reasons of cross-linguistic comparability), we have opted for the Universal Dependencies (UD; de Marneffe et al., 2021) framework, which serves as a de facto international annotation standard, to process the entire corpus. For these purposes, it was necessary to develop both a unified lemmatization system and to adapt synchronous tagging tools (cf. Zeman et al. 2023).

For this purpose, training datasets (etalons) with manually tagged and corrected texts were created that include samples of data from each period (see Tab. 1).

The corpus will be accessible by default via the KonText search interface for standard corpus querying, inspecting concordances or creating frequency distributions. Another output of the project will be freely available UD language models for annotating texts from different periods of Czech language development. The model for contemporary Czech is already available; models for older phases will be based on manually annotated (etalon) samples (for the earliest period up to the end of the 15th century, for Middle Czech from the 16th–18th centuries, and for the 19th century – approx. 100,000 tokens each).

In the following sections, we describe the possibilities of working with the Monitor Corpus within the Timeline Maker application. As has been pointed out, lemmatization and tagging are still in the process of development, so the demonstrations of working with Timeline Maker will be based on working with the beta version of the corpus, which so far contains only word forms and metadata.

3 TIMELINE MAKER

In contrast to synchronic research, which uses standard corpus tools (concordance, collocation, frequency distribution), diachronic research has an additional temporal dimension. At the same time, most of the standard tools for working with language corpora do not have the necessary functionality to capture and visualize phenomena during their change.

This issue is tackled by the Timeline Maker application, which is being developed within the HiČKoK project and is currently in its beta version. Its main advantage over standard corpus tools is the fact that while allowing for standard corpus querying it is designed to be able to work with diachronic data that contains information about the time (year) of creation of the text.

Timeline Maker is designed as a GUI on top of the KonText corpus manager (Machálek 2014) on the R/Shiny platform. Queries (in CQL format) that are entered into the application by the user are transformed into a series of queries to the KonText API so that they cover the entire timeline represented by the corpus (or a subpart of it selected by the user). The results obtained from the KonText API are then visualized via the Plotly library.

One of the key features of Timeline Maker is its ability to work with variable granularity of the timeline. This is a feature of crucial importance as in the case of diachronic data, which is relatively sparse, especially in the older phases of language development, it is often necessary to aggregate the results into larger units (decades, quarter-centuries or half-centuries) in order to be able to spot a development trend.

The application is divided into four modules, each representing a different type of query:

1. simple query: showing the frequency trend for a single phenomenon
2. comparison: represents the proportion of frequencies of two competing variants
3. companions: allows for detecting a similar (correlated) frequency trend of two phenomena, which might suggest an association between them
4. diachronic collocations: visualizes the change of the collocation profile of two words over time

In the following sections, each module will be described and exemplified on a selected phenomenon.

3.1 Frequency trend

The "simple query" module is designed to capture the frequency evolution of a given lexical or grammatical phenomenon. Its main goal is to plot the frequency of a given phenomenon in a graph (with the x-axis being the timeline), both in individual years and in aggregate form (in periods of 10, 25 or 50 years).

In addition to the query itself, which can be in the form of CQL and can contain standard regular expressions, the input form of this module allows the user to specify the time range in which the query will be evaluated.

For the "simple query" module, we show an example illustrating the gradual disappearance of the word *poprávcě* ('judge/executioner'). In earlier periods of Czech, this was a polysemous lexeme related to legal topics, see ESSČ (*Elektronický slovník staré češtiny*, 2006–).

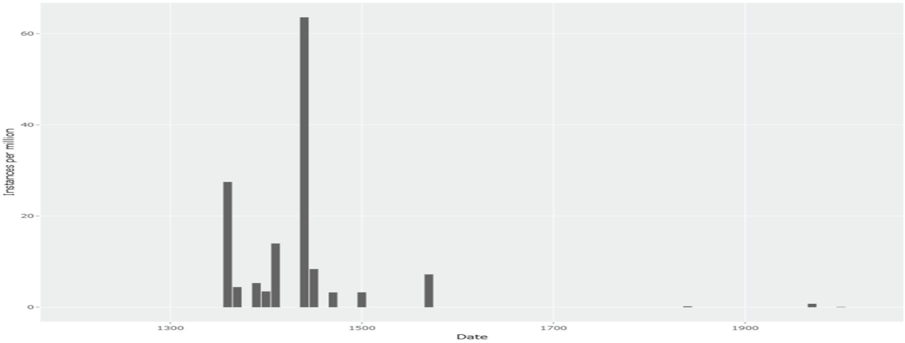


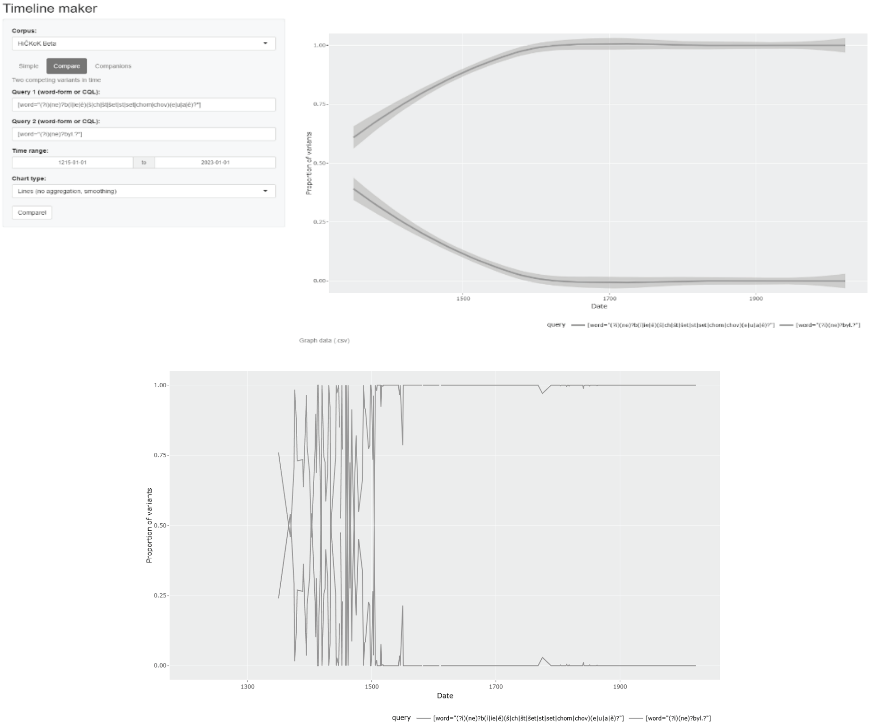
Fig. 1. Occurrence of the word *poprávcě* 'judge/executioner'

The results of a search for *[Pp]oprávc.** by decade (shown in Fig. 1) indicate that the word *poprávcě* ('judge/executioner') appeared consistently in the Old Czech period, namely from the mid-14th century to the end of the 15th century, while in the following periods, it is documented only in isolated instances. This suggests that from the end of the 15th century *poprávcě* ceased to be part of the linguistic usage and was probably replaced by other terms (e.g. *soudce* 'judge', *kat* 'executioner').

3.2 Two competing variants

Similarly to the previous module, for comparing two competing variants, the user enters a CQL query (one for each variant) and selects the time window in which s/he wants to evaluate the query. This module is used to visualize the proportion of morphological, lexical or word-order variants over time. As in the previous case, this module offers the possibility of aggregating results by year, decade, quarter-century or half-century. In addition, it offers smoothing, which provides a better overview of the overall trend using the moving average method.

A suitable example to showcase the possibilities of this could be the gradual disappearance of the simple past tenses in Czech (aorist and imperfect), specifically for the verb *být* (‘to be’). Since the Monitor corpus is not yet annotated, a more complex query was required: `[word="(?i)(ne)?b(í|ie|ě)(š|ch|št|šet|st|set|chom|chov)(e|u|a|ě)?"]` as the first variant versus `[word="(?i)(ne)?byl.?"]` as the second. For the sake of clarity, we have chosen a visualization without any aggregation and the method with smoothing with local polynomial regression and 95% confidence intervals (upper chart of Fig. 2) and without it (lower chart of Fig. 2). The upper-left section of Fig. 2 displays the application’s input form for this module.



Timeline maker

Corpus:
HiČKoK Beta

Simple | Compare | Companions

Two competing variants in time

Query 1 (word-form or CQL):
[word="(?i)(ne)?b(í|ie|ě)(š|ch|št|šet|st|set|chom|chov)(e|u|a|ě)?"]

Query 2 (word-form or CQL):
[word="(?i)(ne)?byl.?"]

Time range:
1215-01-01 to 2023-01-01

Chart type:
Lines (no aggregation, smoothing)

Compare!

Fig. 2. Decline of the simple past tenses in Czech – horizontal axis represents years, vertical axis shows percentage of competing variants

The results (Fig. 2) indicate that the decline of the simple past tenses in Czech has been continuous since the time they were consistently documented in writing, approximately from the 13th century. This process, closely linked to the development of the aspectual system in Czech, was completed in the written language by the early 16th century and likely even earlier in the spoken language (to read more about the decline of simple past tenses in Czech, see, e.g. Kosek 2017). In case of the verb *být*, which is highly frequent (as its aorist and imperfect forms were also used in periphrastic constructions, e.g. *bieše dělal* ‘he was doing’), it can be assumed that these past tense forms persisted longer in written texts than in case of less common or functionally limited verbs.

3.3 Time-based associations

The third module, called Companions, offers a new insight on simultaneous changes in language. It tracks the frequency development of two phenomena in time. Synchronization strength (peaks and valleys of the frequency trend of both phenomena under examination) can be measured by cross-correlation (used in signal processing, for example). It compares the frequency development curves of two words or other phenomena over time and evaluates the similarity of their shapes. In addition to the correlation coefficient (r), it also calculates the lag between curves, by which one phenomenon is delayed in its development relative to another.

Companions, as a method, was originally developed to measure how two words (or other phenomena), triggered by some real-life event, start occurring and gain or lose frequency in the same time slots; in this sense the two words become "companions" in discourse/language (Cvrček and Fidler 2024).

The query input for this module is the same as for the variant comparison. In addition, the user can choose to see both trends in one graph or in graphs below each other. Beyond the visualization, this module also shows the result of the cross-correlation measure (r), which represents the degree of association between the observed phenomena.

The intent of this module is not to explore collocations or other phenomena based on syntagmatic relations, which we could easily inspect by concordancing or by other corpus tools; it is designed to examine the correlation of independent words usually linked by some real-life change. The identification of suitable 'companion' candidates is thus not straightforward. This requires at least a basic historical and socio-cultural awareness of the period under study and an idea of what might be worth looking for. As an example, we chose the words (forms of) *válka* 'war' and *mor* 'plague' to analyze the frequency of their occurrence in the period from 1400 to 1630.

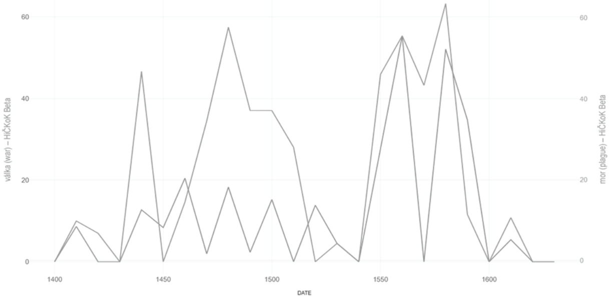


Fig. 3. Time-based association between the words *válka* 'war' and *mor* 'plague'

Both words are presented together for comparison in Fig. 3 (with the respective y-axes on either side), which demonstrates a relatively high degree of time-based association between them (r=0.62). While we can not be 100% sure that these words are true companions (i.e., they are not just randomly correlated), it strongly suggests that *válka* (‘war’) and *mor* (‘plague’) were not merely coincidentally associated during the examined period. By the end of the 17th century, the degree of association had already decreased, indicating that wars were likely no longer as commonly accompanied by plagues as they had been in earlier periods.

3.4 Semantic shift

The last module is designed for detecting the semantic shift of words. For this purpose it uses the comparison of the collocation profiles of a given word in different time periods. Collocation profiles are created for the word under examination at individual time intervals according to the user’s specification (5, 10, 20... years). The user can further specify the association measure that will be used to identify the most significant collocations (log-dice, log-likelihood, MI score, t-score), the context window in which collocations will be evaluated and the minimum collocation frequency (to filter out rare phenomena). Another option in the settings allows ignoring collocates that appear in isolation (in one time period only).

The results for each period are then visualized using a special flow-chart that allows to track the changes in the collocation profile and infer the change in meaning of a word (while maintaining the same form).

To demonstrate the potential of this module for detecting semantic shifts, we chose an example of a word that has undergone substantial semantic change over several centuries. Although there is an abundance of such words, in order to satisfactorily track their semantic change with this application, a sufficient frequency over all the periods studied is required, as this is the only way to obtain statistically reliable results. More than satisfactory results can be presented on the word *páteř* (which in modern Czech means ‘spine’) using the following query settings: time interval 1215–2023, aggregation period 25 years, collocation (association) metric log-dice, context window -3 +3, and a minimum collocation frequency of five.

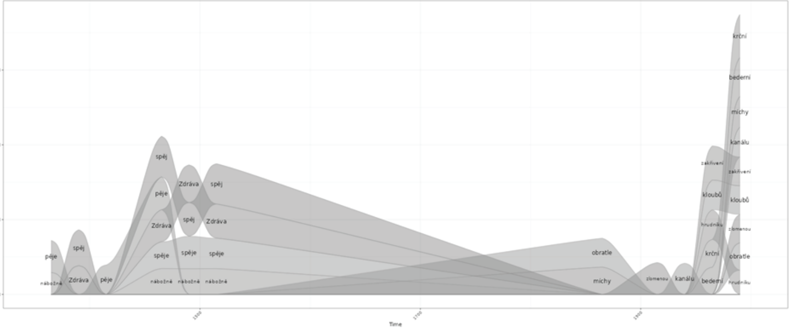


Fig. 4. Semantic shift of the word *páteř* ‘spine’ – the width of the strand for each collocate represents its association strength measured by logDice.

The meaning of the word *páteř* in modern Czech, i.e. the basic part of the skeleton of vertebrae, can be found as early as the 15th century, but at that time it played only a marginal role (see also ESSČ). As can be clearly seen from the collocation profile (Fig. 4), the word *páteř* had a different dominant meaning in Old Czech, that of *Otčenáš* 'Lord's Prayer' (note: Interestingly, the Old Czech dictionary documents another meaning of the word *páteř* in Old Czech – besides 'spine' and 'Lord's Prayer' – namely, 'rosary', which is also related to prayer, as it refers to a string of beads used for counting prayers, particularly repetitions of the Lord's Prayer). Typical collocations for the word *páteř* in Old Czech are words semantically related to prayer: *nábožně* 'piously', *(s)pěti* 'recite', and *Zdráva* 'Hail' (referring to the prayer of the Hail Mary). From around the 18th century, the results show a semantic shift. Collocates from this period are semantically related to the vertebral skeleton, such as *obratel* 'vertebra', *mícha* 'spinal cord', *kloub* 'joint', *bederní* 'lumbar', and *krční* 'cervical'. We can clearly observe a diachronic shift in which the meaning of 'spine' becomes dominant.

4 CONCLUSION

In this paper, we have attempted to illustrate that in order to effectively use corpus methods for diachronic research, two conditions need to be met: 1. corpus data that is representative of the time period, with consistent processing, annotation and metadata, 2. dedicated tools that are capable of evaluating phenomena on a timeline.

For the diachronic study of Czech, the first condition should be met prospectively by the Monitor Corpus of Czech, which is the largest achievement in this field so far, and which fulfills the ambition of a complete coverage of the timeline of Czech language development, uniform annotation and processing.

The second condition is prospectively met by the Timeline Maker application, which is specifically designed with the intent to analyze data in a diachronic perspective. It complements standard corpus tools with visualization capabilities that help interpret developmental phenomena in the language.

ACKNOWLEDGEMENTS

The HiČKoK project (History of Czech in the Corpus Continuum, see https://korpus.cz/hickok) is supported by the Technology Agency of the Czech Republic within the Programme for Support of Applied Research and Innovation SIGMA (no. TQ01000072).

This work has been supported by Charles University Research Centre program No. 24/SSH/009.

References

COHA (Corpus of Historical American English). (n.d.). English Corpora. Accessible at: https://www.english-corpora.org/coha/ [29/03/2025].

Cvrček, V., and Fidler, M. (2024). From News to Disinformation: Unpacking a Parasitic Discursive Practice of Czech Pro-Kremlin Media. Scando-Slavica, 70(1), pp. 32–54. Accessible at: https://doi.org/10.1080/00806765.2024.2317374.

Davidse, K., and De Smet, H. (2020). Diachronic corpora. In: M. Paquot – S. Th. Gries (eds.): A practical handbook of corpus linguistics, pp. 211–233. Springer International Publishing. Accessible at: https://doi.org/10.1007/978-3-030-46216-1_10.

de Marneffe, M.-C., Manning, C. D., Nivre, J., and Zeman, D. (2021). Universal dependencies. Computational Linguistics, 47(2), pp. 255–308. Accessible at: https://doi.org/10.1162/coli_a_00402.

EEBO (Early English Books Online). (n.d.). English Corpora. Accessible at: https://www.english-corpora.org/eebo/ [29/3/2025].

Elektronický slovník staré češtiny [online]. (2006–) Praha: Ústav pro jazyk český AV ČR, v. v. i., oddělení vývoje jazyka [cit. 20/06/2020]. Accessible at: http://vokabular.ujc.cas.cz.

Kosek, P. (2017). IMPERFEKTUM. In: P. Karlík – M. Nekula – J. Pleskalová (eds.): CzechEncy – Nový encyklopedický slovník češtiny. Accessible at: https://www.czechency.org/slovnik/IMPERFEKTUM [last accessed 29/03/2025].

Machálek, T. (2014). KonText – aplikace pro práci s jazykovými korpusy [Cs]. FF UK. Accessible at: https://kontext.korpus.cz.

Rissanen, M., Kytö, M., and Heikkonen, K. (eds.). (1991). The Helsinki Corpus of English Texts: Diachronic and Dialectal. University of Helsinki.

Zeman, D., Kosek, P., Březina, M., and Pergler, J. (2023). Morphosyntactic annotation in universal dependencies for old czech. Jazykovedný časopis/Journal of Linguistics, 74(1), pp. 214–222.

sciendo
DOI 10.2478/jazcas-2025-0015

ANNOTATING MOOD, TENSE AND VOICE IN CZECH CORPORA

TOMÁŠ JELÍNEK
Department of Linguistics, Faculty of Arts, Charles University, Prague,
Czech Republic (ORCID: 0000-0002-8521-4715)

JELÍNEK, Tomáš: Annotating Mood, Tense and Voice in Czech Corpora. Journal of Linguistics, 2025, Vol. 76, No 1, pp. 167 – 176.

Abstract: In the corpora of the Czech National Corpus, the verbtag attribute was introduced for annotating the mood, voice, person, and tense of both simple and compound verb forms. This article presents this attribute and the process of its automatic annotation. We then show an example of the use of verbtag in corpus research by comparing five different text genres in terms of verbal categories expressed in this attribute.

Keywords: Czech morphology, verbs, mood, tense, automatic annotation

1 INTRODUCTION

In contemporary linguistic corpora, users usually have access to lemmatization and morphological annotation of words, which facilitates their work with the corpus. However, morphological annotation is typically limited to individual, isolated forms: while the determination of morphological categories occurs within the context of the entire sentence, the tags apply only to single forms. For example, a token might be marked as a past participle of a verb, but the mood, tense, and voice of compound verb forms are not indicated.

For the corpora of the Czech National Corpus (CNC), we decided to address this shortcoming by introducing a new attribute for tagging the morphosyntactic properties of verb forms, both simple and compound, which was named **verbtag**. Verbtag distinguishes whether a verb is auxiliary or not, and for full verbs, mood, voice, person, number, and tense are annotated. The attribute was first implemented for written corpora starting with the SYN2020 corpus; later it was extended to spoken corpora, beginning with the Ortofon_v3 corpus. This article discusses the motivation for introducing verbtag, the process of its automatic annotation, and demonstrates how verbtag can be used in corpus research on statistics of verb properties in various CNC corpora.

2 RELATED WORK

In recent years, several authors have focused on the automatic annotation of mood, tense, and voice, often with the aim of improving performance in subsequent

NLP tasks such as machine translation. Loáiciga et al. (2014) annotated a small parallel English-French subcorpus from the Europarl corpus with tense, aspect, and mode using a syntactic parser and custom rules. They then trained a tense predictor on this subcorpus, achieving improvements in machine translation. Ramm et al. (2017) also annotated mood, tense, and voice using a syntactic parser and subsequent rules to enhance abstract meaning representation. Myers and Palmer (2019) used the same data with a neural-network-based classifier without prior annotation and achieved significantly better results than Ramm. Recent advances in the use of deep learning in NLP have eliminated the need for verb category annotation as an intermediate step in tasks such as machine translation.

There are few corpora with annotations of mood, tense, and other verb properties, and they are usually small. Ramm worked with the English PropBank corpus (Palmer et al. 2005), which is the Penn Treebank corpus enriched in semantic role labeling and verb categories such as tense, but it only has about 180,000 tokens. The tectogrammatical layer of the Prague Dependency Treebank in Czech is larger, with approximately 675,000 tokens, where full verbs are assigned grammatemes corresponding to verb tense, mood, etc.

In the largest current project of multilingual corpora with comparable annotation, Universal Dependencies (Marneffe et al. 2021), properties such as mood, tense, or voice are considered, but they are assigned only to isolated forms. For example, the participle in English, French, or Czech in the phrase *will be saved* / *sera sauvée* / *bude zachráněna* does not have indicative mood or future tense specified in the feats attribute. In English, it is annotated as past participle used in a passive construction, in French as past participle, and in Czech as passive participle. Only for the English version, the annotation contains a feature (voice) derived from the use of an auxiliary verb.

3 THE VERBTAG ATTRIBUTE IN CNC

3.1 Motivation

Verbs in Czech form both simple (e.g. *jdu* 'I'm going') and compound forms (e.g. *byl bych šel* 'I would have gone'). The full-verb part of a compound form can include active participles (e.g. *přišel* 'came'), passive participles (e.g. *zachráněn* 'saved'), as well as the infinitive in the compound future tense (e.g. *chodit* in *budu chodit* 'I will walk'). In compound forms, some morphosyntactic features of verbs are carried by auxiliary verbs (e.g. person), others by the main verb forms, and still others by the choice of the specific compound form. For instance, in the form *přišla byste* 'you would come', which is a polite form of 2nd person singular of the present conditional, mood and voice (active conditional) follow from the entire form (i.e., the conditional form of the verb *být* 'to be' and the past participle), the tense from the absence of another auxiliary verb in the past tense, the number and gender from the participle form, and the politeness from the number of the auxiliary verb (which differs from the number of the participle). Given the relatively free word order and frequent use of embedded subordinate clauses in written Czech, the individual parts of a compound form can be far apart: it is possible to find comprehensible Czech sentences where thirty other tokens stand between the auxiliary verb and the participle. Without annotation focused on verb categories in compound verb forms, it would be difficult for corpus users to determine these properties using corpus queries alone. Feedback from users indicated a demand for this information. Therefore, a new attribute was conceived which supplements the existing morphological tag with information derived from the entire verb form: it has been named **verbtag**.

3.2 The verbtag attribute

The verbtag attribute has been described elsewhere, e.g. (Jelínek et al. 2021), but for understanding this article, it is necessary to be familiar with it, so we will briefly summarize its properties here. The verbtag attribute is a six-position tag that supplements the original fifteen-position morphological tag. It is relevant only for verbs; for other parts of speech, all positions are empty.[1]

3.2.1 Full verb or auxiliary

The first position of the verbtag specifies whether the verb is auxiliary (A) or full (V). Only forms of the verb *být* 'to be' are considered auxiliary, rather than, e.g. *mít* 'to have'. For auxiliary verbs, all other positions of the verbtag are empty. The auxiliary verb *být* appears in combinations like *četl jsem* 'I have read', *budu číst* 'I will read', *byl bych četl* 'I would have read'; the verb *být* is considered a full verb both as existential (e.g. *Bůh je*. 'God is.') and copular (e.g. *Opak je pravdou.* 'The opposite is the truth').

3.2.2 Mood

The second position distinguishes mood: indicative (D), conditional (C), imperative (I), infinitive (F), transgressive (T), and passive participle not forming a compound verb form (O). The character O stands for "other uses" of the passive participle, such as cases when it stands alone, typically as a predicative complement (e.g. *Hořce* ***zklamán*** *se vrací do ateliéru.* 'Bitterly disappointed, he returns to his studio.'), in sentence segments without a predicate (*Cyklisté* ***vítáni****.* 'Cyclists welcome'), and often in sentences with verbs *mít* 'to have' and *zůstat* 'to stay' (*V této oblasti* ***máme rozpracováno*** *několik iniciativ.* 'We have several initiatives underway in this area.').

3.2.3 Voice

The third position indicates voice: active (A) or passive (P), with passive referring only to the periphrastic passive (*důraz je kladen* 'emphasis is placed'), not to the reflexive passive (*důraz se klade* 'emphasis is placed').

[1] Adjectives derived from passive participles are assigned 'p' on the third position of the verbtag.

3.2.4 Person

The fourth position distinguishes person (1, 2, 3, -). In some rare cases, a syntactic construction may cause a conflict between the person expressed in the morphological tag and the person in the verbtag, such as in the phrase *Bůh suď!* 'God be the judge!' where the imperative has the 2nd person in the tag, whereas the 3rd person in verbtag.

3.2.5 Number

The fifth position indicates number: singular (S), plural (P), and the form of politeness (v); the form of politeness is identified only in the combination of an auxiliary verb with a past or passive participle (e.g. *řekla jste* 'you said'), where the number of the auxiliary verb differs from the number of the participle.

3.2.6 Tense

The last, sixth position indicates tense: pluperfect (Q), past (R), present (P), future (F), and present or future of biaspectual verbs (B). The pluperfect, as found in sentences such as *Víno, jemuž dávno byl odvykl, uvolnilo nyní cenzuru myšlenek i slov.* 'The wine he had given up a long time ago now loosened the censorship of thoughts and words.' is rarely used in Czech. In the texts we annotated automatically, cases where this tense is incorrectly determined due to a text error (such as a typo or a missing comma) are much more frequent than the correct ones.

3.3 Automatic annotation with the verbtag attribute

To allow users access to verbal categories contained in verbtag, it was first necessary to integrate the annotation of verbtag into our annotation process. The annotation process used for both written and spoken corpora of the CNC in the SYN2020 corpus standard has been described for written corpora (Jelínek et al. 2021) and spoken corpora (Jelínek 2023). Here, we focus only on the automatic annotation of verbtag. Extending the original annotation to include verbtag required adding verbtag to the training data (both written and spoken). For written text, where we use a rule-based module for disambiguation, it was necessary to design and test several disambiguation rules. For both written and spoken text, neural tagger models were trained.

3.3.1 Adding verbtag to training data

For training the neural tagger and testing the entire annotation process, we use the training data created in the CNC named Etalon corpus, which consist of approximately 2.25 million tokens of written text and additional 200,000 tokens of spoken text. The written Etalon comprises a balanced selection of texts from the three main genre types of the SYN2020 corpus: fiction, non-fiction (academic and professional literature), and newspapers and magazines. The spoken Etalon was selected from the Ortofon corpus, which consists of transcriptions of spontaneous speech into phonetic and orthographic levels, with the orthographic level used for tagging. These data were previously manually annotated for lemmas and morphological tags by two annotators. With the introduction of verbtag, verbtags were assigned to tokens in the data. For verb forms that are ambiguous in terms of verbtag (e.g. all participles, imperfective infinitives, forms of the verb *být* 'to be'), a set of all potential verbtags was added, from which two annotators independently selected the correct verbtag based on the context of the sentence.

3.3.2 Adding verbtag to data during annotation

In the automatic annotation process, first a set of all possible combinations of lemmas and tags for a given token is assigned to each token. This set is then refined in subsequent disambiguation steps until only one (presumably correct) combination remains. We assign this set based on a version of the MorfFlex dictionary modified for the purposes of the CNC. However, verbtag is not included in this dictionary, as it would multiply the number of dictionary entries and slow down its operation. Instead, an additional step has been included in the processing which expands the set of lemmas and tags (on average, 5.56 tags per token) with verbtag (on average, 7.73 verbtags per verb). The subsequent disambiguation steps then select from the set of lemma-tag-verbtag triplets.

3.3.3 Expanding the linguistic rule module for the verbtag disambiguation

For written texts, a combination of a rule-based module and a neural tagger is used for automatic annotation: we refer to this process as hybrid disambiguation. First, the rule-based module is applied, and for tokens that the rule-based module cannot fully disambiguate, the final combination of lemma and tag is selected by the neural tagger.

With the introduction of verbtag, it was necessary to expand the rule-based module with disambiguation rules focused on verbtag. Generally, the rules work by gradually removing tags, verbtags and lemmas from individual tokens that are not correct in a given context. The rules are applied repeatedly, so the action of one rule can enable the later application of another one. One such verbtag rule is the removal of the conditional interpretation from a participle in a sentence where the conditional form of the verb *být* 'to be', e.g. *bych* 'I would' does not appear, and conversely, the removal of all interpretations except conditional for a past participle located in the same clause with a conditional form of the auxiliary verb. The conditional form can be separated from the participle by an embedded clause; in the case of a subordinate clause, the presence of the conditional form is processed independently. For example, in the sentence *Řekl bych, že moc peněz nevydělal.* 'I would say that he did not earn much money,' the first participle *Řekl* 'said' is undoubtedly a conditional, while the second participle *nevydělal* 'did not earn' is an indicative. Approximately 80 rules focused on verbtag were added.

3.3.4 Training neural tagger models

For disambiguation in spoken corpora and for the second phase of disambiguation in written corpora, a deep-learning-based tagger is used. This is an unpublished, beta version of a tagger, developed as part of the MorphoDiTa family of NLP tools, we call this version MorphoDiTa-research. Its properties are described in (Straka et al. 2019).

After adding verbtag, a model for this tagger for written text was independently trained based on the Etalon corpus data of written language, and another tagger model for spoken text was trained based on the combined data of the Etalon corpus of both written and spoken language, as there is not enough data to train on spoken language alone.

Adding verbtag to the training data increased ambiguity in the text by 39% (the average number of lemma-tag combinations per token in written data is 4.03, the average number of lemma-tag-verbtag combinations per token is 5.60). The accuracy of the tagger during training on the written corpus decreased by only about 0.2% (from 97.69% to 97.47%), indicating that the neural tagger handled the more complex data very well.

3.3.5 Disambiguation accuracy

We measured disambiguation accuracy using the method of ten-fold cross-validation. In the case of spoken data, the tagger was trained on both written and spoken data, but testing was conducted only on spoken data.

Tab. 1 shows disambiguation accuracy. The first column indicates the accuracy of assigning the correct verbtag calculated only for verbs, the second column shows the accuracy of morphological tags calculated for all tokens, the third one shows the accuracy of both tag and verbtag, and the fourth one the accuracy of the combination of lemma, tag, and verbtag (i.e. all attributes). The first row shows the accuracy of tagging written texts using the process used by the CNC for its written corpora, i.e., a combination of linguistic rules and the neural tagger. The second row shows the accuracy of tagging spoken data using the neural tagger, MorphoDiTa-research.

	Verbtag (verbs)	Tag (all tokens)	Tag+Verbtag (all tokens)	All (all tokens)
Written: hybrid approach	99.08	97.76	97.70	97.62
Spoken: neural tagger	96.95	92.95	92.56	92.34

Tab. 1. Disambiguation accuracy

The accuracy of tagging spoken corpora is significantly lower in all measured parameters compared to the accuracy of tagging written corpora. This is primarily due to two reasons: firstly, disambiguation of spoken language is more challenging due to its characteristics (non-standard syntax, word repetition, unfinished statements, etc.), and secondly, we have only a relatively small amount of training data available, with most of the data used to train the model coming from written text.

In written text, verb tag annotation is reliable, with the hybrid process incorrectly assigning verbtags to less than one percent of verbs. The highest error rate is in tag annotation, mainly due to the challenges of case ambiguity in Czech.

4 STATISTICS OF VERB FORMS BASED ON VERBTAG

4.1 Corpora

We provide statistics for three basic text genres of the SYN2020 corpus, a representative corpus of contemporary written Czech: newspapers and magazines (NMG), non-fiction (NFC), and fiction (FIC), and for two corpora of contemporary spoken Czech: the Ortofon_v3 corpus (ORT), consisting of transcripts of spontaneous informal spoken Czech, and the Orator_v3 corpus (ORA), consisting of transcripts of formal, prepared monologic speeches.

4.2 Proportion of auxiliary verbs

Tab. 2 shows the proportion of auxiliary (A) and full verbs (V) in the total number of verbs in the corpus.

	NMG	**NFC**	**FIC**	**ORA**	**ORT**
V	89.47	87.75	87.43	87.88	84.14
A	10.53	12.25	12.57	12.22	15.86
Total	**100**	**100**	**100**	**100**	**100**

Tab. 2. Proportion of auxiliary verbs

The higher proportion of auxiliary verbs in the Ortofon corpus corresponds to a significantly higher proportion of the first person past tense indicative mood in this corpus.

4.3 Proportion of mood

Tab. 3 presents the proportion of mood among full verbs: indicative (D), conditional (C), imperative (I), infinitive (F), transgressive (T) and other uses of passive participle (O).

	NMG	**NFC**	**FIC**	**ORA**	**ORT**
D	80.68	78.22	80.93	80.52	83.82
C	4.34	4.31	5.59	4.67	4.89
I	1.28	1.81	2.01	1.73	2.32

	NMG	NFC	FIC	ORA	ORT
F	13.38	15.15	11.10	12.80	8.88
T	0.04	0.09	0.11	0.02	0.01
O	0.29	0.43	0.26	0.27	0.07
Total	100	100	100	100	100

Tab. 3. Proportion of mood

In all corpora, the indicative mood (D) prevails. In non-fiction, the proportion of the infinitive (F) is noticeably higher, because of a greater representation of modality (modal verbs, modal nouns, the adverb *lze* 'may' etc.) and more complex sentence constructions in this subcorpus. In fiction, the proportion of the conditional (C) is slightly higher.

4.4 Proportion of voice

Tab. 4 presents the proportion of voice: active (A) and periphrastic passive (P) among full verbs.

	NMG	NFC	FIC	ORA	ORT
A	97.18	93.93	98.84	97.98	99.83
P	2.82	6.07	1.16	2.02	0.17
Total	**100**	**100**	**100**	**100**	**100**

Tab. 4. Proportion of voice

The proportion of the periphrastic passive in non-fiction is significantly higher than in other corpora, whereas in the Ortofon corpus, its proportion is negligible. It is noteworthy that in the Orator corpus, which is a corpus of formal spoken discourse, the proportion of the periphrastic passive is higher than in the written corpus of fiction.

4.5 Proportion of person

Tab. 5 shows the proportion of person among full verbs, ignoring cases when person is not expressed (infinitive, transgressive).

	NMG	NFC	FIC	ORA	ORT
1	11.73	12.16	17.37	22.43	29.29
2	4.36	4.73	7.94	9.13	13.99
3	83.91	83.11	74.69	68.44	56.73
Total	**100**	**100**	**100**	**100**	**100**

Tab. 5. Proportion of person

The highest proportion of the first and second person is in the corpus of spontaneous spoken language Ortofon, and the lowest in the subcorpus of newspapers.

4.6 Proportion of number

Tab. 6 shows the proportion of singular (S), plural (P), and form of politeness (v) in the corpora studied.

	NMG	NFC	FIC	ORA	ORT
S	70.23	68.57	81.79	64.66	81.02
P	29.43	31.29	17.70	35.26	18.92
v	0.35	0.14	0.52	0.08	0.07
Total	**100**	**100**	**100**	**100**	**100**

Tab. 6. Proportion of number

The differences in the proportion of singular and plural among the corpora are the largest among the observed attributes. In the Orator corpus, plural is used approximately twice as much compared to the fiction subcorpus (the most frequent in the Orator is the 3rd person plural present, followed by the 1st person plural present, which is largely pluralis modestiae). The fiction subcorpus has a similar proportion of number as the Ortofon corpus, and the non-fiction literature subcorpus is similar to the subcorpus of newspapers.

4.7 Proportion of tense

Tab. 7 presents the proportion of past (R), present (P), and future (F) tense and undifferentiated present or future tense of biaspectual verbs (B). We disregard pluperfect (Q) because its annotation is not reliable.

	NMG	NFC	FIC	ORA	ORT
R	37.87	33.72	55.36	23.28	31.00
P	49.65	55.57	34.92	62.99	55.83
F	12.04	10.12	9.58	13.13	12.94
B	0.43	0.59	0.14	0.61	0.23
Total	**100**	**100**	**100**	**100**	**100**

Tab. 7. Proportion of tense

4.8 Comparison of genre types

The above comparison of five broadly defined genre types is rough; more detailed work with both verbtag and tag values (which exceeds the scope of this article) would better show the differences between genres in the use of verb forms. However, we can draw several conclusions.

It cannot be said that written texts behave in one way and spoken texts in another in terms of verbal categories. In the case of person, the written-text subcorpora of newspapers and of non-fiction stand on one side, and the spoken corpora and the subcorpus of fiction on the other. In terms of number, non-fiction and the corpus of formal spoken language behave similarly, while written fiction is close to the corpus of informal spoken language, with the subcorpus of newspapers standing between them. For mood and voice, the situation is similar; non-fiction and newspapers subcorpora have similar proportions along with the corpus of formal spoken language, while the fiction subcorpus and the corpus of informal spoken language differ from them, being similar to each other.

5 CONCLUSION

The verbtag attribute, which has recently been introduced into the annotation of both written and spoken corpora of the CNC, is a useful tool for searching verb forms in the corpus, regardless of whether these forms are compound or simple. The annotation of verbtag in both written and spoken corpora is relatively reliable. Currently, only a small portion of CNC users utilize verbtag, so the aim of this article was also to raise awareness about it.

ACKNOWLEDGEMENTS

The research has been supported by the Ministry of Education of the Czech Republic through the project Czech National Corpus (LM2023044), and by Charles University through the Cooperatio programme (Linguistics).

References

Jelínek, T., Křivan, J., Petkevič, V., Skoumalová, H., and Šindlerová, J. (2021). SYN2020: A New Corpus of Czech with an Innovated Annotation. Proceedings of TSD 2021, Springer, pp. 48–59.

Jelínek, T. (2023). Morphological Tagging and Lemmatization of Spoken Corpora of Czech. Proceedings of TSD 2023, Springer, pp. 154–163.

Marneffe, M. C., Manning, C., Nivre, J., and Zeman, D. (2021). Universal Dependencies. Computational Linguistics, 47(2), pp. 255–308.

Myers, S., and Palmer, M. (2019, August). ClearTAC: Verb Tense, Aspect, and Form Classification Using Neural Nets. Proceedings of the First International Workshop on Designing Meaning Representations, pp. 136–140.

Palmer, M., Gildea, D., and Kingsbury, P. (2005). The Proposition Bank: An Annotated Corpus of Semantic Roles. Computational Linguistics, 31(1), pp. 71–106.

Ramm, A., Loáiciga, S., Friedrich, A., and Fraser, A. (2017). Annotating tense, mood and voice for English, French and German. Proceedings of ACL 2017, System Demonstrations, pp. 1–6.

Straka, M., Straková, J., and Hajič, J. (2019). Czech Text Processing with Contextual Embeddings: POS Tagging, Lemmatization, Parsing and NER. Proceedings of TSD 2019, Springer, pp. 137–150.

sciendo
DOI 10.2478/jazcas-2025-0016

CHALLENGES IN DEVELOPING A GLAGOLITIC REVERSE DICTIONARY OF CROATIAN CHURCH SLAVONIC

JOSIP MIHALJEVIĆ
Department of Dictionary of the Croatian Redaction of Church Slavonic, Old Church Slavonic Institute, Zagreb, Croatia (ORCID: 0000-0002-7482-7663)

MIHALJEVIĆ, Josip: Challenges in Developing a Glagolitic Reverse Dictionary of Croatian Church Slavonic. Journal of Linguistics, 2025, Vol. 76, No 1, pp. 177 – 191.

Abstract: The Glagolitic script is the oldest Slavic script and one of the two Slavic scripts (the other is Cyrillic). It was actively used in Croatia until the 19th century. Today, the Glagolitic script is a symbol of Croatian national identity. It has significant cultural, artistic, and esthetic value. The goals of the Old Church Slavonic Institute are researching, discovering, recognizing, systematically listing, and editing Glagolitic manuscripts, inscriptions, and printed books. One of the Institute's main tasks is to create a digital version of the *Dictionary of the Croatian Redaction of Church Slavonic*. This is a long-term project, and the dictionary has so far only been published in printed form (A to I). In addition to the creation of the digital dictionary, additional web content is being developed (e.g. games), which will provide an additional source and tool for dictionary compilers and users. One of these additional contents is the reverse dictionary, which consists of the headwords written in Glagolitic from the Dictionary. A reverse dictionary is a dictionary in which the words are sorted alphabetically by their final rather than initial letters, i.e. the words are sorted based on their endings. This allows users to look up words by focusing on their final segments. Reverse dictionaries are useful for scientific research, especially for the study of word-formation (e.g. the study of masculine-feminine pairs), but they can also be used for other purposes, such as finding rhymes and compiling educational games. A reverse dictionary for Croatian Church Slavonic is thus a useful tool for further research. The problem is not only that the word search is not done in a classic matter (where the string is searched from the beginning of the word) but also that the Glagolitic script has some specific letters which have to be mapped to certain characters in the Latin script and that the reverse dictionary has headwords in two scripts (Latin and Glagolitic). Currently, no available software solution can accommodate for these problems, so the reverse dictionary is being developed from scratch using a custom code. This paper presents the challenges and solutions in the development of the Croatian Church Slavonic Reverse Dictionary.

Keywords: Croatian Church Slavonic Language, Glagolitic script, problem-solving, reverse dictionary

1 INTRODUCTION

The Glagolitic script is the oldest Slavic script and predates the Cyrillic script. It was created in the 9th century by St. Cyril (originally named Constantine) from Thessaloniki, Greece, following the initiative of Rastislav of Moravia to promote

Christianity among the people of Great Moravia in Central Europe. Cyril devised a new alphabetic-phonological script, the Glagolitic script. He also developed the literary language called Old Church Slavonic in 863. Cyril's closest collaborator was his older brother, Methodius. Through their efforts, the foundations of Slavic written culture and civilization were established (Bratulić et al. 2009, pp. 36–46). The Glagolitic script and Old Church Slavonic language were subsequently spread by the disciples of Cyril and Methodius. However, following the banishment of their disciples in the late 9th century, the use of Glagolitic gradually declined, as it was increasingly replaced by the Latin and Cyrillic scripts (Mihaljević et al. 2024, p. 17). The Glagolitic script has historically been used in several parts of Europe, including the territories of present-day Slovakia, the Czech Republic, Bulgaria, North Macedonia, Slovenia, and Croatia (Japundžić 1998). However, it persisted the longest in Croatia, where it remained in use for religious and legal documents until the 19th and early 20th centuries (Mihaljević et al. 2024, pp. 17–18). The Glagolitic script is significant for Croatian history as the first script used to record the Croatian language. Its development, adaptation, and preservation by the Croats underscore its importance in the cultural and religious life of the region.

The Old Church Slavonic Institute is the leading institution in Croatia dedicated to the study, preservation, and promotion of the Glagolitic script, the Croatian Church Slavonic language, and other aspects of the Glagolitic heritage. The Institute's recent focus has been on the creation of digital content on its website (stin.hr). Currently, the website has two virtual exhibitions,[1, 2] a program for creating images based on the custom Glagolitic font FSGLA[3] created by the Institute, a database for Glagolitic chants,[4] games for learning the Glagolitic scripts, Croatian Church Slavonic words, and other Glagolitic content.[5] There are also plans to develop an interactive timeline that records the historical facts connected with the usage of the Glagolitic script, an e-grammar, an interactive map for Glagolitic monuments, new virtual exhibitions, and games. This content is developed within the project *Development of the Digital Infrastructure Model of the Old Church Slavonic Institute* (DigiSTIN).[6] *The Dictionary of the Croatian Redaction of Church Slavonic* is a long-term lexicographic project dedicated to documenting the vocabulary of the oldest Croatian literary language. It is based on a corpus of Glagolitic texts spanning from the 11th to the 17th century and includes detailed entries with equivalents in Latin, Greek, Croatian, and English. The content of each dictionary entry clearly and concisely presents orthographic, syntactic, morphological, semantic, stylistic, distributional, usage-related, illustrative, and inter-

[1] https://stin.hr/novakov-misal/ [14/02/2025]

[2] https://stin.hr/zgombicev-zbornik/ [14/02/2025]

[3] https://stin.hr/stvori-sliku-s-tekstom-na-glagoljici/ [14/02/2025]

[4] https://stin.hr/repozitorij-glagoljaskoga-pjevanja/ [14/02/2025]

[5] https://stin.hr/obrazovne-igre/ [14/02/2025]

[6] https://stin.hr/en/content/digistin-en/ [14/02/2025]

lingual information about the word (Badurina-Stipčević et al. 2012, pp. I–II). In addition to the printed volumes, the project now includes the development of a digital edition, as well as interactive content such as educational games and a reverse dictionary to support linguistic research and public engagement.

2 WHAT IS A REVERSE DICTIONARY?

Depending on classification criteria, dictionaries are classified in various ways (Tab. 1). Based on the scope of the lexicon they cover, dictionaries are categorized as general or specialized. General dictionaries contain words that are in common language and frequently used terms. Specialized dictionaries, on the other hand, contain words that are limited to a specific field (e.g. terminological dictionaries that include only terms from a particular profession), a specific functional style (e.g. jargon dictionaries), words of a particular origin (e.g. dictionary of foreign words), or words having a particular relationship (e.g. dictionaries of synonyms, antonyms, or homonyms). Based on the number of languages they include, dictionaries can be monolingual, bilingual, or multilingual. The content in a dictionary is organized into dictionary or lexicographic entries (Mihaljević and Hudeček 2024, p. 448). According to their relationship to the corpus, dictionaries are divided into corpus-illustrated dictionaries (a classical dictionary but illustrated by examples from the corpus), corpus-based dictionaries (where the editor uses the corpus as a reference but has the freedom to decide what to include and can modify the dictionary as needed), and corpus-driven dictionaries (where the dictionary records everything present in the corpus and does not include anything that is not found within it) (Štrkalj Despot and Möhrs 2015, p. 342). Based on the period they cover, dictionaries are divided into historical and contemporary dictionaries. According to the arrangement of the entries, dictionaries can be alphabetical, systematic, or frequency-based. According to the criteria of order, alphabetical dictionaries can be alphabetical (alphabetized from the beginning) and reverse alphabetical (Hudeček, Mihaljević and Jozić 2024, p. 593).

Classification criteria	**Type**			
Scope	general		special	
Number of Languages	monolingual	bilingual		multilingual
Relation to Corpus	classical (not based on a corpus)	corpus-illustrated	corpus-based	corpus-driven

<table>
<tr><td>Time Period Covered</td><td colspan="2">contemporary</td><td colspan="2">historical</td></tr>
<tr><td>Arrangement of Dictionary Entries</td><td>systematic</td><td colspan="2">alphabetical
(alphabetized from the beginning
or from the end)</td><td>frequency-based</td></tr>
</table>

Tab. 1. Criteria for Dictionary Classification (Hudeček, Mihaljević and Jozić 2024, p. 593)

A reverse dictionary or inverted dictionary is a specialized linguistic resource where words are organized in reverse alphabetical order, based on their endings rather than their beginnings (Lewis and Mihaljević 2018, p. 21). For example, in a reverse dictionary, instead of searching for words that start with a given string, we can type -*tic* to find words that end with it, such as *linguistic*, *phonetic*, *didactic*, *mystic*, *arctic*, *hectic*, etc. The term *reverse dictionary* is also used for another type of dictionary, such as the reversedictionary.org[7] and OneLook Dictionary,[8] where a definition, phrase, or example is entered into the search bar, and the search results display a list of words that best match the meaning of the given expression. However, this paper will focus solely on the first type of reverse dictionaries, which orders words by their endings.

This arrangement allows for the analysis of word formation processes, such as derivation and inflection, by facilitating the identification of words sharing common suffixes. Additionally, reverse dictionaries are valuable tools for poets seeking rhyming words and for linguists conducting phonological studies (Hudeček, Mihaljević and Jozić 2024, p. 601). Reverse dictionaries have been used in the study of morphological relationships between words, the analysis of gender noun pairs, and the examination of suffixal derivatives (Mihaljević and Hudeček 2024, p. 448).

The *Dictionary of the Croatian Redaction of Church Slavonic* is an alphabetical, general, historical, corpus-driven multilingual dictionary. The reverse dictionary based on it is a reverse alphabetical, specialized, historical, corpus-driven, monolingual dictionary in two alphabets. It is designed to assist dictionary editors and linguists in finding specific words, analyzing word-formation patterns, and tracing linguistic evolution within the Croatian redaction of Church Slavonic. Currently, it includes only the base forms (lemmas) of the headwords from the source dictionary. Inflected forms are not included at this stage, although such an extension is being considered for the future to improve morphological coverage and search functionality.

[7] https://reversedictionary.org/ [17/02/2025]

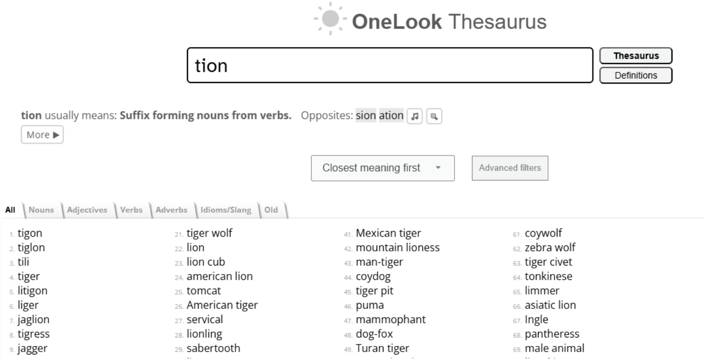
[8] https://www.onelook.com/ [17/02/2025]

3 PREVIOUS WORK

Reverse dictionaries were difficult to compile until the invention of computers, which facilitated their creation, and they are now commonly produced (UKEssays 2018). The first computer-produced reverse dictionary by Stahl and Scavnicky, *Reverse Dictionary of the Spanish Language*, was published in 1973. Reverse dictionaries exist in many languages, such as English, French, German, Russian, Slovenian, Serbian, Italian, and Ukrainian (Grčević 2017, p. 2). However, most of these reverse dictionaries are not available online in the form of a web page where their contents can be directly searched. Reverse dictionaries in electronic form are mostly published as computer files (usually scanned books in PDF format) that need to be opened using pre-installed software (Mihaljević 2022, p. 52). This is also the case with the Croatian language. The first Croatian reverse dictionary is *Rückläufiges Wörterbuch des Serbokroatischen* by Josip Matešić (1965–1967), which is available online on the University of Innsbruck's website as an .mdb (Microsoft Database) file that can be accessed through Microsoft Access. There is also an option to perform a reverse search for words in the Croatian Collocation Database.[9]

Several other reverse dictionaries are available online, e.g. *MyStilus reverse dictionary* for English and Spanish,[10] *Cronopista diccionario de rimas*[11] for Italian, and *Reimlexikon*[12] for German. The *OneLook Thesaurus Search tool*[13] enables users to find English words ending with specific letters by using the asterisk (*) wildcard. For example, searching for **tion* will display words that end in -*tion* (Fig. 1). There are also tools and online resources for finding words that end in a certain sequence, e.g. *Rhymer*,[14] *RhymeZone*[15] can be used to find rhymes for words in English, and the *Word Finder tool* from *YourDictionary*[16] can be used to find words starting and ending with certain letters for a game called Scrabble. Users who are familiar with the programming language Python can also use *Natural Language Toolkit (NLTK)* library[17] for creating their personal reverse dictionaries.

[9] http://ihjj.hr/kolokacije/english/ [26/02/2025]
[10] https://www.mystilus.com/ [17/02/2025]
[11] https://www.cronopista.com/dict-fe/ [26/02/2025]
[12] https://www.reimlexikon.net/ [26/02/2025]
[13] https://www.onelook.com/thesaurus/ [17/02/2025]
[14] https://www.rhymer.com/that.html [26/02/2025]
[15] https://www.rhymezone.com/ [26/02/2025]
[16] https://wordfinder.yourdictionary.com/ [26/02/2025]
[17] https://www.nltk.org/ [26/02/2025]

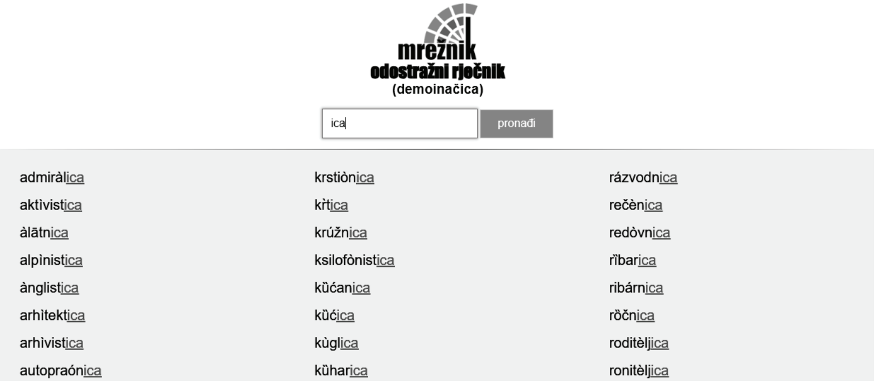


tion usually means: **Suffix forming nouns from verbs.** Opposites: sion ation

1. tigon
2. tiglon
3. tili
4. tiger
5. litigon
6. liger
7. jaglion
8. tigress
9. jagger
21. tiger wolf
22. lion
23. lion cub
24. american lion
25. tomcat
26. American tiger
27. servical
28. lionling
29. sabertooth
41. Mexican tiger
42. mountain lioness
43. man-tiger
44. coydog
45. tiger pit
46. puma
47. mammophant
48. dog-fox
49. Turan tiger
61. coywolf
62. zebra wolf
63. tiger civet
64. tonkinese
65. limmer
66. asiatic lion
67. Ingle
68. pantheress
69. male animal

Fig. 1. Using *OneLook Thesaurus* as a reverse dictionary to find words ending with *-tion*

Within the *Croatian Web Dictionary – Mrežnik* project,[18] a demo version of a reverse dictionary for the Croatian language has been developed and published.[19] This dictionary includes words currently present in *Mrežnik*. At this stage, the reverse dictionary features a functional search bar that retrieves words ending with a specified string. The search results are displayed in a structured list, with the matching suffixes visually highlighted for clarity. Additionally, each word in the results is hyperlinked to its corresponding dictionary entry, enabling seamless access to further lexical information (Fig. 2).



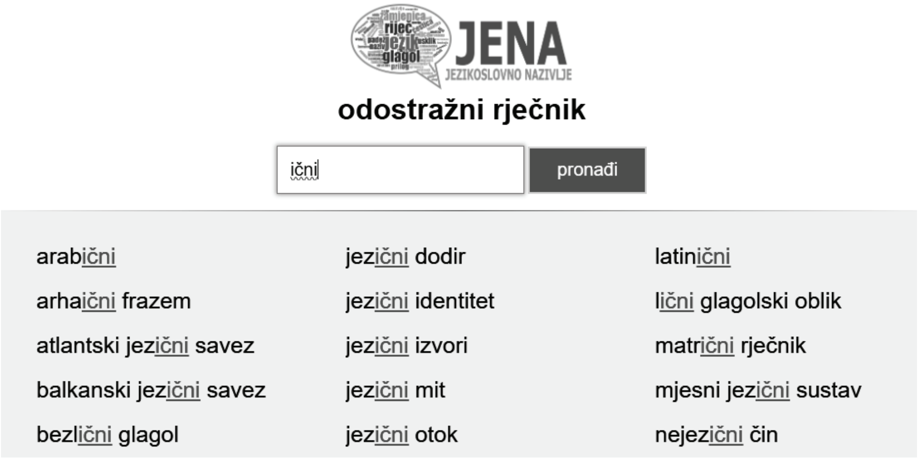
admiràlica
aktìvistica
àlātnica
alpìnistica
ànglistica
arhìtektica
arhìvistica
autopraónica
krstiònica
kȑtica
krúžnica
ksilofònistica
kȕćanica
kȕćica
kùglica
kȕharica
rázvodnica
rečènica
redòvnica
rȉbarica
ribárnica
rȍčnica
roditèljica
ronitèljica

Fig. 2. Demo version of the reverse dictionary for *Mrežnik*, e.g. searching for words that end with *-ica*

[18] The project *Croatian Web Dictionary – Mrežnik* aims at creating a free, monolingual, easily searchable hypertext online dictionary of Standard Croatian. It will be the first web-born dictionary of the Croatian language. More about the dictionary in *Anatomija rječnika Hrvatski mrežni rječnik – Mrežnik* (Hudeček, Mihaljević and Jozić 2024).

[19] https://rjecnik.hr/mreznik/wp-content/odostrazni/ [17/02/2025]

A specialized reverse dictionary has been compiled within the project *Croatian Linguistic Terminology – Jena*. A reverse dictionary has been created based on the alphabetical index of the *Jena* database, allowing users to search for terms based on their endings, including multi-word terms (Fig. 3).[20] This dictionary is used as a tool for making certain normative decisions (e.g. determining whether an adjective derived from a noun ending in *-ica* should take the suffix *-ni* or *-ski*) (Mihaljević et al. 2023, p. 122).



odostražni rječnik

ični

- arabični
- arhaični frazem
- atlantski jezični savez
- balkanski jezični savez
- bezlični glagol
- jezični dodir
- jezični identitet
- jezični izvori
- jezični mit
- jezični otok
- latinični
- lični glagolski oblik
- matrični rječnik
- mjesni jezični sustav
- nejezični čin

Fig. 3. Demo version of the reverse dictionary for *Jena*, e.g. searching for multi-word terms that end with *-ični*

Another specialized reverse dictionary was made for the project *Male and Female in the Croatian Language*.[21] This reverse dictionary is used to gather information that can be useful in standardizing male-female word pairs in the Croatian language.

However, there are no reverse dictionary options available for the Croatian Church Slavic language, making such resources rare and valuable for linguistic research. This is why the reverse dictionary for the Croatian Church Slavic language is useful and innovative.

4 THE PROCESS OF CREATING THE REVERSE DICTIONARY

The creation of a reverse dictionary involves compiling a comprehensive list of words from a language and systematically reversing their letter sequences. These

[20] https://jena.jezik.hr/wp-content/odostrazni-jena/ [17/02/2025]

[21] https://muskozensko.jezik.hr/odostrazni/ [17/02/2025]

reversed words are then sorted alphabetically, enabling users to search for terms based on their suffixes or endings. This method is particularly useful in languages with rich morphological structure and word formation, as it aids in the systematic study of word formation and the identification of patterns within the language system. In the case of the previously mentioned online Croatian reverse dictionaries, they were not developed using standard dictionary systems (e.g. TshwaneLex) or content management systems (e.g. WordPress), as existing frameworks did not provide functionalities for searching and systematizing words in reverse order. Thus, the only solution was to create a custom code that would enable such functionalities. The development process primarily relied on jQuery, .txt files, and HTML5. The reverse dictionary website was designed as a Single-page application,[22] meaning that all necessary components and functionalities were loaded dynamically within a single web page. The design was entirely custom using handwritten HTML and CSS code for structural and stylistic elements. The wordlist for searchable words was placed inside a simple .txt file, where each word occupied a separate line. When a user enters a search query, such as *-ample*, the jQuery script first determines the length of the input string (e.g. *-ample* has the length of five characters). The script then accesses the .txt file and processes its contents line by line. During this process, the algorithm applies a series of filtering steps:

1. **Preliminary filtering** – words shorter than the input string are excluded from further analysis. For instance, in the case of *-ample*, any words containing fewer than five characters are ignored.
2. **Reverse matching** – the script extracts the final segment of each remaining word, corresponding to the length of the input string (e.g. the last five characters in this case). To achieve this, the length of the entire word is first determined, and the length of the input string is subtracted from it. This calculation provides the starting index from which the script begins analyzing the word in reverse order. The substring extracted from this index onward is then compared to the user's query. If a match is found, the word is retained for the final output. In the case of searching through multi-word terms, such as those found in the *Jena* dictionary (Fig. 3), each line of text is first split into individual words using empty spaces as delimiters. Each word is then processed separately, applying the reverse matching steps described above.
3. **Normalization** – to enhance search flexibility and increase the number of results, the system incorporates diacritic and case insensitivity. This ensures that variations of the same letter, such as *À*, *Á*, *Â*, *Ã*, *Ä*, *Å*, *à*, *á*, *â*, *â*, *ã*, *ä*, *å*, *ā*, and *ȁ*, are

[22] A single-page application (SPA) is a web application or website that loads a single HTML page and dynamically updates content as the user interacts with the app, without refreshing the entire page. This approach provides a more fluid user experience, like that of a desktop application (Mozilla Developer Network 2024).

all treated as *a* during the search process when users search for terms ending with -*a*. However, if the user explicitly includes a diacritic in the search query, e.g. -*á*, it will not automatically convert to its simpler form -*a*. This approach provides a balance between broader search inclusivity and precision, allowing users to obtain more comprehensive results while still enabling specific queries when necessary.

The final search results are presented in a structured list format, displaying all matching words. Entries are sorted alphabetically but based on their initial letters rather than their endings. After a new search query, the previous results are cleared, and the search process is restarted from the initial steps, ensuring that each query is processed independently.

5 CHALLENGES IN CREATING A REVERSE DICTIONARY FOR ENTRIES IN THE GLAGOLITIC SCRIPT

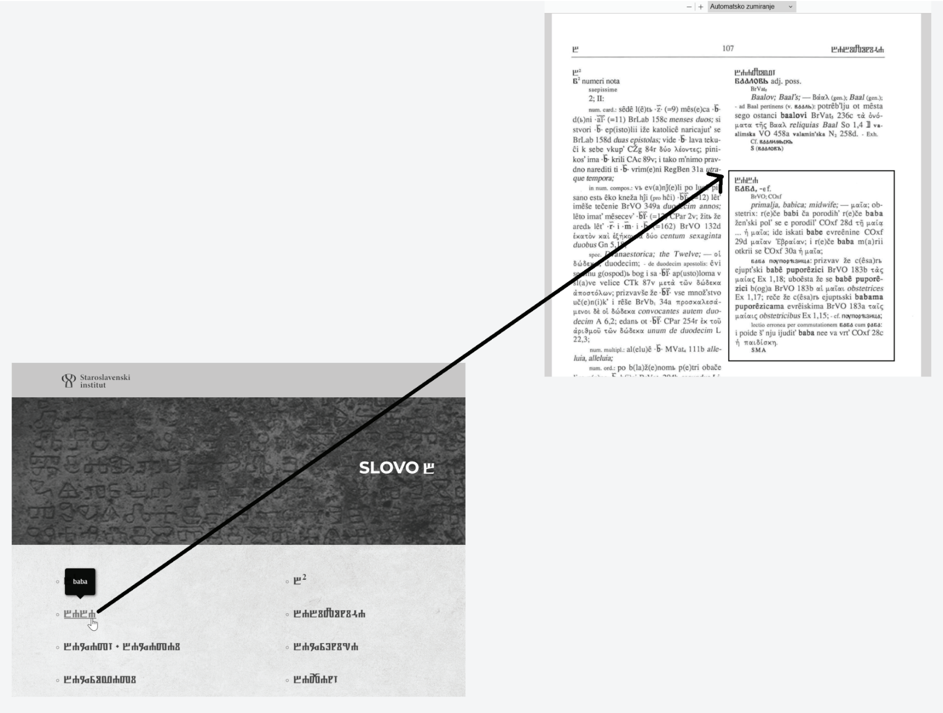
The development of a reverse dictionary for the *Dictionary of the Croatian Redaction of Church Slavonic* posed much greater challenges compared to the previously mentioned reverse dictionaries. While the general dictionary is currently available in PDF format, its web version is still in development. The key complexity lies in the fact that dictionary entries are written in multiple scripts, including Glagolitic, Latin, Cyrillic, and Greek. Headwords are written in Glagolitic and Cyrillic. Cyrillic script is also used for inflectional endings. Greek script is used for Greek equivalents and, in some cases, for original source texts from which translations were made. The remainder of the entry content is written in the Latin script. Although a structured data format for dictionary entries has recently been implemented,[23] the initial phase of online dictionary development adopted a simpler approach, where headwords were linked directly to their corresponding PDF pages. For example, clicking on the entry ⰱⰰⰱⰰ (lat. baba) would redirect users to page 107 of the dictionary PDF version, where the corresponding entry is located (Fig. 4).

During the development of the reverse dictionary, it was decided that headwords would be searchable in both Latin and Glagolitic scripts. In both the printed and digital versions of the *Dictionary of the Croatian Redaction of Church Slavonic*, headwords are written in the Glagolitic script, so the same principle was retained in the reverse dictionary, with the only addition being that the Latin counterpart is displayed alongside the Glagolitic form. One of the challenges in searching the Glagolitic script is that, although it is available in the Unicode format,[24] the existing Unicode Glagolitic symbols are not specific to the Croatian Glagolitic script, which primarily uses the Angular Glagolitic variant. Additionally, another issue with these

[23] More about dictionary content structure in Mihaljević and Mihaljević (2024).

[24] https://en.wikipedia.org/wiki/Glagolitic_(Unicode_block) [20/02/2025]

Unicode symbols is that they are not mapped to standard keyboard layouts, making input and text processing more complex. This further complicates the process of text input, search functionality, and digital representation of Croatian Angular Glagolitic within modern computing systems.



Ⰱ 107 [illegible]

Ⰱ²
Ⰱ² numeri nota
saepissime
2; II:
num. card.: sêdê l(ê)tь ·z· (=9) mês(e)ca ·b· d(ь)ni ·ai· (=11) BrLab 158c *menses duos;* si stvori ·b· ep(isto)lii iže katolicê naricajut' se BrLab 158d *duas epistolas;* vide ·b· lava tekuĉi k sebe vkup' CŽg 84r δύο λέοντες; pinikos' ima ·b· krili CAc 89v; i tako m'nimo pravdno narediti ti ·b· vrim(e)ni RegBen 31a *utraque tempora;*
in num. compos.: vь ev(a)nĵ(e)li po luce pisano estь êko kneža hĵi (pro hĉi) ·bi· (=12) lêt' imêše tečenie BrVO 349a *duodecim annos;* lêto imat' mêsecev' ·bi· (=12) CPar 2v; žitь že aredь lêt' ·r· i ·m· i ·b· (=162) BrVO 132d ἑκατὸν καὶ ἑξήκοντα δύο *centum sexaginta duobus* Gn 5,18;
spec. *Dvanaestorica; the Twelve;* — οἱ δώδεκα, duodecim; - de duodecim apostolis: êvi se nimu g(ospod)ь bog i sa ·bi· ap(usto)loma v sl(a)ve velice CTk 87v μετὰ τῶν δώδεκα ἀποστόλων; prizvavše že ·bi· vse množ'stvo uč(e)n(i)k' i rêše BrVb₁ 34a προσκαλεσάμενοι δὲ οἱ δώδεκα *convocantes autem duodecim* A 6,2; edanь ot ·bi· CPar 254r ἐκ τοῦ ἀριθμοῦ τῶν δώδεκα *unum de duodecim* L 22,3;
num. multipl.: al(elu)ê ·b· MVat₄ 111b *alleluia, alleluia;*
num. ord.: po b(la)ž(e)nomь p(e)tri obače

[illegible]
БААЛОВЬ adj. poss.
BrVat₅
Baalov; Baal's; — Βάαλ (gen.); *Baal* (gen.);
- ad Baal pertinens (v. ВААЛЬ): potrêb'lju ot mêsta sego ostanci **baalovi** BrVat₅ 236c τὰ ὀνόματα τῆς Βααλ *reliquias Baal* So 1,4 ‖ **vaalimska** VO 458a **valamin'ska** N₂ 258d. - Exh.
Cf. ВААЛИМЬСКЬ
S (ВААЛОВЪ)

ⰁⰀⰁⰀ
БАБА, -e f.
BrVO; COxf
primalja, babica; midwife; — μαῖα; obstetrix: r(e)če **babi** ča porodih' r(e)če **baba** žen'ski pol' se e porodil' COxf 28d τῇ μαίᾳ ... ἡ μαῖα; ide iskati **babe** evreênine COxf 29d μαῖαν Ἑβραίαν; i r(e)če **baba** m(a)rii otkrii se COxf 30a ἡ μαῖα;
БАБА ПОУПОРѢЗИЦА: prizvav že c(êsa)rь ejupt'ski **babê puporêzici** BrVO 183b τὰς μαίας Ex 1,18; uboêsta že se **babê puporêzici** b(og)a BrVO 183b αἱ μαῖαι *obstetrices* Ex 1,17; reče že c(êsa)rь ejuptьski **babama puporêzicama** evrêiskima BrVO 183a ταῖς μαίαις *obstetricibus* Ex 1,15; - cf. ПОУПОРѢЗИЦА;
lectio erronea per commutationem БАБА cum РАБА: i poide š' nju ijudit' **baba** nee va vrt' COxf 28c ἡ παιδίσκη.
SMA

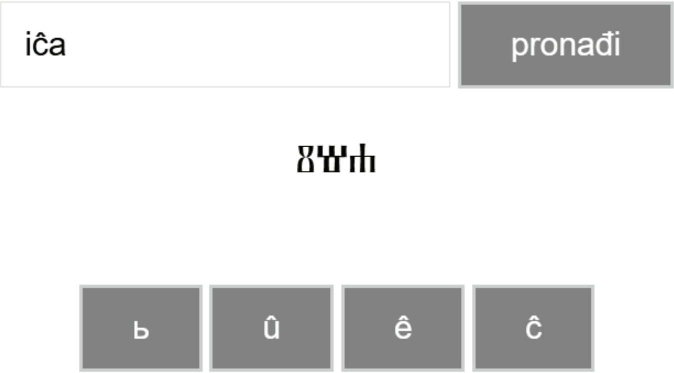
Fig. 4. Linking dictionary headwords to their corresponding PDF pages

5.1 Solutions for the challenges

One solution that effectively addresses most of the major challenges is the use of a custom font for the Glagolitic script, specifically designed to be mapped to a standard keyboard layout. This approach allows for more accessible text input and processing while ensuring accurate representation of Croatian Angular Glagolitic characters. The Old Church Slavonic Institute has developed its own Angular Glagolitic font, known as FSGLA[25], created by Frane Paro. The font was first developed in the early 2000s. However, an initial issue arose as certain letters were incorrectly mapped—for instance, the letter Ⱍ was not mapped to *č* on a Croatian keyboard but instead to */*. This misalignment was likely due to technical constraints

[25] https://stin.hr/wp-content/uploads/2025/01/fsgla.ttf [20/02/2025]

related to specific keyboard layouts of the time. To address this issue, FontForge,[26] an open-source font editing software, was utilized to correctly map characters such as *č*, *ć*, *š*, *ž*, *đ*, and other diacritic signs used in the Croatian language. However, even after enabling users to input symbols into the search bar, a challenge remained with specific letters such as Ⱐ (ь), Ⱓ (û), and Ⱑ (ê), as these characters are not included in standard Croatian keyboard layouts. To resolve this issue, a set of on-screen buttons was implemented below the search bar, allowing users to manually insert these symbols into the search query as needed (Fig. 5). In the search bar, the input text is displayed in Latin script. However, directly below it, a real-time transliteration into the Glagolitic script is also shown, allowing users to visually confirm how their input is being converted and ensuring accuracy in the search process.



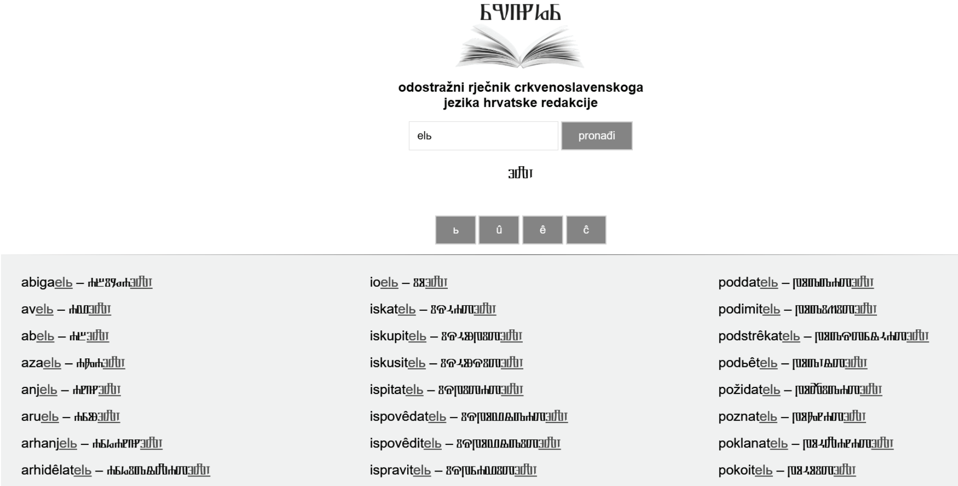
iĉa

ⰹⱋⰰ

Fig. 5. Search bar for the reverse dictionary of the *Dictionary of the Croatian Redaction of Church Slavonic*

Another challenge was that certain dictionary entries contained an apostrophe ('), so an exception was implemented to ignore this symbol during the search process. Additionally, the Latin letters *ć* and *ĉ* are written as Ⱋ in the Glagolitic script. However, to enhance clarity and consistency within the search process, a deliberate decision was made to convert *ć* into *ĉ* during the searches and add an on-screen button for *ĉ*. Since searchable words were displayed in both Latin and Glagolitic scripts, this meant presenting duplicate search results (Fig. 6). However, this did not pose any issues for the search algorithm, as the display of each search result was simply duplicated and the FSGLA font applied to one of them to render it in the Glagolitic script. Since the characters were correctly mapped within the font, this approach ensured accurate representation without causing any display or processing issues.

[26] https://fontforge.org/en-US/ [20/02/2025]



odostražni rječnik crkvenoslavenskoga jezika hrvatske redakcije

elь

ⰵⰾⱐ

abigaelь – ⰰⰱⰹⰳⰰⰵⰾⱐ
avelь – ⰰⰲⰵⰾⱐ
abelь – ⰰⰱⰵⰾⱐ
azaelь – ⰰⰸⰰⰵⰾⱐ
anjelь – ⰰⱀⰼⰵⰾⱐ
aruelь – ⰰⱃⱆⰵⰾⱐ
arhanjelь – ⰰⱃⱈⰰⱀⰼⰵⰾⱐ
arhidêlatelь – ⰰⱃⱈⰹⰴⱑⰾⰰⱅⰵⰾⱐ

ioelь – ⰹⱁⰵⰾⱐ
iskatelь – ⰹⱄⰽⰰⱅⰵⰾⱐ
iskupitelь – ⰹⱄⰽⱆⱂⰹⱅⰵⰾⱐ
iskusitelь – ⰹⱄⰽⱆⱄⰹⱅⰵⰾⱐ
ispitatelь – ⰹⱄⱂⰹⱅⰰⱅⰵⰾⱐ
ispovêdatelь – ⰹⱄⱂⱁⰲⱑⰴⰰⱅⰵⰾⱐ
ispovêditelь – ⰹⱄⱂⱁⰲⱑⰴⰹⱅⰵⰾⱐ
ispravitelь – ⰹⱄⱂⱃⰰⰲⰹⱅⰵⰾⱐ

poddatelь – ⱂⱁⰴⰴⰰⱅⰵⰾⱐ
podimitelь – ⱂⱁⰴⰹⰿⰹⱅⰵⰾⱐ
podstrêkatelь – ⱂⱁⰴⱄⱅⱃⱑⰽⰰⱅⰵⰾⱐ
podьêtelь – ⱂⱁⰴⱐⱑⱅⰵⰾⱐ
požidatelь – ⱂⱁⰶⰹⰴⰰⱅⰵⰾⱐ
poznatelь – ⱂⱁⰸⱀⰰⱅⰵⰾⱐ
poklanatelь – ⱂⱁⰽⰾⰰⱀⰰⱅⰵⰾⱐ
pokoitelь – ⱂⱁⰽⱁⰹⱅⰵⰾⱐ

Fig. 6. Search results for *elь* in the *Reverse Dictionary of the Croatian Redaction of Church Slavonic*

6 FUTURE PLANS

The reverse dictionary based on the *Dictionary of the Croatian Redaction of Church Slavonic* is currently in its demo version, and not all words have been included yet. The current demo version of the dictionary contains 18,078 entries, but the final version is expected to include more entries once the word list for the dictionary is fully complete. As more words are added, it is likely that additional input buttons for special characters (e.g. *ĵ* to be used as ⰼ) will need to be introduced to accommodate for the full range of symbols used in the Glagolitic and Latin scripts. The dictionary is currently publicly available on a GitHub repository[27] and will be available later on the *Dictionary of the Croatian Redaction of Church Slavonic* website[28] through the main navigation menu. The plan also includes linking the words from the reverse dictionary with the processed entries in the web version of the *Dictionary of the Croatian Redaction of Church Slavonic*. Another plan is to use the reverse dictionary to facilitate the acquisition of various morphological content. There are plans to create educational games, which will be available on the Institute website,[29] aimed at learning the lexicon of the Croatian Church Slavonic language. In this process, a reverse dictionary proves to be a valuable resource, as it simplifies the search for and selection of words that belong to the same derivational and declensional/conjugational type.

[27] https://borna12.github.io/odostrazni-gla/ [24/02/2025]

[28] https://stin.hr/crkvenoslavenski-rjecnik [24/02/2025]

[29] https://stin.hr/obrazovne-igre/ [24/02/2025]

7 CONCLUSION

Reverse dictionaries serve as important tools in the field of linguistics, providing unique insights into the structure and formation of words within a language. Their applications extend beyond academic research, offering practical benefits in areas such as poetry, language education, game development, and computational linguistics. The reverse dictionary based on the *Dictionary of the Croatian Redaction of Church Slavonic* may be the first web-based reverse dictionary specifically designed for an ancient script. Having a reverse dictionary written in historical scripts such as Glagolitic is essential for linguistic research, paleography, and lexicography. It enables scholars and researchers to analyze word formation patterns, morphological structures, and suffix-based derivation, which are particularly significant in languages with rich inflectional systems like Church Slavonic. Additionally, it facilitates the deciphering of historical texts, aids in comparative linguistic studies, and supports digital humanities projects by making historical lexicons more accessible and searchable in modern digital environments.

ACKNOWLEDGEMENTS

This work was created within the project Development of a Digital Infrastructure Model of the Old Church Slavonic Institute – DigiSTIN, funded by the European Union – NextGenerationEU. The views and opinions expressed are solely those of the author and do not necessarily reflect the official positions of the European Union or the European Commission. Neither the European Union nor the European Commission can be held responsible for them.

References

Badurina-Stipčević, V., Dürrigl, M.-A., Klenovar, M., Kovačević, A., Mihaljević, M., Miličić, I., Mulc, I., Šimić, M., Turkalj, L., and Vela, J. (2015). Rječnik crkvenoslavenskoga jezika hrvatske redakcije. Svezak II: vrêdьnь – zapovêdnica. Zagreb: Staroslavenski institut.

Bratulić, J., Damjanović, S., Frančić, A., Kuzmić, B., Lisac, J., Matasović, R., Mihaljević, M., and Žagar, M. (2009). Povijest hrvatskoga jezika 1. knjiga: Srednji vijek. Zagreb: Croatica.

Cronopista. (2009). Dict-FE. Accessible at: https://www.cronopista.com/dict-fe/ [accessed 26/02/2025].

FontForge. (2012). FontForge – An Outline Font Editor. Accessible at: https://fontforge.org/en-US/ [accessed 21/02/2025].

GitHub. (2025). Odostražni rječnik RCJHR. Accessible at: https://borna12.github.io/odostrazni-gla/ [accessed 24/02/2025].

Grčević, M. (2017). Hrvatski odostražni rječnik. Hrvatski studiji. Zagreb.

Hudeček, L., Mihaljević, M., and Jozić, Ž. (eds.). (2024) Anatomija rječnika – Hrvatski mrežni rječnik Mrežnik. Zagreb: Institute of Croatian Language.

IHJJ. (2016). Kolokacije – English. Accessible at: http://ihjj.hr/kolokacije/english/ [accessed 26/02/2025].

Japundžić, M. (1998). Hrvatska glagoljica. Zagreb: AGM. Accessible at: https://www.croatianhistory.net/etf/japun.html#*.

JENA. (2022). Odostrazni JENA. Accessible at: https://jena.jezik.hr/wp-content/odostrazni-jena/ [accessed 17/02/2025].

Lewis, K., and Mihaljević, J. (2018). Odostražni rječnik – što je, kako ga izraditi i čemu služi. Hrvatski jezik, 5(2), pp. 21–24.

Matešić, J. (1965–1967). Rückläufiges Wörterbuch des Serbokroatischen, Vols. 1–4. Wiesbaden: Otto Harrassowitz.

Mihaljević, A., and Mihaljević, J. (2024). Mrežna inačica Rječnika crkvenoslavenskoga jezika hrvatske redakcije. Slovo: časopis Staroslavenskoga instituta u Zagrebu, 74, pp. 169–194.

Mihaljević, J. (2022). Igrifikacija hrvatskoga mrežnog rječnika. Zagreb: Institut za hrvatski jezik i jezikoslovlje.

Mihaljević, M., and Hudeček, L. (eds.). (2024). Rječnik jezikoslovnoga nazivlja, Zagreb: Institut za hrvatski jezik.

Mihaljević, M., Hudeček, L., and Jozić, Ž. (eds.). (2023). Hrvatsko jezikoslovno nazivlje. Zagreb: Institut za hrvatski jezik i jezikoslovlje.

Mozilla Developer Network. (2024). Single-page application (SPA). Mozilla. Accessible at: https://developer.mozilla.org/en-US/docs/Glossary/SPA [accessed 19/02/2025].

Mrežnik. (2022). Odostrazni rječnik. Accessible at: https://rjecnik.hr/mreznik/wp-content/odostrazni/ [accessed 17/02/2025].

Muško i žensko u hrvatskome jeziku. (2022). Odostrazni rječnik. Accessible at: https://muskozensko.jezik.hr/odostrazni/ [accessed 17/02/2025].

MyStilus. (2019). MyStilus. Accessible at: https://www.mystilus.com/ [accessed 17/02/2025].

NLTK. (2008). Natural Language Toolkit (NLTK). Accessible at: https://www.nltk.org/ [accessed 26/02/2025].

OneLook. (2010). OneLook Dictionary Search. Accessible at: https://www.onelook.com/ [accessed 17/02/2025].

OneLook. (2016). OneLook Thesaurus. Accessible at: https://www.onelook.com/thesaurus/ [accessed 17/02/2025].

Reimlexikon. (2006). Reimlexikon. Accessible at: https://www.reimlexikon.net/ [accessed 26/02/2025].

Reverse Dictionary. (2006). Reverse Dictionary. Accessible at: https://reversedictionary.org/ [accessed 17/02/2025].

Rhymer. (2017). That Rhymes. Accessible at: https://www.rhymer.com/that.html [accessed 26/02/2025].

RhymeZone. (2006). RhymeZone Dictionary and Thesaurus. Accessible at: https://www.rhymezone.com/ [accessed 26/02/2025].

Stahl, F. A., Scavnicky, G. E. A. (1973). A reverse dictionary of the Spanish language. Urbana, IL: University of Illinois Press.

STIN. (2021). Novakov misal. Accessible at: https://stin.hr/novakov-misal/ [accessed 14/02/2025].

STIN. (2021). Žgombićev zbornik. Accessible at: https://stin.hr/zgombicev-zbornik/ [accessed 14/02/2025].

STIN. (2023). Repozitorij glagoljaškog pjevanja. Accessible at: https://stin.hr/repozitorij-glagoljaskoga-pjevanja/ [accessed 14/02/2025].

STIN. (2024). DigiSTIN. Accessible at: https://stin.hr/en/content/digistin-en/ [accessed 14/02/2025].

STIN. (2024). FSGlA.ttf Font File. Accessible at: https://stin.hr/wp-content/uploads/2025/01/fsgla.ttf [accessed 20/02/2025].

STIN. (2024). Obrazovne igre. Accessible at: https://stin.hr/obrazovne-igre/ [accessed 14/02/2025].

STIN. (2024). Stvori sliku s tekstom na glagoljici. Accessible at: https://stin.hr/stvori-sliku-s-tekstom-na-glagoljici/ [accessed 14/02/2025].

STIN. (2025). Crkvenoslavenski rječnik. Accessible at: https://stin.hr/crkvenoslavenski-rjecnik [accessed 24/02/2025].

Štrkalj Despot, K., and Möhrs, C. (2015). Pogled u e-leksikografiju. Rasprave 41(2), pp. 329–353.

UKEssays. (2018). Efficient Database Driven Reverse Mapping Dictionary. Accessible at: https://www.ukessays.com/essays/computer-science/efficient-database-driven-reverse-mapping-1315.php?vref=1 [accessed 05/03/2025].

Wikipedia. (2020). Glagolitic (Unicode block). Accessible at: https://en.wikipedia.org/wiki/Glagolitic_(Unicode_block) [accessed 20/02/2025].

WordFinder. (2023). WordFinder by YourDictionary. Accessible at: https://wordfinder.yourdictionary.com/ [accessed 26/02/2025].

sciendo
DOI 10.2478/jazcas-2025-0017

MasKIT – ANONYMIZATION AND PSEUDONYMIZATION OF CZECH LEGAL TEXTS

JIŘÍ MÍROVSKÝ[1] – TEREZA NOVOTNÁ[2] – BARBORA HLADKÁ[3]
[1]Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics, Charles University, Prague, Czech Republic (ORCID: 0000-0003-2741-1347)
[2]Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics, Charles University, Prague, Czech Republic (ORCID: 0000-0002-1426-4547)
[3]Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics, Charles University, Prague, Czech Republic (ORCID: 0000-0003-4950-4587)

MÍROVSKÝ, Jiří – NOVOTNÁ, Tereza – HLADKÁ, Barbora: MasKIT – Anonymization and Pseudonymization of Czech Legal Texts. Journal of Linguistics, 2025, Vol. 76, No 1, pp. 192 – 199.

Abstract: MasKIT is a command-line tool, an on-line web application and a REST API service for anonymization and pseudonymization of Czech legal texts. Taking a plain text as input (e.g. a letter sent by a legal authority to a citizen), it runs external services for dependency parsing and named entity recognition and then via a rule-based approach identifies and replaces sensitive information in the text.

Keywords: anonymization, pseudonymization, legal texts, tool, web interface, REST API, Czech

1 INTRODUCTION

1.1 Motivation

Courts in most countries are now obliged to publish their decisions online. Access to case law is a cornerstone of access to justice, which is one of the fundamental principles of the rule of law. Thus, court decisions serve not only as precedents for subsequent cases, but also as a source of a great deal of legal information, interpretations of legal norms or even information about the functioning of society and various political changes. Such analyses then require access to case law as a whole, i.e. full texts and underlying metadata.

Opposite to access to justice in this case is the protection of privacy, which is also a constitutionally guaranteed right. The invasion of privacy caused by the publication of personal data in decisions can be severe, especially in the case of criminal decisions, and can even lead to secondary victimisation and endangerment of victims of crime. The precise anonymization of published judicial decisions is a key element in protecting the privacy of litigants. At the same time, anonymization is also a tool to enable the publication of case law to the general public and thereby strengthen access to justice.

Anonymization is therefore an important step, a *conditio sine qua non*, in the process of publishing legal documents (e.g. court or administrative decisions), but at the same time it must be carried out very precisely and in accordance with the legislation governing the protection of the privacy and personal data of the subjects. Automated anonymization tools therefore have the great potential to save significant time and manual labour in the courts. However, such a tool must be sufficiently precise, comply with data protection legislation and ideally meet the transparency and explainability requirements of both national and European regulations.

In the Czech legal environment, the scope of anonymized data and the method of anonymization are governed by Decree No. 403/2022 Coll., on the publication of court decisions, which implements Act No. 6/2002 Coll., on courts and judges.

In this paper, we present MasKIT, an open-source tool for automatic anonymization and pseudonymization of Czech legal documents. The tool is being developed to comply with the legislative anonymization rules mentioned above. At the same time, it is available under open licences and widely usable not only for anonymization/pseudonymization of legal documents in courts and public administration, but also for the public. The methods on which MasKIT is based meet the strict conditions for transparency and explainability of processes imposed by European legislation (AI Act, GDPR Art. 22).

In the rest of the Introduction, we present recent related work on anonymization/pseudonymization. In Section 2, we describe the system architecture and give a step-by-step example of processing a Czech sentence both in the anonymization and pseudonymization modes. In Section 3, the user interface is shortly introduced. Section 4 is dedicated to evaluating the system and concluding the paper.

1.2 Related work

Csányi et al. (2021) discusses the complexities of anonymizing legal documents, highlighting the need to balance privacy protection with the preservation of information utility. It emphasizes that while Named Entity Recognition methods are crucial, they are insufficient on their own. The authors advocate for integrating machine learning techniques with anonymization models, such as differential privacy, to effectively reduce re-identification risks.

Oksanen et al. (2022) introduces the ANOPPI tool developed for (semi-) automatic anonymization of Finnish texts. The tool can be used both as a web application and programmatically through a REST API. Evaluation shows that ANOPPI performs well with different types of documents, however, further improving the performance of the named entity recognition and disambiguation methods would enhance the usefulness of the software. The tool is being published as open source for public use by the Ministry of Justice in Finland.

Glaser et al. (2021) used the BERT architecture to train an anonymization model that takes into account the context of the anonymized data. Such a method then, according to the authors, does not require non-anonymized data to train, but can be applied to anonymized publicly available legal documents.

Similarly, Licari et al. (2022) presented a model to anonymize Italian judicial decisions, based on transformers and spacy entity recognition.

Recently, the anonymization of legal texts has begun to combine traditional rule-based approaches with fine-tuning and domain-specific pre-training of large language models. For example, Niklaus at al. (2023) explores improvements to the anonymization system used by the Swiss Federal Supreme Court, known as Anom2. The study focuses on enhancing the identification and masking of personal information in legal texts by integrating machine learning techniques. The researchers compiled a large annotated dataset containing entities requiring anonymization, which served as a training and evaluation resource. By pre-training BERT-based models on domain-specific legal data, they achieved an F1-score improvement of over 5% compared to models trained without such in-domain knowledge.

2 THE SYSTEM DESCRIPTION

2.1 The system architecture

MasKIT is a command-line tool and also a client-server application with a web client interface and a REST API server. The tool (the server) is written in Perl and calls two state-of-the-art external services to pre-process the texts:

- UDPipe (Straka 2018) for syntactic analysis of the text, and
- NameTag (Straková et al. 2019) for recognition of named entities.

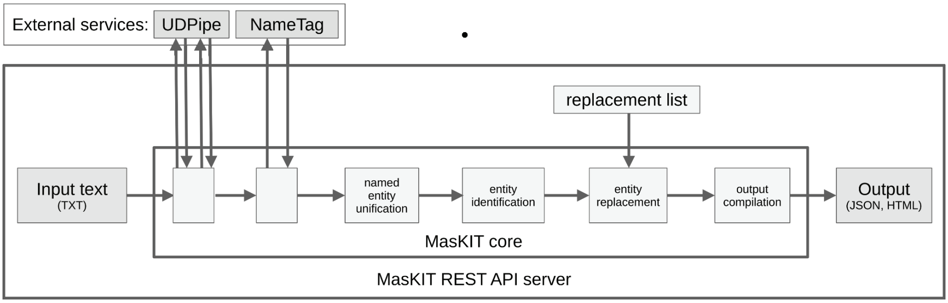


Fig. 1. MasKIT application architecture

The subsequent rule-based analysis uses information from both the external tools. The dataflow of the whole process is given in Fig. 1. The system accepts a plain text as the input and performs the following steps:

(1) **Morphological and syntactic analysis:** UDPipe is called twice with two different models: First, an older model that gives preference to longer sentences is

called for sentence segmentation, which is suitable for legal texts that often contain addresses and various abbreviations with dots in the middle of sentences (wrong segmentation at these places would later harm the performance of named entities recognition). Second, the pre-segmented text is passed to a new model to obtain state-of-the-art morphological tagging and syntactic parsing in the Universal Dependencies framework.[1]

(2) **Named entities recognition:** The parsed data are sent to NameTag for recognition of named entities. Because of time limitations of the NameTag service, long texts are split to shorter segments and processed separately.

(3) **Named entities unification:** As NameTag sometimes fails to recognize all occurrences of the same named entity in a single document, or sometimes mis-classifies some of the occurrences,[2] unification of classification of single-word named entities is performed. For each token[3] in the text that is not a part of a multi-word named entity, all named-entity marks are counted. Unless most of the occurrences of the token are unmarked (without an assigned named-entity mark), all the tokens are (re-)assigned the most frequent named-entity mark. If there are more than one most frequent mark, the class relevant for MasKIT is preferred.[4]

(4) **Rule-based analysis:** The text is analyzed sentence by sentence and token by token, traversing the dependency syntax trees. Expressions that should be anonymized are detected with a series of manually encoded rules, taking advantage of the assigned named-entity marks and the parsed dependency tree structure.

Some types of expressions are best detected using dedicated Perl libraries, which give more reliable results than the named-entity marks (e.g. e-mail addresses). Some hard-to-parse expressions are best detected directly from the surface form of the sentence (such as agenda reference numbers), but for many of the detected types of expressions, using the parsed dependency tree structure is beneficial: For example, detecting dates of birth relies on finding key lemmas (such as *narození* ‘birth’ or *narodit (se)* ‘be born’ in the correct dependency relation with a date, regardless of other words appearing in the surface word order.

(5) **Output generation:** After the whole text is processed, the output is produced in the selected output format (TXT, HTML, CoNLL-U).[5]

[1] https://universaldependencies.org/

[2] Further study would be required to show if and how much it is related to processing longer texts in segments.

[3] More precisely: a combination of its lemma and part of speech.

[4] E.g. a ‘country name’ mark is not relevant for MasKIT, as countries do not get anonymized.

[5] The CoNLL-U format is only available from the command line (not in the web interface or via REST API).

2.2 Anonymization vs. pseudonymization

MasKIT supports both anonymization and pseudonymization. In the anonymization mode, the sensitive expressions are replaced by their classes, while in the pseudonymization mode random words of the same class are used. Let us assume that the sentence from Example (1) is entered in the system.

(1) Paní Marie Nováková z Myslíkovy ulice č. 25 dostala dopis od firmy Škoda Auto, a.s..
[Mrs. Marie Nováková from Myslíkova street No. 25 received a letter from the company Škoda, a.s..]

In the anonymization mode, a woman's surname *Nováková* is replaced by class *M-ŽENA-PŘÍJMENÍ-1* 'M-WOMAN-SURNAME-1', where the numeric index (here *1*) distinguishes replacements for different surnames, see Example (2).

(2) Paní **M-ŽENA-JMÉNO-1** **M-ŽENA-PŘÍJMENÍ-1** z **M-ULICE-1** ulice č. **M-ČÍSLO-ULICE-1** dostala dopis od firmy **M-FIRMA-1**.
[Mrs. **M-WOMAN-NAME-1** **M-WOMAN-SURNAME-1** from **M-STREET-1** street No. **M-STREET-NUMBER-1** received a letter from the company **M-COMPANY-1**.]

In the pseudonymization mode, one of twenty pre-defined surname replacements is used (e.g. *Pospíšilová*) in the correct morphological case, see Example (3).

(3) Paní **Alena** **Pospíšilová** z **Květinové** ulice č. **43** dostala dopis od firmy **UniTechna**.
[Mrs. **Alena** **Pospíšilová** from **Květinová** street No. **43** received a letter from the company **UniTechna**.]

Further occurrences of surname *Nováková* in the same text get replaced by the class with the same index, or with the same surname replacement. Male and female surnames are tied, so if, e.g. also *Mr. Novák* appeared in the text, it would be replaced with *Mr. M-MUŽ-PŘÍJMENÍ-1* or *Mr. Pospíšil*, respectively, to match the corresponding female surname.

The system always tries to replace a multiple-word expression (such as *Škoda Auto, a.s.*) as a whole, i.e. the whole company name is replaced by a single *M-FIRMA-1* 'M-COMPANY-1' or *UniTechna*, resp.

For inspection of the result in comparison with the original text, the user can optionally display also the original expressions in subscript next to the anonymized/pseudonymized replacements, see Example (4) and also Fig. 2.

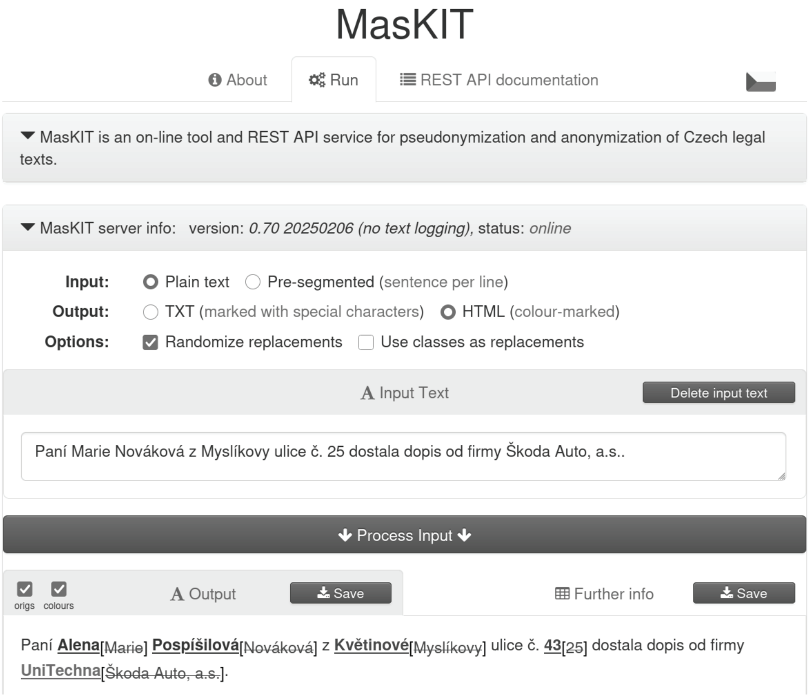
(4) Paní **Alena**[~~Marie~~] **Pospíšilová**[~~Nováková~~] z **Květinové**[~~Myslíkovy~~] ulice č. **43**[~~25~~] dostala dopis od firmy **UniTechna**[~~Škoda Auto, a.s.~~].
[Mrs. **Alena**[~~Marie~~] **Pospíšilová**[~~Nováková~~] from **Květinová**[~~Myslíkova~~] street No. **43**[~~25~~] received a letter from the company **UniTechna**[~~Škoda Auto, a.s.~~].]

3 USER INTERFACE

The MasKIT server can be either run as a command-line utility,[6] or it can be accessed via a web client or a REST API service.

3.1 Web interface

The MasKIT web client is written in PHP[7] and Bootstrap 3[8] and provides a browser-based interface to the server. The user enters a text (directly or as a file) and submits it. The text is passed to the server via REST API, processed by the server and the result is then presented to the user in an interactive way, see Fig. 2.



MasKIT is an on-line tool and REST API service for pseudonymization and anonymization of Czech legal texts.

MasKIT server info: version: *0.70 20250206 (no text logging)*, status: *online*

Input: ● Plain text ○ Pre-segmented (sentence per line)
Output: ○ TXT (marked with special characters) ● HTML (colour-marked)
Options: ☑ Randomize replacements ☐ Use classes as replacements

Paní Marie Nováková z Myslíkovy ulice č. 25 dostala dopis od firmy Škoda Auto, a.s..

☑ origs ☑ colours

Paní **Alena**[~~Marie~~] **Pospíšilová**[~~Nováková~~] z **Květinové**[~~Myslíkovy~~] ulice č. **43**[~~25~~] dostala dopis od firmy **UniTechna**[~~Škoda Auto, a.s.~~].

Fig. 2. MasKIT web interface

[6] See the program documentation: https://ufal.mff.cuni.cz/maskit/users-manual.
[7] https://www.php.net/
[8] https://getbootstrap.com/

3.2 Public REST API anonymous server

The Institute of Formal and Applied Linguistics at Charles University is running a publicly available MasKIT web client[9] and REST API service,[10] free for non-commercial usage. As the submitted texts may be of sensitive nature, the server does not store any part of the processed text. The same anonymity is guaranteed for requests from this server by the external services UDPipe and NameTag.[11]

4 EVALUATION AND CONCLUSION

To evaluate MasKIT, 7 annotators (students of the Faculty of Law at Masaryk University, Brno) have manually annotated 53 legal documents (5,373 sentences, 127 thousand tokens) of the following nature: court decisions (9), citizen's letters to an authority (2), authority's decisions (21), legal advices (12), lawsuits/appeals (3), ombudsman's reports (6). The annotators marked 2,372 sensitive expressions to be anonymized, classifying each occurrence in one of 20 classes recognized by MasKIT plus category 'other' for any other type (not supported by MasKIT). In the same documents, MasKIT anonymized and classified the sensitive information with Recall of 0.8, Precision of 0.64 and F1 measure of 0.7 on recognition of expressions to be anonymized. The classification accuracy on expressions marked both by the annotators and MasKIT was 0.91. The Precision and Recall are comparable to results in Glaser et al. (2021): they evaluated their system on 46 documents from Munich district and financial courts and reported Precision in the range from 0.63 to 0.69 and Recall in the range from 0.52 to 0.79. Their classification accuracy was 0.73.

MasKIT is still under development and does not yet support all classes of sensitive information as defined in Decree No. 403/2022 Coll., on the publication of court decisions. As all these classes have been included in our evaluation, improvements in comparison with the currently reported results are to be expected in future versions. In the present version, MasKIT complies with approx. 20 out of 35 categories designed to be anonymized by the Degree, not yet implementing, most of all, academic titles, account numbers, payment variable symbols, data box numbers and personal ID numbers.

ACKNOWLEDGEMENTS

The work was supported by the project "PONK – Computer-Assisted Clear Legal Writing" (reg. no. TQ01000526, co-funded by the Technology Agency of the Czech Republic) and the project "Human-centred AI for a Sustainable and Adaptive

[9] https://quest.ms.mff.cuni.cz/maskit/

[10] See MasKIT REST API documentation: https://ufal.mff.cuni.cz/maskit/api-reference.

[11] The only information that may be logged: time of usage, size of the processed data, the system configuration and the IP address from where the MasKIT service is accessed.

Society" (reg. no. CZ.02.01.01/00/23_025/0008691, co-funded by the European Union). The research reported in the present contribution has been using language resources developed, stored and distributed by the LINDAT/CLARIAH-CZ project of the Ministry of Education, Youth and Sports of the Czech Republic (LM2023062).

References

Csányi, G. M., Nagy, D., Vági, R., Vadász, J. P., and Orosz, T. (2021). Challenges and Open Problems of Legal Document Anonymization. Symmetry, 13(8), 1490. Accessible at: https://doi.org/10.3390/sym13081490.

Glaser, I., Schamberger, T., and Matthes, F. (2021). Anonymization of German legal court rulings. New York, NY, USA: Association for Computing Machinery. ICAIL 21. Accessible at: https://doi.org/10.1145/3462757.3466087.

Niklaus, J., Mamié, R., Stürmer, M., Brunner, D., and Gygli, M. (2023). Automatic Anonymization of Swiss Federal Supreme Court Rulings. In Proceedings of the Natural Legal Language Processing Workshop 2023, pp. 159–165, Singapore. Association for Computational Linguistics. Accessible at: https://doi.org/10.18653/v1/2023.nllp-1.16.

Licari, D., Romano, M., and Comande, G. (2022). Automatic Anonymization of Italian Legal Textual Documents using Deep Learning. ITA 2022. Accessible at: https://www.iris.sssup.it/handle/11382/548773.

Oksanen, A., Hyvönen, E., Tamper, M., Tuominen, J., Ylimaa, H., Löytynoja, K., Kokkonen, M., and Hietanen, A. (2022). An Anonymization Tool for Open Data Publication of Legal Documents. In Joint Proceedings of ISWC2022 Workshops: the International Workshop on Artificial Intelligence Technologies for Legal Documents (AI4LEGAL) and the International Workshop on Knowledge Graph Summarization (KGSum), pp. 12–21.

Straka, M. (2018). UDPipe 2.0 Prototype at CoNLL 2018 UD Shared Task. In Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies. Association for Computational Linguistics, Brussels, Belgium, pp. 197–207. Accessible at: https://doi.org/10.18653/v1/K18-2020.

Straková, J., Straka, M., and Hajič, J. (2019). Neural Architectures for Nested NER through Linearization. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, Florence, Italy, pp. 5326–5331. Accessible at: https://doi.org/10.18653/v1/P19-1527.

sciendo
DOI 10.2478/jazcas-2025-0018

SYNTACTIC STRUCTURES IN CZECH MULTIWORD UNITS: TYPES AND TOKENS IN THE TOTALITARIAN ERA AND TODAY

VLADIMÍR PETKEVIČ
Department of Linguistics, Faculty of Arts, Charles University, Prague,
Czech Republic (ORCID: 0000-0003-0468-4158)

PETKEVIČ, Vladimír: Syntactic Structures in Czech Multiword Units: Types and Tokens in the Totalitarian Era and Today. Journal of Linguistics, 2025, Vol. 76, No 1, pp. 200 – 211.

Abstract: We compare six selected types of syntactic structures in the Czech multiword units (MWUs) of the language of communist totalitarian journalism, represented by the Frap Totalita corpus, to the same types of structures in the MWUs of the post-totalitarian journalistic language, represented by the Frap SYNv13-PUB corpus. Attention is also paid to their lexical content. For some structures there are relatively large frequency differences in usage, while in others they are (considerably) minor. In terms of the MWU lexical setting, the languages represented by the corpora differ substantially.

Keywords: multiword units, Frap Totalita corpus, Frap SYNv13-PUB corpus, syntactic structures, lexical setting, frequency, comparison

1 INTRODUCTION

We compare six selected types of surface syntactic structures of multiword units (MWUs) in Czech and their lexical setting in two types of Czech journalistic texts. The first type is represented by the Frap Totalita corpus, containing primarily journalistic and political texts from the postwar period (1945–1989), the second type by journalistic texts created after 1989 until today and collected in the Frap SYNv13-PUB corpus. The paper is a contribution to a larger study on the Czech language of three genres:

(a) texts from the period of totalitarianism (1945–1989)
(b) contemporary standard (primarily newspaper) texts
(c) contemporary non-standard and anti-system texts

within the project *Czech phraseology and changes in its use in temporal and genre contexts*.

In the six types of syntactic structures we try to find out whether there are any significant differences in the syntactic structure of MWUs in texts of the (a) and (b) types, or whether the differences are only in the lexical content of MWUs; we do not deal with semantics at all.

2 CORPORA AND SOURCES OF MULTIWORD UNITS

In order to compare the syntax of MWUs[1] and their lexical content in communist totalitarian and post-totalitarian journalistic language, we use the working version of the Frap Totalita corpus and of the Frap SYNv13-PUB one. The Frap Totalita corpus, consisting of 15,947,180 corpus positions, contains journalistic texts from the 1948–1989 period: Rudé právo 'Red Law' newspaper texts published in selected quarters of the years 1952, 1969 and 1977, as well as professional (propaganda) texts, memoirs, speeches, letters and a very small portion of fiction. The corpus contains all texts from the Totalita corpus[2] and some prose texts (ca. 207,000 positions).

The Frap SYNv13-PUB corpus of contemporary journalism, consisting of 74,932,112 corpus positions, contains journalistic texts (mainly from newspapers and magazines) from the 1989–2023 period, every year between 1991–2023 being represented by ca. 2.5 million positions, and the years 1989 and 1990 by ca. 160,000 and 1.5 million positions, respectively.

Thus, the Frap SYNv13-PUB corpus is ca. five times larger. In the following, we will therefore compare the frequency of MWUs and their syntactic structures in terms of the instances-per-million (ipm) measure (mostly rounded).

We use the following sources of MWUs:

- FRANTALEX lexicon of Czech MWUs consisting of ca. 56,000 lemmas[3]
- LEMUR database[4] (LExicon of MUltiword expRessions of Czech) containing ca. 26,000 MWU lemmas (based on FRANTALEX)
- Frap Totalita corpus and Frap SYNv13-PUB corpus, where MWUs from FRANTALEX and LEMUR are tagged as MWU lemmas.

3 SELECTED SYNTACTIC STRUCTURES AND THEIR COMPARISON

The following MWU types of syntactic structures will be studied:

1. Nominal group (Adj, Noun), where the attributive adjective Adj agrees with the immediately following governing Noun in number, gender and case
2. Deverbal noun modified by another deverbal noun[5] in the genitive case
3. Coordination of deverbal nouns

[1] In this paper, the term MWU is conceived in a very broad sense as any collocation representing a single syntactic-semantic unit, not necessarily a phraseme; by phraseme we mean a MWU whose meaning is not determined by the composition of the meanings of its constituents.

[2] https://wiki.korpus.cz/doku.php/cnk:totalita

[3] FRANTALEX is a lexicon containing Czech MWUs based on Čermák et al. (1983–2009), their variants and additional MWUs found in corpora (Skoumalová et al. 2024, p. 2).

[4] Skoumalová et al. 2024.

[5] Deverbal nouns are neuter nouns derived from passive verbal participles and ending in nominative singular with *-ní*/*-tí*.

4. Accumulation of attributive nouns in the genitive case
5. Nominal groups with a nominating nominative and/or a foreign word or an (in)declinable noun
6. Periphrastic comparison of adjectives

Unless otherwise stated for a given structure, we state the following general hypothesis:

Generally, MWUs in totalitarian and post-totalitarian journalistic language do not differ too much in their syntactic structures; however, they differ significantly in their lexical setting.

3.1 Nominal group (Adj, Noun)

A nominal group (Adj, Noun) consists of a governing Noun and an immediately preceding dependent attributive adjective Adj agreeing with Noun in gender, number and case. In Tab. 1a, frequency of this type in both corpora is presented: as a number of occurrences (tokens) and ipm for all 7 cases in Czech:

	Frap Totalita		**Frap SYNv13-PUB**	
Case	**Occurrences / %**	**ipm**	**Occurrences / %**	**ipm**
Nom	306,608 / 20.0%	19,226	1,418,272 / 26.0%	18,927
Gen	593,608 / **38.6%**	**37,205**	1,562,366 / **28.7%**	**20,850**
Dat	65,460 / **4.3%**	**4,105**	186,352 / **3.4%**	**2,487**
Acc	269,954 / 17.5%	16,928	1,168,171 / 21.4%	15,590
Voc	850 / 0.1%	53	1,841 / 0.03%	25
Loc	171,478 / 11.2%	10,753	688,763 / 12.6%	9,192
Ins	127,635 / **8.3%**	**8,004**	427,461 / **7.8%**	**5,705**
Total	1,535,593 / 100.0%	**96,274**	5,453,226 / 100.0%	**72,776**

Tab. 1a. Case frequency of the (Adj, Noun) nominal groups in both corpora (the biggest differences are highlighted in bold)

In Tab. 1b, the most frequent MWUs of this structural type and typical of both types of language represented by their respective corpora are shown. Non-specific, usual MWUs such as příští rok ‘next year’ or the toponym České Budějovice contained in both corpora are omitted.

Frap Totalita		**Frap SYNv13-PUB**	
MWU lemmas: 305,850	**ipm**	**MWU lemmas: 1,127,751**	**ipm**
1. Sovětský svaz ‘Soviet Union’[6]	817	Česká republika ‘Czech Republic’	250
2. komunistická strana ‘communist party’	579	Evropská unie ‘European Union’	101

[6] The glosses are mostly literal translations.

3. dělnická třída ‘working class’	387	životní prostředí ‘environment’	100
4. národní hospodářství ‘national economy’	303	výběrové řízení ‘tender’	55
5. ústřední výbor ‘central committee’	285	tiskový mluvčí ‘spokesperson’	43
6. socialistická země ‘socialist country’	259	městský úřad ‘municipal authority’	39
7. socialistická společnost ‘socialist society’	238	sociální demokrat ‘social democrat’	39
8. národní výbor ‘national committee’	192	lidské právo ‘human right’	35
9. socialistická revoluce ‘socialist revolution’	187	cenný papír ‘security’	33
10. pracující lid ‘working people’	160	mobilní telefon ‘mobile phone’	32

Tab. 1b. The most frequent MWU lemmas corresponding to the (Adj, Noun) structure in both corpora

Given the politicized discourse of totalitarian journalism, it can be stated that:

- The Frap Totalita corpus (ipm=96,274) contains considerably more occurrences of (Adj, Noun) nominal groups than the Frap SYNv13-PUB corpus (ipm=72,776): in totalitarian journalism and post-totalitarian journalism, it is roughly every 11th pair and every 13th pair, respectively.
- Frap Totalita contains considerably more occurrences of genitive structures, which is due to the frequent modification of nouns denoting mainly political events, offices, parties, states: *představitel* ***komunistické strany*** ‘representative of the communist party’, *státy* ***Varšavské smlouvy*** ‘states of the Warsaw pact’
- There are more occurrences of (primarily prepositional) dative and instrumental structures in Frap Totalita: *k pátému* ***pětiletému***$_{\text{dat}}$ ***plánu***$_{\text{dat}}$ ‘to the fifth five-year plan’, *láska k* ***Sovětskému***$_{\text{dat}}$ ***svazu***$_{\text{dat}}$ ‘love for the Soviet Union’, *se* ***Sovětským***$_{\text{ins}}$ ***svazem***$_{\text{ins}}$ ‘with the Soviet Union’, *vztahy mezi* ***socialistickými***$_{\text{ins}}$ ***zeměmi***$_{\text{ins}}$ ‘relations between socialist countries’.
- The differences in MWUs’ lexical content are telling; moreover, the most frequent MWUs in Frap Totalita are considerably more frequent than their counterparts in Frap SYNv13-PUB – a symptom of the thematic poverty of totalitarian journalism.

3.2 Deverbal noun modified by another deverbal noun in the genitive case

We examine nominal groups described by the following structure pattern:

Noun-verb$_1$ Adj$_{\text{gen}}$\{0–3\} Noun-verb$_{\text{2gen}}$

where a deverbal noun Noun-verb$_1$ is modified by a genitive nominal group Adj$_{gen}${0–3} Noun-verb$_{2gen}$, governed by the genitive deverbal noun Noun-verb$_{2gen}$, which is modified by 0–3 agreeing adjectives.

In Tab. 2a, frequency of this structure type is presented:

Frap Totalita			**Frap SYNv13-PUB**		
MWU lemmas	**MWU occurrences**	**ipm**	**MWU lemmas**	**MWU occurrences**	**ipm**
3,027	5,475	**343**	7,018	11,096	**148**

Tab. 2a. Frequency of the Noun-verb$_1$ Adj$_{gen}${0–3} Noun-verb$_{2gen}$ pattern in both corpora in terms of MWU lemmas, their occurrences and ipm

The most frequent MWUs of this type are shown in Tab. 2b.

Frap Totalita		**Frap SYNv13-PUB**	
MWU lemmas: 3,027	**ipm**	**MWU lemmas: 7,018**	**ipm**
1. plnění usnesení ‘implementation of a resolution’	14.4	zahájení trestního stíhání ‘initiation of a criminal prosecution’	1.6
2. zasedání Valného shromáždění ‘meeting of the General Assembly’	13.2	vydání stavebního povolení ‘issuance of a building permit’	1.2
3. snížení zbrojení ‘reductions of arms’	3.8	navýšení základního jmění ‘increase of share capital’	1.0
4. zastavení horečného zbrojení ‘stopping of the feverish arms race’	3.5	podání trestního oznámení ‘filing of a criminal report’	0.7
5. rozvíjení socialistického soutěžení ‘development of socialist competition’	3.0	snížení základního jmění ‘reduction of share capital’	0.7
6. omezení zbrojení ‘limitation of arms’	2.9	zvýšení základního jmění ‘increase of share capital’	0.7
7. splnění usnesení ‘fulfillment of resolutions’	2.5	zastavení trestního stíhání ‘discontinuation of prosecution’	0.6
8. zdokonalování řízení ‘improvement of management’	2.4	sdělení obvinění ‘statement of charges’	0.6
9. splnění stranických usnesení ‘fulfillment of party resolutions’	2.3	zahájení řízení ‘initiation of proceedings’	0.6
10. zlepšení zásobování ‘improvement of supply’	2.1	vydání územního rozhodnutí ‘issuance of a zoning decision’	0.5

Tab. 2b. The most frequent MWUs of the Noun-verb$_1$ Adj$_{gen}${0–3} Noun-verb$_{2gen}$ pattern in both corpora

We see that genitive structures are predominant in the language of totality (ipm 343:148). This is because the texts from the totalitarian period are lexically and thematically much poorer than contemporary ones; they contain (probably intentionally) political clichés to a large extent, cf. the remarkable ipm differences between individual MWUs in Tab. 2b.

The difference in the MWU lexical content is obvious.

It is interesting to note that structures in which Noun-verb$_{2gen}$ is modified by a single adjective are frequent.

3.3 Coordination of deverbal nouns

In the coordination of two deverbal nouns, both nouns are in the same case.

In Tab. 3a we present frequency data on this type of structure:

Frap Totalita			Frap SYNv13-PUB		
MWU lemmas	**MWU occurrences**	**ipm**	**MWU lemmas**	**MWU occurrences**	**ipm**
3,698	6,358	**399**	6,450	8,023	**107**

Tab. 3a. Frequency of the coordination structure of two deverbal nouns in both corpora in terms of MWU lemmas, occurrences and ipm

In Tab. 3b, the ten most frequent MWUs are shown with their respective ipm.

Frap Totalita		Frap SYNv13-PUB	
MWU lemmas: 3,698	**ipm**	**MWU lemmas: 6,450**	**ipm**
1. splnění a překročení ‘meeting and exceeding’	7.53	rozhodnutí a vykázání ‘decision and eviction’[7]	0.81
2. plánování a řízení ‘planning and management’	5.89	padělání a pozměňování ‘forgery and alteration’	0.79
3. plnění a překračování ‘meeting and exceeding’	5.58	padělání a pozměnění ‘forgery and alteration’	0.77
4. řízení a plánování ‘management and planning’	5.52	ubytování a stravování ‘accommodation and catering’	0.75
5. zachování a upevnění ‘preservation and consolidation’	5.02	zajištění a udržení ‘securing and maintaining’	0.37
6. ubytování a stravování ‘accommodation and catering’	4.89	stravování a ubytování ‘catering and accommodation’	0.36
7. vedení a řízení ‘leadership and management’	4.45	chování a jednání ‘behaviour and action’	0.33

[7] Typically in the MWU *maření výkonu úředního **rozhodnutí** a **vykázání*** ‘obstruction of the execution of an official **decision** and **eviction**’.

8. prohlubování a zdokonalování ‘deepening and improving’	2.63	poškození a ohrožení ‘damage and threat’	0.32
9. myšlení a jednání ‘thinking and acting’	2.51	čtení a psaní ‘reading and writing’	0.31
10. rozšíření a prohloubení ‘widening and deepening’	2.20	myšlení a jednání ‘thinking and acting’	0.20

Tab. 3b. The most frequent MWUs constituted by the coordination of two deverbal nouns in terms of MWU lemmas and ipm in both corpora

We see that more structures constituted by the coordination of deverbal nouns are contained in Frap Totalita (399:107 ipm, cf. Tab. 3a). The reasons: Totalitarian language is lexically poor: a small number of lexical types (e.g. *splnění a překročení*) occur very frequently, compare the marked differences between ipm on the same lines in Tab. 3b. Moreover, it turns out that some structures are pleonastic: a single meaning is redundantly expressed by coordination – the meaning of one conjunct includes the meaning of the other (e.g. *splnění a překročení* ‘meeting and exceeding’ of a five-year plan, where *překročení* ‘exceeding’ implies *splnění* ‘meeting’).

The difference in the MWUs’ lexical content is again diametric, the most common MWUs of this type in the Frap SYNv13-PUB corpus belonging to the language of law.

3.4 Accumulation of attributive nouns in the genitive case

This type of structure is formally described as follows:

$\text{Noun}_0\ \text{Adj}_{1\text{gen}}\{0\text{–}3\}\ \text{Noun}_{1\text{gen}}\ \text{Adj}_{2\text{gen}}\{0\text{–}3\}\ \text{Noun}_{2\text{gen}}\ \text{Adj}_{3\text{gen}}\{0\text{–}3\}\ \text{Noun}_{3\text{gen}}\ \text{Adj}_{4\text{gen}}\{0\text{–}3\}\ \text{Noun}_{4\text{gen}}\ldots$

where the substructure $\text{Adj}_{\text{xgen}}\{0\text{–}3\}\ \text{Noun}_{\text{xgen}}$ (x = 1–n) represents a nominal group where the Adj agrees with the following Noun in the genitive case, gender and number.

Consider a nominal group with at least four dependent nouns in the genitive case. In the Frap Totalita corpus, the following are characteristic examples of the accumulation of genitive MWUs:

(1) *Ve vystoupení generálního* ***tajemníka***$_1$ *ústředního* ***výboru***$_2$ *Komunistické* ***strany***$_3$ ***Československa***$_4$*...*
‘In a speech **of** the General **Secretary**$_1$ **of** the Central **Committee**$_2$ **of** the Communist **Party**$_3$ **of Czechoslovakia**$_4$…’

(2) *… z usnesení vyplývají … nové methody* ***sestavování***$_1$ ***státního plánu***$_2$ ***rozvoje***$_3$ *národního* ***hospodářství***$_4$.
‘… arise from the resolutions … new methods **of** the **drawing up**$_1$ **of** the state **plan**$_2$ **of** the **development**$_3$ **of** the national **economy**$_4$.’

In the Frap SYNv13-PUB corpus, the MWU lexical content is, as expected, quite different:

(3) *Zcela nepochybně naplnil skutkovou podstatu* ***přečinu***$_1$ ***maření***$_2$ ***výkonu***$_3$ *úředního* ***rozhodnutí***$_4$.
‘He has undoubtedly fulfilled the offence **of**$_1$ the **obstruction**$_2$ **of** the **execution**$_3$ **of** an official **decision**$_4$.’

(4) *... vypisuje výběrové řízení na obsazení* ***místa***$_1$ ***tajemníka***$_2$ ***sekretariátu***$_3$ *Českého* ***svazu***$_4$ ***vodovodů***$_5$ *a* ***kanalizací***$_5$.
‘…announces a selection procedure for the **position**$_1$ **of Secretary**$_2$ **of** the **Secretariat**$_3$ **of** the Czech **Association**$_4$ **of Water Supply**$_5$ and **Sewerage**$_5$.’

In the examples, the bold nouns are, syntactically, nominal attributes of their respective governing nouns.

Frequency data in Tab. 4 (Frap Totalita ipm=**247** : Frap SYNv13-PUB ipm=**175**) show that in the Frap Totalita corpus, this type of structure is more common.

Frap Totalita			**Frap SYNv13-PUB**		
MWU lemmas	**MWU occurrences**	**ipm**	**MWU lemmas**	**MWU occurrences**	**ipm**
3,470	3,932	**247**	12,418	13,091	**175**

Tab. 4. Frequency – lemmas, occurrences, ipm in both corpora – of nominal groups containing, in addition to the governing noun, at least 4 dependent nouns in the genitive case

It can be stated that in the Frap Totalita corpus, the nominal structure with chained genitive noun attributes is more frequent than in the Frap SYNv13-PUB corpus. However, this structure seems to be a distinctive feature of the official language of both types of language.

3.5 Nominal groups with a nominating nominative and/or a foreign word or an (in)declinable noun

There are three types of nominal groups (A–C) with a nominating nominative, a foreign word or an (in)declinable word:

A. In the nominal group, only the first word is inflected, the second one is in the nominating nominative. The Frap Totalita corpus contains the following typical examples: *v okrese* ***Praha*** 'in the district of Prague', *na dole* ***Zápotocký*** 'at the mine Zápotocký', *v nakladatelství* ***Svoboda*** 'in the Svoboda publishing house'.

Typical examples in the Frap SYNv13-PUB corpus are: *v okrese* ***Praha*** 'in the district of Prague', *v kraji* ***Vysočina*** 'in the Highlands region', *v paláci* ***Akropolis*** 'in Akropolis Palace'.

The structure is common in both corpora.

B. In the nominal group, only the last word (often of foreign origin) is inflected, the first one being a foreign or undeclinable word. In Frap Totalita, the first (bold) word is typically foreign: *v* ***New*** *Yorku* 'in New York', *na* ***Wall*** *Streetu* 'on Wall Street', *po* ***Mao*** *Ce-tungovi* 'after Mao Zedong'. These collocations are almost exclusively proper names.

In Frap SYNv13-PUB, however, a new type of structure appears, where the first word may be a foreign or undeclinable word, but also a declinable word in the nominative case, the second one being an inflected noun (possibly of foreign origin): *na* ***home*** *officu* 'at the home office', *v* ***Sazka*** *aréně* 'in Sazka Arena', *v* ***Tip Sport*** extralize 'in the Tip Sport extraleague'. This type is productive, Czech is clearly influenced by English: the nominal attribute comes first.

C. Not a single word in the nominal structure is inflected. Compared to Frap Totalita (*chargé d'affaires*, *Morning Star*), there are many more such MWUs in the Frap SYNv13-PUB corpus that are adopted from English without change; this type is very productive: *Australian open*, *home credit*, *power play*.

In general, the lexical content of MWUs in the two corpora differs considerably.

3.6 Periphrastic comparison of adjectives

In general, Czech adjectives can be compared in the comparative degree by the synthetic comparative form (*hladší* 'smoother') as well as periphrastically by combining the adverb *více* 'more' or *méně*[8] 'less' with the positive degree of the compared adjective (*více protisrbský* 'more anti-Serbian', *méně hustý* 'less dense'). For some adjectives, only one of the possibilities is grammatically correct. However, in usage, contrary to the standard language codification, we encounter

(i) the form of *více* + comparative form

(ii) the form of *méně* + comparative form.

We distinguish, of course, between the adverbs *více* and *méně* in the comparison function and in the other functions/compounds:

- in the qualitative function, cf.: *bydlení je* ***více***$_{\text{comp}}$ ***dražší***$_{\text{comp}}$ 'housing is **more**$_{\text{comp}}$ **more-expensive**$_{\text{comp}}$' vs. *prodává* ***více***$_{\text{quant}}$ *dražších* ***ledniček*** 'he sells **more**$_{\text{quant}}$ expensive **refrigerators**', i.e. more refrigerators

[8] The constructions with *méně* 'less' are also referred to as comparison.

- in the particle construction *více* [*či*] *méně* ‘more or less’: *to se mu **více méně** podařilo* ‘he **more or less** succeeded’
- in structures where *více* or *méně* modifies a verb rather than an adjective: ***pracoval** stále **více** lepší metodou* ‘he **worked more** using a better method’.

First, we present typical corpus examples of periphrastic comparison:

1. ***více* + positive degree**:

Frap Totalita:

(5) *Budeme ještě **více bdělí***$_{\text{pos}}$.
‘We’ll be even **more vigilant**$_{\text{pos}}$.’

Frap SYNv13-PUB:

(6) *Zdálo se, že jsme byli **více nervózní***$_{\text{pos}}$ *než oni*.
‘We seemed to be **more nervous**$_{\text{pos}}$ than they were.’

2. ***více* + comparative degree**:

Frap Totalita:

(7) *Spojují je pouty mnohem **více papežštějšími***$_{\text{comp}}$ *než sám papež*.
‘They bind them with ties far **more papal**$_{\text{comp}}$ than the Pope himself.’

Frap SYNv13-PUB:

(8) *Lidi na heroinu jsou však **více uzavřenější***$_{\text{comp}}$.
‘People on heroin, however, are **more withdrawn**$_{\text{comp}}$.’

3. ***méně* + comparative degree**:

Frap Totalita:

(9) *Dal se cestou **méně schůdnější***$_{\text{comp}}$ *a méně výnosnou*$_{\text{pos}}$.[9]
‘He took the **less feasible**$_{\text{comp}}$ and less profitable$_{\text{pos}}$ route.’

[9] It is interesting that in (9) the adjective *schůdnější*$_{\text{comp}}$ following the first occurrence of *méně* expresses the comparative degree, whereas the adjective *výnosnou*$_{\text{pos}}$ following the second occurrence of *méně* is in the positive degree.

Frap SYNv13-PUB:

(10) … z *filmařů, kteří jsou možná v tuzemsku* ***méně známější***$_{\text{comp}}$.
‘...of filmmakers who are perhaps **less well known**$_{\text{comp}}$ at home.’

In Tab. 5 we summarize the frequency of adjectival comparison.

	Frap Totalita		**Frap SYNv13-PUB**	
Comparison type	**occurrences**	**ipm**	**occurrences**	**ipm**
synthetic comparative	36,125	2,265	176,810	2,360
více + pos degree	88	5.50	794	10.60
více + comp degree	20	1.25	92	1.23
méně + comp degree	11	0.70	32	0.43

Tab. 5. Frequency of kinds of adjectival comparison in terms of occurrences and ipm in both corpora

Summary:

- *více*+positive: In Frap SYNv13-PUB (ipm=10.6) there are twice as many instances of adjectival comparison as in Frap Totalita (ipm=5.5)
- *více*+comparative: post-totalitarian language does not differ from its totalitarian counterpart (ipm: 1.25 : 1.23)
- *méně*+comparative: Due to insufficient data, it can be said that the frequency of this phenomenon in totalitarian journalism (Frap Totalita ipm=0.7) and in post-totalitarian journalism (Frap SYNv13-PUB ipm=0.43) is similar.

We see that in the Frap SYNv13-PUB corpus, adjectives are compared periphrastically more often than in the Frap Totalita corpus by *více/méně* + positive/ comparative degree. In any case, the synthetic comparative clearly prevails. The reasons for the periphrastical comparison can, generally, be as follows:
(i) the influence of analytical comparison in English
(ii) speakers’ difficulties with forming synthetic comparative forms
(iii) endeavour to emphasize the comparative degree.

4 CONCLUSION

In general, the following conclusions can be drawn:

- The hypothesis stated at the beginning of § 3 has been confirmed. In terms of syntactic structures, totalitarian and post-totalitarian language differ, but not by much: the syntax of Czech has changed very little in 50–70 years.

- In terms of the lexical content of MWUs, the difference between the two types of language studied is considerable, which is, not surprisingly, due to the fact that after Czech society transitioned to freedom and democracy, the language of journalism changed: the post-totalitarian journalistic language has become lexically richer and more varied than its totalitarian counterpart, which was highly politicized, lexically very poor and characterized by a lot of clichés.

ACKNOWLEDGEMENTS

The research was supported by the Grant Agency of the Czech Republic, reg. no. GA ČR 24-10254S.

References

Čermák, F. et al. (1983–2009). Slovník české frazeologie a idiomatiky (SČFI), Vols. 1–4. Praha: Academia/Leda.

Skoumalová, H., Kopřivová, M., Petkevič, V., Jelínek, T., Rosen, A., Vondřička, P., and Hnátková, M. (2024). LEMUR: A lexicon of Czech multiword expressions. In: V. Giouli – V. Barbu Mititelu (eds.): Multiword expressions in lexical resources: Linguistic, lexicographic, and computational perspectives (Phraseology and Multiword Expressions 6). Berlin: Language Science Press, pp. 1–37.

sciendo
DOI 10.2478/jazcas-2025-0019

PHRASEMES AND COLLOCATIONS IN THE CORPUS – HOW TO FIND UNKNOWN VARIANTS

HANA SKOUMALOVÁ[1] – PŘEMYSL VÍTOVEC[2] – MILENA HNÁTKOVÁ[3]
[1]Department of Linguistics, Faculty of Arts, Charles University, Prague, Czech Republic (ORCID: 0000-0002-3519-0233)
[2]Faculty of Arts, Charles University, Prague, Czech Republic (ORCID: 0009-0004-2286-1706)
[3]Department of Linguistics, Faculty of Arts, Charles University, Prague, Czech Republic (ORCID: 0000-0002-4790-9807)

SKOUMALOVÁ, Hana – VÍTOVEC, Přemysl – HNÁTKOVÁ, Milena: Phrasemes and Collocations in the Corpus – How to Find Unknown Variants. Journal of Linguistics, 2025, Vol. 76, No 1, pp. 212 – 222.

Abstract: This paper addresses the identification and annotation of multiword expressions (MWEs) in Czech corpora, focusing on enhancing the search procedure through transformations of existing lexicon entries and the addition of new entries based on syntactic patterns. We discuss the limitations of current annotation systems and introduce a new, efficient annotation system that leverages a comprehensive MWE dictionary. Our methodology includes the use of syntactic patterns to identify new collocations, automatic transformations of known MWEs, and manual searches for creatively varied expressions. The results demonstrate significant improvements in the success rate of corpus annotation, with newly identified collocations and transformed MWEs contributing to a richer and more accurate linguistic resource.

Keywords: emultiword expressions, corpus annotation, syntactic patterns, lexicon transformations, Czech language

1 INTRODUCTION

Collocations and phrasemes are integral to language, studied within phraseology, but their annotation in corpora lags behind other types of annotation. The most comprehensive phraseologically annotated corpus of Czech is SYNv13 (Křen et al. 2024), with a size of about 6.5 billion positions. Another corpus with partial MWE annotation is PDT-C (Hajič et al. 2024), where MWE annotation is part of the annotation at the deep syntactic level. The MWEs were manually annotated and are mostly verb phrases contained in the Vallex dictionary (see Lopatková et al. 2016 and 2022). The size of the syntactically annotated part of PDT-C is about 2.25 mln, but MWEs are annotated only in its part (the original PDT) of about 675,000 words. There is also a pilot corpus resulting from the PARSEME project (Savary et al. 2023), which contains about 830,000 positions.

Another large Czech corpus is csTenTen from the TenTen corpus series (Jakubíček et al. 2013), which contains about 5.7 billion words. The corpus does not directly contain phraseological annotation, but it is possible to use so-called word sketches that reveal the collocation of individual words. Another tool associated with the TenTen corpora is a frequency-ordered list of n-grams, which actually represent MWEs.

When searching for phrasemes and collocations in the corpus, two approaches are possible. One is to use various statistical measures or sketches and n-grams, whereby the user is given a frequency list of the collocations found. These methods are useful for extracting information about individual words and their collocability. The other approach searches and annotates the corpus for collocations based on the dictionary, trying to find all variants of a certain, previously known collocation. This approach is suitable for phraseological research done on corpora. Unlike the first approach, it is possible to annotate (and later retrieve) e.g. proverbs or long sayings that would be difficult to find using statistical methods. Methods based on n-grams or word sketches cannot capture all possible word order variations, different inter-word distances and possible MWE transformations.

2 ANNOTATION OF CORPORA WITH PHRASEMES

In our work, we use an MWE dictionary. For corpus annotation we still use the now obsolete FRANTA system (see Kopřivová and Hnátková 2012). The disadvantages of this solution are (1) the specific format of the phraseological dictionary, which is only machine-readable, and (2) the insufficient speed of annotation. In response to these shortcomings, we are currently working on a new annotation system that works with data from the MWE database LEMUR (see Skoumalová et al. 2024) and implements a very efficient retrieval and annotation algorithm.

Both the dictionary of the FRANTA system (called FRANTALEX) and the dictionary represented by the LEMUR database are based on the Dictionary of Czech phraseology and idiomatics (SČFI, Čermák et al. 1983–2009), but are enriched with other phrasemes and collocations found in corpora (see Hnátková 2006). The dictionary contained in LEMUR is not only machine-readable but it is also suitable for human users (see Skoumalová et al. 2024). It also contains much more information about each entry so it is not only useful for finding collocations in the corpus. A final advantage of the new system is that it can annotate much faster than the previous one, which is mainly due to the fact that the dictionary is compiled into a machine-readable form before being used by the search engine.

FRANTALEX, which serves as a source of entries for the new system, contains about 56,000 entries. A large part of it has already been transferred to the new database, which contains about 26,000 entries, but the two numbers cannot be straightforwardly compared – when the entries are transferred, some variants that were previously separate entries are merged into one entry.

When using either system for corpus annotation, we take care to search for different word-ordered and disjointed variants, or variants with changed lexical content, or fragments (see Kopřivová and Hnátková 2012), e.g.

(1) a.

účel	***světí***	*v boji*	***prostředky***
purpose	sanctifies	in combat	means

‘the end justifies the means in combat’

b.

účel	*mediální propagandy*	***světí***	*jakékoliv*	***prostředky***
purpose	of media propaganda	sanctifies	any	means

c.

účel	*a případný úspěch v politice*	***světí***	*a často omlouvá*	***prostředky***
purpose	and eventual success in politics	sanctifies	and often excuses	means

d.

nepsal nic o	***prostředcích,***	*které by*	***účel***	***světil***
wrote nothing about	means	that.ACC would	the purpose	sanctified

‘he didn’t write anything about the means that would the purpose sanctify’

(2)

vnímat	***jako***	***hrozbu***	⇒	***brát***	***jako***	***hrozbu***
perceive	as	threat		take	as	threat

(3)

Kdo	***jinému***	***jámu...***
Who.NOM	else.DAT	hole.ACC...

‘[He] who [digs] another’s hole [falls into it himself.]’

The word-order and discontinuous variants as well as fragments are described directly in the FRANTALEX dictionary and in the LEMUR database, respectively, and will not be dealt with in this article. We will assume that the newly identified and described MWEs can also occur in such variants.

3 METHODS OF SEARCHING FOR NEW (VARIANTS OF) MWES

However, we have more ambitious goals, namely to create additional variants during compilation – transformations of existing units.

In addition to working with existing units, we are also looking for candidates to be added to the dictionary. This search cannot be done during annotation, but special CQL queries are entered into the annotated corpus, the results are then sorted by frequency and candidates for addition to the MWE dictionary are manually selected.

A final way to search for unknown variants of known collocations is to search for variants that have been creatively varied by speakers of the language. These are various adaptations of proverbs, well-known quotes and sayings. Sometimes two such expressions are contaminated, either deliberately or through ignorance. Example 4 shows one such case.

(4) a. ***mlsný*** ***jazýček*** ***na vahách***
picky tongue/pointer on scales

b. ***mlsný*** ***jazýček***
picky tongue

c. ***jazýček*** ***na vahách***
tongue on scales
'pointer on scales'

The phraseme in 4. a. is a compound of the phrasemes in 4. b. and c. and was used to describe a small political party that could choose which way to lean, and therefore it could make demands.

3.1 New adepts for a dictionary

The basic way to enrich the dictionary with new entries is to search for new collocations based on syntactic patterns. In this way, by which we still enrich FRANTALEX and then transfer the found MWEs to LEMUR, mainly established compounds and terms are found. In the early days of dictionary building, we focused only on semantically idiomatic MWEs. However, we are currently expanding the dictionary to include statistically idiomatic expressions as well. Syntactic patterns such as Adj+Noun, Verb+Noun.ACC, Noun+Noun.GEN, etc. are useful for searching such expressions.

The search is performed by issuing a CQL query to find a certain sequence of tags, e.g. a query

```
1:[tag=”A.*”] 2:[tag=”NN.*”] & 1.c=2.c within <s/>
```

searches for an Adj-Noun sequence in the same case within a single sentence. Other similar queries are

`[tag=”V.*”] [tag=”NN..4.*”] within <s/>`, which searches for verbs with an object in the accusative;

`[tag=”NN..[^2].*”] [tag=”NN..2.*”] within <s/>`, which finds a noun modified by another noun in the genitive case;

`[tag=”NN..[^2].*”] [tag=”A...2.*”] [tag=”NN..2.*”] within <s/>`, which finds a noun modified by an adjective and a noun in the genitive case.

We sort the results of each query by frequency and manually select new entries for the dictionary.

3.2 Identification of MWE transformations during annotation

Another way to search for variants that are not explicitly captured in the dictionary is to create regular transformations of (mainly) verb constructions. For FRANTA, these transformations are created automatically for single phrases and then manually added to the dictionary. For this reason, there are only a limited number of them in FRANTALEX. For a system using LEMUR, they are created automatically when the dictionary is compiled.

The simplest transformations are passivization, nominalization and adjectivization. In these transformations, the base word or its form is changed, and the valency frame may also change, which affects other words in the phraseme. For these diatheses we follow the rules formulated in Rosen and Skoumalová (2018). For example, the saying *hodit flintu do žita* lit. 'to throw rifle into rye(field)', 'to throw in the towel' is found in all moods, tenses and persons in texts, but it is also possible to create the periphrastic passive *flinta je hozena do žita* 'the towel is thrown in' or the reflexive passive *flinta se hodí do žita*. Since the verb *hodit* 'to throw' is transitive in this construction (and has an object in the accusative case), the transformation for the periphrastic passive is as follows:

1. The object in the accusative changes its case to the nominative and becomes the subject of the construction.
2. The rest of the construction is unchanged.
3. The verb can only be in the passive participle form.

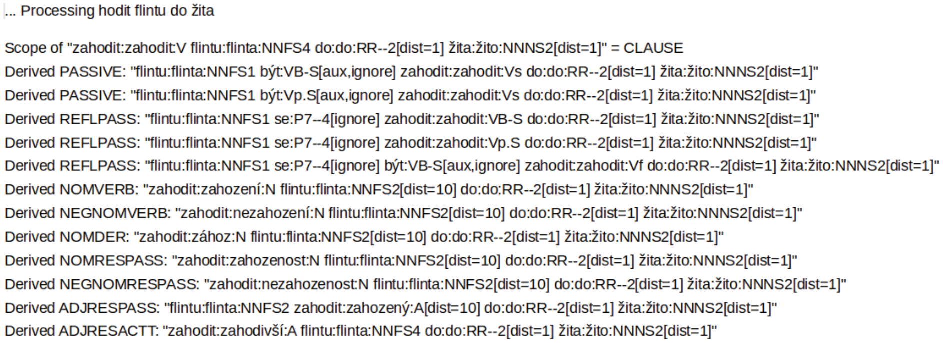
For the reflexive passive, similar rules apply:

1. The object in the accusative changes its case to the nominative and becomes the subject of the construction.
2. The reflexive particle *se* is added to the construction.
3. The rest of the construction does not change.
4. The verb can only be in active forms.

In the algorithm described above, we do not mention the subject of the original construction, because we only work with verb constructions in their basic (dictionary) form, which is the infinitive.

For both kinds of passive, it holds that they cannot be formed from reflexive verbs, so that, for example, in the saying *bojovat/prát se/zahrát si pro čest a slávu* 'to fight/play for honor and glory', only the verb *bojovat* 'to fight' can undergo passive diathesis.

Another possible transformation is nominalization; in our example it would be *hození flinty do žita* 'throwing a rifle into rye', or the adjectivization *flinta hozená do žita* 'rifle thrown into rye'. Some nominalizations and adjectivizations of verbs are word-formationally regular and are captured in the morphological dictionary (see Štěpánková et al. 2020). Other, irregular derivations are retrieved using the Derinet system (Ševčíková and Žabokrtský 2014). From the Derinet network, we retrieve not only nouns derived (according to the traditional view of word formation) from verbs, but we also retrieve words that did not arise by traditional derivation (e.g. *práce* ‘work’ as a derivative of *pracovat* ‘to work’). We can also look for derivations of nouns that fill other positions in the phrase, e.g. diminutives, or feminine nouns. In this way, automatic transformations yield additional variants of the phrases in the dictionary, e.g. *hození flinty do žita* ‘the throwing of a rifle into the rye’, *flinta hozená do žita* ‘a rifle thrown into the rye’, or *házející flintu do žita* ‘[sb] throwing a rifle into the rye’, or possibly *ministryně financí* ‘female-minister of finance’ or *zdravotní sestřička* lit. ‘medical little sister’, ‘nurse’. A partial listing of derivations made during dictionary compilation is shown in Fig. 1.



```
... Processing hodit flintu do žita

Scope of "zahodit:zahodit:V flintu:flinta:NNFS4 do:do:RR--2[dist=1] žita:žito:NNNS2[dist=1]" = CLAUSE
Derived PASSIVE: "flintu:flinta:NNFS1 být:VB-S[aux,ignore] zahodit:zahodit:Vs do:do:RR--2[dist=1] žita:žito:NNNS2[dist=1]"
Derived PASSIVE: "flintu:flinta:NNFS1 být:Vp.S[aux,ignore] zahodit:zahodit:Vs do:do:RR--2[dist=1] žita:žito:NNNS2[dist=1]"
Derived REFLPASS: "flintu:flinta:NNFS1 se:P7--4[ignore] zahodit:zahodit:VB-S do:do:RR--2[dist=1] žita:žito:NNNS2[dist=1]"
Derived REFLPASS: "flintu:flinta:NNFS1 se:P7--4[ignore] zahodit:zahodit:Vp.S do:do:RR--2[dist=1] žita:žito:NNNS2[dist=1]"
Derived REFLPASS: "flintu:flinta:NNFS1 se:P7--4[ignore] být:VB-S[aux,ignore] zahodit:zahodit:Vf do:do:RR--2[dist=1] žita:žito:NNNS2[dist=1]"
Derived NOMVERB: "zahodit:zahození:N flintu:flinta:NNFS2[dist=10] do:do:RR--2[dist=1] žita:žito:NNNS2[dist=1]"
Derived NEGNOMVERB: "zahodit:nezahození:N flintu:flinta:NNFS2[dist=10] do:do:RR--2[dist=1] žita:žito:NNNS2[dist=1]"
Derived NOMDER: "zahodit:zához:N flintu:flinta:NNFS2[dist=10] do:do:RR--2[dist=1] žita:žito:NNNS2[dist=1]"
Derived NOMRESPASS: "zahodit:zahozenost:N flintu:flinta:NNFS2[dist=10] do:do:RR--2[dist=1] žita:žito:NNNS2[dist=1]"
Derived NEGNOMRESPASS: "zahodit:nezahozenost:N flintu:flinta:NNFS2[dist=10] do:do:RR--2[dist=1] žita:žito:NNNS2[dist=1]"
Derived ADJRESPASS: "flintu:flinta:NNFS2 zahodit:zahozený:A[dist=10] do:do:RR--2[dist=1] žita:žito:NNNS2[dist=1]"
Derived ADJRESACTT: "zahodit:zahodivší:A flintu:flinta:NNFS4 do:do:RR--2[dist=1] žita:žito:NNNS2[dist=1]"
```

Fig. 1. Partial list of transformations of the saying *hodit flintu do žita*

We can see that during transformations, overgeneration occurs – we will probably never encounter *nezahozenost flinty* ‘rifle’s un-thrown-ness’ in the corpus, but for this reason, overgeneration is not a problem. A problem can arise if, for example, a diminutive has a different meaning than the base word. For example, *stará panna* ‘old maid’ and *stará panenka* ‘old doll’. We must solve these cases individually and prevent such diminutives from being generated and used.

3.3 Searching for unknown variants in an annotated corpus

As mentioned above, the authors of the texts often creatively modify well-known proverbs, sayings and quotations and their identification in the corpus is difficult. For these purposes, there is no choice but to pick out a possible phraseme and enter CQL queries that might reveal its variations. We illustrate this search with the idiom *vlk se nažral a koza zůstala celá* lit. ‘the wolf has eaten and the goat has remained whole’, ‘an order was formally filled, but practically nothing changed’. If we enter a CQL query (5) that searches for the lemmas *koza* ‘goat’ and *nažrat* ‘eat’ within 10 positions of each other within a single sentence, we get the occurrences shown in Example 6.

(5) `(meet [lemma=”koza” & col_lemma=””] [lemma=”nažrat”] -5 5)`

(6) a. *..., aby* ***se*** *konkurzní hyeny* ***nažraly*** ***a koza zůstala celá***
..., so that bankruptcy hyenas ate and goat remained whole

b. ***vlk*** *poznání* ***se nažere*** ***a*** *klipová* ***koza*** *mečí do éteru dál*
wolf of knowledge eats and clip goat keeps bleating into ether

c. ***Vlk se nažral*** ***a kozy*** ***zůstala*** *půlka.*
Wolf has eaten and of goat remained half.
‘The wolf has eaten and a half of the goat remained.’

d. *... dát* ***nažrat*** ***vlkovi,*** *aby* ***koza*** *přitom* ***zůstala celá.***
... give eat.INF wolf.DAT so that goat at the same time remained whole
‘... to let the wolf eat so that the goat remained whole at the same time’

e. ***Koza se nažere,*** ***vlk zůstane celej,*** *já mám po starostech,...*
Goat eats, wolf remains whole, I have no troubles

It is clear that all of these findings refer to the original saying, but none of them has been identified as an occurrence of it. The variation may consist in an altered lexical setting (6. a. and b.), in a modification of meaning (6. c.), in a change of modality with the corresponding change of case (6. d.), or in a complete reversal of meaning (6. e.). If we modify the CQL query to include two other words from the original saying, we will get additional variations.

The question is whether we should even try to describe and find these variants when annotating. For those that preserve the semantics of the original saying, we need to modify the constraints in the dictionary to allow other lexical settings, or to allow a fragment to suffice for identification. Where the semantics differs, we need to consider a new entry in the dictionary (if the new phraseme is frequent enough), which will be linked to the original entry by a super-lemma.

4 RESULTS OF INDIVIDUAL METHODS

All three methods mentioned in the previous section were really used, although the third method (manual search for variants) was used only to a limited extent. However, the first two methods significantly improved the success rate of corpus annotation. Following is an overview by each method.

4.1 Newly added phrasemes and collocations

The new collocations added by the syntactic patterns search were used in the annotation of a testing corpus of 130 million words, NEWTON2023, a corpus of journalism acquired in 2024, which was annotated by the FRANTA system. Counting the types, the new collocations represent 7.57% of the annotated collocations and for occurrences (tokens) they represent 16.35%. The following table compares the frequency of new collocations with the original ones.

	original types	new types	original tokens	new tokens
Adj-Noun	5,321	2,208	627,272	457,497
Noun-Adj.GEN-Noun.GEN	12	291	133	6,946
Noun-Noun.GEN	2,273	756	56,512	77,284
Verb-Noun.ACC	5,479	362	252,498	12,702

Tab. 1. Comparison of the frequency of new collocations with original collocations

We can see that some syntactic patterns yielded a large number of collocations identified in the corpus, although the newly found types (i.e., individual collocations) were not as numerous. However, these were the most frequent established expressions such as *životní prostředí* 'environment', *hlavní město* 'capital city', or *mistrovství světa* 'world championship', *růst cen* 'price rise', *ministr financí* 'finance minister', etc.

4.2 Collocations and phrasemes identified using transformations

This method has not yet been used for the annotation of any published corpus, we are still testing it. We annotated the same test corpus of 130 million words with a method using a compiled dictionary from LEMUR with automatic transformations, and we found that 5,335 transformations were applied out of 765,518 generated, which is about 0.7% of the proposed transformations. However, there are some very frequent ones among them, such as *zvýšení daně* 'tax increase', which has a higher frequency than the basic form *zvýšit daň* 'to increase tax' (i.p.m. 13.73 versus 5.78), or *odchod do důchodu* 'retirement' (i.p.m. 9.55) versus *jít do důchodu* 'to retire' (i.p.m. 4.91). The distribution of transformations in the corpus by type is shown in Tab. 2.

Type	Occurrences	% of collocations
Without transformations	3,397,366	97,56
PASSIVE (participle ending with *-n/-t*)	6,082	0,17
REFLPASS	12,100	0,35
NOMVERB (nominalization of verb – *-ní/-tí*)	35,694	1,03
NEGNOMVERB (negation of the above)	603	0,02
NOMDER (derived noun – *hrát* 'play' – *herec* 'actor')	17,409	0,50

NOMRESPASS (pass. result – -*nost*/-*tost*)	4	0
NEGNOMRESPASS (negation)	0	0
NOMPOTIMP (possibility – -*telnost*)	19	0
NEGNOMPOTIMP (negation)	1	0
NOMRESACTL (act result – -*lost*)	3	0
NEGNOMRESACTL (negation)	0	0
NPDER (diminutive, feminine... – -*yně*/-*ček*/-*čka*)	7,485	0,21
ADJPOTIMP (possibility – -*telný*)	37	0
ADJPROC (active adj. – -*ící*)	2,933	0,08
ADJRESACTL (act. pres. result – -*lý*)	96	0
ADJRESACTT (act. past result – -*vší*)	1	0
ADJRESPASS (passive result – -*ný*/-*tý*)	2,339	0,07
TOTAL	3,482,172	100

Tab. 2. Frequency of transformations in the corpus

We can see that some transformations have very low representation in the texts. For example, NOMRESPASS denotes derived nouns expressing a resulting state after some action, ending in -*ost*, e.g. *zajištěnost dodávek* ‘supply assurance’, *sehranost komedie* ‘comedy enactment’, etc. On the other hand, derived nouns ending in -*ní*/-*tí* (NOMVERB) represent the most numerous group among the transformations.

5 CONCLUSIONS

In our paper, we have shown three methods that can be used to extend the MWE lexicon and/or improve the success rate of corpus annotation with MWEs. We tried three methods: we retrieved potential collocations according to a syntactic pattern, we used transformations of known collocations and phrasemes, and we tried retrieval of lexically varied and fragmentary variants.

The first method seems to be the most beneficial in terms of the number of subsequently annotated collocations. However, it has a limitation in that it only finds collocations that occur in the canonical word order in the texts and are not split by other words.

The second and third methods do not yield as many newly annotated variants of collocations, but they open up new possibilities in research on the variation of phrasemes and collocations. On the one hand, there are possibilities to investigate what transformations are possible for collocations and how often speakers use them, and on the other hand, it is also possible to investigate the creative variation of established collocations and phrasemes.

In future work, we will develop all three methods of dictionary enrichment and corpus annotation and use them to annotate other corpora.

ACKNOWLEDGEMENTS

The research has been supported by the Czech Science Foundation under the project GA CR 24-10254S, by the Technology Agency of the Czech Republic under the project TQ01000177, and by Ministry of Education, Youth and Sports under the Large Research, Development and Innovation Infrastructures, LM2023044.

References

Čermák, F., et al. (1983–2009). Slovník české frazeologie a idiomatiky 1–4. Praha: Academia/Leda.

Hajič, J., et al. (2024). Prague Dependency Treebank – Consolidated 2.0 (PDT-C 2.0). Data/software, LINDAT-CLARIAH-CZ. Accessible at: http://hdl.handle.net/11234/1-5813.

Hnátková, M. (2006). Typy a povaha komponentů neslovesných frazémů z hlediska lexikálního obsazení. In: F. Čermák – M. Šulc (eds.): Kolokace, Nakladatelství Lidové noviny/Ústav Českého národního korpusu, Praha, pp. 142–167.

Jakubíček, M., Kilgarriff, A., Kovář, V., Rychlý, P., and Suchomel, V. (2013). The TenTen Corpus Family. In 7th International Corpus Linguistics Conference CL 2013, pp. 125–127. Lancaster.

Kopřivová, M., and Hnátková, M. (2012). From Dictionary to Corpus. In Phraseology in Dictionaries and Corpora, pp. 155–168. Maribor.

Křen, M., Cvrček, V., Čapka, T., Hnátková, M., Jelínek, T., Kocek, J., Kováříková, D., Křivan, J., Milička, J., Petkevič, V., Skoumalová, H., Šindlerová, J., and Škrabal, M. (2024). Korpus SYN, v13 from 27/12/2024. Ústav Českého národního korpusu FF UK, Praha. Accessible at: https://www.korpus.cz.

Lopatková, M., Kettnerová, V., Bejček, E., Vernerová, A., and Žabokrtský, Z. (2016). Valenční slovník českých sloves VALLEX. Praha: Karolinum.

Lopatková, M., Kettnerová, V., Mírovský, J., Vernerová, A., Bejček, E., and Žabokrtský, Z. (2022). VALLEX 4.5. LINDAT/CLARIAH-CZ Digital Library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University, Prague. Accessible at: http://hdl.handle.net/11234/1-4756.

Rosen, A., and Skoumalová, H. (2018). No way to have your say out of the frame: specifying valency of multi-word expressions. Prace filologiczne (LXXII), pp. 301–320.

Savary, A., et al. (2023). PARSEME corpora annotated for verbal multiword expressions (version 1.3). LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University. Accessible at: http://hdl.handle.net/11372/LRT-5124.

Ševčíková, M., and Žabokrtský, Z. (2014). Word-Formation Network for Czech. In Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014), pp. 1087–1093. Reykjavík.

Skoumalová, H., Kopřivová, M., Petkevič, V., Jelínek, T., Rosen, A., Vondřička, P., and Hnátková, M. (2024). Lemur: A lexicon of Czech multiword expressions. In: V. Giouli – V. Barbu Mititelu (eds.): Multiword expressions in lexical resources: Linguistic, lexicographic, and computational perspectives. Language Science Press, Berlin, pp. 1–37.

Štěpánková, B., Mikulová, M., and Hajič, J. (2020). The MorfFlex Dictionary of Czech as a Source of Linguistic Data. In Proceedings of XIX EURALEX Congress: Lexicography for Inclusion, pp. 387–392. Democritus University of Thrace, Thrace, Greece.

sciendo
DOI 10.2478/jazcas-2025-0020

DEVELOPMENT OF A DATABASE AND MODELS FOR CHILDREN’S SPEECH IN THE SLOVAK LANGUAGE FOR SPEECH-ORIENTED APPLICATIONS

JÁN STAŠ[1] – STANISLAV ONDÁŠ[2] – MATÚŠ PLEVA[3] –
MATEJ HORVÁTH[4] – RICHARD ŠEVC[5] – PATRIK MICHALANSKÝ[6]
[1]Department of Electronics and Multimedia Communications, Faculty of Electrical Engineering and Informatics, Technical University, Košice, Slovakia
(ORCID: 0000-0001-7403-0235)
[2]Department of Electronics and Multimedia Communications, Faculty of Electrical Engineering and Informatics, Technical University, Košice, Slovakia
(ORCID: 0000-0002-0075-3788)
[3]Department of Electronics and Multimedia Communications, Faculty of Electrical Engineering and Informatics, Technical University, Košice, Slovakia
(ORCID: 0000-0003-4380-0801)
[4]Department of Electronics and Multimedia Communications, Faculty of Electrical Engineering and Informatics, Technical University, Košice, Slovakia
[5]Department of Electronics and Multimedia Communications, Faculty of Electrical Engineering and Informatics, Technical University, Košice, Slovakia
[6]Department of Electronics and Multimedia Communications, Faculty of Electrical Engineering and Informatics, Technical University, Košice, Slovakia

STAŠ, Ján – ONDÁŠ, Stanislava – PLEVA, Matúš – HORVÁTH, Matej – ŠEVC, Richard – MICHALANSKÝ, Patrik: Development of a Database and Models for Children’s Speech in the Slovak Language for Speech-oriented Applications. Journal of Linguistics, 2025, Vol. 76, No 1, pp. 223 – 233.

Abstract: Children’s speech differs significantly from adult speech due to physiological and cognitive developmental factors. Key differences include higher pitch, a shorter vocal tract, greater formant frequencies, slower speaking rates, and greater variability in pronunciation and articulation. These differences result in acoustic mismatches between children’s and adult speech, making traditional automatic speech recognition models trained on adult speech less effective for children. Additionally, linguistic differences, such as limited vocabulary and evolving grammar, further contribute to this challenge. This paper focuses on the creation of a children’s speech database for the low-resource Slovak language. This database has been used to train acoustic models for the automatic recognition of spontaneous children’s speech in Slovak. In this research, we compared three different approaches to speech recognition, with self-supervised learning achieving results comparable to similar studies in this area, despite using relatively small amounts of training data.

Keywords: acoustic model, automatic speech recognition, data augmentation, children’s speech, speech database

1 INTRODUCTION

Automatic speech recognition (ASR) for children remains a challenging area due to fundamental differences in various acoustic and linguistic aspects between children's and adults' speech. Acoustically, children's speech is characterized by a higher fundamental frequency (F0), increased formant frequencies (F1–F3), slower speaking rates, and greater variability in articulation due to their developing vocal tracts and speech motor control (Patel 2014; Shivakumar 2020; Lu 2022). Linguistically, children's speech exhibits greater variability in pronunciation, increased disfluencies, and evolving phonetic structures as they develop (Yeung 2018). Word pronunciation can be inconsistent due to incomplete language acquisition. Additionally, children's shorter vocal tracts result in different resonance characteristics, making the direct adaptation of adult-trained ASR models ineffective (Gerosa 2009; Lu 2022). These factors collectively lead to higher Word Error Rates when adult-trained ASR systems are applied to children's speech (Sobti 2024).

To address these challenges, various methods have been explored in recent years. Traditional Gaussian Mixture Model-Hidden Markov Model (GMM-HMM) approaches have been widely used, but recent advances favor Deep Neural Networks (DNNs), End-to-End architectures, transformer-based or self-supervised learning models such as Wav2Vec 2.0, which have significantly improved children's ASR performance (Bhardwaj 2012; Shivakumar 2020; Lu 2022; Sobti 2024). Adaptation techniques such as Vocal Tract Length Normalization, Feature-space Maximum Likelihood Linear Regression, Subglottal Resonance Normalization, and Speaker Adaptive Training have been applied to reduce acoustic mismatches between adult and child speech (Patel 2014; Yeung 2018). Transfer learning from adult ASR models to child-specific models has also proven effective, particularly when adapting both acoustic and pronunciation models (Shivakumar 2020). Additionally, data augmentation approaches such as Speed Perturbation and Spectral Augmentation have been used to simulate children's speech. Self-supervised learning has also emerged as a promising approach to mitigate the scarcity of children's speech data, with fine-tuning pre-trained adult models on a small amount of children's speech yielding significant improvements (Lu 2022; Sobti 2024).

Despite these advancements, research gaps remain, including persistently high WER, limited robustness to spontaneous speech, and poor performance in low-resource languages. A significant challenge is the lack of large, publicly available multilingual child speech corpora, which hinders model training and limits the generalizability of ASR systems (Yeung 2018; Sobti 2024). Furthermore, most studies focus on read speech, whereas spontaneous speech recognition remains underexplored (Gerosa 2009). Moreover, speech variability across different child age groups suggests that a single ASR model may not be effective for all children

(Yeung 2018). Age-specific models and improved data collection methods are crucial to enhancing ASR systems' accuracy and adaptability.

Since existing adult-trained ASR models struggle with the unique acoustic and linguistic characteristics of children's speech, building new child-specific datasets and developing tailored models are essential steps toward bridging the performance gap in children's speech recognition. For these reasons, we decided to create a children's speech database for the Slovak language to expand the existing range of languages. With the help of this data, we aim to train acoustic models (AMs) applicable to recognizing children's spontaneous speech, as well as to the design and development of other speech-oriented applications, such as children's speech synthesis and human-machine or human-robot interfaces.

2 DATABASE OF CHILDREN'S SPEECH

2.1 Existing databases of children's speech

Children's speech databases are essential for advancing ASR research. This section provides an overview of key corpora, highlighting their design and data collection methods.

The CMU Kids Corpus is a database of children's read-aloud speech recorded in American English. It includes speech from 76 children aged 6 to 11, with 24 male and 52 female speakers. The corpus consists of 5,180 utterances created to train ASR models for the LISTEN project at Carnegie Mellon University. The dataset is divided into two subsets: SUM95, which contains speech from proficient readers recorded in summer camps, and FP, which includes speech from children at risk of developing poor reading skills (Eskenazi 1997).

The OGI Kids' Speech Corpus is a database of children's speech in American English, consisting of recordings from approximately 1,100 children, ranging from kindergarten to grade 10. The corpus includes both prompted and spontaneous speech, with a balanced number of male and female speakers across different grade levels. Data collection involved children reading words and sentences displayed on a screen while synchronized with an animated character (Baldi), whereas spontaneous speech was elicited through conversational prompts (Shobaki 2000).

The PF-STAR Children's Speech Corpus is a multilingual database containing speech recordings from 611 children in British English, German, Italian, and Swedish. It includes both native and non-native speech, featuring read, imitated, spontaneous, and emotional speech recordings, covering an age range from 4 to 14 years. Data collection employed various methodologies, including scripted reading tasks and the AIBO method, in which children interacted with a robot to elicit natural and emotional speech (2005).

The Child Language Data Exchange System (CHILDES) is a database designed for studying child language acquisition. It includes recordings and transcripts in over

20 languages, making it a crucial resource for researchers. The corpus contains data from thousands of children and caregivers, though the distribution of male and female speakers varies across individual subcorpora. It was built by collecting, transcribing, and annotating naturalistic parent-child interactions, recorded in home environments (Sanchez 2019).

The My Science Tutor (MyST) corpus is one of the largest collections of children's conversational speech, containing 393 hours of speech data from 1,371 students in grades 3–5. The corpus is in English and consists of 228,874 utterances recorded during 10,496 virtual tutoring sessions, with equal participation from male and female students. The data was collected through structured spoken dialogues between students and a virtual science tutor, in which students answered science-related questions in a strict turn-taking system (Pradhan 2024).

The study by Claus et al. (Claus 2013) provides a comprehensive survey of children's speech databases, which are essential for ASR research. It identifies and describes a total of 34 databases, primarily in English, with some available in German, Italian, Swedish, and other languages – unfortunately, excluding Slovak. Most databases contain read or spontaneous speech from children aged 6 to 18 years, with significantly fewer resources for younger children. Preschool speech databases are particularly scarce due to recording challenges, as young children cannot read and have short attention spans. The study highlights the need for more extensive and higher-quality speech databases, especially for children under the age of six.

2.2 Building of the Slovak children's speech database

We began working on the creation of the speech database as early as 2018. A portion of the database, consisting solely of excerpts from children's speech taken from the eight parts of the series Dads ('*Oteckovia*'), was published in (Pleva 2019). '*Oteckovia*' was a Slovak family daytime series broadcast on TV Markíza, depicting the lives of four young men—fathers—each struggling with the role of parenthood in his own way. It is an adaptation of the 2014 Argentine telenovela '*Señores Papis*'. Although the speech in the series has a spontaneous character, it is still a spoken script which we do not consider fully authentic.

For this reason, we proceeded to transcribe more authentic speech from child speakers. We focused on two programs: the TV show '*Táraninky*', broadcast by Slovak television RTVS between 2020 and 2023, and the radio show '*Rozhlasové leporelo*', aired on Rádio Regina between 2021 and 2023. Both programs feature speech interactions in which adult speakers ask various types of questions, and child respondents provide answers. '*Táraninky*' was a children's talk show covering various topics, hosted by Marián Čekovský and his little guests. '*Rozhlasové leporelo*' was a children's radio show focused on developing thinking and creativity, encouraging spontaneous reactions and communication skills.

2.3 Data preprocessing and transcription

As already mentioned, the process of acquiring, pre-processing and transcription of the subcorpus of the series ‘*Oteckovia*’ is described in more detail in (Pleva 2019).

Audio recordings of the shows ‘*Táraninky*’ and ‘*Rozhlasové leporelo*’ were obtained from the RTVS online archive[1] and processed in similar way. A total of 70 episodes of the radio show ‘*Rozhlasové leporelo*’ were analyzed, with 36 episodes discarded due to excessive background noise.

Next, only speech segments containing child speech were isolated using the Audacity tool[2], while segments featuring older children, adults, or other irrelevant audio content were removed. A three-second pause was inserted between individual segments of children’s speech, and the recordings were normalized for volume. The processed speech recordings were subsequently down-sampled from 44.1 kHz to 16 kHz with 16-bit resolution and saved in standard WAV (PCM) format.

After processing the audio recordings, it was necessary to create a transcription for each of them. The initial speech-to-text transcription was performed automatically using the SARRA[3] ASR system (Lojka 2018). Further adjustments to the transcription were made manually using the Transcriber tool[4] (Barras 2001), where each speech segment was assigned a unique speaker identifier and a time period indicating the start and end of the speaker’s speech. An example of a transcription in the Transcriber tool is shown in Fig. 1.

2.4 Analysis of the database of children’s speech

As shown in Tab. 1, the latest version of the Slovak children’s database consists of 130 children’s TV and radio shows and contains a total of 2,589 speech segments with 303 speakers (127 males and 176 females), amounting to almost five hours of transcribed speech. The average duration of a speech segment is approximately 6.85 seconds. The total number of words in the database is 35,945, of which 3,441 are unique. The age range of child speakers in the database is between 4 and 12 years.

Next, we analyzed the number of out-of-vocabulary (OOV) words. For this calculation, we used the dictionary from the SARRA ASR system, which contains more than 558,000 unique words. Our findings show that the ‘*Oteckovia*’ subcorpus has the highest percentage of OOV words, reaching up to 4.46%. This is quite an interesting result, considering that the ‘*Táraninky*’ and ‘*Rozhlasové leporelo*’ subcorpora contain fully spontaneous speech. These subcorpora have slightly more than 1% of OOV words. Examples of OOV words in individual subcorpora are shown in Tab. 2.

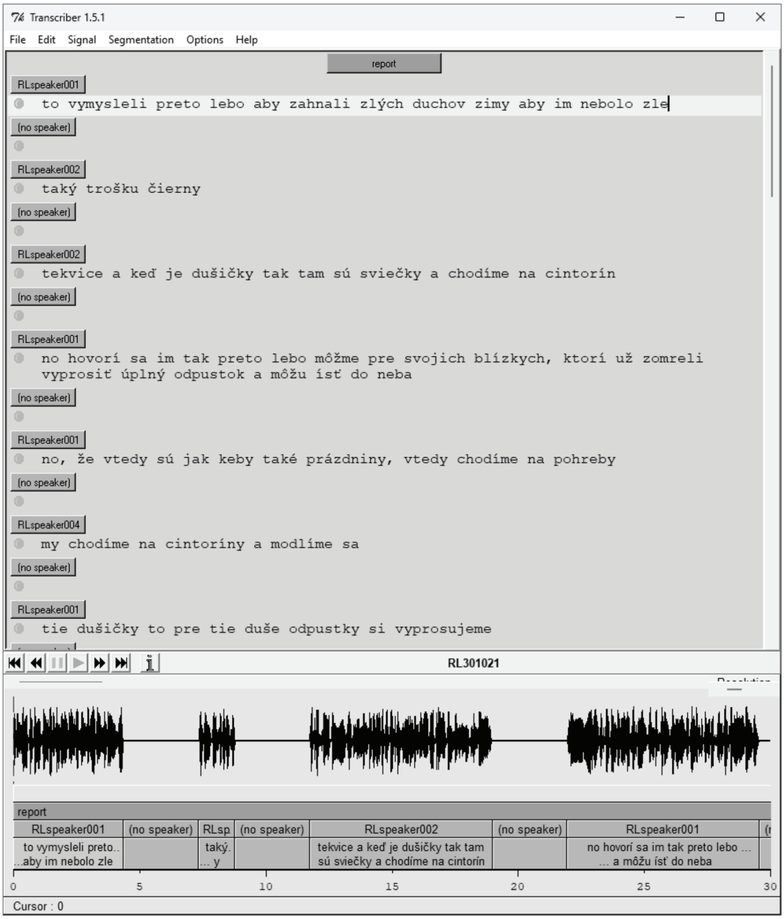
[1] https://www.stvr.sk/archiv

[2] https://www.audacityteam.org/

[3] https://marhula.fei.tuke.sk/sarra/

[4] https://trans.sourceforge.net/

Note that we are still working on transcribing additional sessions and expanding the database. Our goal is to reach at least 30 hours of pure children's speech.



RLspeaker001
to vymysleli preto lebo aby zahnali zlých duchov zimy aby im nebolo zle

(no speaker)

RLspeaker002
taký trošku čierny

(no speaker)

RLspeaker002
tekvice a keď je dušičky tak tam sú sviečky a chodíme na cintorín

(no speaker)

RLspeaker001
no hovorí sa im tak preto lebo môžme pre svojich blízkych, ktorí už zomreli vyprosiť úplný odpustok a môžu ísť do neba

(no speaker)

RLspeaker001
no, že vtedy sú jak keby také prázdniny, vtedy chodíme na pohreby

(no speaker)

RLspeaker004
my chodíme na cintoríny a modlíme sa

(no speaker)

RLspeaker001
tie dušičky to pre tie duše odpustky si vyprosujeme

RLspeaker001	(no speaker)	RLsp.	(no speaker)	RLspeaker002	(no speaker)	RLspeaker001
to vymysleli preto.. ..aby im nebolo zle		taký. ... y		tekvice a keď je dušičky tak tam sú sviečky a chodíme na cintorín		no hovorí sa im tak preto lebo ... ... a môžu ísť do neba

Fig. 1. The process of transcription of the show '*Rozhlasové leporelo*' in the Transcriber tool

3 MODELING OF CHILDREN'S SPEECH

We used the resulting speech database to train AMs for the task of automatic recognition of children's speech in Slovak. We selected three ASR architectures.

The first architecture is based on the freely available Kaldi ASR engine (Georgescu 2021), which performs speech decoding using Weighted Finite State Transducers (WFST). The baseline triphone AM was trained using a standard procedure based on the extraction of Mel-Frequency Cepstral Coefficients (MFCC) with Cepstral Variance and Mean Normalization (CVMN) to eliminate noise in the extracted speech features. To reduce the dimensionality of the acoustic feature vectors, we applied Linear Discriminant Analysis (LDA) in conjunction with Maximum Likelihood Linear Transform (MLLT) and Speaker Adaptive Training (SAT) to increase the model's robustness to speaker variability. This approach follows the standard GMM-HMM framework for speech recognition, incorporating a trigram language model (LM). Additionally, we tested newer, improved approaches in Kaldi based on 'chain' models, which represent a hybrid DNN-HMM approach. These models utilize a Factorized Time-Delay Neural Network (TDNN-F), which effectively captures long-term temporal dependencies in speech data, as well as an extended version, CNN-TDNN-F, which incorporates a Convolutional Neural Networks as an input layer to extract robust acoustic features. The main advantage of a framework built on the Kaldi recognizer lies in its control over the recognition dictionary, a feature that current architectures based on transformers or self-supervised learning do not provide.

Subcorpus	**'*Oteckovia*'**	**'*Táraninky*'**	**'*Rozhlasové leporelo*'**
Genre	TV family comedy	Kid's TV talk show	Radio sessions for children
Type	narrated scripts	spoken dialogues	spoken dialogues
Years of broadcasting	2018–2019	2020–2023	2021–2023
Age range of speakers	7–12	5–8	4–11
Number of sessions	8	88	34
Number of speakers	6 M / 6 F	43 M / 60 F	78 M / 110 F
Number of utterances	305 (154 M / 151 F)	1,667 (798 M / 869 F)	617 (229 M / 388 F)
Average utterance duration	5.80 seconds	7.20 seconds	6.42 seconds
Number of words	3,970	22,760	9,215
Number of unique words	1,377	4,981	2,521
OOV rate [%]	4.46	1.31	1.36
Total duration (hh:mm:ss)	00:33:13	03:19:45	01:06:00
	04:58:58		

Tab. 1. Statistics on subcorpora of the Slovak children's speech database

'*Oteckovia*'	**'*Táraninky*'**	**'*Rozhlasové leporelo*'**
Ajštaj, baragraft, bratránko, dákam, fotrovi, furt, kakauko, lakťovať, Peťuľka, ocí, potvorko, Prdelákovce, rucpaku, superpes, tato, tatí, tatino, tatuš, tatušó, trampoška, trampošku, vyšokovaný	*barz, Bebík, dáku, dinosaure, elzovský, gaťuše, jaké, jakeby, jakému, jakú, kakaničky, kakauko, našuchne, nesnežná, oblečko, očkatý, odznačky, prečarovať, rozburdať, súrodencovia, tatík, tatíkom, stamaď, sudokmeň, tuna, šlifka, videofotiek, vybáca, zabambuší, zaležaný, zbúraninka, zbúraniny, zlatokopia*	*akrobácie, baletkovú, balzy, docela, drúhe, jaké, jakú, lietatiel, makanie, náťaž, neni, pozauna, snižec, šalený, tehálmi, tote, toten, tuná, zverací*

Tab. 2. Examples of out-of-vocabulary words

To expand the training set, we applied data augmentation, including Speech Perturbation (SP) – modifying the speech rate by factors of 0.9 (slower) and 1.1

(faster) – and Spectral Augmentation (SA). These techniques increased the training set to four times its original size. Other data augmentation techniques, such as perturbation of pitch, volume, tempo, or vocal tract length, did not yield further improvements.

As a final step, after decoding the speech, we also applied post-processing using a LM based on Recurrent Neural Networks (RNN LM).

We chose Whisper[5] (Radford 2023) as the second ASR architecture. It is a closed, End-to-End recognition system implemented as an encoder-decoder Transformer. The decoder is trained to predict the corresponding text caption, interspersed with special tokens that enable the model to perform multilingual ASR, speech translation, and language identification. Whisper's AMs have been trained on a large and diverse multilingual dataset comprising 680,000 hours of audio[6]. Whisper offers several pre-trained models suitable for further fine-tuning. In this research, we evaluated the base, small, medium, and turbo models, with the medium model yielding the best ASR results. Fine-tuning the large model was beyond our computational capabilities.

The third and final architecture we used is Wav2Vec 2.0 (Bayevski 2020), a self-supervised framework for speech representation learning. It is based on a Transformer architecture and learns speech representations by masking parts of raw audio waveforms and predicting them from context. The model is pre-trained on large amounts of unlabeled audio data and can be fine-tuned for ASR with minimal labeled data. There are numerous pre-trained models based on the Wav2Vec 2.0 architecture that are suitable for fine-tuning with Slovak data. Among them, we applied:

- XLS-R-300M[7] (Babu 2022) – a large-scale multilingual ASR model pre-trained on 436,000 hours of unlabeled speech in 128 languages, including Slovak, with 300M parameters;
- MMS-1B-All[8] (Pratap 2024) – a large-scale multilingual ASR model pre-trained on one billion parameters across over 1,000+ languages.

4 EXPERIMENTS AND RESULTS

At the beginning, we divided the database of children's speech in Slovak into a training set and a test set. The test set contained approximately one-third of randomly selected speech segments from the '*Táraninky*' subcorpus, while the remaining data was used for training and validating the models. As a result, the test set included 389 speech segments from 29 speakers (10 males and 19 females), with

[5] https://github.com/openai/whisper

[6] The exact amount of Slovak audio data is not known.

[7] https://huggingface.co/facebook/wav2vec2-xls-r-300m

[8] https://huggingface.co/facebook/mms-1b-all

a total duration of 45 minutes and 57 seconds. The training and validation set was used either to train models from scratch or to fine-tune pre-trained models based on the Whisper or Wav2Vec 2.0 architectures.

We used the standard Word Error Rate (WER) to evaluate the model performance. WER is a common metric for assessing speech recognition performance, calculated as the ratio of the total number of substitutions, deletions, and insertions to the total number of words in the reference transcript.

The results summarized in Tab. 3 show that fine-tuning on children's speech data and applying data augmentation significantly improved ASR performance across all architectures. Kaldi models achieved a WER reduction from 46.10% to 24.19% with CNN-TDNN-F and data augmentation. The fine-tuned Whisper model achieved a WER of 18.29%, outperforming Kaldi. Wav2Vec 2.0 models demonstrated strong performance, with XLS-R-300M fine-tuned on augmented data and a trigram LM achieving a WER of 16.38%. MMS-1B-ALL performed best, reaching the lowest WER of 15.10% when fine-tuned on augmented data with a trigram LM, highlighting the effectiveness of self-supervised learning for child speech recognition.

ASR architecture	**Acoustic model setup**	**WER [%]**
Kaldi	MFCC+CMVN + LDA+MLLT+SAT + trigram LM	**46.10**
	TDNN-F + trigram LM	37.32
	CNN-TDNN-F + trigram LM	37.57
	TDNN-F + RNN LM	35.41
	CNN-TDNN-F + RNN LM	36.05
	TDNN-F + data augmentation (SP+SA) + RNN LM	27.27
	CNN-TDNN-F + data augmentation (SP+SA) + RNN LM	**24.19**
Whisper	medium	44.10
	medium fine-tuned on children's speech data	18.96
	medium fine-tuned on augmented dataset (SP+SA)	**18.29**
Wav2Vec 2.0	XLS-R-300M	41.14
	XLS-R-300M fine-tuned on children's speech data	26.48
	XLS-R-300M fine-tuned on augmented dataset (SP+SA)	25.55
	XLS-R-300M fine-tuned on augmented dataset + trigram LM	**16.38**
	MMS-1B-ALL	34.13
	MMS-1B-ALL fine-tuned on children's speech data	23.86
	MMS-1B-ALL fine-tuned on augmented dataset (SP+SA)	22.45
	MMS-1B-ALL fine-tuned on augmented dataset + trigram LM	**15.10**

Tab. 3. Summary of the results of newly trained or fine-tuned models for children's speech

5 CONCLUSION

Improving children's ASR requires a combination of age-specific adaptation techniques, data augmentation, and self-supervised learning to address data scarcity.

In this research, we compared three different approaches to speech recognition, with self-supervised learning achieving a WER of 15.10%, which is comparable to similar studies (Bhardwaj 2022; Sobti 2024), despite using less than 4 hours of training data.

Future research should focus on expanding child speech corpora, collecting more diverse speech samples from children across various age groups, refining transfer learning techniques, and developing more effective domain adaptation strategies to bridge the performance gap between adult and child ASR. These advancements will enable more accurate and inclusive speech recognition systems for educational, assistive, and interactive speech-oriented applications.

ACKNOWLEDGEMENTS

The research in this paper was partially supported by Ministry of Education, Research, Development and Youth of the Slovak Republic under the research projects KEGA 049TUKE-4/2024 and 041TUKE-4/2025, and by the Slovak Research and Development Agency under the research projects APVV-22-0261 and APVV-22-0414.

References

Babu, A., Wang, Ch., Tjandra, A., Lakhotia, K., Xu, Q., Goyal, N., Singh, K., von Platen, P., Saraf, Y., Pino, J., Baevski, A., Conneau, A., and Auli, M. (2022). XLS-R: Self-supervised cross-lingual speech representation learning at scale. In Proc. of INTERSPEECH 2022, Incheon, Korea, pp. 2278–2282.

Baevski, A., Zhou, H., Mohamed, A., and Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. In Proc. of NISP 2020, Vancouver BC, Canada, pp. 12449–12460.

Barras, C., Geoffrois, E., Wu, Z., and Liberman, M. (2001). Transcriber: Development and use of a tool for assisting speech corpora production. Speech Communications, Special Issue on Speech Annotation and Corpus Tools, 33(1–2), pp. 5–22.

Batliner, A., Blomberg, M., D'Arcy, Sh., Elenius, D., Giuliani, D., Gerosa, M., Hacker, Ch., Russell, M., Steidl, S., and Wong, M. (2005). The PF_STAR children's speech corpus. In Proc. of INTERSPEECH 2005, Lisbon, Portugal.

Bhardwaj, V., Othman, M.T.B., Kukreja, V., Belkhier, Y., Bajaj, M., Goud, B. S., Rehman, A. U., Shafiq, M., and Hamam, H. (2022). Automatic speech recognition (ASR) systems for children: A systematic literature review. Applied Sciences, 12(9), paper 4419.

Claus, F., Rosales, H. G., Petrick, R., Hain, H.-U., and Hoffmann, R. (2013). A survey about databases of children's speech. In Proc. of INTERSPEECH 2013, Lyon, France.

Eskenazi, M., Mostow, J., and Graff, D. (1997). The CMU kids corpus. LDC97S63. Philadelphia: Linguistic Data Consortium.

Georgescu, A.-L., Pappalardo, A., Cucu, H., and Blott, M. (2021). Performance vs. hardware requirements in state-of-the-art automatic speech recognition. EURASIP Journal on Audio, Speech, and Music Processing, 2021(28), pp. 1–30.

Gerosa, M., Giuliani, D., Narayanan, Sh., and Potamianos, A. (2009). A review of ASR technologies for children's speech. In Proc. of WOCCI 2009, Cambridge, MA, USA.

Huber, J. E., and Stathopoulos, E. T. (1999). Formants of children, women, and men: The effects of vocal intensity variation. Journal of Acoustical Society of America, 106(3 Pt 1), pp. 1532–1542.

Lojka, M., Viszlay, P., Staš, J., Hládek, D., and Juhár, J. (2018). Slovak broadcast news speech recognition and transcription system. In: L. Barolli – N. Kryvinska – T. Enokido – M. Takizawa (eds.): Advances in Network-Based Information Systems, LNDECT 22, Springer, Cham, pp. 385–394.

Lu, R., Shahin, M. A., and Ahmed, B. (2022). Improving children's speech recognition by fine-tuning self-supervised adult speech representations. arXiv Preprint. Accessible at: https://arxiv.org/abs/2211.07769.

Patel, T., and Scharenborg, O. (2024). Improving end-to-end models for children's speech recognition. Applied Sciences, 14(6), paper 2353.

Pradhan, S. S., Cole, R. A., and Ward, W. H. (2024). My Science Tutor (MyST) – A large corpus of children's conversational speech. In Proc. of LREC-COLING 2024, Torino, Italia, pp. 12040–12045.

Pleva, M., Ondáš, S., Hládek, D., Juhár, J., and Staš, J. (2019). Building of children speech corpus for improving automatic subtitling services. In Proc. of ROCLING 2019, New Taipei City, Taiwan, pp. 325–333.

Pratap, V., Tjandra, A., Shi, B., Tomasello, P., Babu, A., Kundu, S., Elkahky, A., Ni, Z., Vyas, A., Fazel-Zarandi, M., Baevskyi, A., Adi, Y., Zhang, X., Hsu, W.-N., Conneau, A., and Auli, M. (2024). Scaling speech technology to 1,000+ languages. Journal of Machine Learning Research, 25, pp. 1–52.

Radford A., Kim, J. W., Xu, T., Brockman, G., McLeavy, Ch., and Sutskever, I. (2023). Robust speech recognition via large-scale weak supervision. In Proc. of ICML 2023, Honolulu, Hawai, USA, pp. 28492–28518.

Sanchez, A., Meylan, S. C., Braginsky, M., MacDonald, K. E., Yurovsky, D., and Frank, M. C. (2019). childes-db: A flexible and reproducible interface to the child language data exchange system. Behavior Research Methods, 51, pp. 1928–1941.

Shivakumar, P. G., and Georgiou, P. (2020). Transfer learning from adult to children for speech recognition: Evaluation, analysis, and recommendations. Computer Speech & Language, 63, paper 101077.

Shobaki, K., Hosom, J.-P., and Cole, R. A. (2000). The OGI kids' speech corpus and recognizers. In Proc. of ICSLP 2000, Beijing, China, pp. 1–4.

Sobti, R., Guleria, K., and Kadyan, V. (2024). Comprehensive literature review on children automatic speech recognition system, acoustic linguistic mismatch approaches and challenges. Multimedia Tools and Applications, 83, pp. 81933–81995.

Yeung, G., and Alwan, A. (2018). On the difficulties of automatic speech recognition for kindergarten-aged children. In Proc. of INTERSPEECH 2018, Hyderabad, India, pp. 1661–1665.

NATURAL LANGUAGE PROCESSING

sciendo
DOI 10.2478/jazcas-2025-0021

REDEFINING TERMINOLOGY WORK: THE ROLE OF GLOSSARIES AND TMS IN CUSTOMIZED AI-DRIVEN MACHINE TRANSLATION

JAKUB ABSOLON
Department of British and American Studies, Faculty of Arts, University of Ss. Cyril and Methodius, Trnava, Slovakia (ORCID: 0009-0008-2365-1048)

ABSOLON, Jakub: Redefining Terminology Work: The Role of Glossaries and TMs in Customized AI-driven Machine Translation. Journal of Linguistics, 2025, Vol. 76, No 1, pp. 237 – 246.

Abstract: This study investigates whether traditional terminology work in the customization of machine translation (MT) systems can be effectively replaced by translation memories (TMs) alone. Given the growing reliance on AI-driven translation tools, we evaluated three MT configurations using English–Slovak technical documentation: a baseline (non-customized system), a system customized with both TMs and a glossary, and a system customized with TMs only. Since the text corpora for the given area were sufficient, we used the LLM model to generate additional training data. Results show that TM-only customization can achieve terminology translation accuracy nearly equivalent to setups that include glossaries—particularly when supported by high-quality, domain-specific bilingual data. Nonetheless, glossary-based customization further improves consistency, and terminology errors persist across all systems. This suggests that although automation of translation processes can reduce dependence on traditional terminology building, terminology databases remain essential for ensuring the quality (QA) of the output text. The study offers practical guidance for translators, terminologists, and developers of translation tools by emphasizing the importance of collaboration between automated and human-driven translation processes. It also underscores both the promise and limitations of LLM-generated data for domain adaptation in low-resource language settings.

Keywords: terminology, machine translation, customization, translation memory, glossary, AI, domain adaptation, low-resource languages, quality assurance, post-editing

1 INTRODUCTION

The rapid evolution of artificial intelligence (AI) has transformed machine translation (MT), reshaping traditional translation workflows. The translation process used to be linear and the privilege of translators, human beings. Nowadays, translation increasingly utilizes artificial intelligence, which speeds up and streamlines the process; however, it also introduces the risk of unpredictability in the quality of the final text and the possibility of critical errors in high-risk areas. Consequently, customized MT solutions, leveraging domain-specific adaptations, are emerging as a practical middle-ground to balance automation with quality.

Customizable MT engines allow users to incorporate domain-specific resources, such as translation memories (TMs) and glossaries, potentially challenging the traditional role of terminological work in ensuring translation accuracy and consistency—particularly in technical and scientific contexts.

1.1 Theoretical framework

The theoretical foundation for our research is grounded in two dominant schools of terminology theory:

1. Socioterminology, pioneered by (Gaudin 1993; as cited in Temmerman 2000), positions terminology as an inherently socially situated phenomenon. Terms emerge and evolve within expert communities, reflecting usage variation and social context.
2. Sociocognitive terminology, introduced by Temmerman (2000), emphasizes the cognitive and contextual dimensions of term usage. This approach emphasizes the role of contextual, cognitive, and discursive influences in the creation of meanings and variations of concepts.

Furthermore, terminology developed by Faber (2012), based on a cognitive framework, integrates cognitive semantics and terminology management.

1.2 Research focus

Building on this theoretical grounding, our study examines whether TMs alone can effectively customize MT systems, potentially substituting traditional glossary-based terminology practices. Since glossary creation requires considerable resources and the use of translation memory-based solutions is growing, research on this issue provides important practical and academic insights.

2 RELATED WORK

2.1 Traditional terminology work

Terminology work ensures consistency in technical domains through the collection and management of terms. The dynamics of language are shaped by technological, social, and cultural factors, and therefore require approaches to terminology work that are capable of responding to changing contexts and shifts in the meaning of terms. Translators must handle:

- **Conceptual deviation**: Term meanings may diverge across domains or cultures.
- **False friends**: Lexical homonyms with different meanings in different languages, which increases the risk of misinterpretation in translation.

- **Cultural/Contextual gaps**: MT systems often lack nuance, requiring human input.

This complexity drives a shift toward integrating AI with traditional term workflows.

2.2 AI-enhanced terminology strategies

Modern research suggests that combining AI technologies with traditional terminology workflows can mitigate many of the above limitations:

- **Terminology-aware MT**: Techniques like constrained decoding improve term accuracy (Bogoychev and Chen 2023).
- **WMT 2023 Shared Task**: Term injection during training/inference improved accuracy, though BLEU gains varied (Semenov et al. 2023).
- **Human-AI synergy**: Translators now focus on creativity and QA while machines handle routine tasks (Gao 2022).

The above-mentioned strategies underscore the importance of designing work processes that consider both cognitive mechanisms and technical efficiency.

2.3 LLM-based synthetic terminology training

Large language models (LLMs) such as GPT-4 are increasingly used to augment training data when bilingual corpora are scarce. A study by Moslem et al. (2023) during WMT 2023 demonstrated that synthetic parallel sentences generated by LLMs, followed by fine-tuning and human post-editing, led to notable improvements in terminology translation accuracy, from 37% to over 70% for domain-specific terms (ACL Anthology).

However, the synthetic data often requires rigorous human curation, as LLM outputs may introduce semantic simplifications or hallucinate context.

2.4 Domain-specific terminology: The volcanology case

A study by Harris et al. (2017) explored the translation of volcanological terms across multiple languages, emphasizing that terminological consistency and scientific accuracy are essential in high-risk fields. Their work confirmed that machine translation alone cannot guarantee conceptual clarity or cross-cultural appropriateness without human oversight.

3 MACHINE TRANSLATION (MT) CUSTOMIZATION

Machine translation (MT) customization is a critical area of both applied practice and ongoing research, particularly when aiming to improve translation quality in specialized domains. The process of adapting machine translation typically involves the use of domain-specific resources, primarily translation memories and glossaries, to increase the accuracy and relevance of the output. This section reviews current customization techniques and outlines best practices based on recent empirical findings.

3.1 Customization techniques

3.1.1 Fine-tuning and data selection

Fine-tuning MT models using in-domain bilingual data has been shown to significantly improve both terminology translation accuracy and overall contextual fidelity. In our English–Slovak case study, fine-tuning yielded measurable performance gains. Selecting high-quality training segments—especially with the aid of document classification tools—enables more efficient domain adaptation, often outperforming generic MT systems trained on larger but less relevant datasets.

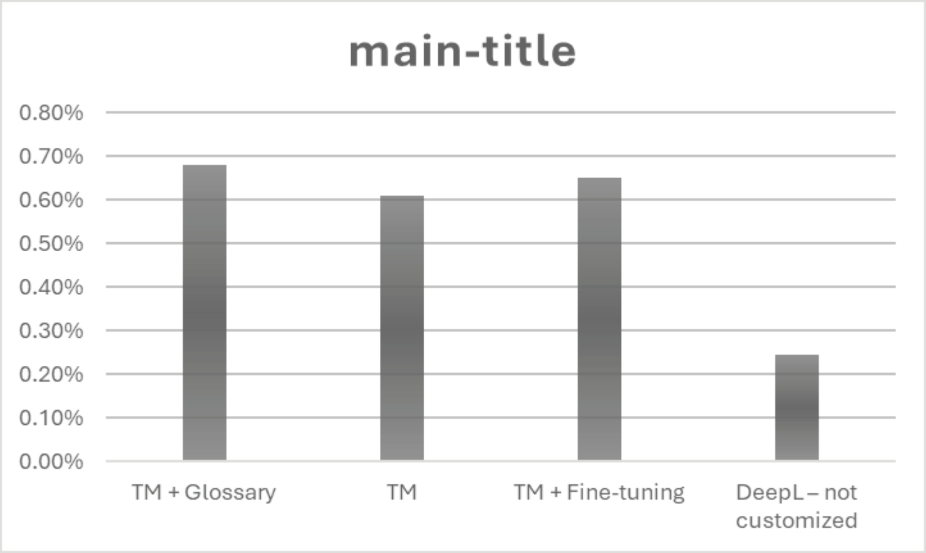


Fig. 1. Terminology translation accuracy

3.1.2 Terminology integration

Integrating user-defined glossaries into MT engines ensures that domain-critical terminology is translated consistently and accurately. The process of adapting machine translation typically involves utilizing specific domain resources, primarily translation memories and glossaries, to enhance the accuracy and relevance of the output text. By defining preferred translation equivalents, glossaries guide MT system output towards consistency and compliance with industry standards.

3.2 Implications and future directions

Integrating TMs and glossaries into MT customization is a proven strategy for enhancing translation quality, particularly in specialized domains. These approaches not only increase accuracy and consistency but also the ability of MT systems to respond to specific user needs. Future research should focus on developing scalable, cost-effective solutions such as AI-supported glossary generation and adaptive MT systems with real-time customization capabilities.

4 METHODOLOGY

4.1 Source text description

The source text used for evaluation is a specialized technical manual for industrial packaging equipment. Although originally authored in English, its lexical patterns and syntactic structures indicate influence from Italian, making it representative of multilingual industrial documentation. The manual is intended for use by technicians and maintenance personnel in manufacturing environments.

It includes detailed operational instructions, safety guidelines, component specifications, and references to EU regulations (e.g. Directive 2006/42/EC). The text is characterized by high terminological density, frequent use of compound noun phrases, imperative forms, and structurally consistent formatting. Key terms include film tensioning system, pre-stretch carriage, photocell sensor, vacuum chamber, and emergency stop function — all of which pose a challenge for accurate machine translation and make the material well-suited for evaluating terminology handling and consistency in customized MT systems.

4.2 Machine translation configurations

According to Akhulkova (2023), the Language Technology Atlas identifies 111 MT solutions that are currently available, with 33 offering customization capabilities. The Intento "State of Machine Translation 2024" report evaluated 52 MT engines and LLMs. Among 28 MT engines, 7 supported both TM and glossary customization, 2 supported glossary-only customization, and 1 supported TM-only customization. Among the LLMs assessed, 9 supported only glossary customization and 15 supported both glossary and TM customization via techniques such as fine-tuning, retrieval-augmented generation (RAG), or prompt engineering.

Despite the broader contextual understanding of LLMs, their tendency to hallucinate factual content made them less suitable for high-precision translation in this study. Therefore, we focused on customizable neural MT (NMT) systems, evaluating three widely used platforms: DeepL, Microsoft Translator, and Google Translate.

For the English–Slovak language pair, Microsoft Custom Translator was selected due to its technical feasibility, affordability, and robust language support. Notably, the platform supports both inference-time glossary integration and weakly supervised fine-tuning using translation memory (TM) data—an approach also mirrored by several top-performing systems in the WMT 2023 Terminology Shared Task (Semenov et al. 2023). In collaboration with ASAP-translation.com, s.r.o., we prepared a domain-specific dataset, which includes a TMX file containing 29,334 bilingual sentence pairs and a glossary of 39 verified English–Slovak term pairs.

4.3 MT system configurations tested

To assess the impact of TMs and glossaries on translation quality, we tested three MT system configurations:

- **Non-Customized MT (Baseline)**: A generic MT system with no domain-specific adaptation.
  → DeepL was chosen for this configuration, based on its empirical performance in the EN–SK language pair.
- **Customized MT with TM + Glossary**: A system trained with both a domain-specific translation memory and a glossary containing the target terminology.
  → Implemented using Microsoft Custom Translator, model MT Custom 1.0.
- **Customized MT with TM Only**: A system trained solely on the translation memory, without an integrated glossary.
  → Implemented using Microsoft Custom Translator, models MT Custom 1.1 and 1.2.

For Model MT Custom 1.2, the dataset was extended to include additional translation units (TUs) containing five selected test terms, ensuring sufficient exposure during fine-tuning. Where authentic parallel data was insufficient, we generated synthetic training data using GPT-4.0 (OpenAI). To investigate the impact of term frequency on translation accuracy, we varied the number of training instances per term as follows:

- pulley – remenica: 25 TUs
- transpallet – paletový vozík: 50 TUs
- transit – posun: 100 TUs
- drawbar – ťahadlo: 100 TUs
- carriage – unášač: 200 TUs

This variation was designed to assess whether increased exposure to specific terms enhances their translation accuracy across various system configurations.

4.4 Evaluation and assessment criteria

The evaluation focused on both terminology-specific performance and overall translation quality. We employed three core evaluation dimensions:

1. **Terminology Handling**: Accuracy of term translation in context, fidelity to the intended technical meaning, and alignment with glossary entries (where applicable).
2. **Terminology Consistency**: Consistent use of terminology throughout the translated text, minimizing synonym variation or inconsistent renderings.
3. **Overall Translation Quality**: General fluency, adequacy, and faithfulness of the translations, assessed via both human and automated methods.

Human evaluation was conducted independently by two professional translators with expertise in technical translation. They assessed the accuracy and consistency of terminology on a subset of translated segments. In addition, the **BLEU (Bilingual Evaluation Understudy)** score was used as an automatic metric to supplement human judgments and facilitate comparison across MT configurations.

5 RESULTS AND DISCUSSIONS

5.1 Terminology handling and consistency

The effectiveness of different MT configurations in handling domain-specific terminology was evaluated based on the accurate rendering of five target terms across 69 segment occurrences. The results are as follows:

- **MT Custom 1.0 (TM + glossary)**: 68.1% accuracy (47/69)
- **MT Custom 1.1 (TM only)**: 60.9% accuracy (42/69)
- **MT Custom 1.2 (TM with fine-tuned synthetic data)**: 65.2% accuracy (45/69)
- **DeepL (non-customized baseline)**: 24.6% accuracy (17/69)

The highest accuracy was achieved using the configuration that incorporated both a domain-specific translation memory and glossary (MT Custom 1.0). Surprisingly, the fine-tuned model (MT Custom 1.2) achieved slightly lower accuracy, despite being supplemented with additional LLM-generated examples. This suggests that while synthetic data may help bridge gaps in terminology coverage, it cannot fully replace curated, human-validated content.

The TM-only configurations still performed reasonably well, confirming that a high-quality translation memory can support robust terminology handling even without a glossary. In contrast, the non-customized DeepL system struggled with specialized terms, reinforcing the need for domain adaptation.

All systems demonstrated vulnerabilities when contextual cues were insufficient, even those with glossary integration. This observation supports the notion that glossaries alone are insufficient for ensuring terminology accuracy and that leveraging full-sentence parallel data remains critical for robust customization.

Inconsistent term usage, even within customized models, further highlights the importance of post-editing and terminological quality assurance (QA). Moreover, the synthetic data generated by LLMs showed a tendency to simplify terminology, necessitating human review for effective integration into training workflows.

5.2 Overall translation quality

BLEU scores provide a supplementary measure of overall translation performance across the four system configurations:

- **MT Custom 1.2 (TM + synthetic fine-tuning)**: 75.18
- **MT Custom 1.0 (TM + glossary)**: 71.99
- **MT Custom 1.1 (TM only)**: 71.05
- **DeepL (baseline)**: 52.69

These results confirm that MT customization—particularly when augmented with fine-tuning on in-domain data—significantly enhances translation quality. While all customized configurations outperformed the baseline, the highest BLEU score was achieved by the system using LLM-generated supplemental training data, suggesting that targeted augmentation can improve general fluency and lexical adequacy, even if terminology fidelity remains a challenge.

Microsoft Custom Translator's fine-tuning mechanism proved effective, especially when trained with adequate domain-specific content. However, the marginal difference between the TM-only and TM+glossary configurations suggests that in some contexts, high-quality TMs alone can achieve near-equivalent performance.

5.3 Implications for terminology work

The results underscore the critical role of high-quality, human-validated data in effective MT customization. While glossaries enhance precision, their creation and maintenance remain resource-intensive. TM-only approaches, particularly when paired with synthetic data augmentation, offer a cost-effective alternative with reasonable performance.

This challenges the traditional view of terminologies as the core carriers of meaning in translation, as noted by Semenov et al. (2023), who argue that such assumptions may be overstated, especially given the comparable performance of systems relying solely on high-quality TMs.

Nevertheless, terminology errors persist, especially in complex technical texts. Consistent, expert-driven terminology work remains essential for both training data quality and post-editing workflows. The increasing availability of AI-driven tools, such as automated term extraction, can help alleviate some of the manual burden. However, these tools require careful human curation to ensure that termbases remain accurate, contextually appropriate, and aligned with evolving domain standards.

Ultimately, scalable and high-quality localization will depend not on replacing traditional terminology work but on transforming it into a curation-centered, collaborative process, with translators, terminologists, and AI systems working in concert.

6 CONCLUSION AND FUTURE WORK

This study examined the interplay between translation memories (TMs), glossary integration, and traditional terminology work in the context of customized machine translation (MT) for the English–Slovak language pair. By evaluating multiple MT configurations, including TM-only customization, TM combined with a glossary, and TM fine-tuned with LLM-generated data, we explored the extent to which TMs can replace or complement conventional terminological resources in domain-specific translation workflows.

Our findings indicate that systems combining TMs with glossaries achieved the highest terminology translation accuracy. However, TM-only configurations delivered a comparable performance, particularly when enhanced with synthetic training data from large language models (LLMs). This suggests that well-constructed translation memories may, in some cases, reduce the need for exhaustive glossary compilation—especially in cost-sensitive or time-constrained settings.

Despite these gains, terminology inconsistencies persisted across all configurations. General-purpose MT systems like DeepL performed poorly with specialized terms, underscoring the importance of domain adaptation. Interestingly, the MT engine often prioritized TM-derived patterns over glossary entries, emphasizing the continued value of validated and well-curated termbases, particularly for post-editing and quality assurance (QA) processes.

Future research should further investigate how factors such as TM quality, term frequency, and domain variability influence terminology handling across language pairs. LLM-generated bilingual data for fine-tuning appears promising but requires rigorous human validation due to risks of semantic simplification and context loss. Additionally, the integration of AI-based term extraction and dynamic glossary adaptation during translation represents a key area for innovation.

As the scale and speed of localization increase, the field is gradually shifting from terminology creation to terminology curation. Supporting this shift will require deeper collaboration between human experts and machine learning systems, ensuring that automation enhances, rather than compromises, translation quality.

ACKNOWLEDGEMENTS

The author would like to thank ASAP-translation.com, s.r.o., for their collaboration and for providing domain-specific translation data and technical infrastructure essential to the successful execution of this research.

References

Akhulkova, Y. (2023). The 2023 Nimdzi Language Technology Atlas. Nimdzi Insights. Accessible at: https://www.nimdzi.com/language-technology-atlas/.

Buysschaert, J., and Kovács, L. (2017). Challenges encountered during the compilation of a multilingual termbase in the domain of communication. Terminology, 23(1), pp. 1–18. Accessible at: https://doi.org/10.18460/ANY.2017.1.001.

Faber, P. (ed.). (2012). A Cognitive Linguistics View of Terminology and Specialized Language (Applications of Cognitive Linguistics, Vol. 20). De Gruyter Mouton. Accessible at: https://doi.org/10.1515/9783110277203.

Forcada, M. L. (2017). Making sense of neural machine translation. Translation Spaces, 6(2), pp. 291–309. Accessible at: https://doi.org/10.1075/ts.6.2.04for.

Gao, J. (2022). The impact of digital technologies on the structure of translation activities. Litera, 10, pp. 72–86. Accessible at: https://doi.org/10.25136/2409-8698.2022.10.39067.

Harris, A. J. L., Belousov, A., Calvari, S., Delgado-Granados, H., Hort, M., Koga, K. T., Wulan Mei, E. T., Harijoko, A., Pacheco, J., Prival, J.-M., Solana, C., Þórðarson, Þ., Thouret, J.-C., and van Wyk de Vries, B. (2017). Translations of volcanological terms: Cross-cultural standards for teaching, communication, and reporting. Bulletin of Volcanology, 79(7), 57 p. Accessible at: https://doi.org/10.1007/S00445-017-1141-9.

Intento. (2024). The State of Machine Translation 2024: Independent Evaluation of MT Engines and LLMs. Accessible at: https://www.inten.to.

Li, B. (2023). Conceptual deviation in terminology translation. Terminology. Accessible at: https://doi.org/10.1075/term.00073.li.

Microsoft. (2024). Microsoft Custom Translator Documentation. Microsoft Learn. Accessible at: https://learn.microsoft.com/en-us/azure/ai-services/translator/custom-translator/overview [30/03/2025].

Moslem, Y., Romani, G., Molaei, M., Haque, R., Kelleher, J. D., and Way, A. (2023). Domain Terminology Integration into Machine Translation: Leveraging Large Language Models. In: P. Koehn – B. Haddow – T. Kocmi – C. Monz (eds.): Proceedings of the Eighth Conference on Machine Translation, pp. 902–911. Association for Computational Linguistics. Accessible at: https://doi.org/10.18653/v1/2023.wmt-1.82.

Semenov, K., Zouhar, V., Kocmi, T., Zhang, D., Zhou, W., and Jiang, Y. E. (2023). Findings of the WMT 2023 Shared Task on Machine Translation with Terminologies. In: P. Koehn – B. Haddow – T. Kocmi – C. Monz (eds.): Proceedings of the Eighth Conference on Machine Translation, pp. 663–671. Association for Computational Linguistics. Accessible at: https://doi.org/10.18653/v1/2023.wmt-1.54.

Temmerman, R. (2000). Towards New Ways of Terminology Description: The Sociocognitive Approach. Terminology and Lexicography Research and Practice, Vol. 3. Amsterdam/Philadelphia: John Benjamins Publishing Company.

sciendo
DOI 10.2478/jazcas-2025-0022

FINANCIAL QUESTION-ANSWERING DATASET FOR SLOVAK LANGUAGE MODEL EVALUATION

DANIEL HLÁDEK[1] – KRISTIÁN SOPKOVIČ[2] – JÁN STAŠ[3] – ZUZANA SOKOLOVÁ[4] – MATÚŠ PLEVA[5]
[1]Department of Electronics and Multimedia Communications, Faculty of Electrical Engineering and Informatics, Technical University, Košice, Slovakia
(ORCID: 0000-0003-1148-3194)
[2]Department of Electronics and Multimedia Communications, Faculty of Electrical Engineering and Informatics, Technical University, Košice, Slovakia
(ORCID: 0009-0007-0835-3491)
[3]Department of Electronics and Multimedia Communications, Faculty of Electrical Engineering and Informatics, Technical University, Košice, Slovakia
(ORCID: 0000-0001-7403-0235)
[4]Department of Electronics and Multimedia Communications, Faculty of Electrical Engineering and Informatics, Technical University, Košice, Slovakia
(ORCID: 0000-0002-2337-8749)
[5]Department of Electronics and Multimedia Communications, Faculty of Electrical Engineering and Informatics, Technical University, Košice, Slovakia
(ORCID: 0000-0003-4380-0801)

HLÁDEK, Daniel – SOPKOVIČ, Kristián – STAŠ, Ján – SOKOLOVÁ, Zuzana – PLEVA, Matúš: Financial Question-answering Dataset for Slovak Language Model Evaluation. Journal of Linguistics, 2025, Vol. 76, No 1, pp. 247 – 257.

Abstract: The limited availability of language resources for Slovak presents a significant challenge for the development and evaluation of language models. In this paper, we introduce a multiple-choice question-answering dataset specifically designed for the financial domain in Slovak. The dataset contains 1,334 questions, each with one correct answer and four incorrect ones. It is systematically organized by topic and difficulty level to facilitate structured evaluation. Using this dataset, we assess the performance of several Slovak generative language models and compare their results against a general question-answering dataset to analyze domain-specific model capabilities. The best-performing model is a monolingual Slovak model. Furthermore, the observed performance differences between financial-domain and general question-answering tasks suggest that domain-specific language modeling requires further research.

Keywords: question answering, financial domain, large language model, evaluation, Slovak language resource

1 INTRODUCTION

The development of high-performing natural language processing (NLP) models heavily depends on the availability of high-quality datasets and evaluation benchmarks. While significant progress has been made in creating language resources for widely spoken languages such as English, low-resource languages, including Slovak, remain under-represented. This lack of resources poses challenges in training, fine-tuning, and evaluating Slovak generative language models, particularly for specialized domains like finance and law. Without domain-specific benchmarks, it is difficult to measure model performance accurately and ensure its practical applicability.

Existing Slovak language models are often evaluated on machine-translated or general-purpose datasets that do not sufficiently capture the complexity of real-world applications. Financial and legal texts, for example, involve specialized terminology and structured reasoning, which may not be well-represented in commonly available corpora.

To address these challenges, we introduce a question-answering dataset specifically designed for the legal and financial domain in the Slovak language. The dataset is structured according to topic and difficulty level, allowing for targeted assessment and benchmarking of language models. By providing a structured and domain-specific evaluation resource, we enable more precise measurement of model capabilities and facilitate further advancements in Slovak NLP. The dataset contains 1,334 questions from the financial domain. Each question has 5 possible answers; exactly one is correct. The language model can calculate the probability of each question-answer pair and select the best. This method of evaluation does not require specific fine-tuning; thus it is useful for the assessment of foundation models, trained only on unannotated data. Using the *lm-evaluation-harness* framework (Sutawika 2025), we evaluate multiple Slovak generative language models on our dataset and compare their results with those obtained on a general fact question-answering dataset.

2 STATE OF THE ART

There are multiple surveys of language model evaluation. The recent "Survey on evaluation of Large Language Models" (LLMs) (Chang 2024) claims that, as LLMs are becoming larger with more emergent abilities, existing evaluation protocols may not be enough to evaluate their capabilities and potential risks.

Guo (2023) categorizes the evaluation of LLMs into three major groups:

1. knowledge and capability evaluation,
2. alignment evaluation,
3. and safety evaluation.

In addition, it collates a compendium of evaluations pertaining to LLM performance in specialized domains, and discusses the construction of comprehensive evaluation platforms that cover LLM evaluations on capabilities, alignment, safety, and applicability.

2.1 General language model benchmarks

The Holistic Evaluation of Language Models (HELM) (Liang 2023) is a comprehensive framework developed to enhance the transparency and understanding of large language models (LLMs). HELM addresses the vast array of potential use cases and evaluation metrics by establishing a taxonomy that identifies and categorizes these scenarios and desiderata. By acknowledging existing gaps and under-represented areas, HELM provides a more inclusive and thorough assessment of LLMs. It evaluates LLMs across seven key metrics: accuracy, calibration, robustness, fairness, bias, toxicity, and efficiency. In a large-scale evaluation, HELM assessed 30 prominent LLMs across 42 scenarios, half of which were novel to mainstream evaluation.

The best language model in a general evaluation benchmark might not be the best Slovak language model. Lai et al. (2023) introduced Okapi, a system with instruction-tuned LLMs based on reinforcement learning with human feedback (RLHF) in 26 diverse languages to facilitate experiments and the development of future multilingual research. It also created benchmark datasets to enable the evaluation of generative LLMs in multiple languages including Slovak. The Okapi evaluation benchmark machine-translated is widely used English datasets. It leverages four datasets in the HuggingFace Open LLM Leaderboard, i.e., AI2 Reasoning Challenge (ARC) (Clark 2018), HellaSwag (Zellers 2019), MMLU (Hendrycks 2020), TruthfulQA (Lin 2022) to evaluate multilingual fine-tuned LLMs.

The *lm-evaluation-harness* framework (Sutawika 2025) is a widely used open-source tool designed to systematically evaluate language models across various tasks. It provides a standardized methodology for assessing model performance by enabling direct comparisons across different architectures, datasets, and evaluation metrics. The framework supports a diverse range of question-answering, reasoning, and text generation tasks, making it particularly useful for benchmarking generative language models. By using *lm-evaluation-harness*, researchers can ensure consistency in evaluation, reducing biases introduced by ad-hoc testing procedures. Its compatibility with multiple pre-trained models allows for seamless integration and fair performance assessment across various NLP applications.

2.2 Financial language model benchmarks

General language model benchmarks often fail to accurately reflect a model's performance within specialized domains, such as finance. Besides that, they are often specific to the English-speaking cultural domain. To address this limitation, several financial domain-specific benchmarks have been developed. For instance, the Financial Language Understanding Evaluation (FLUE) (Shah 2022) offers a comprehensive suite of tasks tailored to financial contexts, including sentiment analysis and named entity recognition. Labrak (2023) evaluates four state-of-the-art instruction-tuned large language models on a set of 13 real-world clinical and biomedical natural NLP tasks in English, such as named-entity recognition, question-answering or relation extraction.

Z. Guo et al. (2023) present FinLMEval, a framework for Financial Language Model Evaluation, comprising nine datasets designed to evaluate the performance of language models in English. X. Guo et al. (2023) present FinEval – a benchmark to evaluate Chinese language models in the financial domain. The dataset contains 8,351 multiple choice questions categorized into four different key areas: Financial Academic Knowle.g. Financial Industry Knowle.g. Financial Security Knowle.g. and Financial Agent.

2.3 Language model benchmarks with Slovak support

There are only a couple of publicly available monolingual Slovak datasets suitable for fine-tuning or evaluation of a generative language model.

GEST (Pikuliak 2024) is a manually created dataset designed to measure gender-stereotypical reasoning in language models and machine translation systems. GEST contains samples for 16 gender stereotypes about men and women in the English language and 9 Slavic languages, including Slovak.

The largest manually annotated set of questions and answers from Wikipedia in Slovak is SkQuAD (Hládek 2023). It consists of more than 91k factual questions and answers from various fields. Each question has an answer marked in the corresponding paragraph. It also contains negative examples in the form of "unanswered questions" and "plausible answers".

The same authors automatically translated the original SQUAD (Rajpurkar 2018) into Slovak (Staš 2023). The dataset was automatically translated from the original English SQuAD v2.0 using the Marian neural machine translation together with the Helsinki-NLP Opus English-Slovak model.

SlovakSum is a Slovak news summarization dataset consisting of over 200,000 news articles with titles and short abstracts obtained from multiple Slovak newspapers. The abstractive approach, including mBART and mT5 models, was used to evaluate various baselines in by Ondrejová (2024).

Annotation of the named entities is one of the tasks common for generative model evaluation, for example in mT5 family of models (Xue 2021). WikiGoldSK (Šuba 2023) is a dataset consisting of 10,000 manually annotated named entities in over 400 Wikipedia pages.

2.4 Generative models with Slovak support

Mistral (Jiang 2023) is a series of advanced language models developed by Mistral AI, designed to handle complex multilingual tasks with robust reasoning capabilities. The flagship model, Mistral Large, is fluent in multiple languages, including Slovak. With a context window of 32,000 tokens, it can accurately recall information from extensive documents, facilitating precise and contextually relevant text generation, advanced reasoning and agentic capabilities. Slovak Mistral (mistral-sk-7b) is a Slovak language model created by full fine-tuning of the Mistral-7B-v0.1 large language model (Jiang 2023) with data from the Araneum Slovacum VII Maximum web corpus (Benko 2024). The model was developed in collaboration with the Technical University of Košice and the Slovak Academy of Sciences. Presently, the model is not fine-tuned to follow instructions.

LLaMA (Large Language Model Meta AI) is a series of open-source language models developed by Meta, designed to advance natural language understanding and generation. These models are pre-trained on a diverse range of languages, enabling them to perform effectively across multiple linguistic contexts (Grattafiori 2024).

Qwen (Yang 2024), developed by Alibaba, is another prominent series of language models emphasizing multilingual proficiency. These models have demonstrated strong performance in multilingual tasks, making them suitable for applications requiring cross-lingual understanding and generation.

RWKV (Peng 2023) is a language model architecture different from the classic transformer encoder-only models. It combines the strengths of recurrent neural networks (RNNs) and transformers, aiming to offer efficient training and inference capabilities. Its design inherently supports sequential data processing, which can be advantageous for modeling languages with complex syntactic structures.

3 THE PROPOSED DATASET

The proposed dataset consists of 1,334 questions from the financial advisor certification of the Slovak National Bank, according to § 22 Act. no. 186/2009 Z. z., valid until 5.8.2023. The questions are published in XML and PDF format on the website of the Slovak National Bank (NBS). We parsed the test, extracted meta-information about the difficulty level and the category. The test evaluates knowledge of the applicant in the areas in Tab. 1. Tab. 2 displays the detailed table of contents of the dataset within formation about the question category, difficulty level and identification number; the example questions and answers are in Tab. 3 and Tab. 4.

Acronym	**Name**	**Translation**
VSE	*Všeobecná časť*	General questions
PaZ	*Sektor poistenia alebo zaistenia*	Insurance or reinsurance
KT	*Sektor kapitálového trhu*	Capital market
Vkl	*Sektor prijímania vkladov*	Deposits
Uv	*Sektor poskytovania úverov*	Credit granting
DDS	*Sektor doplnkového dôchodkového sporenia*	Supplementary pension
DSS	*Sektor starobného dôchodkového sporenia*	Retirement pension savings

Tab. 1. The categories of the dataset

Acronym	**Topic**	**Difficulty Level**	**Last ID**	**Count**
VSE	General questions	1	236	236
PaZ	Insurance or reinsurance	2	373	137
PaZ	Insurance or reinsurance	3	438	65
KT	Capital market	2	606	168
KT	Capital market	3	646	40
Vkl	Deposits	2	792	146
Vkl	Deposits	3	850	58
Uv	Credit granting	2	1005	135
Uv	Credit granting	3	1032	27
DDS	Supplementary pension	2	1171	139
DDS	Supplementary pension	3	1228	57
SDS	Retirement pension savings	2	1309	81
SDS	Retirement pension savings	3	1334	25

Tab. 2. Detailed table of contents of the dataset

Question	*Blízkou osobou v priamom rade je:*	A close person in the direct line is:
A	*Bratranec*	Cousin
B	*Otcov brat*	Father’s brother
C	*Druh-družka*	Spouse
D	***Syn***	**Son**
E	*Neter*	Niece

Tab. 3. Example general question and answers (The correct answer is D.)

Question	*Národná evidencia vozidiel je:*	The National Vehicle Registry is:
A	*Evidencia všetkých poistených vozidiel v poistnom kmeni poisťovne vedená Národnou bankou Slovenska*	A registry of all insured vehicles in the insurance portfolio of an insurance company maintained by the National Bank of Slovakia
B	*Evidencia všetkých vozidiel, ktoré predajca predal v kalendárnom roku vedená Slovenskou obchodnou inšpekciou*	A registry of all vehicles sold by a seller in a calendar year maintained by the Slovak Trade Inspection
C	***Informačný systém o motorových vozidlách evidovaných v Slovenskej republike evidovaný Ministerstvom vnútra Slovenskej republiky***	**An information system on motor vehicles registered in the Slovak Republic maintained by the Ministry of the Interior of the Slovak Republic**
D	*Evidencia vozidiel ktoré sú v premávke na pozemných komunikáciách v Slovenskej republike vedený Ministerstvom vnútra Slovenskej republiky*	A registry of vehicles in traffic on land roads in the Slovak Republic maintained by the Ministry of the Interior of the Slovak Republic

E	*Elektronický informačný systém o vlastníkoch motorových vozidiel v Slovenskej republike ktorý spravuje Slovenská kancelária poisťovateľov.*	An electronic information system on motor vehicle owners in the Slovak Republic administered by the Slovak Insurers' Office.

Tab. 4. Example question and answers from the insurance category (The correct answer is C.)

4 EXPERIMENTS

In this study, we evaluate the performance of widely used LLMs across two distinct datasets. To ensure comparability and reproducibility, we select open-source models with approximately 7 billion parameters. The foundation models are trained solely for next-token prediction and are not further adapted to understand instructions. In contrast, instruction fine-tuned models are evaluated to assess the impact of fine-tuning on task performance. The selected generative models are used as-is, without additional fine-tuning.

The primary research questions in this study are:

1. Does performance on the financial-domain dataset correlate with that on the SKQuad dataset?
2. Does instruction fine-tuning improve performance in both tasks?

We investigate the extent to which instruction fine-tuning enhances both multiple-choice and generative question-answering tasks, providing insights into its effectiveness across different evaluation settings.

4.1 Evaluation metrics

The models are tested on the presented financial multiple-choice dataset, as well as on a general open-domain question-answering dataset, SKQuad. The financial dataset consists of manually curated multiple-choice questions where the model is tasked with selecting the most probable answer from a set of options. In contrast, the SKQuad dataset follows a generative question-answering paradigm, where the model generates an answer after reading a provided context and question. The generated response is then compared to the expected answer to evaluate accuracy.

To measure model performance, we employ different evaluation metrics tailored to each dataset type. For the multiple-choice financial dataaset, we compute normalized accuracy to account for the length of the possible answer. The evaluation system takes the question and possible answer together and calculates the probability of an answer.

The answer with the highest probability is chosen as the generated answer. Furthermore, the answer probability is divided by the number of its words to mitigate too long answers.

In the SKQuad dataset, performance is measured using standard text similarity metrics such as F1-score, ensuring a fair comparison between generated and expected answers. The question is used as a prompt and the language model generates the answer. The generated and expected answer can consist of several words. The F1 metric calculates the overlap between the generated and expected answers. This metric is considered the standard for this dataset.

4.2 Evaluation results

Results of the evaluation of the foundation models are presented in Tab. 5; the instruction models in Tab. 6. According to the experiments, the correlation between performance on the financial-domain dataset and the SKQuad dataset appears weak. While some models, such as Gemma 7B, maintain relatively strong performance across both datasets, others, such as Slovak Mistral 7B, achieve high accuracy in the financial domain but comparatively lower scores in SKQuad. This fact shows that the proposed financial dataset evaluates different abilities of the language model rather than the database of general facts.

Instruction fine-tuning of multilingual LLMs does not show a clear and consistent improvement across both tasks. These findings suggest that instruction fine-tuning has variable effects depending on the model architecture and task, highlighting the need for task-specific tuning strategies.

The best model for answering questions from the financial domain is fine-tuned with a large corpus of the Slovak web data. Its normalized accuracy is 46.6, which is much better than pure random selection, but the model still answers more than half of the questions incorrectly. Taking the current rules into the account, the best Slovak language model still can not become a certified financial advisor. For that, we would need a better model and more data for the model fine-tuning.

Dataset and metric	Slovak Financial Normalized Acc	SKQuad F1
Gemma 7B	40.70	42.52
Qwen 2.5 7B	33.58	48.62
LLama 3.2 3B	34.63	38.40
Slovak Mistral 7B	46.62	41.01
Mistral 7B 0.3	33.80	40.27

Tab. 5. Foundation models evaluation

Dataset and metric	Slovak Financial Normalized Acc	SKQuad F1
Gemma 7B	32.95	48.76
Mistral 7B 0.3	34.63	45.09
Qwen 2.5 7B	34.63	35.57
RWKV-6-finch-7B	39.80	25.07

Tab. 6. Instruct models evaluation

ACKNOWLEDGEMENTS

The research was partially supported by the Ministry of Education, Development, Research, and Youth of the Slovak Republic under the projects KEGA 049TUKE-4/2024, by the Slovak Research and Development Agency under the projects APVV-22-0414 & APVV-22-0261, as well as by the Erasmus+ MetaCog 2024-1-FI01-KA220-HED-000247530 project.

References

Benko, V. (2024). The Aranea Corpora Family: Ten+ Years of Processing Web-Crawled Data. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 15048 LNAI, pp. 55–70. Accessible at: https://doi.org/10.1007/978-3-031-70563-2_5/TABLES/4.

Chang, Y., Wang, X. U., Yi, X., Wang, Y., Ye, W., Yu, P. S., Chang, Y., et al. (2024). A Survey on Evaluation of Large Language Models. Journal of the ACM, 37(3), 39 p. Accessible at https://doi.org/10.1145/3641289.

Clark, P., Cowhey, I., Etzioni, O., Khot, T., Sabharwal, A., Schoenick, C., and Tafjord, O. (2018). Think you have Solved Question Answering? Try ARC, the AI2 Reasoning Challenge. Accessible at: https://arxiv.org/abs/1803.05457v1.

Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., Yang, A., Fan, A., Goyal, A., Hartshorn, A., Yang, A., Mitra, A., Sravankumar, A., Korenev, A., Hinsvark, A., … and Ma, Z. (2024). The Llama 3 Herd of Models. Accessible at: https://arxiv.org/abs/2407.21783v3.

Guo, X., Xia, H., Liu, Z., Cao, H., Yang, Z., Liu, Z., Wang, S., Niu, J., Wang, C., Wang, Y., Liang, X., Huang, X., Zhu, B., Wei, Z., Chen, Y., Shen, W., and Zhang, L. (2023). FinEval: A Chinese Financial Domain Knowledge Evaluation Benchmark for Large Language Models. Accessible at: https://arxiv.org/abs/2308.09975v2.

Guo, Y., Xu, Z., and Yang, Y. (2023). Is ChatGPT a Financial Expert? Evaluating Language Models on Financial Natural Language Processing. Accessible at: https://arxiv.org/abs/2310.12664v1.

Guo, Z., Jin, R., Liu, C., Huang, Y., Shi, D., Supryadi, Yu, L., Liu, Y., Li, J., Xiong, B., and Xiong, D. (2023). Evaluating Large Language Models: A Comprehensive Survey. Accessible at: https://arxiv.org/abs/2310.19736v3.

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., and Steinhardt, J. (2020). Measuring Massive Multitask Language Understanding. ICLR 2021 – 9th International Conference on Learning Representations. Accessible at: https://arxiv.org/abs/2009.03300v3.

Hládek, D., Staš, J., Juhár, J., and Koctúr, T. (2023). Slovak Dataset for Multilingual Question Answering. IEEE Access, 11, pp. 32869–32881. Accessible at: https://doi.org/10.1109/ACCESS.2023.3262308.

Staš J., Hládek, D., and Koctúr, T. (2023). Slovak Question Answering Dataset Based on the Machine Translation of the SQuAD v2.0. Jazykovedný Časopis, 74 (1), pp. 381–390.

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., Casas, D. de las, Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L. R., Lachaux, M.-A., Stock, P., Scao, T. le, Lavril, T., Wang, T., Lacroix, T., and Sayed, W. el. (2023). Mistral 7B. Accessible at: https://arxiv.org/abs/2310.06825v1.

Labrak, Y., Rouvier, M., and Dufour, R. (2023). A Zero-shot and Few-shot Study of Instruction-Finetuned Large Language Models Applied to Clinical and Biomedical Tasks. 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC-COLING 2024 – Main Conference Proceedings, pp. 2049–2066. Accessible at: https://arxiv.org/abs/2307.12114v3.

Lai, V. D., van Nguyen, C., Ngo, N. T., Nguyen, T., Dernoncourt, F., Rossi, R. A., and Nguyen, T. H. (2023). Okapi: Instruction-tuned Large Language Models in Multiple Languages with Reinforcement Learning from Human Feedback. EMNLP 2023–2023 Conference on Empirical Methods in Natural Language Processing, Proceedings of the System Demonstrations, pp. 318–327. Accessible at: https://doi.org/10.18653/v1/2023.emnlp-demo.28.

Liang, P., Bommasani, R., Lee, T., Tsipras, D., Soylu, D., Yasunaga, M., Zhang, Y., Narayanan, D., Wu, Y., Kumar, A., New-Man, B., Yuan, B., Yan, B., Zhang, C., Cosgrove, C., Manning, C. D., Ré, C., Acosta-Navas, D., Hudson, D. A., … and Koreeda, Y. (2023). Holistic Evaluation of Language Models. Accessible at: https://doi.org/10.48550/arXiv.2211.09110.

Lin, S., Hilton, J., and Evans, O. (2022). TruthfulQA: Measuring How Models Mimic Human Falsehoods. Proceedings of the Annual Meeting of the Association for Computational Linguistics, 1, pp. 3214–3252. Accessible at: https://doi.org/10.18653/V1/2022.ACL-LONG.229.

NBS National Bank of Slovakia. Accessible at: https://regfap.nbs.sk/static/otazky/otazky-2023-08-05.pdf.

Ondrejová, V., and Šuppa, M. (2024). SlovakSum: A Large Scale Slovak Summarization Dataset, pp. 14916–14922. Accessible at: https://aclanthology.org/2024.lrec-main.1298/.

Open LLM Leaderboard – a Hugging Face Space by open-llm-leaderboard. (n.d.). Accessible at: https://huggingface.co/spaces/open-llm-leaderboard/open_llm_leaderboard#/quote [24/02/2025].

Pikuliak, M., Hrčková, A., Oreško, S., and Šimko, M. (2023). Women Are Beautiful, Men Are Leaders: Gender Stereotypes in Machine Translation and Language Modeling. Accessible at: https://arxiv.org/abs/2311.18711v3.

Peng, B., Alcaide, E., Anthony, Q., Albalak, A., Arcadinho, S., Biderman, S., Cao, H., Cheng, X., Chung, M., Du, X., Grella, M., Kranthi Kiran, G. v., He, X., Hou, H., Lin, J., Kazienko, P., Kocon, J., Kong, J., Koptyra, B., … and Zhu, R. J. (2023). RWKV: Reinventing RNNs for the Transformer Era. Findings of the Association for Computational Linguistics: EMNLP 2023, pp. 14048–14077. Accessible at: https://doi.org/10.18653/v1/2023.findings-emnlp.936.

Rajpurkar, P., Jia, R., and Liang, P. (2018). Know what you don't know: Unanswerable questions for SQuAD. ACL 2018 – 56th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers), 2, pp. 784–789. Accessible at: https://doi.org/10.18653/v1/p18-2124.

Shah, R. S., Chawla, K., Eidnani, D., Shah, A., Du, W., Chava, S., Raman, N., Smiley, C., Chen, J., and Yang, D. (2022). When FLUE Meets FLANG: Benchmarks and Large

Pretrained Language Model for Financial Domain. Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, pp. 2322–2335. Accessible at: https://doi.org/10.18653/V1/2022.EMNLP-MAIN.148.

Šuba, D., Šuppa, M., Kubík, J., Hamerlik, E., and Takáč, M. (2023). WikiGoldSK: Annotated Dataset, Baselines and Few-Shot Learning Experiments for Slovak Named Entity Recognition. Accessible at: https://arxiv.org/abs/2304.04026v1.

Sutawika L, Schoelkopf H. , Gao L, Abbasi B. , Biderman S., Tow J. et al. (2025). 'Eleutherai/lm-evaluation-harness: V0.4.8'. Zenodo (March 5, 2025). Accessible at: https://doi.org/10.5281/zenodo.14970487.

Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., Barua, A., and Raffel, C. (2020). mT5: A massively multilingual pre-trained text-to-text transformer. NAACL-HLT 2021 – 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference, pp. 483–498. Accessible at: https://doi.org/10.18653/v1/2021.naacl-main.41.

Yang, A., Yang, B., Hui, B., Zheng, B., Yu, B., Zhou, C., Li, C., Li, C., Liu, D., Huang, F., Dong, G., Wei, H., Lin, H., Tang, J., Wang, J., Yang, J., Tu, J., Zhang, J., Ma, J., … and Group, A. (2024). Qwen2 Technical Report. Accessible at: https://arxiv.org/abs/2407.10671v4.

Zellers, R., Holtzman, A., Bisk, Y., Farhadi, A., and Choi, Y. (2019). HellaSwag: Can a Machine Really Finish Your Sentence? ACL 2019 – 57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference, pp. 4791–4800. Accessible at: https://doi.org/10.18653/V1/P19-1472.

sciendo
DOI 10.2478/jazcas-2025-0023

AI AND THE TRANSLATION OF IDIOMS – CHALLENGES, SUCCESS, AND A CORPUS PERSPECTIVE

FILIP KALAŠ[1] – PAVOL LIPTÁK[2]
[1]Department of Linguistics and Translation, Faculty of Applied Languages, University of Economics and Business, Bratislava, Slovakia (ORCID: 0009-0009-1802-1700)
[2]Department of Marketing, Faculty of Commerce, University of Economics and Business, Bratislava, Slovakia (ORCID: 0000-0003-2554-9465)

KALAŠ, Filip – LIPTÁK, Pavol: AI and the Translation of Idioms – Challenges, Success, and a Corpus Perspective. Journal of Linguistics, 2025, Vol. 76, No 1, pp. 258 – 267.

Abstract: This study explores the capabilities of artificial intelligence in translating English idioms into Slovak, with and without contextual information. Using a dataset of 100 idioms and evaluating AI-generated translations against a validated bilingual dictionary of idioms, both qualitative and statistical analyses were employed. The results show an unexpectedly high accuracy in context-free translations, while context occasionally led to deterioration. McNemar's test and a t-test confirmed a statistically significant shift in performance. The study highlights key advantages and limitations of AI, suggesting further research into reverse and cross-linguistic translation as well as employment of corpus-based methods.

Keywords: AI translation, idiom translation, phraseology, machine translation evaluation, corpus linguistics

1 INTRODUCTORY REMARKS

The interpreting profession has undergone unprecedented transformation in the last few years, due to the impact of the COVID-19 pandemic and the introduction of generative artificial intelligence (Wang and Fantinuoli 2024). The pandemic triggered the widespread adoption of remote communication technologies, and remote interpreting became a mandatory practice. The crisis notwithstanding, the continued popularity of virtual and hybrid events has sustained the demand for such tools. Dong et al. (2019) underline that interpreting industry has been revolutionized by NLP in facilitating automatic text-based operations. These advancements have reshaped the interpreting industry, inspired by innovations among interpreters, software developers, and researchers (Rodriguez 2024).

Artificial intelligence (AI) has played a foundational role and been an impactful contribution towards the field of linguistics, particularly in translation. The progression and utilization of many AI-instrumented software, as Google Translate, DeepL, OpenAI, Mistral AI, Trados Studio etc., have helped streamline translation across languages to a significantly greater extent. This has contributed towards real-world scenarios and scholarly pursuits within scientific works (Lund 2023). These advancements have not only improved machine translation to be more efficient and accurate but have also enabled the processing of complex linguistic structures, such as idiomatic expressions. Additionally, AI-powered translation software keeps evolving, with the integration of deep learning and large language models to improve contextual understanding and guarantee translation accuracy across different languages.

According to Rodriguez (2024, p. 118), phraseology serves as a fundamental intraparameter that ensures the proper transfer of the source language and its respective specialist terminology. For this purpose, whether or not AI is able to properly process and handle domain-specific phraseology is among the major parameters for evaluating AI-based translation vis-à-vis human interpreters. Yet, AI systems use pre-trained models and large databases to identify and generate fixed phrases, technical vocabulary, or idiomatic expressions in a specific context.

1.1 Research aims

The objective of this study is to examine the impact of context on the accuracy of AI-produced idiom translation. Specifically, it investigates whether context at the sentence level leads to more accurate and idiomatically better Slovak translations of English idioms. The following working hypotheses were postulated:

H_0: The presence of context does not significantly affect the quality of AI-generated idiom translation from English to Slovak.

H_1: The presence of context improves the quality of AI-generated idiom translation from English to Slovak.

The study combines quantitative statistical analysis with qualitative linguistic evaluation to test these hypotheses.

1.2 Methodology

This quantitative-qualitative analysis was conducted on a dataset of randomly selected 100 English idioms. To evaluate the accuracy of translations produced by AI (more specifically subscripted GPT-4o), a reliable source of Slovak equivalents was required. For this purpose, the *Prekladový anglicko-slovenský frazeologický slovník* (Kvetko 2014) was selected. This bilingual dictionary contains approximately 8,000 English idioms, accompanied by around 16,000 Slovak equivalents. A key criterion for its selections was the inclusion of real-context examples for each entry.

Given the stylistic variation typical of phraseological units, only stylistically neutral idioms were included in the analysis. The stylistic characteristics are explicitly indicated in each entry in the dictionary.

From both morphological and syntactic perspectives, phraseological units can vary significantly. To ensure representativeness and to focus on forms most commonly occurring in spoken language, the dataset was limited to idioms proper – divided equally into 50 verbal phrases and 50 nominal phrases. Sentence-like phraseological units (e.g. proverbs, sayings), similes, and binomials were excluded.

The idioms were compiled in an Excel spreadsheet structured into the following columns: *Idiom without context, Translated by AI, Human evaluation, Idiom in context, Translated by AI, Human evaluation*, and *Improvement*. The meaning of the individual column titles is largely self-explanatory. However, some clarification may be required for the column *Improvement*. In certain cases, the AI correctly interprets and renders the idiom even without context; however, this equivalent may not be preserved in the contextual translation. Conversely, the AI may provide an improved rendering when context is available, offering more accurate or idiomatically appropriate Slovak equivalent. Thus, the *Improvement* column captures not only correction of previously inaccurate translations but also enhancements in idiomatic precision or naturalness.

A custom translation prompt was used to generate AI translations of the idioms, both in isolation and within contextual sentences. The prompt goes as follows:

Translate 100 English idioms into Slovak using the following spreadsheet structure. The idioms are located in cells B2 through B101. Write the Slovak translations in the corresponding cells D2 through D101 (i.e., the translation of B2 goes into D2, and so on). After translating all idioms, proceed to the contextual sentences in cells F2 through F101. Translate these sentences into Slovak and place the results in cells G2 through G101. Once all translations are complete, prepare the updated spreadsheet for download.

Following translation, a manual (human) evaluation of each output was conducted to assess the quality of the renditions. Subsequently, statistical methods were applied to determine whether the presence of context produced a statistically significant difference in translation quality and to address the hypothesis under investigation.

As for statistical analysis, McNemar's test was selected, as it is specifically designed for paired nominal data. This test is particularly suitable for assessing changes in categorical outcomes (e.g. correct vs. incorrect) before and after an intervention – in this case, the addition of context. It is commonly used when the same subjects (idioms) are evaluated under two different conditions, which makes it ideal for detecting shifts in translation accuracy.

$$x^2 = \frac{(b-c)^2}{b+c}$$

Fig. 1. Formula for McNemar's test

In addition, a one-tailed t-test was employed to examine whether the presence of context had a statistically significant effect on the overall quality of idiom translation. This allowed for a comparison of translation scores across conditions to determine the magnitude and direction of change.

2 THEORETICAL BACKGROUND

Translation has become an integral part of daily communication, frequently used in conversations with ChatGPT. Language consists of various expressions, inter alia, idioms, whose translation poses a significant challenge for both traditional neural machine translation systems and modern large language models due to the figurative nature of idiomatic expressions. In this paper, we proceed from Sinclair's (1991, p. 172) definition of an idiom, which is a "group of two or more words which are chosen together in order to produce a specific meaning or effect in speech or writing". As Baziotis et al. (2023) point out, "literal translation errors of idioms remain a major issue in automated translation, requiring novel evaluation metrics to assess their accuracy." In contrast to a literal translation, an idiom involves far more than a perfect semantic relation; it necessitates the integration of context and cultural customization that any AI powered translation systems must incorporate.

Since the emergence of AI chatbots and related technologies, a substantial body of research has focused on evaluating the effectiveness of AI-driven tools in translating idiomatic expressions (Hamood 2024; Mughal et al. 2024; Hakami and Abomoati 2024; Abjalova and Sharipova 2024; Obeidat et al. 2024). Some scholars concentrate on semantic and grammatical aspects of the translation process, while others focus more on its computational and informatics foundations.

Recent studies demonstrate that LLMs, or more specifically ChatGPT, have achieved significant improvements on idiomatic translation over baseline NMT models (Zhu et al. 2024). Castaldo and Monti (2024) also emphasize the importance of effective prompting strategies, stating that "the quality of LLM-generated translations is highly dependent on the structure and clarity of user prompts." This suggests that user interaction is crucial in guiding LLMs to produce more precise and context-aware translations of idioms.

From a computational perspective, the inclusion of knowledge bases like IdiomKB in translation models has been found to enhance the accuracy of translation by bringing back the figurative meanings of idioms rather than their literal meanings (Li et al. 2023). Donthi et al. (2024) highlight the potential of cosine similarity scoring in bringing idiomatic expressions in languages into alignment, with the point that "such methods enable LLMs to maintain linguistic style while guaranteeing semantic fidelity". Moreover, multilingual instruction tuning has been found to induce translation ability in LLMs even for low-resource languages (Li et al. 2024).

As AI-based translation models advance, the synthesis of linguistic and informatics methodologies is critical to improving idiomatic translation. Although LLMs have shown incredible ability, their probabilistic modeling foundation ensures that some mistranslations and biases continue to arise, calling for improvements in training practices and assessment frameworks.

While there are scholars (such as Jiao 2023) who see chatbots like ChatGPT as important tools for real-time and automated translation of texts, alongside machine translation, they acknowledge the frequent errors in their output. Nevertheless, more scholars (Derner and Batistič 2023; Sison et al. 2023; Artamonova 2023) contend that ChatGPT is extremely risky, citing its capacity to create misleading translations, disseminate misinformation, and cause ethical problems in language processing.

3 RESULTS

This study examined the impact of context on the quality of AI idioms translation. A total of 100 English idioms were evaluated under two conditions: without context and with context. Each translation's correctness was assessed, and statistical tests were conducted to determine the significance of differences obtained.

Quantitative results reveal that the addition of context led to a notable decline in translation accuracy. Without context, 91% of the idioms (n=91) were correctly translated. With context, accuracy declined to 77% (n=77).

A two-sample t-test with equal variance revealed that the difference was statistically significant, t(196) = 2.82, p = 0.0054 (two-tailed). The null hypothesis, which predicted no difference in translation quality between the two conditions, was therefore rejected.

In addition, McNemar's test was conducted to assess categorical change in translation accuracy. The test revealed that 17 idioms were correct without context but incorrect with context (Tab. 1), while only 3 showed clear-cut binary improvement.

Idiom	Translation by Kvetko	Translated by AI	Idiom in context	Translated by AI
cry wolf	robiť planý poplach; predstierať nebezpečenstvo	spôsobovať planý poplach	Moody had to be here somewhere. If he weren't, Judd knew what McGreavy would think. It would be the boy who ***cried wolf***.	Moody musel byť niekde nablízku. Ak nie, Judd vedel, čo by si McGreavy pomyslel. Že je to ako s chlapcom, ktorý ***kričal "vlk"***!

an old flame	stará láska	stará láska	You were luck to run into this girl. Who is she? Some ***old flame*** of yours?	Mal si šťastie, že si stretol to dievča. Kto to je? ***Starý plameň***?

Tab. 1. Examples of mistranslations with provided context

The remaining 75 idioms retained the same accuracy. The resulting chi-square value ($x^2 = 8.45$, $p = 0.0036$) also leads to the conclusion that the introduction of context had an effect on translation quality, but not in a positive direction. A detailed breakdown of translation outcomes is presented in Fig. 2.

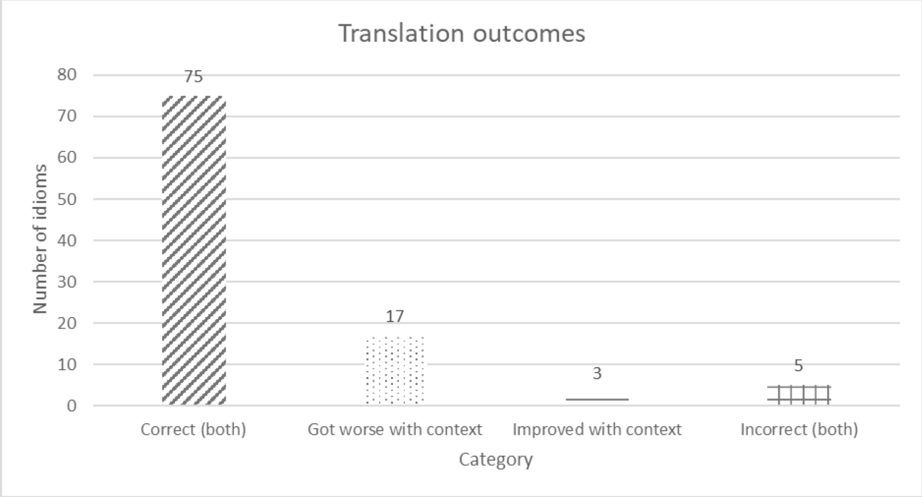


Fig. 2. Translation outcomes divided into four categories

A closer look at the individual examples supports the quantitative findings by showing how contextual information sometimes caused the AI to shift from accurate idiomatic expressions to less appropriate or overly literal renderings. For instance, the idiom *cry wolf* was initially rendered correctly by the AI and as such it was a semantically acceptable and idiomatic equivalent. However, when placed in a contextual sentence, the AI produced a literal back-translation. This rendering fails to convey the figurative meaning and results in a loss of communicative intent. Interestingly, it demonstrates the model's tendency to prioritize surface-level lexical matching over pragmatic interpretation when embedded in context.

As for the second example, *an old flame,* the situation was the same. The translation without context is correct, however, there is a literal translation whencontext is provided. According to SSSJ (Jarošová et al. 2021), the Slovak word *plameň* 'flame' denotes figuratively to passion or zeal and is linked also to love in the collocation *plameň lásky* 'flame of love'. As a result, the AI would have ended up with a correct translation if it had added the word *lásky* to *plameň*, however, again it failed to deliver the semantic information within its translation proposal.

In five cases, the AI produced incorrect translations regardless of whether contextual information was present or not (Tab. 2).

Idiom	Translation by Kvetko	Translated by AI	Idiom in context	Translated by AI
agree to differ	myslieť si svoje	zhodnúť sa, že sa nezhodneme	Sometimes in a close friendship, where important matters are concerned, people ***agree to differ***, and fall silent.	Niekedy v blízkom priateľstve, kde ide o dôležité veci, sa ***ľudia zhodnú, že sa nezhodnú***, a zmĺknu.

Tab. 2. Examples of mistranslations in both stages

This example illustrates a case in which the translation produced by the AI may initially appear acceptable, as it conveys a meaning that a Slovak reader can understand both in isolation and within context. However, it reveals two key shortcomings: it is fully literal and lacks the natural fluency characteristic of idiomatic Slovak. The resulting expression, while intelligible, is awkward and stylistically marked, deviating from conventional usage. Instances of this nature were observed sporadically and should be considered exceptions rather than representative of the overall translation patters.

Qualitative analyses revealed several patterns. In some cases, the AI system initially proposed a correct Slovak equivalent out of context but selected a less idiomatic or excessively literal rendition when context was added. In other instances, it could not integrate the idiom meaningfully into the wider sentence structure. This suggests limitations on contextual reasoning or phraseological awareness.

Nonetheless, particular attention was given to the factor of improvement (Tab. 3). If the translation with context was identical to the translation without context, it was deemed acceptable and adequate. However, qualitative analysis showed that in 20 cases, the contextual translation was considerably improved in terms of idiomaticity and naturalness. This highlights that while binary statistical methods identify only a small number of improvements, a more nuanced linguistic analysis reveals a greater degree of positive change due to context.

Idiom	Translation by Kvetko	Translated by AI	Idiom in context	Translated by AI
beat the air	hádzať hrach na stenu	márne sa snažiť; robiť niečo zbytočne	"Aren't we ***beating the air***, Blythe?"	***"Nehovoríme len do vetra***, Blythe?"
a body blow	ťažký úder, vážny neúspech	tvrdá rana	"That's your new assignment." It was ***a body blow*** and Liz winced.	"To je tvoja nová úloha." Bola to ***rana pod pás*** a Liz sa strhla.

Tab. 3. Examples of improved translations

4 DISCUSSION AND CONCLUSION

This study reveals several promising and surprising insights regarding the capabilities of AI in idiom translation. Interestingly, a substantial proportion of the 100 English idioms were rendered correctly by the AI, even without contextual support. This result challenges common assumptions that non-compositional, figurative expressions fall outside the reach of computational models. Given the longstanding view that idioms resist rule-based or literal translation, the high baseline accuracy demonstrated here is a compelling indication of the progress made in LLMs.

Nevertheless, the study also exposed critical limitations. The AI often produced literal, stylistically awkward, or semantically mismatched translations when context was introduced. These cases suggest that while surface-level idiomatic retrieval may be successful, deeper contextual and pragmatic integration remains a challenge. Furthermore, a major technical drawback emerged during batch translation attempts: when prompted with a list of idioms in spreadsheet format, the system processed only five, requiring the rest to be input manually. This underpins inefficiencies in AI interaction design for linguistic research.

While the present study was not corpus-driven in design, future work could benefit from a closer integration with corpus linguistics. For example, idiom translations generated by AI could be compared with those found in parallel corpora. However, this approach would be limited by the availability and structure of idioms in such corpora, because identifying and aligning idiomatic expressions remains complex.

It is also important to note that while statistical analysis showed only three improvements due to context, qualitative assessment found 20 cases with considerably improved idiomaticity. This suggests that broader evaluation criteria can offer a fuller picture of translation quality.

Future research could extend the current findings by exploring idiom translation in reverse direction – from Slovak into English – and further across other language pairs, such as Slovak-German or English-German. Such studies would allow comparative insights into whether AI systems perform differently depending on the source and target language, especially in the case of structurally distant or closely related languages. In addition, future experiments could incorporate low-resource idioms, culturally bound expressions, or idioms with multiple transferred layers, which would further test the model's semantic awareness. Research on how prompt engineering and fine-tuning influence idiomatic output also remains a promising avenue.

ACKNOWLEDGEMENTS

The research has been elaborated within the project A-25-106/3020-18 *Artificial intelligence: challenges for linguistics and marketing*.

References

Abjalova, M., and Sharipova, S. (2024). Semantic and Grammatical Issues in Translating Idioms with Automatic Translation Systems. In 2024 9th International Conference on Computer Science and Engineering, pp. 58–63.

Artamonova, M. V. et al. (2023). Chatbot as a translation tool. In Litera 8, pp. 235–253.

Baziotis, Ch. et al. (2023). Automatic Evaluation and Analysis of Idioms in Neural Machine Translation. In Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, Dubrovnik: Association for Computational Linguistics, pp. 3682–3700. Accessible at: https://aclanthology.org/2023.eacl-main.267/.

Derner, E., and Batistič, K. (2023). Beyond the Safeguards: Exploring the Security Risks of ChatGPT. Accessible at: https://arxiv.org/abs/2305.08005.

Dong, Y. et al. (2019): Acquisition of interpreting strategies by student interpreters. The Interpreter and Translator Trainer 13(4), pp. 408–425.

Donthi, S. (2025). Improving LLM Abilities in Idiomatic Translation. In Proceedings of the First Workshop on Language Models for Low-Resource Languages. Abu Dhabi: Association for Computational Linguistics, pp. 175–181. Accessible at: https://arxiv.org/abs/2407.03518.

Hakami, A. H., and Abomoati, G. S. (2024). Exploring the Impact of Prompt Formulation in AI Chatbots on the Translation of English-to-Arabic and Arabic-to-English Idioms: A Case-Study. Pakistan Journal of Life and Social Sciences 22(2), pp. 21371–21381.

Hamood, M. I. (2024). The Translation of English Food Idioms into Arabic Through ChatGPT: Problems and Solutions. Accessible at: https://shorturl.at/9TcUT.

Jarošová, A. et al. (2021). Slovník súčasného slovenského jazyka. Bratislava: Veda.

Jiao, W. et al. (2023). Is ChatGPT a Good Translator? Yes With GPT-4 As The Engine. Accessible at: https://arxiv.org/abs/2301.08745.

Li, J. et al. (2024). Eliciting the Translation Ability of Large Language Models via Multilingual Finetuning with Translation Instructions. In Transactions of the Association for

Computational Linguistics 12, pp. 576–592. Accessible at: https://aclanthology.org/2024.tacl-1.32/.

Li, S. et al. (2023). Translate Meanings, Not Just Words: IdiomKB's Role in Optimizing Idiomatic Translation with Language Models. In Computation and Language. Accessible at: https://arxiv.org/abs/2308.13961.

Lund, B. (2023). ChatGPT and a New Academic Reality: Artificial Intelligence-Written Research Papers and the Ethics of the Large Language Models in Scholarly Publishing. In Journal of the Association for Information Science and Technology. Accessible at: https://arxiv.org/abs/2303.13367.

Mughal, U. A. et al. (2024). The intersection of linguistics and artificial intelligence: A corpus-based study of idiom translation. Journal of applied linguistics and Tesol 7(4), pp. 1453–1460.

Obeidat, M. M. et al. (2024). Analyzing the Performance of Gemini, ChatGPT, and Google Translate into Rendering English Idioms into Arabic. Journal of Social Sciences 18(4), pp. 1–18.

Rodriguez, P. R. (2024). Phraseological evaluation of automatic interpretation assisted by Yandex. Translation Matters 6(2), pp. 115–130.

Sinclair, J. (1991). Corpus, Concordance, Collocation. Oxford: Oxford University Press.

Sison, A. J. et al. (2023). ChatGPT: More than a Weapon of Mass Deception, Ethical challenges and responses from the Human-Centered Artificial Intelligence (HCAI) perspective. Accessible at: https://arxiv.org/abs/2304.11215.

Wang, X., and Fantinuoli, C. (2024). Exploring the correlation between human and machine evaluation of simultaneous speech translation. In Proceedings of the 25th Annual Conference of the European Association for Machine Translation, Sheffield: EAMT, pp. 327–336. Accessible at: https://aclanthology.org/2024.eamt-1.28/.

Zhu, W. (2024). Multilingual Machine Translation with Large Language Models: Empirical Results and Analysis. In Findings of the Association for Computational Linguistics: NAACL 2024. Mexico City: Association for Computational Linguistics, pp. 2765–2781. Accessible at: https://arxiv.org/abs/2304.04675.

sciendo
DOI 10.2478/jazcas-2025-0024

TAILORED FINE-TUNING FOR COMMA INSERTION IN CZECH

JAKUB MACHURA[1] – HANA ŽIŽKOVÁ[2] – PATRIK STANO[3] – TEREZA VRABCOVÁ[4] – DANA HLAVÁČKOVÁ[5] – ONDŘEJ TRNOVEC[6]

[1]Department Czech Language, Faculty of Arts, Masaryk University, Brno, Czech Republic (ORCID: 0000-0002-6623-3064)

[2]Department of Czech Language, Faculty of Arts, Masaryk University, Brno, Czech Republic (ORCID: 0000-0002-6483-6603)

[3]Department of Machine Learning and Data Processing, Faculty of Informatics, Masaryk University, Brno, Czech Republic (ORCID: 0009-0001-8339-6084)

[4]Department of Machine Learning and Data Processing, Faculty of Informatics, Masaryk University, Brno, Czech Republic (ORCID: 0009-0009-5674-3827)

[5]Department of Czech Language, Faculty of Arts, Masaryk University, Brno, Czech Republic (ORCID: 0000-0002-9918-0958)

[6]Department of Czech Language, Faculty of Arts, Masaryk University, Brno, Czech Republic (ORCID: 0009-0009-7756-9661)

MACHURA, Jakub – ŽIŽKOVÁ, Hana – STANO, Patrik – VRABCOVÁ, Tereza – HLAVÁČKOVÁ, Dana – TRNOVEC, Ondřej: Tailored Fine-tuning for Comma Insertion in Czech. Journal of Linguistics, 2025, Vol. 76, No 1, pp. 268 – 278.

Abstract: Transfer learning techniques, particularly the use of pre-trained Transformers, can be trained on vast amounts of text in a particular language and can be tailored to specific grammar correction tasks, such as automatic punctuation correction. The Czech pre-trained RoBERTa model demonstrates outstanding performance in this task (Machura et al. 2022); however, previous attempts to improve the model have so far led to a slight degradation (Machura et al. 2023). In this paper, we present a more targeted fine-tuning of this model, addressing linguistic phenomena that the base model overlooked. Additionally, we provide a comparison with other models trained on a more diverse dataset beyond just web texts.

Keywords: comma, Czech, Fine-tuning, Large Language Model (LLM)

1 INTRODUCTION

Punctuation, along with other graphical markers, plays a crucial role in ensuring the accurate comprehension of any text. The automatic insertion of sentence commas is typically addressed in two key tasks: (1) punctuation restoration in speech transcripts generated by automatic speech recognition (ASR), where reinstating punctuation significantly enhances readability, and (2) grammatical error correction in written texts, where commas may be either missing or redundant. In Czech, automatic punctuation correction is one of the most critical aspects of grammatical error correction, as the comma is not only the most frequently used punctuation mark (see Švec et al. 2021) but also a fundamental indicator of refined and structured writing.

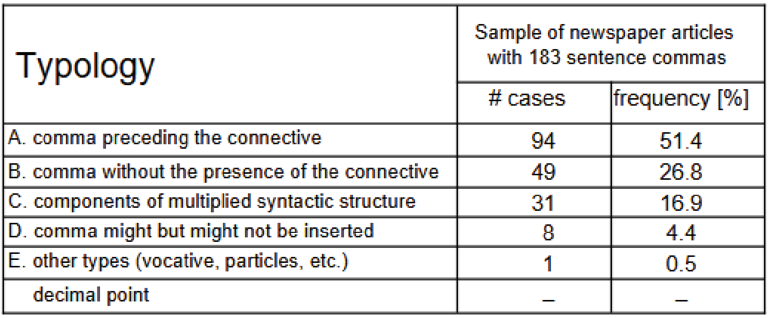
For an extended period, the rule-based approach introduced by Kovář et al. (2016) demonstrated the highest precision in comma insertion. However, its recall did not exceed 60% of all commas. In recent years, advancements in machine learning have led to significant improvements in precision. The transformer-based approach proposed in (Machura et al. 2022) highlights the prevailing trend of training language models, analyzing their errors, and exploring potential refinements. This approach achieves precision comparable to or surpassing that of rule-based methods, while its recall exceeds 80%. A notable advantage of rule-based methods is their interpretability, as it is relatively straightforward to identify the specific rule responsible for a false positive. In contrast, neural network models function as black boxes, making it challenging not only to determine the cause of a particular error but also to implement targeted corrections.

This paper first extends previous experiments on re-training the RoBERTa model (see Section 2.1) and then introduces additional models fine-tuned on data from the SYN v9 corpus (Křen et al. 2021) available from the LINDAT/CLARIAH-CZ digital library. Finally, the study includes a comprehensive evaluation of all models using texts that have served as benchmarks for this task for nearly a decade (see Section 3).

2 COMMA INSERTION USING PRE-TRAINED MODELS

This section introduces various models pre-trained for the task of comma insertion. One set of experiments was conducted using a model trained and fine-tuned on web texts, many of which had not undergone any proofreading. Despite this, the results exceeded expectations, prompting the question of how models pre-trained or fine-tuned on higher-quality texts would perform.

In (Machura et al. 2022) the typology of the comma insertion place was comprehensively described. This allows 1) to specify the place (boundary) in the sentence structure where comma is inserted, 2) to analyze the type of commas that users of the language omit or overuse, or 3) to evaluate the results of language models that are pre-trained, namely for the task of inserting commas into text, and then subsequently improve these models.



Typology	Sample of newspaper articles with 183 sentence commas	
	# cases	frequency [%]
A. comma preceding the connective	94	51.4
B. comma without the presence of the connective	49	26.8
C. components of multiplied syntactic structure	31	16.9
D. comma might but might not be inserted	8	4.4
E. other types (vocative, particles, etc.)	1	0.5
decimal point	–	–

Tab. 1. Estimated general distribution of commas in Czech texts according to typology

2.1 RoBERTa – Training on web texts

Machura et al. (2023) examined the feasibility of re-training the RoBERTa language model, which had been pre-trained on a collection of web data, including Common Crawl[1] and texts from the Czech Wikipedia, and fine-tuned with a random data set from the Czech Common Crawl. The study focused specifically on improving RoBERTa's ability to detect commas in Czech vocatives by utilizing example sentences where the model had previously made errors. To this end, researchers extracted 170,000 sentences from the csTenTen17 corpus (Suchomel 2018) and employed two re-training strategies: (i) additional fine-tuning and (ii) an expanded training dataset, wherein the original large corpus was merged with a specialized corpus containing vocative phrases. While precision improved, recall declined significantly, likely due to overfitting to a specific comma type. The findings underscore the importance of training data distribution, highlighting the necessity of a broader dataset to preserve the model's overall functionality. The study ultimately demonstrates that while re-training RoBERTa is feasible, it requires careful structuring of the dataset to ensure balanced performance.

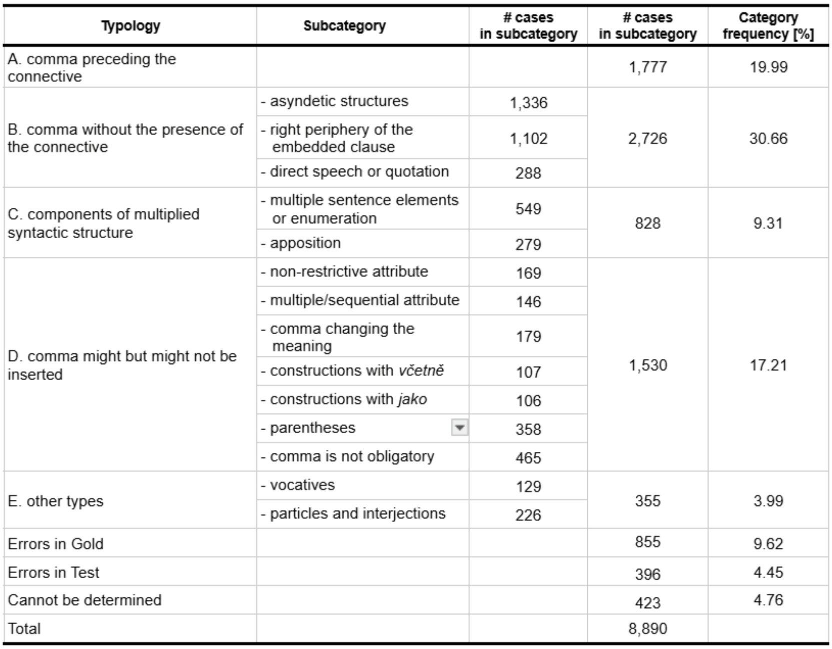
Following the vocative experiment, we have developed a fine-tuning dataset intended to enhance the overall performance of the RoBERTa base model for comma insertion – not solely for a specific type of comma. To identify RoBERTa's strengths and weaknesses, we conducted a thorough analysis on a dataset comprising 67,378 sentences and 87,379 commas extracted from news articles. The original texts (referred to as Gold) were presumed to be error-free following proofreading, although some errors did persist. By providing the model with these texts devoid of commas, it subsequently inserted 78,146 commas. The output (referred to as Test) was then compared with Gold using a script, revealing approximately 10,000 sentences where the model's comma placement diverged from the Gold standard.

Subsequently, we compiled a dataset of these sentences featuring mismatched commas (see the following Tab. 2), aligning each pair from Gold and Test side by side for annotation based on the aforementioned typology. Although the number of sentences identified by the script (10,000) slightly exceeded those annotated (8,890), this discrepancy likely arises from human annotation error and imperfect sentence separation by the script, particularly when segregating sentences in category A based on the connective following the comma.

The table below indicates that while categories A and C pose minimal challenges for the model, category B presents a moderate level of difficulty. In contrast, categories D and E are the most problematic. Comparing Tab. 1 with Tab. 2, the challenges associated with categories D and E are expected, given their lower relative distribution in the text, which results in a reduced amount of training data and consequently limits the model's ability to generalize effectively in these cases. Additionally, category

[1] https://commoncrawl.org/

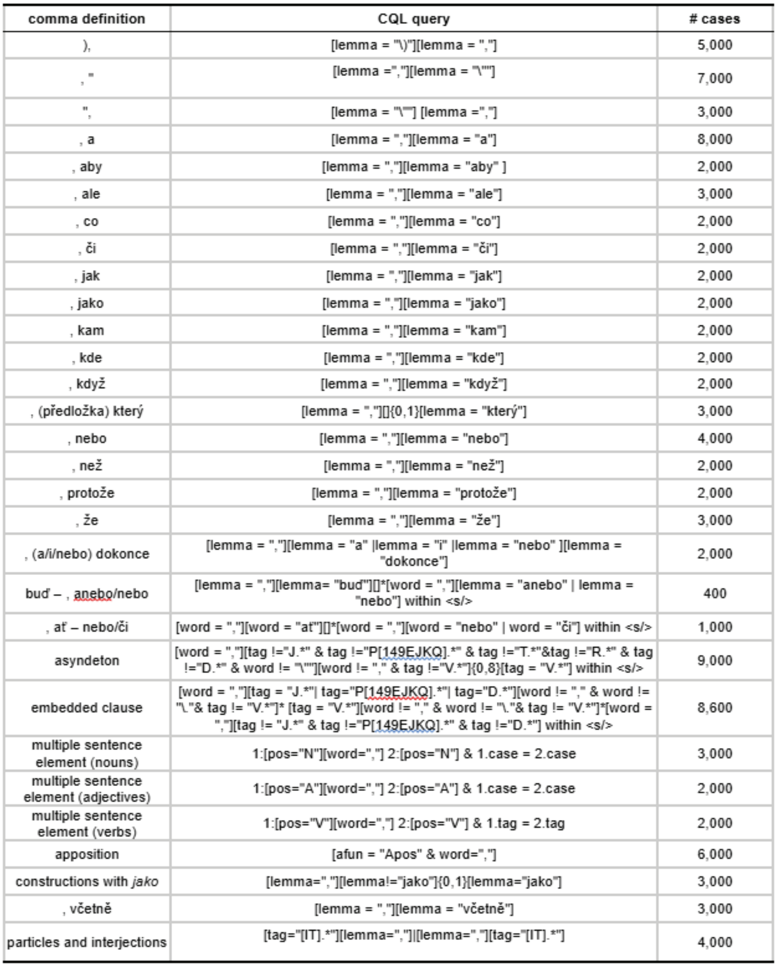
D necessitates a deeper semantic and pragmatic understanding for accurate comma insertion. Based on these observations, we prioritized our fine-tuning efforts on more frequently occurring categories that offer greater potential for improvement.



Typology	Subcategory	# cases in subcategory	# cases in subcategory	Category frequency [%]
A. comma preceding the connective			1,777	19.99
B. comma without the presence of the connective	- asyndetic structures	1,336	2,726	30.66
	- right periphery of the embedded clause	1,102		
	- direct speech or quotation	288		
C. components of multiplied syntactic structure	- multiple sentence elements or enumeration	549	828	9.31
	- apposition	279		
D. comma might but might not be inserted	- non-restrictive attribute	169	1,530	17.21
	- multiple/sequential attribute	146		
	- comma changing the meaning	179		
	- constructions with *včetně*	107		
	- constructions with *jako*	106		
	- parentheses ▾	358		
	- comma is not obligatory	465		
E. other types	- vocatives	129	355	3.99
	- particles and interjections	226		
Errors in Gold			855	9.62
Errors in Test			396	4.45
Cannot be determined			423	4.76
Total			8,890	

Tab. 2. Estimated distribution of the mismatched commas of the RoBERTa base model (Machura et al. 2022)

With this insight in mind, we compiled two datasets for fine-tuning RoBERTa. The first dataset, consisting of 1,313 sentences, was constructed using CQL queries on the internet corpus csTenTen2023 (Suchomel 2018) in Sketch Engine. Each sentence in this dataset was manually verified to ensure that it contained the correct type of comma as required by the CQL query. To identify sentences containing apposition, we utilized the syntactic function Apos in the syntactically annotated corpus SYN2020 (Křen et al. 2020) accessible via KonText (Machálek 2020). The second dataset, comprising 100,000 sentences, was entirely sourced from the SYN2020 corpus. This choice was motivated by the assumption that SYN2020 – composed solely of printed texts (fiction, non-fiction, newspapers, and magazines) – exhibits a higher linguistic standard compared to an internet corpus such as csTenTen2023, despite the absence of human verification for comma type accuracy. The CQL queries used to compile this larger dataset, along with the sentence counts for each query, are detailed in Tab. 3. Although the relative distribution of sentences and the queries for the smaller dataset are largely consistent, minor differences exist due to the disparate morphological tagsets employed by each corpus manager. Again, two training strategies were used – (i) additional fine-tuning and (ii) an expanded training dataset, wherein the original large corpus was merged with a specialized corpus containing 1,313 or 100,000 sentences – yielding four model variants.



comma definition	CQL query	# cases
),	[lemma = "\)"][lemma = ","]	5,000
, "	[lemma =","][lemma = "\""]	7,000
",	[lemma = "\""] [lemma =","]	3,000
, a	[lemma = ","][lemma = "a"]	8,000
, aby	[lemma = ","][lemma = "aby"]	2,000
, ale	[lemma = ","][lemma = "ale"]	3,000
, co	[lemma = ","][lemma = "co"]	2,000
, či	[lemma = ","][lemma = "či"]	2,000
, jak	[lemma = ","][lemma = "jak"]	2,000
, jako	[lemma = ","][lemma = "jako"]	2,000
, kam	[lemma = ","][lemma = "kam"]	2,000
, kde	[lemma = ","][lemma = "kde"]	2,000
, když	[lemma = ","][lemma = "když"]	2,000
, (předložka) který	[lemma = ","][]{0,1}[lemma = "který"]	3,000
, nebo	[lemma = ","][lemma = "nebo"]	4,000
, než	[lemma = ","][lemma = "než"]	2,000
, protože	[lemma = ","][lemma = "protože"]	2,000
, že	[lemma = ","][lemma = "že"]	3,000
, (a/i/nebo) dokonce	[lemma = ","][lemma = "a" |lemma = "i" |lemma = "nebo"][lemma = "dokonce"]	2,000
buď – , anebo/nebo	[lemma = ","][lemma= "buď"][]*[word = ","][lemma = "anebo" | lemma = "nebo"] within <s/>	400
, ať – nebo/či	[word = ","][word = "ať"][]*[word = ","][word = "nebo" | word = "či"] within <s/>	1,000
asyndeton	[word = ","][tag !="J.*" & tag !="P[149EJKQ].*" & tag !="T.*"&tag !="R.*" & tag !="D.*" & word != "\""][word != "," & tag !="V.*"]{0,8}[tag = "V.*"] within <s/>	9,000
embedded clause	[word = ","][tag = "J.*"| tag="P[149EJKQ].*"| tag="D.*"][word != "," & word != "\."& tag != "V.*"]* [tag = "V.*"][word != "," & word != "\."& tag != "V.*"]*[word = ","][tag != "J.*" & tag !="P[149EJKQ].*" & tag !="D.*"] within <s/>	8,600
multiple sentence element (nouns)	1:[pos="N"][word=","] 2:[pos="N"] & 1.case = 2.case	3,000
multiple sentence element (adjectives)	1:[pos="A"][word=","] 2:[pos="A"] & 1.case = 2.case	2,000
multiple sentence element (verbs)	1:[pos="V"][word=","] 2:[pos="V"] & 1.tag = 2.tag	2,000
apposition	[afun = "Apos" & word=","]	6,000
constructions with *jako*	[lemma=","][lemma!="jako"]{0,1}[lemma="jako"]	3,000
, včetně	[lemma = ","][lemma = "včetně"]	3,000
particles and interjections	[tag="[IT].*"][lemma=","]|[lemma=","][tag="[IT].*"]	4,000

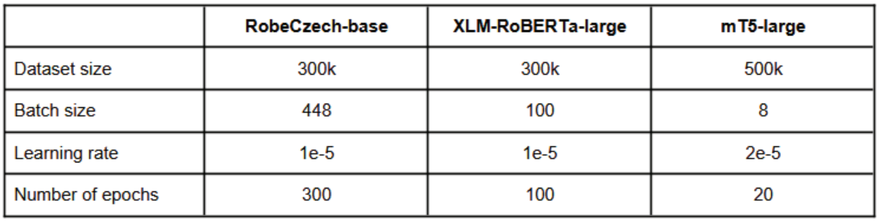
Tab. 3. List of CQL queries for compilation of 100,000 sentence dataset

2.2 Fine-tuning with SYN v9

To investigate the effectiveness of fine-tuning for automatic comma insertion in Czech text, we trained three different transformer-based models: RobeCzech-base (Straka et al. 2021), XLM-RoBERTa-large (Conneau et al. 2020), and mT5-large

(Xue et al. 2021). The RobeCzech-base and XLM-RoBERTa-large models were fine-tuned as token classification models, where the objective was to predict whether a given token should be followed by a comma. The mT5-large model was fine-tuned as a text-to-text model with the objective of adding commas to a text without any commas.

Training Setup*:* Each model was trained using the SYN v9 dataset (Křen et al. 2021), available in the LINDAT repository, which was filtered to include only lines containing at least one comma. SYN v9 was chosen because the training of the RoBERTa baseline model was done on random texts from the internet and achieved quite good results, so the idea was to use texts that had mostly undergone some proofreading and might contain a wider variety of comma types. The dataset from SYN v9 was selected for its diverse curated content, as prior research (Machura et al. 2023) demonstrated that fine-tuning on an unfiltered Common Crawl dataset yielded significant results. However, even here, the comma type is random and may not match the frequency distribution of each comma type. Models were trained on datasets of 100,000, 300,000, and 500,000 lines from SYN v9, with experiments conducted using various numbers of training epochs. The best-performing hyperparameters for each model are listed below. Training and evaluation were performed on a single Nvidia A40 GPU, employing the AdamW optimizer and cross-entropy loss function.



	RobeCzech-base	XLM-RoBERTa-large	mT5-large
Dataset size	300k	300k	500k
Batch size	448	100	8
Learning rate	1e-5	1e-5	2e-5
Number of epochs	300	100	20

Tab. 4. The best hyperparameters for individual models

Preprocessing steps included tokenization using the respective model's tokenizer, as well as ensuring that quotation marks were tokenized as a separate token, and an optional transformation during evaluation where quotation marks were removed from the text. The impact of this transformation was analyzed in the evaluation phase (see Section 3).

2.3 Grammatical Error Correction (GEC)

In this experiment, we explore the application of the sequence-to-labels approach to grammatical error correction (GEC) for restoring missing commas in the text. This approach was inspired by the sequence labeling methods often used for the named entity recognition (NER) task (Kumar et al. 2023), as well as parts of the

GEC implementation of the grammarly/GeCToR architecture (Omelianchuk et al. 2020). Unlike in the more common sequence-to-sequence approach where the output is only the corrected input text, this approach returns both the corrected text and the labels showing where the changes have occurred, making it easier to interpret the model's decision.

We have prepared the training and evaluation datasets by introducing synthetic mistakes in the text, namely removing all commas from the text. Output of our preprocessing were pairs of documents:

- plain-text document (all commas were removed)
- label document where each word is tagged with a corresponding label:
  $KEEP: The word is correct and should not be changed.
  $MISSING_PUNCT_,: A comma should be inserted after this word.

Using the prepared training and evaluation datasets, we have fine-tuned a pre-trained RobeCzech-base model (Straka et al. 2021), tokenizing our datasets using the base model's tokenizer. To properly align the tokens with the reference word-level labels, the original word's label is duplicated across all corresponding tokens. During the fine-tuning process we evaluate the models' performance using the precision, recall, and F1-score for the $MISSING_PUNCT_, label class. At the end of the fine-tuning we evaluate the model with the highest F1-score during training on the test dataset. As the model predicts labels per token, during post-processing we convert the token-level predictions back into word-level labels, aggregating predictions for each word and selecting the predicted label with the highest frequency. If multiple labels have the same frequency, one is arbitrarily selected.

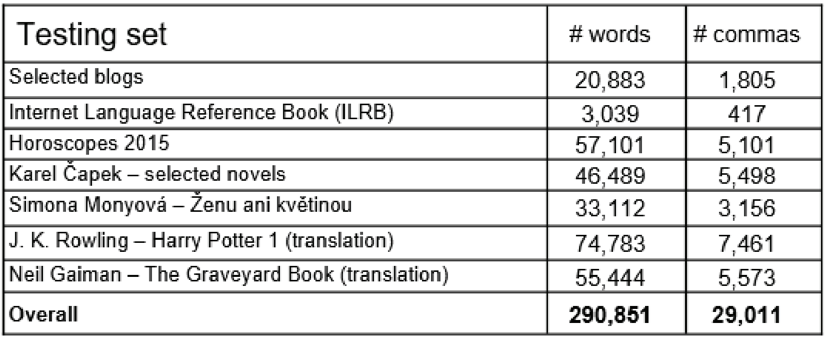
2.4 GPT-4o

For comparison, we also conducted an initial experiment in comma insertion using a generative language model GPT-4o-2024-08-06 (OpenAI 2024)[2]. Employing a temperature setting of 0.1 and a prompt instructing the model – "You are an expert in writing sentence commas in Czech and always respond in JSON format. Your task is to add missing commas to sentences" – the model demonstrated promising performance. A notable issue with this approach, however, was that the model occasionally modified the sentences beyond merely adding commas (e.g. altering or inserting words, correcting grammar), thereby complicating direct sentence comparisons. Modified sentences accounted for about 3%. This challenge could potentially be mitigated by refining the prompt or implementing a feedback loop to ensure that only commas are modified.

[2] https://chat.openai.com/

3 EXPERIMENTAL RESULTS

The dataset presented in Kovář et al. (2016) was utilized to evaluate and compare the methods described above. These texts were specifically designed for automatic comma insertion. As the dataset remains unchanged, the current results can be directly compared with previous evaluations. In total, seven texts of varying nature and style were used, as shown in Tab. 5.



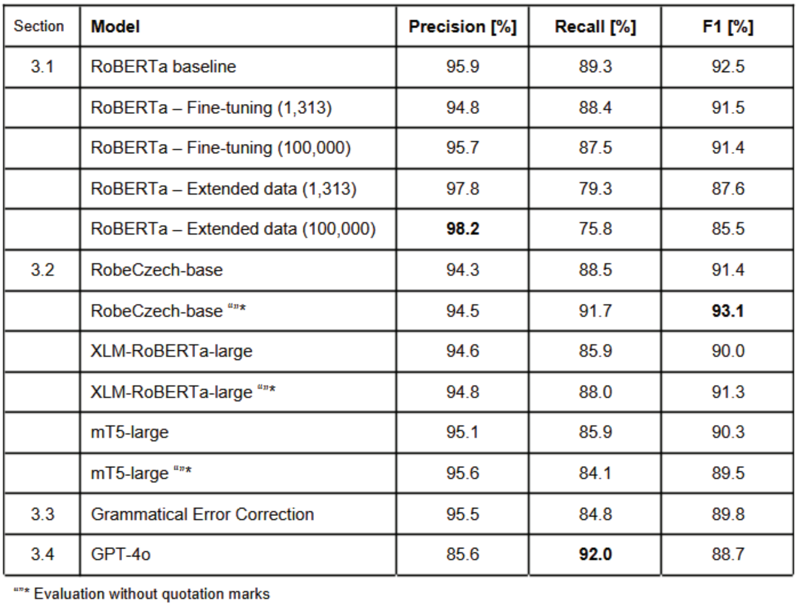
Testing set	# words	# commas
Selected blogs	20,883	1,805
Internet Language Reference Book (ILRB)	3,039	417
Horoscopes 2015	57,101	5,101
Karel Čapek – selected novels	46,489	5,498
Simona Monyová – Ženu ani květinou	33,112	3,156
J. K. Rowling – Harry Potter 1 (translation)	74,783	7,461
Neil Gaiman – The Graveyard Book (translation)	55,444	5,573
Overall	**290,851**	**29,011**

Tab. 5. Statistics of the test data for automatic comma insertion

The highest F1 score (93.1%) was achieved by the fine-tuned RobeCzech-base model when quotation marks were removed in preprocessing. The model outperformed the RoBERTa baseline model in terms of recall but exhibited lower precision. It is worth noting that in all RoBERTa baseline model experiments, post-processing was required for fiction texts, as the model consistently placed a comma after closing quotation marks in direct speech, despite the correct placement being before them. Overall, GPT-4o achieved the highest recall (92.0%); however, this came at the cost of precision, as it produced nearly 4,500 false positives (85.6%).

In the RoBERTa experiments (Section 3.1), an increase in training data consistently improved precision, reaching up to 98.2%; however, recall decreased significantly. The incorporation of additional datasets likely disrupted the frequency distribution of different comma types, leading the model to insert fewer commas with greater confidence. Notably, fine-tuning with the selected dataset, which was specifically designed to target phenomena ignored by the RoBERTa baseline model, yielded unexpected results, as all evaluation metrics declined.

Results of models from Section 3.2 – the RobeCzech-base and XLM-RoBERTa-large models showed improved performance when quotation marks were removed in preprocessing, while mT5-large achieved a better result with quotations included. A plausible hypothesis is that quotation marks can serve as useful syntactic cues for larger language models, aiding in the recognition of grammatical structures. For smaller models with more limited capacity, such as RobeCzech-base, they may act as a source of noise or distraction. Despite being the smallest model, RobeCzech-base outperformed both XLM-RoBERTa-large and mT5. Its best performance surpasses a result reported in (Machura et al. 2022), while the other models failed to surpass this benchmark. The superior performance of RobeCzech-base suggests that a model specifically designed for Czech text may be more effective for this task than larger multilingual models. Further analysis could explore whether additional fine-tuning techniques or architectural modifications might enhance the performance of the larger models.



Section	Model	Precision [%]	Recall [%]	F1 [%]
3.1	RoBERTa baseline	95.9	89.3	92.5
	RoBERTa – Fine-tuning (1,313)	94.8	88.4	91.5
	RoBERTa – Fine-tuning (100,000)	95.7	87.5	91.4
	RoBERTa – Extended data (1,313)	97.8	79.3	87.6
	RoBERTa – Extended data (100,000)	**98.2**	75.8	85.5
3.2	RobeCzech-base	94.3	88.5	91.4
	RobeCzech-base “”*	94.5	91.7	**93.1**
	XLM-RoBERTa-large	94.6	85.9	90.0
	XLM-RoBERTa-large “”*	94.8	88.0	91.3
	mT5-large	95.1	85.9	90.3
	mT5-large “”*	95.6	84.1	89.5
3.3	Grammatical Error Correction	95.5	84.8	89.8
3.4	GPT-4o	85.6	**92.0**	88.7

“”* Evaluation without quotation marks

Tab. 6. Results of all mentioned models

4 CONCLUSION

The primary objective of this study was to develop a tailored dataset that incorporates linguistic phenomena overlooked by the RoBERTa baseline model. However, selecting the most frequently missing comma types to construct a retraining dataset did not lead to an improvement in the model’s original performance.

The second objective was to compare models trained on web-based data – which, not having been proofread, often might contain false positives – with models trained on texts from the SYN v9 corpus, which are presumed to be of higher quality. The RobeCzech-base model fine-tuned on SYN v9 data outperformed the previous

RoBERTa model overall, but achieved a slightly lower precision. Further improvement could be achieved by filtering the SYN v9 dataset to be more representative of the natural frequency distribution of commas in Czech.

Additionally, an interesting comparison was made with GPT-4o and Grammatical Error Correction (GEC), both of which demonstrated comparable or superior performance in certain metrics. Nevertheless, their overall F1 scores remained relatively average.

In the next phase of this research, we will seek to identify the optimal composition of training data that encompasses all comma types in accordance with their natural frequency distribution, thereby maximizing recall. Simultaneously, the dataset must be balanced to achieve the highest possible precision, as the model must learn not only where to insert a comma—such as before a connective or other relevant token—but also where a comma should not be placed. For instance, while more than 4% of all commas in the SYN2020 corpus precede the conjunction *ale* 'but', in over one-quarter of all instances where *ale* 'but' appears, a comma is not required. Since neural networks function as a black box, we cannot determine with certainty whether this approach will produce the desired results. However, we believe that precisely constructing a balanced training dataset from SYN corpora could improve the functionality of the tested models.

ACKNOWLEDGEMENTS

The authors acknowledge that this work was supported by the OSCARS project, funded by the European Commission's Horizon Europe Research and Innovation program (grant agreement No. 101129751), led by the five Science Clusters: ENVRI, ESCAPE, LS RI, PaNOSC, and SSHOC.

References

Conneau, A. et al. (2020). Unsupervised Cross-lingual Representation Learning at Scale. In: D. Juravsky et al. (eds.): Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, pp. 8440–8451.

Kovář, V. et al. (2016). Evaluation and improvements in punctuation detection for Czech. In: P. Sojka et al. (eds.): Text, Speech, and Dialogue. Springer International Publishing, pp. 287–294.

Křen, M. et al. (2020). SYN2020: A representative corpus of written Czech. Institute of the Czech National Corpus, Faculty of Arts, Charles University, Prague. Accessible at: http://www.korpus.cz.

Křen, M. et al. (2021). SYN v9: large corpus of written Czech, LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of

Mathematics and Physics, Charles University. Accessible at: http://hdl.handle.net/11234/1-4635.

Kumar, P. et al. (2023). Transformer-Based Models for Named Entity Recognition: A Comparative Study. 14th International Conference on Computing Communication and Networking Technologies (ICCCNT), Delhi, India, 2023, pp. 1–5. Accessible at: https://doi.org/10.1109/ICCCNT56998.2023.10308039.

Machálek, T. (2020): KonText: Advanced and Flexible Corpus Query Interface. In Proceedings of LREC 2020, pp. 7005–7010.

Machura, J. et al. (2022). Automatic Grammar Correction of Commas in Czech Written Texts: Comparative Study. In: P. Sojka et al. (eds): Text, Speech, and Dialogue. TSD 2022. Lecture Notes in Computer Science, Vol. 13502. Springer. Accessible at: https://doi.org/10.1007/978-3-031-16270-1_10.

Machura, J. et al. (2023). Is it possible to re-educate RoBERTa? Expert-driven machine learning for punctuation correction. In Slovko (October 18 – 20, 2023) Bratislava. Accessible at: https://dx.doi.org/10.2478/jazcas-2023-0052.

Omelianchuk, K. el al. (2020). GECToR – Grammatical Error Correction: Tag, Not Rewrite. In Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications, Seattle, WA, USA → Online. Association for Computational Linguistics, pp. 163–170.

OpenAI. (2024). ChatGPT-4o. Accessible at: https://chat.openai.com.

Straka, M. et al. (2021). RobeCzech: Czech RoBERTa, a Monolingual Contextualized Language Representation Model. In: K. Ekštein et al. (eds): Text, Speech, and Dialogue. TSD 2021. Lecture Notes in Computer Science, Vol. 12848. Springer, Cham. Accessible at: https://doi.org/10.1007/978-3-030-83527-9_17.

Suchomel, V. (2018). csTenTen17, a Recent Czech Web Corpus. In Twelveth Workshop on Recent Advances in Slavonic Natural Language Processing. Brno: Tribun EU, pp. 111–123.

Švec, J. et al. (2021). Transformer-based automatic punctuation prediction and word casing reconstruction of the ASR output. In: Ekštein, K. et al. (eds.): Text, Speech, and Dialogue, Springer International Publishing, pp. 86–94.

Xue, L. et al. (2021). mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 483–498, Online. Association for Computational Linguistics.

sciendo
DOI 10.2478/jazcas-2025-0025

ANNOTATED SLOVAK DATASETS FOR TOXICITY, HATE SPEECH, AND SENTIMENT ANALYSIS

ZUZANA SOKOLOVÁ[1] – MAROŠ HARAHUS[2] – DANIEL HLÁDEK[3] – JÁN STAŠ[4]
[1]Department of Electronics and Multimedia Communications, Faculty of Electrical Engineering and Informatics, Technical University, Košice, Slovakia
(ORCID: 0000-0002-2337-8749)
[2]Department of Electronics and Multimedia Communications, Faculty of Electrical Engineering and Informatics, Technical University, Košice, Slovakia
(ORCID: 0000-0002-1756-123X)
[3]Department of Electronics and Multimedia Communications, Faculty of Electrical Engineering and Informatics, Technical University, Košice, Slovakia
(ORCID: 0000-0003-1148-3194)
[4]Department of Electronics and Multimedia Communications, Faculty of Electrical Engineering and Informatics, Technical University, Košice, Slovakia
(ORCID: 0000-0001-7403-0235)

SOKOLOVÁ, Zuzana – HARAHUS, Maroš – HLÁDEK, Daniel – STAŠ, Ján: Annotated Slovak Datasets for Toxicity, Hate Speech, and Sentiment Analysis. Journal of Linguistics, 2025, Vol. 76, No 1, pp. 279 – 289.

Abstract: The rise of social media has led to an increase in toxic language, hate speech, and offensive content. While extensive research exists for widely spoken languages like English, Slovak remains underrepresented due to the lack of high-quality datasets. This gap limits the development of effective models for toxicity detection and sentiment analysis in Slovak. To address this, we introduce three new annotated Slovak datasets focused on toxic language, offensive language, hate speech detection, and sentiment analysis. These native datasets provide a more reliable foundation for automated moderation compared to machine-translated alternatives. Our research also highlights the real-world impact of online toxicity, including social polarization and psychological distress, emphasizing the need for proactive detection systems on social media platforms. This paper reviews existing Slovak datasets, presents our newly developed resources, and provides a comparative analysis. Finally, we outline key contributions and suggest future directions for improving toxic language detection in Slovak.

Keywords: datasets, hate speech, natural language processing, sentiment analysis, Slovak language, toxic language

1 INTRODUCTION

Natural language processing (NLP) is gaining popularity, driven by the increasing online presence of people and their active use of social media to discuss various topics like politics, the climate crisis, celebrity manners, movie reviews and the like. Unfortunately, these platforms, instead of fostering constructive discussions, are becoming toxic environments filled with hate speech and hostility.

Rising rivalry, arrogance, and resentment among users contribute to social polarization. Social media comments often turn into arguments, insults, and attempts to prove superiority. Detecting toxicity is crucial to mitigating these negative effects and promoting a healthier online space.

Negative online behaviour can have serious consequences, including mental health issues, self-harm, and substance abuse (Chen 2023; Park 2024; Stroińska 2020). Addressing this issue is essential to reducing harmful speech and creating a safer digital environment. Our research focuses on detecting hate speech and offensive language on social media, aiming to foster respectful and constructive discussions online primarily in the Slovak language.

The rapid expansion of online communication and social media has led to a surge in toxic language, hate speech, and offensive expressions. While numerous studies have focused on detecting such language in widely spoken languages like English, research on Slovak remains scarce. The absence of high-quality Slovak datasets significantly limits the development and evaluation of models for detecting toxicity, offensive speech, and hate speech, as well as sentiment analysis in this language.

Social media platforms, such as Facebook and X, primarily rely on user-reported content to handle harmful language, rather than proactively addressing the issue. However, with access to vast amounts of textual data, these platforms have the potential to implement more effective automated detection systems. Detecting harmful language is essential not only for reducing online toxicity but also for mitigating its real-world consequences, including social polarization, psychological distress, and hate-driven violence. At the same time, we focus on creating native Slovak datasets because machine-translated datasets still do not achieve the same level of effectiveness, as discussed in Sokolová et al. (2023).

To address this gap, we introduce three new annotated Slovak datasets focused on toxic language, offensive language, and hate speech detection, as well as sentiment analysis. These datasets are designed to support the development of robust models tailored to the Slovak language, enabling more effective moderation and analysis of online discourse. Our paper contributes to the growing need for multilingual NLP resources and aims to foster a healthier and safer online environment.

In Section 1 we briefly outline the motivation for creating datasets in the Slovak language and emphasized why machine translation of datasets remains inefficient. Section 2 focuses on existing publicly available Slovak datasets related to toxic language, hate speech, offensive language, and sentiment analysis. As part of study Sokolová (2024), two annotated datasets were created—one for toxic language and another for sentiment analysis. Additionally, annotated a hate speech dataset was developed as part of a bachelor thesis (Ferko 2024). All datasets are introduced and compared in detail in Section 3, while Section 4 summarizes the scientific contributions of this paper and suggests future directions in detecting toxic language, hate speech, offensive language, and sentiment analysis.

2 RELATED WORK

2.1 Comparison of global datasets

Datasets of textual data worldwide focus on multiple categories of hate speech. In Tab. 1, we present the most well-known, widely used, and verified datasets intended for hate speech detection. However, examples of hate speech in some datasets are not entirely clear, such as the text dataset by Waseem and Hovy (2016) or hierarchical datasets. Moreover, these datasets are of low quality because they are not regularly updated, even though X users adopt new phrases or abbreviations. Additionally, approximately 60% of dataset creators found agreement among annotators (Poletto et al. 2021). Therefore, a useful predictive detection model for hate speech requires relevant and up-to-date datasets. The maturity of datasets is considered a unique challenge for top-quality systems.

According to Kocoń et al. (2021), the separation of annotator groups has a significant impact on the performance of hate detection systems. They also stated that group consensus affects recognition quality. It has been demonstrated that the identity of people who publish tweets introduces bias into the dataset, making it difficult to compile and ensure the quality of negative data. This means that implicit hate speech is therefore difficult to measure (Wiegand et al. 2021). Additionally, many datasets overlap between class labels, as shown by Waseem (2016), who found an overlap of 2,876 tweets between the Waseem and Hovy dataset.

In their analysis, Alkomah and Ma (2022) showed that research requires more robust, reliable, and extensive datasets due to the broad applications of hate speech detection. Vashistha and Zubiaga (2020) created a robust and massive dataset by combining four well-known datasets. Their merged dataset included HASOC (Mandl et al. 2019) and SemEval, which are among the most popular datasets. HASOC is divided into three sub-tasks:

- the first focuses on identifying hate speech and offensive language,
- the second focuses on identifying the type of hate speech,
- the third focuses on identifying the target group (or individuals) of hate speech.

Basile et al. (2019) focused on multilingual hate speech detection against immigrants and women on the X platform using the SemEval Task 5 dataset. Zampieri et al. (2019a), in their study addresses the identification and categorization of offensive language on social media using the SemEval Task 6 dataset. The latest OLID dataset (Zampieri et al. 2019b) for offensive language identification contains over 14,000 English tweets and is aimed at similar tasks as the HASOC dataset. The HASOC 2020 dataset (Mandl et al. 2020b) contains only 3,708 English tweet samples, but is considered substantial and competitive.

Mishra et al. (2020) achieved an F1 score of 51.52% in the first task for English when classifying tweets into two categories: whether a tweet is hateful and offensive or the opposite. In the second task, they achieved an F1 score of 23.41%, where tweets (labelled as hateful and offensive in the first task) were classified into three categories: hateful, offensive, and disrespectful.

ElSherief et al. (2018), in their study, compiled a dataset for hate speech containing 27,330 tweets. They also managed to extract 25,278 instigators of hate speech and 22,287 target accounts. Their research focused on comparing hate speech instigators, their targets, and general X users. They found that hate instigators tend to target more visible users and that participation in hateful discussions is associated with higher visibility. Additionally, it was shown that both instigators and targets of hate have unique personality traits that may contribute to hate speech, such as anger or depression.

Davidson et al. (2017), in their study, classified textual data into three categories (hateful, offensive, neutral). They found that racist and homophobic tweets are more likely to be classified as hate speech, whereas sexist tweets are generally classified as offensive. Other studies that also focus on dataset creation and classification are listed in Tab. 1, along with the corresponding categories and the number of tweets.

2.2 Comparison of Slovak datasets

The detection of toxicity, meaning the identification of hate speech and offensive language in the Slovak language, has so far been the subject of very few scientific studies. In Tab. 2 we have listed the available corpora of textual data in Slovak, where the focus of individual datasets and their size can also be seen. Most commonly, authors have classified hate speech into two categories (hateful, neutral). Alternatively, datasets have been divided into three categories (positive, negative, neutral) or even four categories (neutral, mildly toxic, moderately toxic, and highly toxic).

Author / Dataset Name / Reference	Dataset Size (No. Tweets)	Dataset Categories
Waseem and Hovy (2016a)	16,000	Racism, Sexism, Neither
Waseem et al. (2016b)	6,909	Racism, Sexism, Neither, Both
Davidson et al. (2017)	24,783	Hateful, Offensive, Neither
Harassment (Golbeck et al. 2017)	35,000	Harassing, Neutral
Twitter & Reddit SA (Shen and Rudzicz 2017)	162,980 & 37,249	Positive, Neutral, or Negative

ElSherief et al. (2018)	27,330	Archaic, Class-based, Disability, Ethnicity, Gender, Religion, Sexual Orientation
Founta et al. (2018)	80,000	Offensive, Abusive, Hateful, Aggressive, Cyberbullying, Spam, Normal
Amievalita (Fersini et al. 2018)	4,000	Misogynistic, Discrediting, Sexual Harassment, Stereotype, Dominance
Women (Fersini et al. 2018)	3,977	Misogyny, Stereotype, Dominance, Sexual Harassment, Discrediting, Misogyny Target
OLID (Zampieri et al. 2019a)	14,000	Offensive, Non-offensive, Targeted Insults. Individual, Group
L-HSAB (Mulki et al. 2019)	5,846	Hateful, Offensive, Normal, Targeted
HASOC (Mandl et al. 2019)	5,335	Hateful and Non-offensive
	7005	Hateful, Offensive, Vulgar
Ousidhoum et al. (2019)	5,647	Hateful, Offensive, Neither, Directness, Hostility, Target
MMHS150K (Winter et al. 2019)	150,000	Neutral, Religion, Sexism, Racism, Homophobia, Other Hate
AbusEval (Caselli et al. 2020)	18,740	Offensive, Non-offensive, Targeted, Non-targeted, Explicitly Insulting, Implicitly Insulting, Non-insulting
HatEval (Yang et al. 2020)	13,000	Hateful, Neutral, Individual Target, Group Target
HateXplain (Mathew et al. 2021)	20,148	Hateful, Offensive, Normal
Sentiment Analysis (Shrivastava 2023)	905,874	Positive, Negative
Flipkart (Vaghani et al. 2023)	205,053	Positive, Neutral, or Negative
Youtube Statistics (Patil 2023)	19,658	Positive, Negative, Neutral

Tab. 1. Comparison of Global Corpora

Author / Dataset Name / Reference	**Dataset Size (No. Tweets)**	**Dataset Categories**
Sentigrade (Krchnavy and Simko 2017)	1,584	Positive, Negative, Neutral
Švec et al. (2018)	80,000	Hateful, Neutral
Machová et al. (2022a)	24,000	Positive, Negative, Neutral
Machová et al. (2022b)	3,092	Neutral, Mildly Toxic, Moderately Toxic, Very Toxic
Mojžiš and Kvassay (2022)	2,283	Hateful, Neutral
	10,000	Hateful, Neutral
Papcunová et al. (2023)	283	Hateful, Neutral

Tab. 2. Comparison of Slovak Corpora

3 DATASETS

In machine learning tasks, a dataset is required to train a model for performing various machine learning or deep learning tasks. The reason why a dataset is necessary is that machine learning heavily depends on data. Without data, artificial intelligence cannot learn, making it the most important aspect that enables the training of machine learning algorithms. Regardless of the skills or knowledge of the team and the size of the dataset, if the dataset is not of sufficient quality, the entire artificial intelligence project will not achieve satisfactory results.

Criteria	Value
Number of Annotators	7
Age	25–40
Gender	Women and Men
Education	PhD Students and Research Assistants From DEMC

Tab. 3. Basic characteristics of the annotators of ToxicSK and SentiSK datasets

Criteria	Value
Number of Annotators	60
Age	18–22
Gender	Women and Men
Education	1st and 2nd Year Bachelor's Students

Tab. 4. Basic characteristics of the annotators of hate_speech_slovak dataset

When working with artificial intelligence, we largely rely on the dataset. From training, tuning, model selection, to testing, we use a dataset divided into three sets: training, validation, and test sets. The training set is used to train the model, the validation set is used to adjust weights and fine-tune the model, and the test set is used to evaluate the trained model. Often, simply gathering data is not enough; on the contrary, in most artificial intelligence tasks, classifying and annotating the dataset takes the majority of the time, especially for corpora that are sufficiently accurate to reflect a realistic vision of the world.

In this section, we present the created datasets SentiSK, ToxicSK, and hate_speech_slovak. In Tab. 3, we provided the basic characteristics of the annotators who participated in annotating the created ToxicSK and SentiSK datasets. In Tab. 4, we outlined the key characteristics of the annotators involved in labeling the hate_speech_slovak dataset. All comments contained in these datasets were obtained through our custom-developed web scraping tool and were publicly accessible at the time of collection. The preprocessing pipeline involved the removal of duplicate entries and URLs.

3.1 Dataset: ToxicSK

The ToxicSK dataset (TUKE-KEMT/toxic-sk 2024) was created as part of a research task focused on detecting toxicity on social media. We focused on the Slovak language. The comments is a collection of public posts on the Facebook social network.

The collected comments were annotated using the Prodigy tool into two categories: toxic (1) and non-toxic (0). The ToxicSK dataset is class-balanced and contains 4,420 toxic and 4,420 non-toxic comments.

Dataset	ToxicSK
Number of Comments	8,840
Number of Categories	2
Type of Categories	Toxic (1), Non-toxic (0)
Number of Negative Comments	4,420
Number of Positive Comments	4,420
Number of Words	89,756
Number of Characters	476,170
Average Number of Words per Sentence	10.15
Number of Unique Words	18,883
Number of Unique Words	11,602
Number of Stopwords	20,958
Data Source	Facebook

Tab. 5. Specification of the ToxicSK dataset

3.2 Dataset: hate_speech_slovak

The hate_speech_slovak dataset (TUKE-KEMT/hate_speech_slovak 2024) contains posts from a social network that have been annotated by humans. Each post is labelled by 1, if contains hateful or offensive language, and by 0 if not. The data was collected from a variety of public pages on topics such as sports, politics, and general discussions. To ensure the quality of the data, the collected posts underwent a cleaning process using text clustering. The annotations were provided by a group of students from the Technical University of Košice in Slovakia.

To maintain reliability, the dataset underwent a filtering process to remove annotations from users who showed a low level of agreement with others. Annotations were evaluated based on a scoring system: annotators received positive points when their annotations aligned with others and negative points when they differed. Any annotator with a low agreement ratio (below 70%) was excluded from the dataset. Additionally, for each post, votes for the positive, neutral, and negative categories were calculated from the remaining reliable annotators, with posts where the neutral class was the majority being discarded. Despite these efforts, some bias remains in the dataset due to the personal opinions of the annotators. For most items, the class was determined by the votes of trustworthy annotators, but in some cases, items had only a single vote.

Dataset	**hate_speech_slovak**
Number of Comments	13,189
Number of Categories	2
Type of Categories	Hate Speech (1), Neutral (0)
Number of Hate Speech Comments	3,605
Number of Neutral Comments	9,584
Number of Sentences	11,870
Number of Words	218,984
Number of Characters	1,130,860
Average Number of Words per Sentence	18.45
Number of Unique Words	42,031
Number of Unique Words	28,649
Number of Stopwords	50,151
Data Source	Facebook

Tab. 6. Specification of the hate_speech_slovak dataset

3.3 Dataset: SentiSK

The SentiSK dataset (TUKE-KEMT/senti-sk 2024) was created as part of research focused on sentiment analysis in the Slovak language. The SentiSK dataset contains 34,006 comments from the social media platform Facebook. The comments were collected using a Python tool for extracting data from websites, specifically comments under posts by three Slovak politicians. Data preprocessing involved cleaning the text of unwanted characters, as well as removing empty lines, extra spaces, periods, etc. The NLTK library was used for preprocessing. The dataset was annotated using the Prodigy annotation tool provided by the Department of Electronics and Multimedia Communications (DEMC). The SentiSK dataset was annotated into three sentiment categories: 20,668 negative comments, 9,581 neutral comments, and 3,779 positive comments. The distribution of comments in these categories indicates that the SentiSK dataset is class-imbalanced. Since the data were taken from the posts by Slovak politicians, there was a high number of negative comments.

Dataset	**SentiSK**
Number of Comments	34,006
Number of Categories	3
Type of Categories	Negative, Neutral, Positive
Number of Negative Comments	20,668
Number of Neutral Comments	9,581
Number of Positive Comments	3,779
Number of Words	401,937

Number of Characters	2,213,773
Average Number of Words per Sentence	11.82
Number of Unique Words	65,049
Number of Unique Words	43,365
Number of Stopwords	90,376
Data Source	Facebook

Tab. 7. Specification of the SentiSK dataset

4 CONCLUSION

This research highlights the growing need for Slovak-specific datasets in the field of toxic language, hate speech, and sentiment analysis. While significant progress has been made in detecting harmful language in widely spoken languages, Slovak remains underexplored, limiting the effectiveness of moderation systems (Cao et al. 2023; Jaggi et al. 2024; Lee et al. 2024; Hee et al. 2024). By analyzing 26 existing datasets and introducing three new annotated datasets—ToxicSK, SentiSK, and hate_speech_slovak—we contribute to closing this gap and provide a solid foundation for future advancements in Slovak NLP.

Our findings emphasize that native datasets significantly improve detection accuracy compared to machine-translated alternatives. Furthermore, we underscore the importance of automated detection systems in combating online toxicity and its real-world consequences. Moving forward, future research should focus on expanding dataset size, improving annotation consistency, and integrating advanced machine learning techniques to enhance detection models.

Given recent advances in large language models, future research should consider leveraging pre-trained and instruction-finetuned LLMs for toxicity detection in Slovak, as these approaches may offer improved performance even in under-resourced settings.

By fostering a more robust NLP ecosystem for Slovak, this work aims to support safer and healthier online interactions while contributing to multilingual NLP advancements.

ACKNOWLEDGEMENTS

The research in this paper was supported by the Ministry of Education, Research, Development, and Youth of the Slovak Republic under the research projects KEGA 049TUKE-4/2024 and KEGA 041TUKE-4/2025, and by the Slovak Research and Development Agency under the research projects APVV-22-0261 and APVV-22-0414. There was no additional external funding received for this study.

References

Alkomah, F., and Ma, X. (2022). A literature review of textual hate speech detection methods and datasets. Information, 13(6), 273 p.

Basile, V., Bosco, C., Fersini, E., Nozza, D., Patti, V., Pardo, F. M. R., ... and Sanguinetti, M. (2019). Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In Proceedings of the 13th international workshop on semantic evaluation, pp. 54–63.

Cao, Y. T., Domingo, L. F., Gilbert, S. A., Mazurek, M., Shilton, K., and Daumé III, H. (2023). Toxicity detection is not all you need: Measuring the gaps to supporting volunteer content moderators. Accessible at: arXiv preprint arXiv:2311.07879.

Caselli, T., Basile, V., Mitrović, J., Kartoziya, I., and Granitzer, M. (2020, May). I feel offended, don't be abusive! implicit/explicit messages in offensive and abusive language. In Proceedings of the twelfth language resources and evaluation conference, pp. 6193–6202.

Chen, M. B., Lau, J. H., and Frermann, L. (2023). The uncivil empathy: Investigating the relation between empathy and toxicity in online mental health support forums. In Proceedings of the 21st Annual Workshop of the Australasian Language Technology Association, pp. 136–147.

Davidson, T., Warmsley, D., Macy, M., and Weber, I. (2017). Automated hate speech detection and the problem of offensive language. In Proceedings of the international AAAI conference on web and social media, 11(1), pp. 512–515.

ElSherief, M., Nilizadeh, S., Nguyen, D., Vigna, G., and Belding, E. (2018). Peer to peer hate: Hate speech instigators and their targets. In Proceedings of the International AAAI Conference on Web and Social Media, 12(1).

Ferko, V., (2024). Anotácia a vyhodnotenie slovenskej databázy nenávistnej reči. Košice: Technická univerzita v Košiciach, Fakulta elektrotechniky a informatiky, 55 p. Vedúci práce: doc. Ing. Daniel Hládek, PhD.

Fersini, E., Nozza, D., and Rosso, P. (2018). Overview of the evalita 2018 task on automatic misogyny identification (ami). In CEUR workshop proceedings, Vol. 2263, pp. 1–9. CEUR-WS.

Founta, A., Djouvas, C., Chatzakou, D., Leontiadis, I., Blackburn, J., Stringhini, G., ... and Kourtellis, N. (2018). Large scale crowdsourcing and characterization of twitter abusive behavior. In Proceedings of the international AAAI conference on web and social media, 12(1).

Golbeck, J., Ashktorab, Z., Banjo, R. O., Berlinger, A., Bhagwan, S., Buntain, C., ... and Wu, D. M. (2017, June). A large labeled corpus for online harassment research. In Proceedings of the 2017 ACM on web science conference, pp. 229–233.

Hee, M. S., Sharma, S., Cao, R., Nandi, P., Nakov, P., Chakraborty, T., and Lee, R. (2024). Recent advances in online hate speech moderation: Multimodality and the role of large models. Findings of the Association for Computational Linguistics: EMNLP 2024, pp. 4407–4419.

Jaggi, H., Murali, K., Fleisig, E., and Bıyık, E. (2024). Accurate and Data-Efficient Toxicity Prediction when Annotators Disagree. Accessible at: arXiv preprint arXiv:2410.12217.

Kocoń, J., Figas, A., Gruza, M., Puchalska, D., Kajdanowicz, T., and Kazienko, P. (2021). Offensive, aggressive, and hate speech analysis: From data-centric to human-centered approach. Information Processing & Management, 58(5), 102643.

Krchnavy, R., and Simko, M. (2017). Sentiment analysis of social network posts in Slovak language. In 2017 12th International Workshop on Semantic and Social Media Adaptation and Personalization (SMAP), pp. 20–25.

Kvassay, M. (2022). New Public Dataset for Classification of Inappropriate Comments in Slovak language. In 2022 20th International Conference on Emerging eLearning Technologies and Applications (ICETA), pp. 437–441.

Lee, N., Jung, C., Myung, J., Jin, J., Camacho-Collados, J., Kim, J., and Oh, A. (2023). Exploring cross-cultural differences in English hate speech annotations: From dataset construction to analysis. Accessible at: arXiv preprint arXiv:2308.16705.

Machová, K., Mach, M., and Vasilko, M. (2022). Recognition of toxicity of reviews in online discussions. Acta Polytechnica Hungarica, 19(4).

Machová, K., Mach, M., and Adamišín, K. (2022). Machine learning and lexicon approach to texts processing in the detection of degrees of toxicity in online discussions. Sensors, 22(17), 6468.

Mandl, T., Modha, S., Majumder, P., Patel, D., Dave, M., Mandlia, C., and Patel, A. (2019). Overview of the hasoc track at fire 2019: Hate speech and offensive content identification in indo-european languages. In Proceedings of the 11th annual meeting of the Forum for Information Retrieval Evaluation, pp. 14–17.

Mandl, T., Modha, S., Kumar M, A., and Chakravarthi, B. R. (2020). Overview of the hasoc track at fire 2020: Hate speech and offensive language identification in tamil, malayalam, hindi, english and german. In Proceedings of the 12th annual meeting of the forum for information retrieval evaluation, pp. 29–32.

Mathew, B., Saha, P., Yimam, S. M., Biemann, C., Goyal, P., and Mukherjee, A. (2021). Hatexplain: A benchmark dataset for explainable hate speech detection. In Proceedings of the AAAI conference on artificial intelligence, 35(17), pp. 14867–14875.

Mishra, A. K., Saumya, S., and Kumar, A. (2020). IIIT_DWD@ HASOC 2020: Identifying offensive content in Indo-European languages. In FIRE (working notes), pp. 139–144).

Mulki, H., Haddad, H., Ali, C. B., and Alshabani, H. (2019). L-hsab: A levantine twitter dataset for hate speech and abusive language. In Proceedings of the third workshop on abusive language online, pp. 111–118.

Ousidhoum, N., Lin, Z., Zhang, H., Song, Y., and Yeung, D. Y. (2019). Multilingual and multi-aspect hate speech analysis. Accessible at: arXiv preprint arXiv:1908.11049.

Papcunová, J., Martončik, M., Fedáková, D., Kentoš, M., Bozogáňová, M., Srba, I., ... and Adamkovič, M. (2023). Hate speech operationalization: a preliminary examination of hate speech indicators and their structure. Complex & intelligent systems, 9(3), pp. 2827–2842.

Park, K., Baik, M. J., Hwang, Y., Shin, Y., Lee, H., Lee, R., ... and Park, S. (2024). Harmful Suicide Content Detection. Accessible at: arXiv preprint arXiv:2407.13942.

Patil, A., (2023). Youtube Statistics, Accessible at: https://www.kaggle.com/datasets/advaypatil/youtube-statistics.

Poletto, F., Basile, V., Sanguinetti, M., Bosco, C., and Patti, V. (2021). Resources and benchmark corpora for hate speech detection: a systematic review. Language Resources and Evaluation, 55, pp. 477–523.

sciendo
DOI 10.2478/jazcas-2025-0026

FROM RULE-BASED PROOFREADER BETA OPRAVIDLO TO AI-POWERED OPRAVIDLO 2.0

HANA ŽIŽKOVÁ[1] – ZUZANA NEVĚŘILOVÁ[2] – JAKUB MACHURA[3] – ALEŠ HORÁK[4] – DANA HLAVÁČKOVÁ[5] – PATRIK STANO[6]

[1]Department of Czech Language, Faculty of Arts, Masaryk University, Brno, Czech Republic (ORCID: 0000-0002-6483-6603)
[2]Department of Czech Language, Faculty of Arts, Masaryk University, Brno, Czech Republic (ORCID: 0000-0002-7133-9269)
[3]Department of Czech Language, Faculty of Arts, Masaryk University, Brno, Czech Republic (ORCID: 0000-0002-6623-3064)
[4]Department of Machine Learning and Data Processing, Faculty of Informatics, Masaryk University, Brno, Czech Republic (ORCID: 0000-0001-6348-109X)
[5]Department of Czech Language, Faculty of Arts, Masaryk University, Brno, Czech Republic (ORCID: 0000-0002-9918-0958)
[6]Department of Machine Learning and Data Processing, Faculty of Informatics, Masaryk University, Brno, Czech Republic (ORCID: 0009-0001-8339-6084)

ŽIŽKOVÁ, Hana – NEVĚŘILOVÁ, Zuzana – MACHURA, Jakub – HORÁK, Aleš – HLAVÁČKOVÁ, Dana – STANO, Patrik: From Rule-based Proofreader Beta Opravidlo to AI-powered Opravidlo 2.0. Journal of Linguistics, 2025, Vol. 76, No 1, pp. 290 – 299.

Abstract: The demand for accurate and error-free written communication in Czech has led to the development of automated proofreading tools. Beta Opravidlo, a rule-based online proofreader launched in 2022, demonstrated high precision and recall in correcting Czech texts. However, its reliance on predefined linguistic rules limited recall and processing speed. With advancements in machine learning and large language models (LLMs), a transition toward AI-powered proofreading became necessary. This article explores the evolution from Beta Opravidlo to Opravidlo 2.0, integrating deep neural networks to enhance correction capabilities. We discuss proofreading as a machine learning task, compare traditional rule-based and neural approaches, and challenges such as explainability, system integration or computational requirements. The most effective solution is a hybrid approach combining rule-based precision with AI-driven adaptability. Opravidlo 2.0 aims to improve recall, optimize inference time, and extend support to other Slavic languages. This interdisciplinary effort highlights the potential of AI-powered proofreading to set new standards in language correction and usability.

Keywords: Opravidlo, proofreader, Czech language, AI-powered proofreader

1 INTRODUCTION

The ability to express oneself in written and spoken language without linguistic errors is required and positively evaluated in the Czech environment. Readers often look at authors of texts who commit spelling mistakes with derision and distrust, and texts with spelling mistakes reduce the author’s credibility. Texts matter greatly for their correctness, even today are checked by human proofreaders. However, a human proofreader is not always convenient for various reasons: it is expensive and slow. Thus, in recent decades, all sorts of automatic tools have been developed to proofread text. This was not an easy task because Czech is an inflectional language with a lot of homonymy. One proofreader presented to the public is the rule-based online proofreader Opravidlo (Hlaváčková et al. 2022). Its beta version was released in mid-2022, and at that time, it was the proofreader with the highest precision and best recall. However, the situation changed with the advent of large language models (LLMs) and machine learning, whose concepts proved functional for correcting languages like Czech. The following article describes the starting point for Beta Opravidlo and the possibilities for AI-powered Opravidlo 2.0.

2 BETA OPRAVIDLO

Beta Opravidlo is an online proofreader freely available at www.opravidlo.cz. It was published in May 2022 and was created thanks to a project funded by TA ČR. The project was carried out in cooperation with three academic departments: Masaryk University, the Institute for Czech Language of the CAS, and Charles University. The team also included business partners Seznam.cz and Wikimedia ČR.

The project course was based on the experts’ cooperation, knowledge-sharing, language data, software, hardware and some existing solutions to partial problems in developing the language proofreader. It also had the advantage of involving experienced experts (mostly linguists) and promising PhD students in computational linguistics in one project. The project’s interdisciplinary nature brought together the knowledge of linguists, computational linguists and programmers, forming an ideal basis for developing the language proofreader.

The freely accessible web interface allows users to type or insert Czech text, and the right-hand side contains suggestions for corrections. The user can decide whether or not to accept the correction. Some typographical corrections are made automatically without user intervention. The correction refers to an explanation of the phenomenon found in the Internet Language Reference Book (2025) for complex topics like agreement or punctuation.

Since May 2022, the public has used the proofreader, with more than 200,000 correction requests per month, however, the low recall is perceived as a drawback by its users.

2.1 How Beta Opravidlo works

The Beta Opravidlo works by first decomposing the inserted text into tokens. The Unitok tool is used for this. The MorphoDita tagger or majka tagger performs the subsequent morphological analysis. The choice of tagger depends on the subsequent processing. Considering that Beta Opravidlo is a rule-based system, a total of 7500 rules are included in the tool, divided into several thematic modules. In most modules, the detection phase is performed by the SET parser (Syntax Elements of Text) (Kovář et al. 2011). This parser is primarily designed as a universal, language-independent syntactic parser. It processes the input text into a tree structure according to a specific error-detection grammar. This grammar is defined in a text file containing various rules written in a specific format. However, as the example below demonstrates, the rules created for SET are not limited to syntactic parsing alone. The following rule captures the situation where the word *"více"* '*more'* is, incorrectly, directly followed by a comparative adjective without the conjunction *"než" 'than',* unless this is remedied by a noun in the genitive plural as a third word in the prepositional phrase:

```
TMPL:(word více)(tag k2.*d2.*)
MARK 0 DEP 1 LABEL <komparativ-nok> PROB 100
TMPL:(word více)(tag k2.*d2.*)(tag kl.*nP.*c2.*)
MARK 0 DEP 1 LABEL <komparativ-ok> PROB 400
```

The first rule marks the word *"více"* '*more*' as redundant in following sentence: *"Bylo to více horší, než jsme čekali."* '*It was more worse than we expected.'* The second rule applies to grammatically correct sentences of the type: *"Dostal více těžších úkolů." 'He was given more difficult tasks.'*

The Beta Opravidlo contains the following modules:

- punctuation,
- non-grammatical structures,
- spelling,
- spelling in context,
- agreement,
- typography,
- dependent clauses,
- capital letters,
- preposition vocalisation,
- pronouns,
- other errors.

This modular system's advantage is that it can analyse the text in parallel. However, evaluating an error still takes several tens of seconds, which is a significant disadvantage.

The exact list of language phenomena that Beta Opravidlo can correct is available on this page https://www.opravidlo.cz/co-korektor-umi.html. The precision of the individual modules ranges from 91% to 96%, which we consider an excellent result. Regarding recall, given the nature of the Czech language, the lowest coverage is 40%, and the best is 80%. We have found that the success rate of recall is highly dependent on the nature of the text and the number of errors in the text.

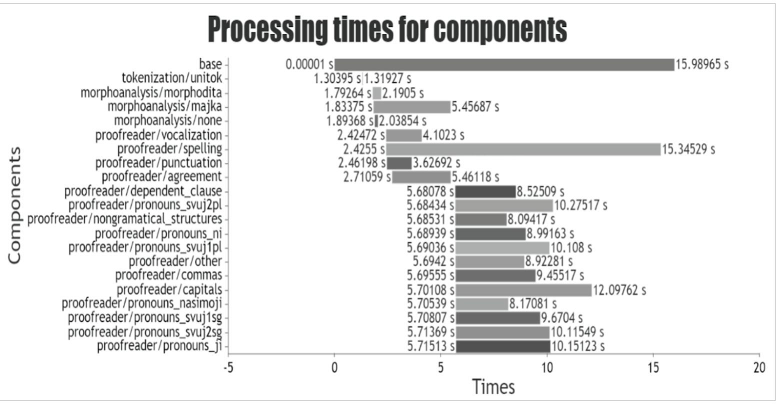


Fig. 1. Efficiency of Beta Opravidlo (Mrkývka 2024)

3 FROM BETA OPRAVIDLO TO OPRAVIDLO 2.0

We received valuable feedback from users while testing the Beta Opravidlo before its release. On the one hand, it was very positively evaluated that we follow the rule-based path in developing the proofreader. On the other hand, the question of incorporating machine learning and using neural networks was discussed, which was not planned in the project. When we started developing the Beta Opravidlo in 2019, it was still unclear what progress machine learning would make, yet we considered incorporating machine learning into some modules. Experimentally, we performed comparisons of punctuation insertion, and it turned out that neural networks showed higher recall and, on some texts, higher precission (Machura et al. 2022). In mid-2022, we believed the issue stemmed from specific errors that neural networks failed to correct, assuming they simply required retraining. In the following years, and thanks to experimental studies, it turned out that neural networks could be very effective in correcting Czech texts (Machura et al. 2023; Medková and Horák 2022).

All these findings have resulted in the creation of a new interdisciplinary team, working on integrating deep neural networks into the Beta Opravidlo, with the aim of increasing its ability to correct language errors first in the Czech language, subsequently in other Slavic languages.

4 PROOFREADING AS A MACHINE LEARNING TASK

Proofreading is a complex task combining several subtasks; from the functional point-of-view, proofreading consists of:

- error detection,
- suggestions for error correction,
- explanation of the error.

From the language point-of-view, proofreading can be seen as a combination of:

- spellchecking,
- grammar correction,
- typography correction,
- style improvement.

The evolution of proofreading consists of changes in methods, coverage of text phenomena, and performance metrics used for comparison.

In the 1980s, natural language processing (NLP) addressed the "ill-formed input", where parsers struggled to output a parse tree. With the rise of machine learning, grammar correction was formulated as an NLP task: grammar error correction (GEC). In (Wang et al. 2020), GEC is described as "Errors that violate rules of English and expectation usage of English native speakers in morphological, lexical, syntactic and semantic forms are all treated as a target to be corrected." GEC systems usually get an ungrammatical sentence as input and output the corrected sentence. This approach is a sister task of machine translation. Therefore, GEC systems followed a similar path in their methods: early approaches used statistical methods, and later, neural architectures such as long short-term memory (LSTM), and convolutional neural networks (CNN). Current studies predominantly examine transformer-based models, including BERT variants and LLMs like PaLM2-XS (Liu et al. 2024).

Early proofreading systems focused only on spellchecking, solving the task with a "dictionary" – a simple wordlist for a particular language. Later, GEC systems focused on fewer errors, such as correcting prepositions (Prokofyev et al. 2014). Recent GEC systems attempt to perform comprehensive error correction. The coverage of the text phenomena is connected with the used methods: for tasks more related to language rules, rule-based systems or n-gram statistics perform well, for punctuation or fluency-related issues, a larger context is needed, and therefore, transformer-based methods yield better results.

Evaluation metrics also changed from accuracy measures to metrics for fluency and overall text quality. Particularly, GEC systems commonly use F0.5 score, GLEU, BLEU, METEOR, precision, recall, and F1 score as performance metrics, with recent work incorporating broader evaluation frameworks.

The majority of GEC systems are trained and evaluated in English. More recently, with the emergence of multilingual models, GEC for non-English texts is achieving plausible results. Multilingual transformer models, particularly mT5, show promise in handling non-English languages due to their pre-training on multilingual datasets. Successful GEC systems are developed e.g. for Arabic, a more recent work for a Slavic language is (Kholodna and Vysotska 2023). A recent and comprehensive survey on GEC can be found in (Bryant 2023).

4.1 Challenges in transition to machine learning system

While the rule-based method is effective in targeted error correction (e.g. the correct form of pronouns), it struggles with long-range dependencies or syntactic ambiguity. Transformer-based approaches are better at text understanding so that they can handle the text in a more comprehensive way, potentially with much higher recall. The targeted error correction has a strong advantage we would like to keep and develop: the errors are classified and can be explained easily. For example, if the rule-based system detects an incorrect pronoun form, it can label the error, provide a correct form, and explain the rules for pronominal inflection. With a simple machine translation-like approach, we would lose the system's ability to explain errors and teach the language users.

Currently, we perform experiments in several streams:

- filling in the punctuation,
- edit-based approaches,
- grammar error explanation methods.

4.2 Filling in the punctuation

Since missing punctuation is one of the most common errors that is also difficult to capture by rules, we focus on punctuation errors in the first phase, similarly to (Machura et al. 2023). A Czech variant of the RoBERTa model has been modified by the addition of a classifier head and fine-tuned to classify tokens based on the presence or absence of a comma. This approach can be extended to include other punctuation marks (e.g. a full stop, a question mark, an exclamation point). Other grammatical phenomena, such as casing, can be resolved similarly, possibly by the same model, by including a second classification head and fine-tuning both tasks. This approach simplifies the grammar correction task, enhancing the achieved precision and recall. However, it lacks explainability, which is critical for a reliable grammar correction service, as it provides credibility. Consequently, a classifier model is a useful secondary option in addition to an approach that provides explanations. The confidence of classification can be utilised to assess its credibility, and the confidence score could be displayed to the end user, enabling them to decide whether to accept the suggestion to add punctuation or not. This approach is to be tested and evaluated.

4.3 Edit-based approaches

In (Omelianchuk et al. 2020), the authors generate a sequence of token-level edits to perform grammatical error corrections. The advantages of such an approach are: 1) minimum intervention in users' text, and 2) explainability of the errors.

Our experimental setup is a sequence to edit architecture, where each token of the input sequence gets labels such as KEEP, DELETE, APPEND, REPLACE, and TRANSFORM. The last label is enriched with the type of transformation. Each label can be enriched by the explanation.

So far, we have used synthetic data. We introduced errors into the Czech part of the WMT dataset (Bojar et al. 2014) – we removed punctuation, added punctuation after random tokens, and converted capitalised tokens to lowercase in various combinations.

Gold standard	Pamatujte: kdo rychle dává, dvakrát dává.
Input	pamatujte: kdo, rychle dává dvakrát dává
Output labels	$MAKE_CASE_UPPER $EXTRA_COMMA $KEEP $MISSING_PUNCT_, $KEEP $MISSING_PUNCT_.

Tab. 1. Example of the training data

For the task, we fine-tuned the RobeCzech-base (Straka et al. 2021) model with 953,620 sentences, for evaluation, we used 2,999 sentences. The system achieved F1=96.7. We know punctuation and capitalisation are only a small part of the proofreading task; however, the results seem very promising, and we plan to continue with this approach.

4.4 Grammar error explanation methods

In Song et al. (2024) the authors performed a series of experiments with ChatGPT-4 to explain grammar errors in natural language. They developed a two-step pipeline that leverages fine-tuned and prompted LLMs to perform structured atomic token edit extraction, followed by prompting GPT-4 to explain each edit.

We do not plan to use generative LLMs in the production version. The main reasons are deployment costs and prediction time. However, using generative LLMs for comparison and evaluation is desirable.

4.5 A hybrid approach to proofreading

Currently, a hybrid approach seems to be beneficial. We plan to keep the rule-based approach for phenomena well covered by the rules (high precision and high recall) and at the same time, the inference time is not longer than that of a neural model. In the later development phase of the neural approaches, we will decide whether the neural model outperformed the rule-based system in some aspects. Also, an ensemble model could be made, possibly including even multiple models.

While the rule-based system is not scalable, the hardware can influence the prediction time of neural models. Currently, we plan to deploy the proofreading service on GPU servers at the Faculty of Informatics at Masaryk University. It depends on the possibilities of large research infrastructures such as CLARIN whether a future GPU deployment would be possible.

5 PREPARATION OF TESTING DATA

For the purpose of training models for automatic comma insertion, a comprehensive analysis of comma distribution was conducted using the SYN2020 corpus (Křen et al. 2020). This corpus contains just over 8 million commas. The study focused on the following aspects:

a. The most common lexical contexts, specifically:

a) the most frequent expressions that appear immediately after commas,

b) the most common particles and interjections that occur before commas, and

c) the most frequent vocative phrases, which must be separated by a comma as they lie outside the syntactic structure of the sentence.

The aim of this part of the analysis was to determine the actual distribution of commas in relation to specific lexical expressions—specifically, how frequently a given expression appears with a preceding comma and in what proportion it occurs without one. For instance, the word "že" 'that' appears 951,302 times in the SYN2020 corpus, with 902,351 of these instances (94.85%) following a comma. The remaining 5.15% occur without a preceding comma.

The goal is to construct a testing dataset that accurately reflects these proportions. Hypothetically, if the dataset contained 100,000 commas, 11.25% of those commas would be followed by the expression *"že" 'that'*, and additionally, 5.15% of the total occurrences of *"že" 'that'* would appear without a preceding comma. This approach ensures realistic representation of the expression *"že" 'that'* and allows the model to learn both its typical usage with commas and less frequent instances without them.

b. Proportion of sentences of a given token length with and without commas

This parameter examines the ratio of sentences of a specific length (in number of tokens) that contain at least one comma versus those of the same length that contain none. The analysis is based on the SYN2020 corpus, which includes 6,791,880 sentences ending with a period, exclamation mark, or question mark.

The most frequent sentence length in the corpus is 9 tokens (identified using the CQL query: `<s> [word!=″\.|\!|\?″]{9} [word=″\.|\!|\?″] </s>)`, totaling 323,369 sentences—representing 4.76% of all sentences. Of these, 129,130 sentences contain at least one comma (1.90% of all sentences), while 194,239 contain no commas (2.86%).

To ensure the testing dataset reflects these proportions, the same ratios are applied. Hypothetically, if the testing dataset includes 100,000 sentences, 1.90% should be 9-token sentences that include a comma, and 2.86% should be 9-token sentences without any comma. These ratios were similarly calculated for all sentence lengths ranging from 1 to 45 tokens.

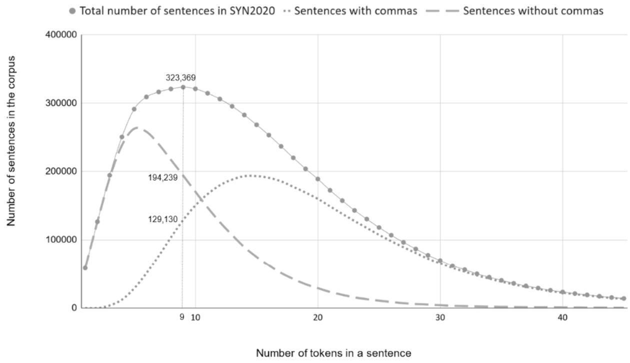


Fig. 2. Proportion of sentences of a given token length with and without commas

6 CONCLUSION

The development of Beta Opravidlo has demonstrated the strengths of rule-based proofreading, particularly in precision and linguistic transparency. However, the advent of machine learning and neural networks offers significant potential to improve recall and overall correction effectiveness. By integrating AI-driven approaches, Opravidlo 2.0 aims to enhance recall, provide high precision, and expand its capabilities beyond Czech to other Slavic languages.

Despite challenges such as data availability, explainability, and system integration, a hybrid approach combining rule-based and machine-learning methods appears to be the most promising solution. Future research and development efforts will focus on refining this hybrid model, optimising deployment infrastructure, and ensuring a seamless user experience. With continued interdisciplinary collaboration, Opravidlo 2.0 has the potential to set a new standard in automated proofreading for complex languages like Czech.

ACKNOWLEDGEMENTS

The authors acknowledge that this work was supported by the OSCARS project, funded by the European Commission's Horizon Europe Research and Innovation programme (grant agreement No. 101129751), led by the five Science Clusters: ENVRI, ESCAPE, LS RI, PaNOSC, and SSHOC.

References

Bojar, O. et al. (2014). Findings of the 2014 Workshop on Statistical Machine Translation. In Proceedings of the Ninth Workshop on Statistical Machine Translation, Baltimore, Maryland, USA. ACL, pp. 12–58.

Bryant, Ch. et al. (2023). GEC: A Survey of the State of the Art. Computational Linguistics 2023; 49(3), pp. 643–701. Accessible at: https://doi.org/10.1162/coli_a_00478.

Internet Language Reference Book (2025). Praha: ÚJČ AV ČR.

Hlaváčková D. et al. (2022). Opravidlo.

Kholodna, N., and Vysotska, V. (2023). Technology for grammatical errors correction in Ukrainian text content based on machine learning methods. Radio Electronics, Computer Science, Control, (1), 114. Accessible at: https://doi.org/10.15588/1607-3274-2023-1-12.

Kovář, V. et al. (2011). Syntactic Analysis Using Finite Patterns: A New Parsing System for Czech. In Human Language Technology. Challenges for Computer Science and Linguistics. Berlin/Heidelberg: Springer, pp. 161–171. Accessible at: http://dx.doi.org/10.1007/978-3-642-20095-3_15.

Křen, M. et al. (2020). SYN2020: A representative corpus of written Czech. UCNK FF UK. Accessible at: http://www.korpus.cz.

Liu, R. et al. (2024). Proofread: Fixes All Errors with One Tap. Accessible at: arXiv preprint arXiv:2406.04523.

Machura, J. et al. (2022). Automatic Grammar Correction of Commas in Czech Written Texts. Online. In: P. Sojka et al. (eds): TSD 2022. Cham (CH): Springer, pp. 113–124. Accessible at: https://dx.doi.org/10.1007/978-3-031-16270-1_10.

Machura, J. et al. (2023). Is it Possible to Re-educate RoBERTa? Jazykovedný časopis, 74(1), pp. 357–368. Accessible at: https://dx.doi.org/10.2478/jazcas-2023-0052.

Medková, H., and A. Horák. (2022). Distinguishing the Types of Coordinated Verbs with a Shared Argument by means of New ZeugBERT Language Model and ZeugmaDataset. In: A. Dimou et al. (eds.): Towards a Knowledge-Aware AI: SEMANTiCS 2022. Amsterdam: IOS Press, pp. 206–218. Accessible at: https://dx.doi.org/10.3233/SSW220022.

Mrkývka, V. (2023). Webový korektor jako prostředek formalizace pravidel českého jazyka. PhD Thesis, Brno: MU.

Omelianchuk, K. et al. (2020). GECToR – Grammatical Error Correction: Tag, Not Rewrite. In Proceedings of the 15th Workshop on Innovative Use of NLP for Building Educational Applications, Seattle, WA, USA, pp. 163–170.

Prokofyev, R. et al. (2014). Correct Me If I'm Wrong: Fixing Grammatical Errors by Preposition Ranking. In Proceedings of CIKM'14. Association for Computing Machinery, New York, NY, USA, pp. 331–340. Accessible at: https://doi.org/10.1145/2661829.2661942.

Song, Y. et al. (2024). GEE! Grammar Error Explanation with Large Language Models. In Findings of the Association for Computational Linguistics: NAACL 2024, Mexico City, Mexico. ACL, pp. 754–781.

Straka, M. et al. (2021). RobeCzech: Czech RoBERTa, a Monolingual Contextualized Language Representation Model. In: K. Ekštein et al. (eds): TSD 2021. Lecture Notes in Computer Science, Vol. 12848. Springer, Cham. Accessible at: https://doi.org/10.1007/978-3-030-83527-9_17.

Wang, Y. et al. (2020). A comprehensive survey of grammar error correction. Accessible at: arXiv preprint arXiv:2005.06600.

CREATION AND USE
OF LANGUAGE RESOURCES

sciendo
DOI 10.2478/jazcas-2025-0027

DO FREQUENCY TYPES MATTER IN LEXICOGRAPHY?

MAREK BLAHUŠ[1] – VOJTĚCH KOVÁŘ[2] – FRANTIŠEK KOVAŘÍK[3]
[1]Lexical Computing CZ, s.r.o., Brno, Czech Republic
(ORCID: 0009-0009-4096-4158)
[2]Lexical Computing CZ, s.r.o., Brno, Czech Republic & Department of Machine Learning and Data Processing, Faculty of Informatics, Masaryk University, Brno, Czech Republic (ORCID: 0009-0005-0307-9046)
[3]Lexical Computing CZ, s.r.o., Brno, Czech Republic & Faculty of Arts, Masaryk University, Brno, Czech Republic (ORCID: 0009-0003-8002-8360)

BLAHUŠ, Marek – KOVÁŘ, Vojtěch – KOVAŘÍK, František: Do Frequency Types Matter in Lexicography? Journal of Linguistics, 2025, Vol. 76, No 1, pp. 303 – 311.

Abstract: Word frequency in a corpus can be calculated in several different ways. Amongst the most common frequency types are the absolute frequency, the document frequency, ALDF and ARF. This paper focuses on comparing these four types in terms of "word correctness." For determining whether a word is correct or not, we use the data gathered for the Czech lexicon used for the recent Czech Dictionary Express project. In this project, each of the top 100,000 most frequent headwords was reviewed by several Czech native speakers, who decided whether the word should be accepted or rejected or has some minor issues. The quality of the "word correctness" is further discussed in the paper.

Keywords: corpus annotation, semi-automatic dictionary drafting, Dictionary Express, word frequency, frequency type, absolute frequency, document frequency, ALDF, ARF, Czech

1 INTRODUCTION

Word frequency is a number heavily used in corpus linguistics for statistics. It represents the word count across the corpus. The frequency of a word, a lemma or a token illustrates its distribution, determines the score of a collocation, and constitutes the base for frequency wordlists.

Frequency wordlists are lists of words (lemmas, tokens, etc.) sorted from the most frequent ones down to the least frequently used words (typically with one occurrence).

There are different strategies for counting a word's frequency. This paper revolves around four of the most typically used word frequency types, and examines how differences in word frequency can correlate with the occurrences of typos, non-words, words of a different language than the corpus, non-standard words and incorrectly lemmatized and/or POS-tagged words, as well as the rest – the "correct" words, in the Czech Web (csTenTen12+17+19) corpus (Suchomel 2018). For distinguishing whether a word is "correct" or faces some issues, we use annotation data gathered manually from Czech native speakers in the Czech Dictionary Express project.

Chapter 2 briefly introduces the principle of Dictionary Express projects, the manual annotation of Czech headwords, and the criteria of "correctness" of headwords. Chapter 3 the purpose of frequency types, their differences and their usage. Chapter 4 presents the correlation statistics between higher or lower frequency of each type and the "correctness" rate of headwords of these frequencies.

2 HEADWORD ANNOTATION

2.1 Dictionary Express

Dictionary Express (DE) is a series of dictionary making projects, which focus on rapid semi-automatic dictionary drafting methods (Kovařík et al. 2024). Each DE project concentrates on a different language and divides the dictionary making process into simple tasks such as building the vocabulary, selecting proper word forms for every headword, word sense disambiguation etc. As opposed to the "traditional dictionaries", created one entry at a time, the DE dictionaries are done in stages matching the tasks: the first stage includes going through the whole set of headwords and creating a proper vocabulary; the next stage includes going through the whole vocabulary and choosing the proper forms; etc.

Each stage is prepared automatically, using data from large language corpora (with tens of billions of tokens), preferably lemmatized and POS-tagged. The data is then manually annotated by a team of native speakers without academic education in linguistics, called the *annotators*.

2.2 Annotation

In the first stage, the annotators go through a list of headwords (i.e. pairs of lemma and part of speech), which are automatically lemmatized and POS-tagged by specialized tools. The annotators assign each headword one of these possible "flags":

- *don*'*t know the word* if they do not understand the word;
- *not Czech* if they know of the word but the word isn't part of the Czech language (based on their native speakers' intuition);
- *non-standard* if the word is not part of the standard Czech language (we take Czech *spisovný jazyk* as the standard, although again the annotators' language intuition is determinant);
- *wrong lemma* if the lemma is incorrect (including words with incorrect lemmatization, words in their non-lemma form and words with typos);
- *wrong POS* if the POS is incorrect;
- *ok* if the lemma and the POS are correct;
- *name* if the lemma and the POS are correct and the word is a proper name.

The annotators don't see the context of the words and are not allowed to look up the word in any other dictionary or on the internet.

This way, each headword has got at least two flags from two different annotators.

2.3 Revision

The headwords that were annotated with a variety of flags (i.e. with an insufficient inter-annotator agreement) and the ones whose majority flag was *non-standard*, *wrong lemma* and *wrong POS* had to be revised.

A group of experienced annotators (called "inspectors") went through each of these headwords, and according to the flags previously assigned to them and their corpus context, they decided whether the word is correct or incorrect or should be revised to another lemma or POS.

2.4 "Correctness" criteria

The wordlist for Czech Dictionary Express was created using document frequency. It contained the 100,000 most frequent headwords of the Czech Web corpus. After the revision, each of the headwords was either considered correct (marked *ok* or *name*) or incorrect (marked *don't know the word* or *not Czech* or revised to a correct headword).

3 FREQUENCY TYPES

Word frequency can be counted in a number of ways. This paper examines four of the most commonly used frequency types: absolute frequency, document frequency, ALDF and ARF.

Absolute frequency is the number of occurrences a word has in a corpus (Sketch Engine 2024). For smaller corpora with a specific topic, this can be an effective and simple way to count the words and analyze the vocabulary statistically. Absolute frequency, however, can be easily manipulated if a single word is used a lot of times in a single document or in a narrow area of texts. In other words, it ignores the word burstiness.

Word burstiness is the quality of the distribution of a word, i.e. whether it is used only in a closed area (it "bursts" somewhere) or whether it is spread throughout the corpus (or the language) (Rychlý 2011). Some words can be used many times in only a few documents. Absolute frequency of these words is high, but their distribution over the whole corpus or language use is narrow.

For taking word burstiness into consideration, the lexicographer can use other frequency types, such as document frequency, ALDF and ARF.

Document frequency is the number of documents a word occurs in at least once. This makes the widely distributed words more frequent than the ones that are only used in a few documents.

ALDF, or average logarithm distance frequency, reflects the average distance between the occurrences of the word. For two words with the same absolute frequency, ALDF is lower for the word only used in a small number of texts or text areas (Sketch Engine 2022). ARF, or average reduced frequency, though counted in a different manner, serves a similar goal.

Choosing a proper frequency type that does or does not take word burstiness into account can make a big difference when examining a small area of words or differences between particular words or their usage. But what about bigger tasks, such as choosing words for a mono-lingual dictionary? The next chapter discusses the role of the frequency types in building a dictionary lexicon.

4 FREQUENCY WORDLIST DIFFERENCES

4.1 Relation between word frequency and its "correctness"

As suggested in chapter 3, we consider a word "correct" if most of the annotators agreed it is a standard part of the language or if an inspector revised it to be correct after seeing its previous annotations and its context. We mark the "correctness" with quotations, since this is not a measure of whether a word should or shouldn't be considered a stable and directive part of the language system, but only a consideration based upon the intuition of several native speakers.

The 100,000 most frequent headwords according to the document frequency have been differentiated this way. In Fig. 1, we see how the percentage of "correct" words is related to higher frequency. (For easier calculation, the frequency wordlist of 100,000 headwords has been divided into "percentiles" of 1,000 words. The numbers on the X axis represent these groups. To get the document frequency rank of a headword in a particular area, multiply the number by 1,000.) On the left are the headwords at the top of the frequency wordlist, on the right the words with the frequency rank up to 100,000.

We can see that the more frequent a word is, the more likely it is going to be considered "correct". This relation is very linear, at least for the 100,000 most frequent headwords.

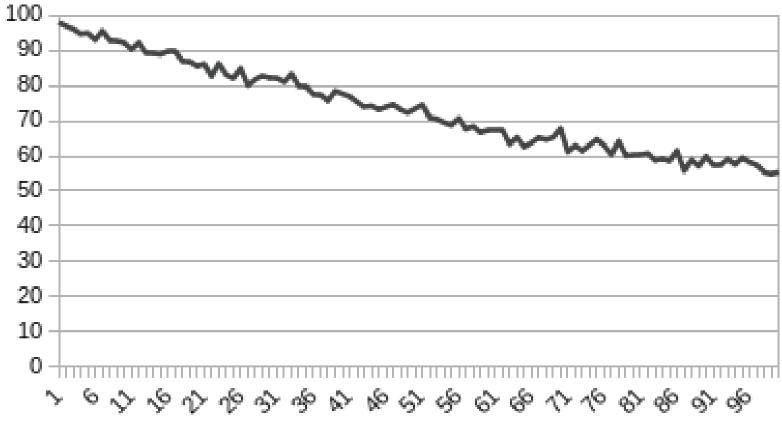


Fig. 1. Relation between the rank in document frequency divided by 1,000 (X axis) and "correctness" percentage (Y axis)

Fig. 2 shows a similar graph, but wordlists of all four frequency types are present now, represented by a separate color. The lines copy a very similar trajectory, except for the right ends of the wordlists. The data of the wordlists other than that of the document frequency are getting more scarce, because only the words from the 100,000 document frequency wordlist have been used, so some of the words from the ends of other wordlists are missing (as explained further, see Tab. 1), and thus more noise can be expected.

This means for the 100,000 most frequent headwords, there aren't many differences between the frequency types considering the "correctness" of the headwords.

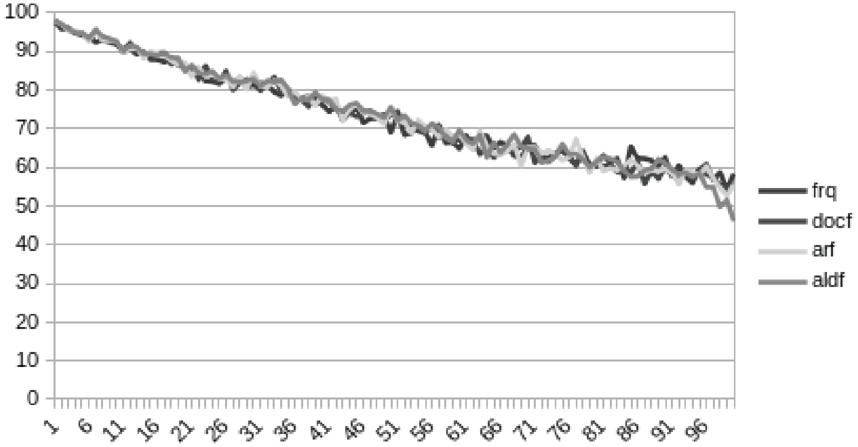


Fig. 2. Relation between the rank in the wordlist of frequency of a given type divided by 1,000 (X axis) and "correctness" percentage (Y axis)

The lexicon of the Czech DE project is based on the document frequency wordlist. Tab. 1 presents the word differences between the 100,000 document frequency wordlist and the wordlists of the other frequency types. Each number represents the number of words that are in the document frequency wordlist and are missing from the wordlist of a particular frequency, and vice versa.

As we can see in Tab. 1, the ALDF and ARF frequency wordlists are more similar to the document frequency wordlists than the one of absolute frequency. This should come as no surprise since both ALDF and ARF as well as the document frequency reflect not only the word count of a headword, but also its burstiness.

	Words missing in Doc. F.
Abs. F.	4962
ARF	1722
ALDF	1927

Tab. 1. Differences in wordlists of document frequency and of other frequency types

Tab. 2 presents the percentage of "correct" headwords within the 10,000, 50,000, 80,000 and approximately 100,000 most frequent headwords based on absolute frequency, document frequency, ARF and ALDF. (Since only the 100,000 most frequent headwords based on document frequency have been annotated, the statistics of headwords from the ends of the 100,000 wordlists of absolute frequency, ALDF and ARF are missing. Only the 95,038 most frequent words from the absolute frequency wordlist, the 98,278 most frequent words from the ARF wordlist and the 98,073 most frequent words from the ALDF wordlist have been decided to be "correct" or "incorrect". The ends of these 100,000 wordlists are still waiting to be properly annotated and revised by the annotators.)

	10,000	50,000	80,000	cca 100,000
Abs. F.	94.08%	83.33%	76.78%	74.07%
Doc. F.	94.65%	83.95%	77.07%	73.28%
ARF	94.70%	84.11%	77.29%	73.82%
ALDF	94.85%	84.44%	77.62%	74.09%

Tab. 2. The percentage of "correct" headwords in different frequency wordlists

We do not see a big difference between absolute frequency and the other types, even though absolute frequency seemed to be different from the other types considering the words of its 100,000 frequency wordlist (Tab. 1).

From the 100,000 document frequency wordlist, 73,278 have been marked "correct" and 6,518 headwords have been added as the result of the correction of "incorrect" headwords in the revision phase. This means that based on the quality of the corpus, the word lemmatization, POS tagging and language factors, a dictionary of 80,000 "correct" headwords needs approximately a 100,000-word wordlist. Considering the curve of the frequency-"correctness" relation in Fig. 1 and Fig. 2 and its predictable continuation, a dictionary of 100,000 "correct" headwords could require some 150,000 words from the frequency wordlists. Although there are differences between the wordlists, as shown in Tab. 1, these do not exceed 5% of the wordlists.

There could be, however, a bigger difference in the less frequent headwords, i.e. the headwords after the rank 100,000 of the document frequency wordlist. This is to be examined in future research focusing on the headwords after the frequency rank 100,000 and whether these headwords show different frequency-"correctness" relations than the more frequent ones.

4.2 Wordlist difference examples

For each wordlist, the words can be separated into 5 categories based on our research:

- *present accepted* are words that are in the 100,000 wordlist of a frequency type and are considered "correct";

- *present rejected* are words that are in the 100,000 wordlist of a frequency type and are considered "incorrect";
- *missing accepted* are words that are not in the 100,000 wordlist of a frequency type and are considered "correct";
- *missing rejected* are words that are not in the 100,000 wordlist of a frequency type and are considered "incorrect";
- and *missing from document frequency* are words that are present in the 100,000 wordlist of a frequency type other that document frequency and are not in the 100,000 document frequency wordlist – these words have not been yet marked "correct" or "incorrect" since only the document frequency wordlist has been annotated and revised, and are subject to further research.

The main subject of quality comparison between the wordlists has become the *present accepted* category, since these are the words a lexicographer would prefer to have in the dictionary yet are not included in some of the wordlists. Most of these are words from the end of the 100,000 most frequently used headwords.

The absolute frequency wordlist contains more company names and web page URLs than the other types, e.g. *Vareni.cz (noun)*, *Echo24 (noun)*, *Skyscanner (noun)*, *Ulož.to (noun)* and *ČSDF.cz (noun)*, whereas it is lacking many less frequent words such as *vypoklonkovat (verb)* – "to bow sb. out", *libující (adjective)* – "relishing", *utuchat (verb)* – "to weaken (literary)", *třímající (adjective)* – "holding (literary)", or *polovičatě (adverb)* – "halfway, poorly". This should come as no surprise, since company names and URLs can be very frequent in a small number of texts (their frequency is high, yet their overall distribution is low) and the common Czech words the absolute frequency wordlist is lacking are distributed more evenly across the whole corpus, although their frequency isn't as high.

As mentioned in 3.1, the absolute frequency wordlist is more different than all the other wordlists, although the "correctness" of its words is similar. The words missing from the other wordlists that are present in the absolute frequency wordlists, however, are not of the same quality as vice versa. In a dictionary, it would be preferable to include the less frequent words which are missing from the absolute frequency wordlist over the company names and web page, i.e. proper names of various origin.

Comparing the ARF and ALDF wordlists, the ARF wordlist does seem to have more company names and web page URLs, include more proper names in general, and also include more words of a foreign origin, such as *crowdfunding (noun)*, *selfíčko (noun)* – "selfie" and *magenergie (noun)* – "mana (fantasy)", whereas the ALDF wordlist has more originally Czech words similar to the ones missing from absolute frequency, e.g. *skotačící (adjective)* – "frolicking", *setrvávající (adjective)* – "remaining (literary)", *usekat (verb)* – "to cut off" or *brždění (noun)* – "braking (e.g. with brakes)".

This leads to the conclusion that the ALDF could be the preferred frequency type for building a dictionary lexicon if choosing from absolute frequency, ALDF and ARF. However, the research cannot be considered complete until the headwords from the end of ALDF, ARF and absolute frequency wordlist (the ones *missing from document frequency*) are annotated and marked "correct" or "incorrect". After this, conclusions can be made about the differences between all the wordlists, including document frequency, which has been used as the base for the DE lexicon.

5 CONCLUSION

We have examined four frequency wordlists containing the 100,000 most frequent headwords, calculated using absolute frequency, document frequency, ALDF and ARF. We have found some differences between the wordlists which could have a small impact on dictionary drafting and on building a dictionary lexicon.

The annotations, revisions and the quality of "correctness" were only gathered for the 100,000 most frequent headwords of the document frequency wordlist. A complete statistic of "correctness" in the 100,000 wordlists for all four types of frequencies should be a matter of subsequent research. Further research should be also made for the words after the 100,000 ranks and whether these words show different frequency-"correctness" relations than the more frequent words.

From examining the example differences between wordlists of different frequency types, it seems ALDF could be the preferred frequency type for building a Czech dictionary from a large web corpus. However, a vocabulary of good quality could also be achieved combining wordlists of all frequency types, and annotating words from the 100,000 wordlists of all frequency types. Considering the low rate of differences between the frequency wordlists of the 100,000 most frequent words, which do not exceed 5% of the wordlists, this would not make the dictionary making process noticeably more complex or time-consuming.

References

Kovařík, F., Kovář, V., and Blahuš, M. (2024). On Rapid Annotation of Czech Headwords: Analysing the First Tasks of Czech Dictionary Express. Online. In: Kristina Š. Despot – A. Ostroški Anić – I. Brač: Lexicography and Semantics: Proceedings of the XXI EURALEX International Congress. Cavtat: Institut za hrvatski jezik, 2024, pp. 336–344. Accessible at: https://euralex.jezik.hr/wp-content/uploads/2021/09/Euralex-XXI-proceedings_1st.pdf [cit. 27/03/2025].

Rychlý, P. (2011). Words' Burstiness in Language Models. In: A. Horák, P. Rychlý: Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2011. Brno: Tribun EU, 2011, pp. 131–137. Accessible at: https://nlp.fi.muni.cz/raslan/2011/paper17.pdf [cit. 27/03/2025].

Sketch Engine. ALDF – Average Logarithm Distance Frequency. Online. Sketch Engine. 2022, [28/02/2023]. Accessible at: https://www.sketchengine.eu/aldf-average-logarithmic-distance-frequency/ [cit. 27/03/2025].

Sketch Engine. Frequency. Online. Sketch Engine. 2024, [12/11/2024]. Accessible at: https://www.sketchengine.eu/glossary/frequency/ [cit. 27/03/2025].

Suchomel, V. (2018). csTenTen17, a Recent Czech Web Corpus. In Twelveth Workshop on Recent Advances in Slavonic Natural Language Processing. Brno: Tribun EU, 2018, pp. 111–123. Accessible at: https://www.sketchengine.eu/wp-content/uploads/cstenten17.pdf [cit. 27/03/2025].

sciendo
DOI 10.2478/jazcas-2025-0028

MODELLING VALENCY FRAMES USING INHERITANCE: THE CASE OF CZECH ADJECTIVES *-telný* ‘able’

VÁCLAVA KETTNEROVÁ[1] – JIŘÍ MÍROVSKÝ[2] – MICHAL OLBRICH[3]
[1]Department of Formal and Applied Linguistics, Faculty of Mathematics and Physics, Charles University, Prague, Czech Republic (ORCID: 0000-0001-9694-1304)
[2]Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics, Charles University, Prague, Czech Republic (ORCID: 0000-0003-2741-1347)
[3]Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics, Charles University, Prague, Czech Republic (ORCID: 0009-0008-2832-3211)

KETTNEROVÁ, Václava – MÍROVSKÝ, Jiří – OLBRICH, Michal: Modelling Valency Frames Using Inheritance: The Case of Czech Adjectives *-telný* ‘able’. Journal of Linguistics, 2025, Vol. 76, No 1, pp. 312 – 321.

Abstract: We introduce a first step to modelling valency frames of selected types of nominals. We work on the assumption that nominals inherit – at least to some extent – valency from their base verbs. We illustrate this task in a case study focused on modelling valency frames of Czech deverbal adjectives *-telný* ‘able’. First, the valency frames of the adjectives *-telný* contained in NomVallex are compared with the valency frames of their base verbs in VALLEX. Based on this comparison, two formal rules describing valency changes in the valency frames of adjectives *-telný* are formulated. Second, for each lexical unit of a verb that satisfies the conditions imposed by some of the rules, the derived adjective *-telný* is extracted from DeriNet, if such an adjective is available. Third, the valency frame of the adjective is derived from the valency frame of the verb based on the respective rule. Lastly, the accuracy of both rules is verified in the corpus data. The experiment has shown that the valency of these adjectives can be modeled on the rule basis. However, if this task is to be accurate, it requires advanced linguistic information, namely the information on semantic class membership of verbs and on compound adjectives.

Keywords: inheritance of valency, deverbal adjectives, valency lexicons, derivational relations, suffixes, Czech

1 INTRODUCTION

Valency, a number and type of complements a verb (or a noun and an adjective) combines with, is a lexicosyntactic property, which is asserted not to be automatically predictable, and as such it has to be described in a lexicon (e.g. Žabokrtský 2005). However, a significant number of words share their valency properties with their base words. For example, the number and type of valency complements of the adjective *obvinitelný* ‘accusable’ are identical to those of its base verb *obviňovat* ‘to accuse’, cf. examples (1) and (2). The same applies to the adjective *polepšitelný* ‘corrigible’ and its base verb *polepšit se* ‘to improve oneself’, cf. (3) and (4).[1] Changes occur in the surface expression of adjectival complements, as morphemic forms of some of the complements and the surface ellipsis show.

(1) *obviňovat* ‘to accuse’: $\text{ACT}_{1}^{\text{obl}}$ $\text{ADDR}_{4}^{\text{obl}}$ $\text{PAT}_{\text{z+2,že}}^{\text{obl}}$
Otec matku obviňoval z absurdní přecitlivělosti (SYN v13)
‘The father accused the mother of absurd hypersensitivity’

(2) *obvinitelný* ‘accusable’: $\text{ACT}_{7}^{\text{obl}}$ $\text{ADDR}_{\uparrow}^{\text{obl}}$ $\text{PAT}_{\text{z+2,že}}^{\text{obl}}$
matka obvinitelná otcem z absurdní přecitlivělosti
‘the mother accusable by the father of absurd hypersensitivity’

(3) *polepšit se* ‘to improve oneself’: $\text{ACT}_{1}^{\text{obl}}$
Polepší se lidstvo?
‘Will humankind improve?’

(4) *polepšitelný* ‘corrigible’: $\text{ACT}_{\uparrow}^{\text{obl}}$
Je lidstvo polepšitelné? (SYN v13)
‘Is humankind corrigible?’

The fact that deverbal adjectives can inherit valency properties from their base verbs, as illustrated with examples (2) and (4),[2] can be exploited to automatically model their valency frames, the manual annotation of which is both time-consuming and demanding in terms of human resources.

2 VALENCY OF ADJECTIVES

2.1 Systemic ellipsis of complements in adjectives

Most investigation into valency is concerned with verbs. The valency of non-verbal predicates, nouns and adjectives, is still under-researched. The valency of Czech adjectives has been outlined by Daneš et al. (1987), and described in the light of corpus data by Kopřivová (2006) and recently by Najbrtová (2017). Systematic attention has been paid to this issue in the Functional Generative Description (FGD), which serves here as the theoretical background as well (see esp. Panevová 1998; Kolářová et al. 2021). Moreover, the valency of selected nouns and adjectives is captured by Svozilová et al. (2005) and recently in the valency lexicons NomVallex

[1] The examples come from the Czech National Corpus, SYN v13, acessible at https://www.korpus.cz/. Examples that are not indicated are modified. The numbers stand for cases (1=Nom, 2=Gen, 3=Dat, 4=Acc, 6=Loc, and 7=Ins). The conjunction marks dependent clauses. The sign ↑ indicates that a valency complement cannot be expressed on the surface although it is present in the deep structure (Sect. 2.1).

[2] The inheritance of valency applies to deverbal, deadjectival or denominal nouns, too.

(Kolářová et al. 2024) and PDT-Vallex (Urešová et al. 2024). However, the numbers of nouns and adjectives included in these lexicons are still limited.

In the valency theory of FGD, five actants are distinguished, based on the syntactico-semantic criteria: ACTor, ADDRressee, PATient, EFFect, and ORIGin. The information on valency is captured in the valency frame, which is modeled as a sequence of valency slots. Each slot stands for one valency complement. For each complement, the valency frame provides the information on its type and obligatoriness. Morphemic forms then indicate surface realization of the complement. The number and type of valency complements determine the deep valency whereas morphemic forms indicate the surface valency.

The surface valency of adjectives is specific, as one of adjectival complements is systematically elided from the surface, despite being present in the deep valency of the adjective. The antecedent of this complement is expressed out of the adjectival structure either as the governor of the adjective, see the antecedent of ADDRessee *matka* 'mother' in (2), or as the subject of the copular construction, see the antecedent of ACTor *lidstvo* 'humankind' in (4). As examples (2) and (4) show, the complement that is subject to the systemic ellipsis can vary in type.

2.2 Inheritance of valency in adjectives

"Inheritance is the phenomenon that complex words have properties which are identical to properties of one of their morphological constituents" (Booij 2000, p. 857). Involved in many morphological processes, inheritance concerns (among other properties) valency as well, see, e.g. Bierwisch (2015). Accordingly, we can observe that the valency of deverbal adjectives corresponds, at least to some extent, to that of their base verbs and can be thus considered inherited from verbs.

In the case of the *deep valency*, two situations occur. First, a deverbal adjective fully inherits all the complements from its base verb. As a result, it has the same number and type of complements, see the adjectives *obvinitelný* 'accusable' (2) and *polepšitelný* 'corrigible' (4) above. Second, a deverbal adjective inherits only some of complements from its base verb, see the adjective *znalý* 'knowledgeable' (6) that lacks the complement ORIGin compared to its base verb *znát* 'know' (5).

(5) *znát* 'to know'
$\mathrm{ACT}_{1}^{\mathrm{obl}}$ $\mathrm{PAT}_{4,\mathrm{zda},\mathrm{cont}}^{\mathrm{obl}}$ $\mathrm{ORIG}_{\mathrm{od}+2,\mathrm{z}+2}^{\mathrm{opt}}$
... byl vzdělaným mužem, který znal řečtinu a hebrejštinu od svých předků.
'... he was an educated man, who knew Greek and Hebrew from his ancestors.'

(6) *znalý* 'knowledgeable'
$\mathrm{ACT}_{\uparrow}^{\mathrm{obl}}$ $\mathrm{PAT}_{2,\check{\mathrm{z}}\mathrm{e},\mathrm{cont}}^{\mathrm{obl}}$
... byl vzdělaným mužem, znalým řečtiny a hebrejštiny. (SYN v13)
'... he was an educated man, knowledgeable in Greek and Hebrew.'

In contrast to the deep valency, the *surface valency* of adjectives differs from that of the verbal one each time. As the surface realization of complements is indicated by morphemic forms (Sect. 2.1), changes in the surface expression of complements can be detected based on changes of their morphemic forms. Two situations in deverbal adjectives can occur concerning morphemic forms of their complements. First, morphemic forms remain the same or undergo so-called systemic changes, i.e. those changes that allow complements of deverbal adjectives to be expressed in the adjectival structure. These involve (i) the change of Nom into Instr or *od*+Gen,[3] and (ii) the surface ellipsis of one of adjectival complements (Sect. 2.1), see (2) and (4) above. Second, some complements of deverbal adjectives exhibit non-systemic changes of their morphemic forms, cf. the form of ACTor of the adjective *čtivý* 'readable' *pro*+Acc (8) with Nom of ACTor of its base verb (7).

(7) *číst* 'to read'
$\mathrm{ACT}_{1}^{\mathrm{obl}}\ \mathrm{PAT}_{4,o+6,zda,že,cont}^{\mathrm{obl}}$
Knihu čtou rádi i čtenáři, kteří nejsou odborníci na právo.
'Even readers who are not experts in law enjoy reading the book.'

(8) *čtivý* 'readable'
$\mathrm{ACT}_{pro+4}^{\mathrm{obl}}\ \mathrm{PAT}_{\uparrow}^{\mathrm{obl}}$
kniha čtivá i pro čtenáře, který není odborníkem na právo (SYN v13)
'the book readable even for readers who are not experts in law'

2.3 Modelling valency frames of deverbal adjectives using inheritance

We suppose the following mechanism to account for the valency of deverbal adjectives: a verb provides its valency frame, and the suffix used in the derivation of a deverbal adjective modifies it, resulting in the valency frame of the deverbal adjective. Changes in the valency of deverbal adjectives are thus attributed to the suffixes used in their derivation. To model the valency frames of deverbal adjectives based on the valency frames of their base verbs thus presupposes to correctly identify the changes brought about by each suffix. This task, however, poses a challenge due to the fact that there is no one-to-one correspondence between individual suffixes and valency changes: the same suffix can give rise to different valency changes and, conversely, one and the same change can be produced by different suffixes (Bierwisch 2015). Modelling valency frames of deverbal adjectives thus requires several steps:

[3] Kolářová et al. (2021) categorize the change of Acc into Gen in adjectival complements as systemic as well. This view is justified by the fact that in deverbal nouns Acc systematically changes into Gen. However, in deverbal adjectives, this change is attested only in the deverbal adjectives with the partial inheritance of the deep valency, see the complement PATient in (6) above. This issue thus deserves further investigation.

(i) A sufficiently large sample of deverbal adjectives with the same suffix has to be compared with their base verbs in order to determine the valency changes the suffix produces.

(ii) Those suffixes that induce the same valency changes are clustered together.

(iii) Formal rules describing the valency changes in deverbal adjectives are formulated for individual clusters.

(iv) Based on these rules, the valency frames of deverbal adjectives are derived from the valency frames of their base verbs.

(v) The accuracy of the derived valency frames is verified in the corpus data and the rules are modified, if necessary.

3 A CASE STUDY: ADJECTIVES *-TELNÝ* ‘ABLE’

We illustrate tasks (i), (iii), (iv) and (v), introduced in Sect. 2.3, with the deverbal adjectives with the suffix *-telný*, see (2) and (4) above, focusing on challenges that arise in each step.[4] These adjectives denote potential affectedness by an event expressed by their base verbs. We select this type as these deverbal adjectives are expected to fully inherit the deep valency from their base verbs, and their surface valency is supposed to be subject to systemic changes (Sect. 2.2). Moreover, these deverbal adjectives are part of derivational morphology, whereas, e.g. verbal adjectives (e.g. *obviňující* ‘accusing’ and *obviněný* ‘accused’, see esp. Jelínek 2003), which exhibit the full inheritance of valency complements displaying surface systemic changes, can still be viewed from a certain perspective as part of verbal inflection.

To identify the valency changes produced by the suffix *-telný* we compared the valency frames of the adjectives with this suffix contained in NomVallex[5] (47 in total) with the valency frames of their base verbs in VALLEX.[6] Based on this comparison, two rules are formulated. The first rule describes the changes in the valency of the adjectives *-telný* that represent the so-called passive type (e.g. *dělitelný* ‘divisible’, *přemístitelný* ‘movable’, *využitelný* ‘usable’), where the systemic surface ellipsis affects the complement expressed in base verbs by Acc (Sect. 3.1). The second rule applies to those adjectives that are of the active type (e.g. *polepšitelný* ‘corrigible’, *přizpůsobitelný* ‘adaptable’, *rozptýlitelný* ‘distractible’), where the nominative complement of base verbs is elided (Sect. 3.2). This split of rules is conditioned by the advanced semantic information on the type of reflexive verbs.

[4] As a preliminary study, we limit ourselves to one suffix, thus leaving aside step (ii).

[5] http://hdl.handle.net/11234/1-3420

[6] http://hdl.handle.net/11234/1-4756

3.1 Rule 1: Passive type

Passive type of adjectives *-telný*	
conditions	¬reflexverb: decaus|autocaus & SE ACT$_1$ & X$_4$[ADDR|PAT|EFF]
ACT	* → 7
X	* → ↑
Y	jako+4 → jako+1
obligatoriness	X

Fig. 1. Rule 1 determining the changes in the valency frames of adjectives *-telný* of the passive type

The first row of the rule determines the conditions on which the rule is applied to a valency frame of a verb. The conditions specify that the verb is not a decausative or autocausative reflexive verb[7] with the reflexive *se* in its lemma (¬ rules out a certain value and & indicates ‘at the same time’). Further, it requires the nominative ACTor and at the same time ADDRessee, PATient or EFFect in Acc in its frame (represented by the variable X).

The second row captures changes in the valency frame of the verb needed to derive the valency frame of the adjective. First, it determines that all the forms of ACTor (represented by the sign *) are changed into Instr. Second, it specifies that all the forms of the complement in Acc, represented by the variable X, are replaced by ↑, indicating its surface ellipsis. Third, it states that the complement expressed by the form *jako*+Acc, represented by Y (typically EFF or COMPL), changes its form into *jako*+Nom. Lastly, the rule determines that the complement X must be obligatory (even if it is optional in the valency frame of the verb).

Other complements from the valency frame of the verb, including their forms, remain preserved.[8] See examples below.

- X=PAT

napravit ‘to correct’ → *napravitelný* ‘corrigible’
ACT$_1^{obl}$ PAT$_4^{obl}$ → ACT$_7^{obl}$ PAT$_{\uparrow}^{obl}$
Ve většině pracovních kolektivů je případná chyba napravitelná samotným pracovníkem
‘In most teams, a potential mistake is corrigible by the worker themselves’

[7] For decausative and autocausative reflexive verbs see Geniušienė (1987).

[8] We leave aside that the complement X can be marginally expressed by Gen as well (cf. *sotva si povšimnout detailů*$_{gen}$ ‘to hardly notice details’ and *sotva povšimnutelné detaily* ‘hardly noticeable details’).

- X=PAT, Y=EFF

vnímat ‘to perceive’ → *vnímatelný* ‘perceivable’
ACT$_{1}^{obl}$ PAT$_{4,že}^{obl}$ EFF$_{jako+4}^{obl}$ → ACT$_{7}^{obl}$ PAT$_{\uparrow}^{obl}$ EFF$_{jako+1}^{obl}$
Deseti Oscary oceněná West Side Story, byť dnes vnímatelná jako přece jen rozvleklá
‘West Side Story awarded ten Oscars, though now perceivable as somewhat lengthy’

Remark on Rule 1. ACTor of adjectives *-telný* can be marginally expressed by the form *od*+Gen (e.g. *je od něho ovlivnitelná* ‘she is influenceable by him’), by Dat that can alternate with *pro*+Acc (e.g. *Můj pracovní zápřah ... je asi mnohým lidem/pro mnohé lidi těžko vůbec představitelný* ‘My workload ... is probably hard for many people even to imagine’). The latter forms typically occur in the adjectives *-telný* derived from mental verbs. The occurrence of *od*+Gen is more tricky. We can observe that this form is typically accepted by the adjectives *-telný* derived from base verbs whose Acc complement is filled with an animate participant (e.g. *vydíratelný* ‘blackmailable’).

3.2 Rule 2: Active type

<table>
<tr><th colspan="2">Active type of adjectives <i>-telný</i></th></tr>
<tr><td>conditions</td><td>reflexverb: decaus|autocaus & SE
ACT$_{1}$</td></tr>
<tr><td>ACT
obligatoriness</td><td>* → ↑
ACT</td></tr>
</table>

Fig. 2. Rule 2 determining the changes in the valency frames of adjectives *-telný* of the active type

The conditions determine that the rule is applied to decausative or autocausative reflexive verbs with the reflexive *se* in their lemmas and that these verbs have the nominative ACTor. In the valency frame of the derived adjectives, all the forms of the ACTor are overwritten by ↑, indicating that this complement is elided from the surface. The rule further states that ACTor is obligatory in the valency frame of adjectives (regardless of its possible optionality in the valency frame of the base verb). Other complements remain unchanged. See example below.

napravit se ‘to correct oneself’ → *napravitelný* ‘corrigible’
ACT$_{1}^{obl}$ → ACT$_{\uparrow}$
nenapravitelný hříšník
‘incorrigible sinner’

3.3 Ambiguity and compound adjectives

It should be pointed out that those adjectives to which Rule 2 relates are ambiguous: they are either of the active type or the passive type (see, e.g. the adjective *napravitelný* ‘corrigible’ in Sect. 3.1 and 3.2). The source of ambiguity is as follows. First, non-reflexive verbs with the accusative complement undergo reflexivization, resulting in decausative and/or autocausative reflexive verbs. Then both non-reflexive and reflexive verbs are base verbs from which adjectives *-telný* are derived: the non-reflexive verbs motivate the passive type of adjectives *-telný* with the valency frames produced by Rule 1, while reflexive verbs are the basis for the active type of adjectives, the valency frames of which result from Rule 2.

Further, the same valency frames as those derived by Rule 2 underlie the valency of a specific type of adjectives *-telný*, namely compounds of the Pron+Adj type (e.g. *samořiditelný* 'self-driving'). The accusative complement of their base verbs is occupied by the reflexive pronoun (e.g. *auta*$_{\text{nom}}$ *řídí sebe*$_{\text{acc}}$ *sama* 'the cars drive themselves') that becomes part of the adjectival lemma and is thus dropped from the valency frame of the adjective (e.g. *samořiditelná auta* 'self-driving cars'). The only complement remaining in its frame is the ACTor that undergoes the surface ellipsis. As a result, these adjectives can be attributed the same valency frame as the adjectives of the active type, produced by Rule 2, although their base verbs satisfy the conditions introduced by Rule 1.[9]

3.4 Derivation of valency frames and their verification

Based on Rule 1 and 2, introduced in Sect. 3.1 and 3.2, respectively, 2,937 valency frames were derived for adjectives *-telný*. First, verb lemmas representing lexical units of verbs satisfying the conditions set in Rule 1 and 2 were extracted from VALLEX (3,040 lexical units represented by 2,655 verb lemmas, for Rule 1, and 507 lexical units and 718 verb lemmas for Rule 2; lexical units describing idioms were filtered out). For each verb lemma, an adjective *-telný* was searched in DeriNet:[10] 1,234 and 350 adjectival lemmas *-telný* derived from the verb lemmas identified by Rule 1 and 2, respectively, were obtained (only adjectival lemmas attested in corpus data were taken into account). For these adjectives, 2,482 and 455 valency frames based on Rule 1 and Rule 2, respectively, were derived from the valency frames of their respective base verbs. See Tab. 1.

	Number of lemmas	Number of valency frames
Rule 1 all	2,731	6,019
Rule 1 excl. idioms	2,669	5,253
Rule 1 attested in corpus	1,234	2,482
Rule 2 all	710	937
Rule 2 excl. idioms	703	905
Rule 2 attested in corpus	350	455

Tab. 1. Adjectival lemmas *-telný* and valency frames derived for them based on Rule 1 and 2

[9] However, these compound adjectives are rare (in SYN v13 there are 33 lemmas with 5,490 occurrences of the type *samo.*telný*, and 7 lemmas with 15 occurrences of the type *sebe.*telný*).

[10] http://hdl.handle.net/11234/1-3765

The accuracy of the derived valency frames was verified in the corpus data. For 40 adjectives *-telný,* 100 corpus sentences were randomly selected from SYN v13 and annotated with respect to the systemic surface ellipsis and other valency complements. 88% of instances represent the passive type governed by Rule 1, while 12% are of the active type described by Rule 2. The annotation shows that both rules correctly determine the systemic surface ellipsis. Further, only less than 3% of valency complements of selected adjectives were expressed on the surface; their forms were identified with almost 88% accuracy. The lower figure results from the inaccurate determination of the form of ACTor, due to the form *pro*+Acc which alternates with Instr. We thus propose to integrate this form of ACTor as an alternative to Instr into Rule 1.

4 CONCLUSION

We have proposed the procedure of modelling the valency frames of adjectives based on the valency frames of their base verbs. In the case study focused on the adjectives -*telný*, we have shown that this task is feasible. However, to achieve high precision it requires rich linguistic information. Last but not least, the case study has shown that the assertion that valency is unpredictable is not generally valid and that the valency of some derived words can be modeled based on the valency of their base words. This fact can be further used in building valency lexicons.

ACKNOWLEDGEMENTS

The research reported in this paper was supported by the Czech Science Foundation under the project 25-16716S. This work has been using language resources developed, stored and distributed by the LINDAT/CLARIAH-CZ project (LM2023062) of the Ministry of Education, Youth and Sports of the Czech Republic.

References

Bierwisch, M. (2015). Word-formation and argument structure. In: P. O. Müller – I. Ohnheiser – S. Olsen – R. Franz (eds.): Word-Formation: An International Handbook of the Languages of Europe. Vol. 2. Berlin/Boston: Walter de Gruyter, pp. 1056–1099. Accessible at: https://doi.org/10.1515/9783110246278-016.

Booij, G. (2000). Inheritance. In: G. Booj – Ch. Lehmann – J. Mugdan (eds.): Morphology. An International Handbook of Inflection and Word-Formation. Berlin/New York: Walter de Gruyter, pp. 1057–1099. Accessible at: https://doi.org/10.1515/9783110111286.1.11.857.

Daneš, F., Hlavsa, Z., and Grepl. M. (1987). Mluvnice češtiny 3. Skladba. Praha: Academia, 748 p.

Geniušienė, E. (1987). The Typology of Reflexives. Berlin/New York/Amsterdam: Mouton de Gruyter, 435 p.

Jelínek, M. (2003). Transpoziční verbální adjektiva aktivní. In Sborník prací Filozofické fakulty brněnské univerzity, 52(A51), pp. 113–123.

Kolářová, V., Vernerová, A., and Klímová, J. (2021). Systemic and non-systemic valency behavior of Czech deverbal adjectives. Jazykovedný časopis, 72(2), pp. 371–382. Accessible at: https://doi.org/10.2478/jazcas-2021-0034.

Kolářová, V. et al. (2024). NomVallex. LINDAT/CLARIAH-CZ digital library at ÚFAL, MFF UK. Accessible at: http://hdl.handle.net/11234/1-5826.

Kopřivová, M. (2006). Valence českých adjektiv. Praha: NLN, 125 p.

Najbrtová, K. (2017). Korpusová analýza přejímání valenčních rámců u adjektiv derivovaných sufixem -telný. Ph.D. thesis. Brno: Masarykova univerzita, 226 p.

Panevová, J. (1998). Ještě k teorii valence. Slovo a slovesnost, 59(1), pp. 1–14.

Svozilová, S., Prouzová, H., and Jirsová, A. (2005). Slovník slovesných, substantivních a adjektivních vazeb a spojení. Praha: Academia, 579 p.

Urešová, Z. et al. (2024). PDT-Vallex: valenční slovník češtiny propojený s korpusy 4.5. LINDAT/CLARIAH-CZ digital library at the ÚFAL, MFF UK. Accessible at: http://hdl.handle.net/11234/1-5814.

Žabokrtský, Z. (2005). Valency Lexicon of Czech Verbs. Ph.D. thesis. Prague: Faculty of Mathematics and Physics, 130 p.

sciendo
DOI 10.2478/jazcas-2025-0029

ADVANCED SYNTACTIC PHENOMENA IN THE NOMVALLEX LEXICON

VERONIKA KOLÁŘOVÁ[1] – VÁCLAVA KETTNEROVÁ[2] – JIŘÍ MÍROVSKÝ[3]

[1]Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics, Charles University, Prague, Czech Republic (ORCID: 0000-0001-5184-579X)
[2]Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics, Charles University, Prague, Czech Republic (ORCID: 0000-0001-9694-1304)
[3]Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics, Charles University, Prague, Czech Republic (ORCID: 0000-0003-2741-1347)

KOLÁŘOVÁ, Veronika – KETTNEROVÁ, Václava – MÍROVSKÝ, Jiří: Advanced Syntactic Phenomena in the NomVallex Lexicon. Journal of Linguistics, 2025, Vol. 76, No 1, pp. 322 – 332.

Abstract: NomVallex 2.5 is a valency lexicon of Czech nouns and adjectives that besides valency of particular noun and adjectival lexical units captures several valency-related syntactic phenomena, namely (i) active and passive syntax, (ii) systemic and non-systemic valency behavior, (iii) reflexivity and reciprocity, and (iv) negation. These phenomena, represented in the lexicon mostly by deverbal and deadjectival derivatives, are described here with respect to the derivational category of denominal adjectives. Though being regarded as less typical representatives of valency bearers in the non-verbal domain, denominal adjectives turned out to be involved in all of the studied syntactic phenomena. Their syntactic behavior thus can be viewed as similar to other derivational categories, especially to deverbal adjectives.

Keywords: denominal adjectives, negation, reciprocity, valency, valency lexicon, syntax

1 INTRODUCTION

Valency, forming the syntactic core of the sentence, is involved in various syntactic phenomena, some of which are specific to the particular parts of speech; e.g. while diatheses (e.g. active and passive voice) are characteristic of verbs (Lopatková et al. 2016), active or passive syntax are typical of nouns and adjectives (Kolářová 2024). In this paper, we focus on valency-related syntactic phenomena in which nouns or adjectives are involved.

Valency properties of Czech nouns and adjectives are covered in several Czech valency lexicons, namely in a printed dictionary by Svozilová, Prouzová and Jirsová (2005), and in two electronic valency lexicons, PDT-Vallex (Urešová et al. 2024) and NomVallex (Kolářová and Vernerová 2022). While the former two lexical resources only describe the valency of particular nouns and adjectives, NomVallex (Section 2) aims to provide users with further valency-related syntactic phenomena relevant to these parts of speech (Section 3).

In this paper, we focus on denominal adjectives; compared to deverbal or deadjectival derivatives, denominal adjectives are regarded as less typical representatives of valency bearers in the non-verbal domain, and it is thus worth investigating whether the syntactic phenomena studied here are also relevant to them. Though this derivational category is only marginally represented in NomVallex 2.5 (Tab. 1), it will be, besides denominal nouns, the category focused on in the next version of the lexicon (Section 3.5).[1]

2 THE NOMVALLEX LEXICON

NomVallex is a manually created valency lexicon of Czech nouns and adjectives, adopting the Functional Generative Description (FGD) as its theoretical basis. Its newest version, NomVallex 2.5 (available as publicly accessible web-pages[2] and as downloadable and machine readable data; Kolářová et al. 2024), comprises 1,337 lexical units contained in 730 lexemes. As for derivational categories, it covers deverbal, deadjectival and denominal nouns, and deverbal, denominal, deadjectival and primary adjectives (Tab. 1).

Part-of-speech category	Derivational category	Lexical units	Lexemes
Nouns	deverbal	682	328
	deadjectival	301	197
	denominal	9	3
Adjectives	deverbal	244	151
	denominal	31	15
	deadjectival	8	8
	primary	62	28
Total		1,337	730

Tab. 1. The structure of the NomVallex 2.5 lexicon

NomVallex adopts and further modifies, where necessary, the annotation scheme of the valency lexicon of Czech verbs, VALLEX (Lopatková et al. 2022) (Fig. 1). The lexicon entry contains a lexeme, an abstract unit associating lexical forms with their lexical units (LUs), i.e., word senses. NomVallex applies the valency theory of the FGD (Panevová 1980): Valency properties of a lexical unit

[1] However, in some cases, it is not clear whether the adjective is motivated by a noun, or by a verb, cf. *podobný* ‘similar, resembling’ < *podoba* ‘similarity’ / *podobat se* ‘resemble’.

[2] https://ufal.mff.cuni.cz/nomvallex/2.5/

are captured in a valency frame, a sequence of valency slots, each supplemented with a list of morphemic forms. The following types of complements may be a part of valency frames: obligatory or optional actants (i.e., ACTor, PATient, ADDRessee, EFFect, and ORIGin, e.g. *Petrův*$_{\text{ACT}}$ *smysl pro humor*$_{\text{PAT}}$ ‘Peter’s sense of humor’, *výrobek prodejný mládeži*$_{\text{ADDR}}$ ‘product marketable to young people’, *vyhláška pochybná svou legalitou*$_{\text{EFF}}$ ‘regulation questionable as to its legality’), and obligatory free modifications (e.g. *muž povolaný do armády*$_{\text{DIR3}}$ ‘a man drafted into the army’). In NomVallex, valency properties of a lexical unit are illustrated with examples from the Czech National Corpus (corpora SYNv12 and Araneum Bohemicum Maximum).[3]

It is typical of adjectival valency structures, unlike the verbal and noun ones, that one valency complement of the adjective is systematically elided from the surface and thus cannot be expressed on the surface as a modification of the adjective. Instead, it refers to its antecedent expressed outside the adjectival structure either as the noun governing the adjective, see (2–3) for constructions with the deverbal adjectives *snášenlivý* ‘tolerant’ and *snesitelný* ‘tolerable’, or as the subject of the copula verb that the adjective forms a predicate with, see (1), cf. also Kettnerová and Kolářová (2023). In the valency frames of adjectives, this valency complement is marked by an upward arrow (4–5); this sign is also used in (1–3) to pinpoint the antecedents of the systematically elided adjectival valency complements.

(1) *plodina*$_{\uparrow}$ *je snášenlivá vůči suchu*$_{\text{PAT-vůči ‘to’+Dat}}$
‘the crop is tolerant of drought’

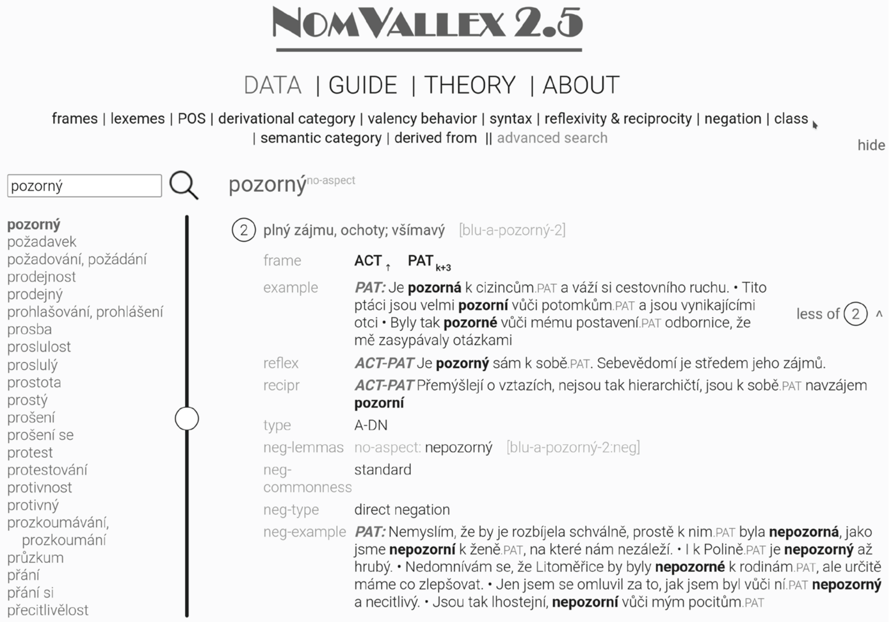
(2) *plodina*$_{\uparrow}$ *snášenlivá vůči suchu*$_{\text{PAT-vůči ‘to’+Dat}}$
‘the crop tolerant of drought’

(3) *podnebí*$_{\uparrow}$ *snesitelné pro Evropany*$_{\text{ACT-pro ‘for’+Acc}}$
‘climate tolerable to Europeans’

(4) *snášenlivý* ‘tolerant’ ACT$_{\uparrow}$ PAT$_{\text{k ‘to’+Dat}}$

(5) *snesitelný* ‘tolerable’ ACT$_{\text{Ins,pro ‘for’+Acc}}$ PAT$_{\uparrow}$

[3] Accessible at: https://www.korpus.cz/.



pozorný^no-aspect

2 plný zájmu, ochoty; všímavý [blu-a-pozorný-2]

frame $\text{ACT}_{\uparrow}$ PAT_{k+3}

example ***PAT:*** Je **pozorná** k cizincům.PAT a váží si cestovního ruchu. • Tito ptáci jsou velmi **pozorní** vůči potomkům.PAT a jsou vynikajícími otci • Byly tak **pozorné** vůči mému postavení.PAT odbornice, že mě zasypávaly otázkami

reflex ***ACT-PAT*** Je **pozorný** sám k sobě.PAT. Sebevědomí je středem jeho zájmů.

recipr ***ACT-PAT*** Přemýšlejí o vztazích, nejsou tak hierarchičtí, jsou k sobě.PAT navzájem **pozorní**

type A-DN

neg-lemmas no-aspect: nepozorný [blu-a-pozorný-2:neg]

neg-commonness standard

neg-type direct negation

neg-example ***PAT:*** Nemyslím, že by je rozbíjela schválně, prostě k nim.PAT byla **nepozorná**, jako jsme **nepozorní** k ženě.PAT, na které nám nezáleží. • I k Polině.PAT je **nepozorný** až hrubý. • Nedomnívám se, že Litoměřice by byly **nepozorné** k rodinám.PAT, ale určitě máme co zlepšovat. • Jen jsem se omluvil za to, jak jsem byl vůči ní.PAT **nepozorný** a necitlivý. • Jsou tak lhostejní, **nepozorní** vůči mým pocitům.PAT

Fig. 1. The NomVallex entry for the adjective pozorný ‘thoughtful/attentive’

3 VALENCY-RELATED SYNTACTIC PHENOMENA

Though primarily focused on valency, NomVallex aims to provide language material and lexicographic software, allowing for linguistic research of various language phenomena, including derivational relations, lexical semantics (e.g. semantic classes, semantic categories and various semantic shifts reflected in different lexical units), or valency-related syntactic behavior. In this section, we present four advanced syntactic phenomena treated in NomVallex that are closely related to valency of Czech nouns or adjectives, namely (i) active- and passive-like syntax (Section 3.1), (ii) systemic and non-systemic valency behavior (Section 3.2), (iii) reciprocity and reflexivity (Section 3.3), and (iv) negation (Section 3.4). These phenomena can be searched for or filtered using the lexicon’s web pages (see Fig. 1 capturing the entry for the denominal adjective *pozorný* ‘thoughtful/attentive’).

3.1 Active- and passive-like syntax

Kolářová (2024) deals with the way verbal active and passive constructions are reflected in the syntactic structures of adjectives and nouns and shows that the type of syntax a verbal or non-verbal predicate can use (i.e., both active and passive syntax, or only one of these) represents a notable difference between verbs and deverbal nouns on the one hand and adjectives and deadjectival nouns on the other: While nouns directly derived from transitive verbs usually display both active and passive syntax (cf. *snášet útrapy* ‘endure hardships’ > *Petrovo snášení útrap* ‘Peter’s enduring of hardships’ vs. *snášení útrap Petrem* ‘enduring of hardships by Peter’), adjectives and deadjectival nouns (even those motivated by a transitive verb) use either active or passive syntax only; for example, the adjective *snášenlivý* ‘tolerant’ can only use the active syntax, cf. (6) and (8). Typically, while in active-like adjectival constructions Actor is systematically elided from their surface valency structure, cf. (6), in passive-like adjectival constructions, Actor is one of the valency complements expressed on the surface, modifying the given adjective, in which case one of the forms of Actor usually is the prepositionless instrumental, see (9) for the adjective *snesitelný* ‘tolerable’.[4]

(6) *plodina*$_{\uparrow}$ (je) *snášenlivá vůči suchu*$_{\text{PAT-vůči ‘to’+Dat}}$
‘the crop (is) tolerant of drought’

(7) *plodina, která*$_{\text{ACT}}$ *snáší sucho*$_{\text{PAT}}$
‘the crop that is able to endure drought’

(8) **sucho*$_{\uparrow}$ *snášenlivé plodinou*$_{\text{ACT-Ins}}$
‘drought tolerant of by the crop’

(9) *podnebí*$_{\uparrow}$ *snesitelné Evropany*$_{\text{ACT-Ins}}$
‘climate tolerable by Europeans’

(10) *podnebí, které*$_{\text{PAT}}$ *Evropané*$_{\text{ACT-Nom}}$ *mohou snést*
‘climate that Europeans are able to tolerate’

Unlike verbal constructions, adjectives are predetermined to arrange their complements by adopting either the active or the passive syntax of their base predicates, not both, according to whether they systematically elide Actor (6) or Patient (9) / Addressee (see Section 2). Concerning deverbal adjectives, active and passive syntax usually depends on their derivational type (Kolářová 2024).

As for denominal adjectives, the research question is what type of syntax this derivational category may display. In order to determine valency characteristics of denominal adjectives, including ‘N+A’ compounds such as *pozoruhodný* ‘noteworthy’, we analyze the adjectival constructions paraphrasing them with an attributive clause, see (11) and (14) for the adjectives *pozorný* ‘thoughtful’ and

[4] Alternative forms are *pro* ‘for’+Acc and prepositionless dative, characteristic of Czech *-able* adjectives.

pozoruhodný 'noteworthy', respectively, both motivated by the noun *pozor* 'attention'. Comparing these paraphrases with those of deverbal adjectives, cf. *snášenlivý* 'tolerant' (7) and *snesitelný* 'tolerable' (10), we suppose that while most of denominal adjectives display active syntax, cf. the adjective *pozorný* 'thoughtful' in (12) and its valency frame in (13), in isolated cases even denominal adjectives display passive syntax, see the adjective *pozoruhodný* 'noteworthy' in (15) and its valency frame in (16).[5]

(11) *pozorný člověk*$_{\uparrow}$ 'thoughtful man':
'a man who shows consideration for the needs of other people'

(12) *člověk*$_{\uparrow}$*pozorný k cizincům*$_{\text{PAT}}$
'a man thoughtful to foreigners'

(13) *pozorný* 'thoughtful' ACT$_{\uparrow}$ PAT$_{\text{k 'to'+Dat}}$

(14) *pozoruhodný nález*$_{\uparrow}$ 'noteworthy find':
'a find that is worthy of notice/to be noticed'

(15) *nález*$_{\uparrow}$*pozoruhodný pro vědce*$_{\textit{ACT-pro 'for'+Acc}}$
'a find noteworthy/interesting for scientists'

(16) *pozoruhodný* 'noteworthy' ACT$_{\text{pro 'for'+Acc}}$ PAT$_{\uparrow}$ EFF$_{\text{Ins}}$

3.2 Systemic and non-systemic valency behavior

One of the main goals of the NomVallex lexicon is to make it possible to study changes in valency across part-of-speech categories and derivational types, with emphasis on their systemic and non-systemic valency behavior (see below).

To enable analysis of the relationship between the valency behavior of base words and their derivatives, lexical units of nouns and adjectives in NomVallex are linked to their respective base lexical units (contained either in NomVallex itself or, in the case of verbs, in the VALLEX lexicon) by two attributes, namely (i) the attribute *derivedFrom* (providing a link from a particular LU to its base LU), and (ii) the attribute *derivedLUs* (capturing a set of links to all LUs derived from the base LU). Each lexical unit of an adjective or a noun with a link to its respective base LU is automatically supplemented with information on differences between the valency frames of the two LUs; namely, the number and types of valency complements and

[5] Adjectives corresponding to verbs with non-prototypical frames (containing ACT in a form other than Nom, e.g. *Jana*$_{\text{ACT-Acc}}$ *dojala vzpomínka*$_{\text{PAT-Nom}}$ *na matku* 'Jan was touched by the memory of his mother'~ *pro Jana*$_{\text{ACT}}$ *dojemná vzpomínka* 'a touching memory for Jan') are to be examined in future research.

their morphemic forms are automatically compared. The changes (if any) are specified in the *valdiff* attribute.

As a result, with all the derived lexical units, systemic and non-systemic valency behavior can be distinguished (e.g. for deverbal adjectives, see Kolářová et al. 2021). Nouns or adjectives displaying systemic valency behavior inherit all valency complements from their base words and their morphemic forms do not change or result from regular changes. Non-systemic valency behavior involves changes in the number and type of valency complements, and non-systemic morphemic forms, i.e., forms that cannot be regularly derived from the forms of complements of the base word.

For example, there are two lexical units of the denominal adjective *žádostivý* (< *žádost* 'desire', see its valency frame in (17)), namely *žádostivý*$_1$ 'desirous, eager', and *žádostivý*$_2$ 'curious', see their valency frames in (18) and (20), respectively. While the adjective *žádostivý*$_1$ 'desirous, eager' displays systemic valency behavior, see (19), the adjective *žádostivý*$_2$ 'curious' undergoes both a shift in meaning and non-systemic forms of its Patient, e.g. the prepositional phrase introduced by *na* 'on' in (21), thus showingnon-systemic behavior.

(17) *žádost* 'desire' ACT$_{\text{Gen,poss}}$ PAT$_{\text{Gen,po 'for'+Loc,inf}}$

(18) *žádostivý*$_1$ 'desirous, eager' ACT$_{\uparrow}$ PAT$_{\text{Gen,po 'for'+Loc,inf}}$

(19) *Lidé*$_{\uparrow}$*, žádostiví nových bytů*$_{\text{PAT-Gen}}$*, si začali rezidence na Palmovce rezervovat*
'People, eager for new apartments, have started to reserve residences in Palmovka'

(20) *žádostivý*$_2$ 'curious' ACT$_{\uparrow}$ PAT$_{\text{na 'on'+Acc,zda 'if',cont}}$

(21) *Já*$_{\uparrow}$ *budu velmi žádostivý na důvody*$_{\text{PAT-na 'on'+Acc}}$ *ministerstva obrany*
'I will be very curious about the reasons of the Ministry of Defense'

3.3 Reflexivity and reciprocity

While reflexivity and reciprocity of verbs belong to intensively studied language phenomena, reflexivity and reciprocity of nouns and adjectives still call for a systematic analysis (Kettnerová and Kolářová 2023). Reflexivity refers to the situation where two valency complements of a noun or an adjective have the same referent, such as ACT and PAT of the denominal adjective *pozorný* 'attentive', which share the referent *Peter*, an antecedent of ACT, see (22). Reciprocity involves two complements that share two referents, see (23) for the reciprocal structure of the adjective *pozorný* 'attentive', in which both PAT and elided ACT have two referents, *Peter* and *Jane*.

(22) *Petr*$_{\uparrow}$ *je pozorný sám k sobě*$_{\text{PAT}}$
‘Peter is attentive to himself’

(23) *(Petr a Jana)*$_{\uparrow}$ *jsou k sobě*$_{\text{PAT}}$ *navzájem pozorní*
‘Peter and Jane are attentive to each other’

NomVallex, following the representation of these phenomena proposed in VALLEX, captures the information on reflexivity and reciprocity in the attributes *reflex* and *recipr*, respectively, assigned to those lexical units of nouns and adjectives that allow for reflexive and reciprocal structures. These attributes contain the pair of the complements (only actants and obligatory free modifications are covered) involved in reflexivity and reciprocity, respectively (e.g. the ACT-PAT pair in the case of the adjective *pozorný* ‘attentive’ in Fig. 1). If more than one pair of complements can be affected, the lexical unit is assigned more than one attribute, distinguished by Arabic numerals, each comprising different pairs of complements. The attributes then provide corpus evidence of reflexive and reciprocal constructions of the involved lexical units.

Inherently reciprocal nouns or adjectives, e.g. *společný* ‘common’ in (24), are captured in the attribute *reciprnoun* or *recipradj*, respectively, containing the value inherent.

(24) *témata společná konzervativcům a liberálům*
‘topics common to conservatives and liberals’

Moreover, the representation of reflexivity and reciprocity makes it possible to identify ambiguous reflexive and reciprocal constructions. For example, the adjective *pozorný* ‘attentive’ features both attributes *reflex* and *recipr*, containing the same pair of complements, namely ACT and PAT. This adjective can thus form ambiguous constructions that are either reflexive or reciprocal, see the construction in (25), which has two meanings.

(25) *(Petr a Jana)*$_{\uparrow}$ *jsou k sobě*$_{\text{PAT}}$ *pozorní*
‘Peter and Jane are attentive to themselves’ or
‘Peter and Jane are attentive to each other’

3.4 Negation

In Czech, the universal exponent of word-level (lexical) negation is the morpheme *ne-* (Pavlovič 2015; e.g. *neprodejný* ‘unmarketable’), used with autosemantic word classes. When the negative prefix *ne-* is combined with verbs, it typically only denies the affirmative content, i.e., the original predication, without specifying new features (Lotko 1975; cf. *prodat* ‘to sell’ vs. *neprodat* ‘not to sell’).

In contrast, with nouns and adjectives, the situation is more complex: in addition to direct negation, e.g. *(ne)pohodlí* '(dis)comfort' and *(ne)prodejný* '(un)marketable', the prefix *ne-* may lead to a semantic shift (Pavlovič 2015; cf. *smysl* 'sense' vs. *nesmysl* 'nonsense', *pohodlný* 'comfortable' vs. *nepohodlný* 'unwanted'), referred to here as lexicalized negation. The difference in meaning can be accompanied by a difference in valency, e.g. *nedůtklivost* 'touchiness' has two complements, see (26), whereas *důtklivost* 'urgency' only has one, see (27).

(26) *Petrova*$_{\text{ACT}}$ *nedůtklivost vůči kritice*$_{\text{PAT}}$
'Peter's touchiness on criticism'

(27) *důtklivost jeho výkladu*$_{\text{ACT}}$
'the urgency of his presentation'

In the case of direct negation, negative prefixes do not normally affect the valency of the base (Curiel 2015), however, some exceptions exist, cf. Eng. *dependence on* and *independence of/from*. In NomVallex, all lexical units of nouns and adjectives are examined with respect to whether or not they can be used in negative forms expressing direct negation of the affirmative forms, and whether they keep the same valency (Kolářová and Mírovský 2024). If yes, such as in the case of the affirmative and the negative form of the denominal adjective *(ne)pozorný* '(un thoughtful' in (28) and (29), the negative form is treated within the entry for the corresponding affirmative form (Fig. 1), and both forms share the same valency frame, cf. (13) in Section 3.1.

(28) *Navíc byl nebývale pozorný k ženám*$_{\text{PAT}}$
'Moreover, he was extraordinarily thoughtful to women'

(29) *I k Polině*$_{\text{PAT}}$ *je nepozorný až hrubý.*
'He is unthoughtful to Polina too, even rude.'

If the meaning of the negative form of a word shifts away from its affirmative form, it is assigned a separate entry, represented by the negative lemma(s) (e.g. *nepohodlný svědek* 'unwanted witness' vs. *(ne)pohodlné cestování* '(un)comfortable traveling').

Further, NomVallex differentiates between the so-called negativum tantum (the term used in Czech terminology for words that start with the string *ne-* but that in present-day Czech have no meaning without the string, e.g. *nenávistný* 'hateful') and the words that cannot have the negative forms for various reasons (e.g. *mocný (hlas)* 'powerful (voice)'; the value inapplicable is used in this case).

3.5 Denominal adjectives in the NomVallex data

In the current working version of the NomVallex data, denominal adjectives are represented by 66 lexical units in 37 lexemes, including 6 lexemes of ‘N+A’ compounds such as *pozoruhodný* ‘noteworthy’. All the adjectives were annotated with respect to the syntactic phenomena described in this paper. The simplified results presented in Tab. 2 show that in many aspects, syntactic behavior of denominal adjectives can be viewed as similar to deverbal adjectives.

Syntactic phenomenon		Number of LUs	Example
syntax	active	47	*pozorný* ‘thoughtful’
	passive	9	*pozoruhodný* ‘noteworthy’
	non-prototypical frame	10	*dojemný* ‘touching’
valency behavior	systemic	4	*žádostivý*$_1$ ‘desirous, eager’
	non-systemic	36	*žádostivý*$_2$ ‘curious’
	NA (the base noun is not present in NomVallex)	26	*vlastní* ‘own’
reflexivity	reflexive	21	*náročný* ‘demanding’
	non-reflexive	45	*vinný* ‘guilty’
reciprocity	inherently reciprocal	1	*společný* ‘common’
	reciprocal	21	*bezcitný* ‘heartless’
	non-reciprocal	44	*nemocný* ‘ill’
negation	direct	41	*(ne)pozorný* ‘(in)attentive’
	lexicalized	2	*neúnavný* ‘tireless’
	negativum tantum	1	*nenávistný* ‘hateful’
	inapplicable	22	*bezradný* ‘helpless’

Tab. 2. Denominal adjectives in the NomVallex working data

4 CONCLUSION

We have presented the annotation strategies adopted in NomVallex for four advanced valency-related syntactic phenomena, namely (i) active and passive syntax, (ii) systemic and non-systemic valency behavior, (iii) reflexivity and reciprocity, and (iv) negation. We have demonstrated that all the introduced phenomena are also relevant for denominal adjectives, one of the least represented derivational categories in NomVallex so far. A detailed description of valency and syntactic behavior of denominal adjectives, including, for example, forms of their Actor in passive-like constructions, appears to be an interesting area for further research.

ACKNOWLEDGEMENTS

The research reported in the paper was supported by the Czech Science Foundation under the project 25-16716S. The work described herein has been using data and tools provided by the LINDAT/CLARIAH-CZ Research Infrastructure (https://lindat.cz), supported by the Ministry of Education, Youth and Sports of the Czech Republic (Project No. LM2023062).

References

Curiel, M. M. (2015). Negation. In: P. O. Müler et al. (eds.): Word-Formation. An International Handbook of the Languages of Europe. Volume 2. De Gryuter Mouton, pp. 1351–1359.

Kettnerová, V., and Kolářová, V. (2023). K reciprocitě adjektiv v češtině. Slovo a slovesnost 84(3), pp. 179–200. Accessible at: https://doi.org/10.58756/s1138449.

Kolářová, V. (2024). Active and passive syntax of Czech deverbal and deadjectival nouns. Lingua, 307, 103686. Accessible at: https://doi.org/10.1016/j.lingua.2024.103686.

Kolářová, V. et al. (2024). NomVallex 2.5. LINDAT/CLARIAH-CZ digital library at ÚFAL MFF UK. Accessible at: http://hdl.handle.net/11234/1-5826.

Kolářová, V., and Mírovský, J. (2024). Looking for sense in nonsense: Valency of negative forms of nouns and adjectives in the NomVallex lexicon. In: K. Š. Despot – A. Ostroški Anić – I. Brač (eds.): Lexicography and Semantics. Proceedings of the XXI EURALEX International Congress. Zagreb, pp. 485–496.

Kolářová, V., and Vernerová, A. (2022). NomVallex: A Valency Lexicon of Czech Nouns and Adjectives. In Proceedings of LREC 2022, pp. 1344–1352, ELRA, Marseille, France.

Kolářová, V., Vernerová, A., and Klímová, J. (2021). Systemic and Non-systemic Valency Behavior of Czech Deverbal Adjectives. Jazykovedný časopis, 72(2), pp. 371–382.

Lopatková, M. et al. (2016). Valenční slovník českých sloves VALLEX. Praha, Karolinum, 700 p.

Lopatková, M. et al. (2022). VALLEX 4.5. LINDAT/CLARIAH-CZ digital library at ÚFAL MFF UK. Accessible at: http://hdl.handle.net/11234/1-4756.

Lotko, E. (1973). Lexikální negace v současné češtině. Státní pedagogické nakladatelství, 101 p.

Panevová, J. (1980). Formy a funkce ve stavbě české věty. Praha: Academia, 222 p.

Pavlovič, J. (2015). Negation in the Slavic and Germanic languages. In: P. O. Müler et al. (ed.): Word-Formation. An International Handbook of the Languages of Europe. Vol. 2. De Gryuter Mouton, pp. 1360–1373.

Svozilová N., Prouzová, H., and Jirsová, A. (2005). Slovník slovesných, substantivních a adjektivních vazeb a spojení. Praha: Academia, 579 p.

Urešová, Z. et al. (2024). PDT-Vallex: Valenční slovník češtiny propojený s korpusy 4.5. LINDAT/CLARIAH-CZ digital library at ÚFAL MFF UK. Accessible at: http://hdl.handle.net/11234/1-5814.

DIGITAL HUMANITIES

sciendo
DOI 10.2478/jazcas-2025-0030

A CORPUS-BASED ANALYSIS OF COLOR IMAGERY IN THE POETRY COLLECTION *DAR* BY MAŠA HAĽAMOVÁ

KATARÍNA GAJDOŠOVÁ[1] – IVICA HAJDUČEKOVÁ[2] – PETER MALČOVSKÝ[3]
[1]Slovak National Corpus, Ľudovít Štúr Institute of Linguistics, Slovak Academy of Sciences, Bratislava, Slovakia (ORCID: 0009-0005-2995-1146)
[2]Department of Slovak Studies, Slavonic Philologies and Communication, Faculty of Arts, Pavel Jozef Šafárik University, Košice, Slovakia (ORCID: 0000-0001-7718-335X)
[3]Slovak National Corpus, Ľudovít Štúr Institute of Linguistics, Slovak Academy of Sciences, Bratislava, Slovakia (ORCID: 0009-0001-2970-789X)

GAJDOŠOVÁ, Katarína – HAJDUČEKOVÁ, Ivica – MALČOVSKÝ, Peter: A Corpus-based Analysis of Color Imagery in the Poetry Collection *Dar* by Maša Haľamová. Journal of Linguistics, 2025, Vol. 76, No 1, pp. 335 – 344.

Abstract: The study presents a methodological approach that employs an annotated corpus to track the frequency of colors and their shades, as well as to visualize color clusters and their preferences. Corpus analysis reveals that the highest frequencies in the debut collection *Dar* [The Gift] are found in red, yellow with gold, white, but also gray and green. The results indicate that the color palette in the collection functions as an integral component of lyrical structure and actively contributes to the process of spiritualization. The poetic symbolization of color and its compositional use in expressing spirituality are part of the author's artistic strategy.

Keywords: color imagery, corpus, Maša Haľamová, poetry of the heart, the poetry collection *Dar* [The Gift], spiritualization

1 INTRODUCTION

In contemporary literary criticism, particular emphasis was placed on the subtlety and lyricism of Máša Haľamová's poetry, which scholars have sought to align with modernist tendencies, such as impressionism, neo-romanticism, or symbolism. As her poetry, in some respects, eluded these poetological categories, it became more and more evident that she was a solitary voice, positioned outside of modernist experimentations.

Current literary research confirms that the expressive qualities of Haľamová's poetry are subordinated to its spiritual meaning (Hajdučeková and Bónová 2024). Sensually inflected natural motifs play an important role in this poetry, becoming key elements in the processes of symbolization and emblematization. However, the sensual effect of natural imagery is further accentuated by the deliberately chosen use of color. This color scale reflects not only the emotional fluctuation of the lyrical subject but also deliberate allusions pointing to the processes of spiritualization and the sacralization of poetic expression (Hajdučeková and Bónová 2024, pp. 121–171).

As shown in the research of I. Hajdučeková a I. Bónová, the function of the nature-psychological parallel with spiritual significance is already strongly developed in the emblematic poem The Gift. On white silk, within the flowers of a wild bouquet, a symbolic scale of several colors is represented (red, pale violet, blue, green with the dominance of white). The emblematic poem The Gift, with its suggestive color scale, marked the poetic beginning and anticipated the core of the themes of Haľamová's later work. The study notes that, in terms of color motifs found in selected poems, white, red, gold, and silver appear most frequently. These colors intensify the spiritual dimension of meaning and the sacrality of expression (ibid. p. 152). Within research on the nuanced spectrum of spirituality in M. Haľamová's work, color connotations and their sequencing have so far been observed only secondarily as a thematic thread within subtle lyricism.

However, color can also be examined as an independent component of aesthetic structure, revealing deeper meanings across the poet's entire body of work. The integration of corpus linguistics and literary studies, which is gaining increasing ground within the field of digital humanities (Změlík 2019) can, through poems processed in corpora and appropriately visualized, offer new impulses for broader interpretations of a specific poetry collection or the author's entire oeuvre. In the present study, we test selected methodological approaches and the hypothesis that the color scale in the debut collection *The Gift* functions as a meaningful component of lyrical structure and participates in the process of spiritualization.

2 RESEARCH METHODOLOGY

To examine color usage, we adopted a combined approach in which literary texts processed within a corpus are visualized[1] and subsequently interpreted. From the three poetry collections by M. Haľamová The Gift (1928); Red Poppy (1932), and I Live Your Death (1966), we created an internal text corpus (6,777 tokens). The corpus is lemmatized and morphologically annotated by MorphoDiTa, trained and tuned on the Slovak National Corpus tagset. In addition to the attribute lemma, we manually added another attribute lemma1, in which we always recorded the color in the form of an adjective, e.g. *red*. We explicitly assigned this value to all adjectives denoting the given color in the lemma attribute, as well as to all tokens that carry an implicit association with that color in the poem, e.g. *krvavý* 'bloody', *srdce* 'heart', *oheň* 'fire'. The color attribute was also assigned to multiple words within a construction when the adjective conveyed a broader color reference that extended to the modified noun, e.g. *red poppy*. Some colors were grouped under a broader, overarching color category, which we indicated in the annotation using an underscore, e.g. words like *ružička* 'little rose', *ružové slnko* 'pink (neutral) sun',

[1] We would like to express our gratitude to colleague Jakob Horsch for his willingness and help in creating an R script for data visualization, which we further modified according to our needs.

ružový máj ‘pink (masculine) May’ were assigned the color attribute *red_pink*. In the collection *Gift*, 33 tagged[2] colors and 23 overarching[3] color categories were identified. Given the nature of the annotated text and the specificity of literary interpretation, some color designations – such as *light-dark, white-black, transparent* – are considered non-standard colors.

In this study, we worked with the primary color labels in the visualization – e.g. the color *red_lilac* was processed as *red*. The annotated corpus data was placed into various types of graphical representations. Among them, the most suitable for our purposes proved to be a complex visualization of individual poems in the collection arranged vertically, as well as figures showing the distribution of colors within individual poems. The colors in poems are displayed in the figures in sequence according to the numerical position of the token as it appears in the corpus, with the *y*-axis representing the title of the poem.

The size of the color dots represents the accumulation of identical colors within the textual space. For example, in the poem Gift, Fig. 2 shows a third red color dot with a value of 5, indicating that five instances of red color references occur consecutively in the text without interruption by another color.

3 ANALYSIS OF KEY COLORS

Based on the analysis of overarching color frequency (see Note 2) in the collection *Gift*[4], several colors compete for prominence at the center of the color spectrum. This configuration of colors suggests that the somber, balladic atmosphere – despite being present in the thematic focus of several poems – is not overtly emphasized.

[2] red (40), white (37), dark_black (29), yellow (22), green (18), grey (15), dark_grey (15), golden (12), colorful (11), transparent (11), brown (10), dark (10), light (9), blue (6), greenish brown (6), colorless (5), red_ pink (4), dull grey (4), pitch dark (3), red_dark red (3), bright (2), silvery (2), light-dark_grey (2), goldish (2), red_violet (2), red_bloody (2), red_purple (2), spectrum (1), light-dark_ colorful (1), greenish blue (1), green, light green (1), green_dark green (1), black and white (1)

[3] dark (54), red (53), white (37), yellow (22), green (20), grey (15), golden (12), colorful (11), transparent (11), brown (10), light (9), blue (6), greenish brown (6), colorless (5), dull grey (4), light dark (3), pitch dark (3), bright (2), silvery (2), goldish (2), spectrum (1), greenish blue (1), black and white (1)

[4] The collection is divided into four parts: Dar (Gift), Balada o veľkom žiali (Ballad of Great Sorrow), Pieseň (Song), List (Letter), Láska (Love), Z večera (From the Evening), Balada o hre slnka a vetra (Ballad of the Play Between Sun and Wind), Západ marcového slnka (March Sunset), Sekera v lese (Ax in the Forest), Horniaky (Highlands), V marci (In March), Jiřímu Wolkrovi (To Jiří Wolker), Za Dušanom Kardossom (For Dušan Kardoss), Z kancionálu (From the Hymnal), Riadok z kancionálu (A Line From the Hymnal), Z knihy žalmov (From the Book of Psalms), Zo sanatoria (From the Sanatorium), Agonia (Agony), Balada o klamných ružiach (Ballad of Deceptive Roses), “Buď wůle Twá” (Thy Will Be Done), Milému (To My Beloved), Epilog (Epilogue), Legenda (Legend).

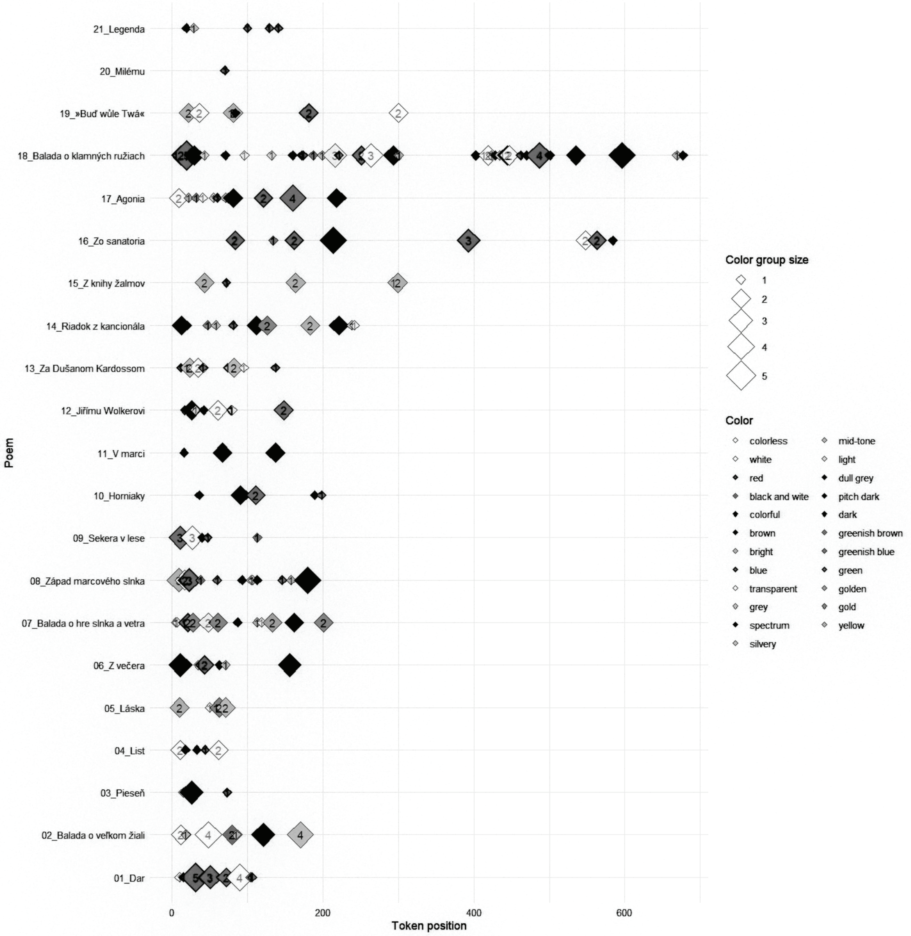


Fig. 1. Visualization of the colors in all the poems in the poetry collection *The Gift*

In the following section, we analyze the preferred colors identified in the debut collection *Gift* – namely red, gold and yellow, white, dark black[5], green and gray.

The graphical representation of individual poems shows that among the dominant color combinations, those featuring the overarching red color[6] stand out. This red is connected to either explicit expression or figurative and symbolic representation, as for example:

[5] Although this color is not an overarching one, we have chosen to pay special attention to it and its surrounding colors, particularly in contrast with white.

[6] In addition to explicitly stated red, the attribute red also includes the following: violet red, bloody red, crimson red, pink red, red dark red.

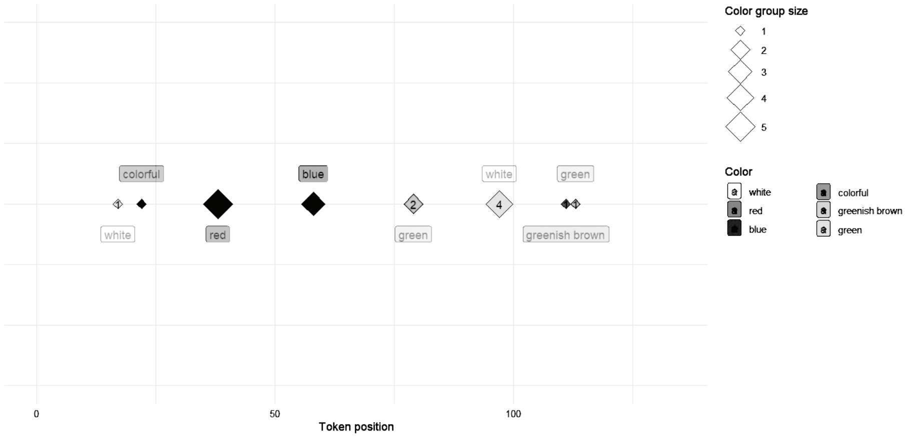


Fig. 2. Visualization of the colors in the poem *The Gift*

a) in the motif of the heart (13): Song (1), For Dušan Kardoss (3), From the Book of Psalms (1), From the Sanatorium (1), Agony (2), Thy Will be Done (2), To My Beloved (1), Legend (2);
b) in the motif of blood: *výčitka krvavá* 'bloody reprach' (Gift), *krvavá ruža* 'bloody rose' (Ballad of Deceptive Roses), *krv z rozťatej rany* 'blood from an open wound'*,* (March Sunset), *halúzky krvácali* 'twigs were bleeding' (Ax in the Forest);
c) in the motifs of natural phenomena and flowers: *červený mak* 'red poppy', *lilavé sirôtky* 'lilac pansies' (Gift), *ružové slnko* 'pink sun' (Ballad of the Play Between Sun and Wind), *rudá zora* 'crimson dawn'*, rudý klinčok* 'crimson carnation' (From the Sanatorium), *ruža* 'rose', *temnorudý kalich kvetov* 'dark crimson flower chalice' (Ballad of Deceptive Roses);
d) in representations of physicality and emotionality: *oheň oka* 'fire of the eye' (From the Evening), *rumeň líc* 'blush of the cheeks', *záblesk plameňa* 'flash of flame', *purpur rudý v tvári* 'crimson flush in the face' (Ballad of Deceptive Roses), and similar expressions.

The frequency of red forms a meaning-bearing axis throughout the entire collection. It is interesting to compare the presence of the color red in the framing poems Gift (5) and Legend (2), where it appears in the symbolism of the red poppy and the heart. Both poetically developed symbols are semantically linked to the gift of Divine love, which in the opening poem is sacred but in the last one is antithetically profaned.

It was only through the graphical representation of overarching colors that the distinct coloration of certain poems – serving as focal points within individual sections of the collection – became evident. For example, in the poem Love, yellow and gold colors dominate: yellow in the sky and the gold color of letters in the declaration of love *„Milujem!"* 'I Love!'; yellow underscores the spiritual dimension of transcendence, while gold emphasizes the sacredness of love. A similar character can be observed in the poem Ballad of the Play Between Sun and Wind, which was not identified as a semantic focal point within the analysis of the collection's polyfocal architecture (Hajdučeková and Bónová 2024, pp. 214–217). However, based on the analysis of color-related motifs – sun implying yellow color (3), *zlatá/zlatistá kader* 'golden/golden-hued lock' (4/2) and zlaté *vlasy* 'golden hair' (2) – as well as the poem's position in the collection, it becomes evident that it serves as a pendant to the poem Love, forming a polyfocal pair along the axis of individually experienced Božej L/lásky (Divine L/love) in an eternally present parable.

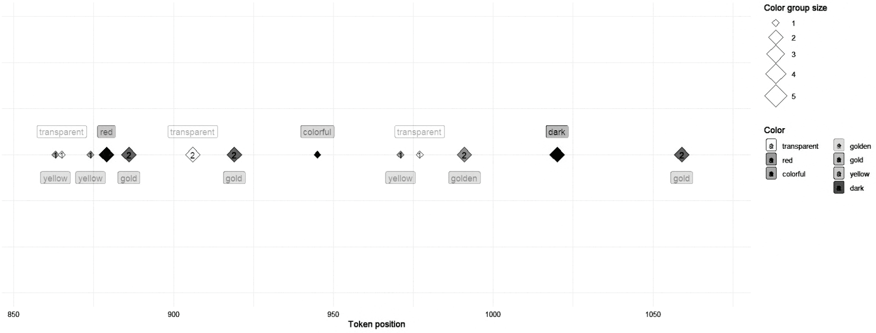


Fig. 3. Visualization of the colors in the poem *The Ballad of the Play Between Sun and Wind*

Yellow and gold also dominate in the poems A Line from the Hymnal and From the Book of Psalms, which form a focal point of the poetic whole. The yellow color here is also an iconic expression of the sacredness of the chalice and the yellow page in the psalter emphasizes the tradition of spiritual song and dialogue with God. *Žltá stuha* 'yellow ribbon' (3) also marks the place of a mother's final prayer. Both poems thus create an initiation point where a personal spirituality of the heart is born and unfolds within the poetic unit with the symbolic title Gift. The yellow color also appears in the third part: in the opening image of the golden sun as an accent marking sacralized time and space (Agony), as well as in the central poem Thy Will Be Done, where the yellow color of the hospital becomes a sign of mystery – a struggle for life in contact the Transcendent.

Through these semantic associations and new findings, it becomes evident that the specific colors – yellow and gold highlight the polyfocal structure of the collection's composition and symbolically reinforce the focal points of spiritualization. Yellow represents the suprasensory dimension of (auto)t/Transcendence, while gold iconically conveys the sacrality of expression.

The strong presence of white is already developed in the emblematic poem Gift (5), where its frequency matches that of red. White serves as a pendant to the red color in its evocation of spirituality, which is also seen in the graphical representation of several poems. In Gift, it forms the foundation of silk *'hodváb'* for the embroidered image of a bouquet and is also connected to the central symbol of the white dove *'holubice bielej'* – a biblical representation of the Holy Spirit. In contrast to red, which dominates primarily in floral motif, white – when in focal position – implies a penetration into the transcendent. It highlights the initiation point of sensuous perception of spirituality, that is, a diaphanously doubled reality (both revealed and concealed). It connects predominantly with:

a) natural imagery: *biela priepasť* 'white abyss', *biele stromy* 'white trees', *snežný rad* 'white row of snow' (Ballad of Great Sorrow); *obláčik ľahký, biely* 'a small, light white cloud' (2) in a combination with colorful blossoms (Letter); *snežné hory ako krv z rozťatej rany* 'snowy mountains like blood from an open wound' (March Sunset);

b) human fate and death: *biela poduška* 'white pillow' – here contrasted with the absent red carnation which serves as a sign of death (From the Sanatorium), *biela izba a záclony* 'white room and curtains' as delimited space which, when combined with red in the image of the transcending heart evokes the mystery of death (Agony), *krvavá ruža na bielej tvári a rumeň líc* 'bloody rose and blush of cheeks' or *purpur rudý* 'crimson flush', are signs of illness and also death (Ballad of Deceptive Roses), *biela posteľ v priestore sanatória* 'white bed in the sanatorium' – combined here *budovou žltou* 'yellow building' – signifies a spiritual struggle at the threshold of life and death (Thy Will Be Done).

A semantic analysis of the color white reveals that it becomes established as a symbolic marker of the perception of spirituality and meta-empirical dimension of reality. It is associated with intuitive and suprasensory experiences of auto-transcendence, faith, and existential situations. In the context of diaphony, it appears in a dual color pairing most often of white and red[7], or white and gray. In complementation with red – which in the collection is meaningfully tied to motifs of blossoms, the heart, blood and sacrifice – white contributes to a sensuously perceivable marker of the doubled threshold between life and death, which constitutes a semiotic precondition

[7] Colors in the poem From the Sanatorium represented by a particular token and its order in corpus: blue (siné 'dusky blue'; 85) – blue (mraky 'clouds'; 86) – greenish blue (more 'sea'; 135) – red (rudej 'crimson'; 163) – red (zore 'dawn'; 164) – dark_black (Čierne 'Black'; 211) – dark_black (nebo 'sky'; 213) – dark_black (černejšia 'blacker'; 218) – dark_black (zem 'soil'; 220) – red (ret 'lip'; 264) – red (retom 'with lip'; 415) – red (srdca 'of the heart'; 504) – white (bielu 'of white'; 549) – white (podušku 'of pillow'; 550) – red (klinčok 'carnation'; 563) – red (rudý 'crimson'; 565) – brown (hrudy 'of soil'; 585).

for the poetic shaping of the lyrical parable (Turner 2005). Thus, the color white intervenes in the composition of the lyrical structure and assumes the function of a compositional component (Sabolová 2000, pp. 500–503).

The color dark_black stands in contrast to white. Its occurrences, categorized under the overarching color dark[8], suggest that connotations of a balladic atmosphere – usually expressed explicitly through black – are relatively rare in this collection. The presence of the color black can be observed as follows:

a) in the attribute of initiatory objects: *kancionál čierny* 'black hymnal', *čierne dosky* 'black planks' (A Line From the Hymnal);
b) as a temporal mark: *čierne noci* 'black nights' (Song), *(pod)večer* '(from) evening)' (From Evening), *noc* 'night' (Ballad of the Play Between Sun and Wind), *ruže večerné* 'evening roses' (Ballad of Deceptive Roses);
c) as an attribute of topos: *brázdy zčernelých orníc* 'furrows of blackened soil' (March Sunset), *čierne nebo, černejšia zem* 'black sky, even blacker soil' (From the Sanatorium).
d) In the color spectrum, the appearance of dark_black (29) is surpassed by the combined presence of grey, dark_grey, light_gray and dull grey. With a total of 36 occurrences, these colors weaken and counterbalance the semantic effect of dark_black. Instead of reinforcing a tragic tension, they establish a gradual scale (while, dull grey, grey, light grey_grey, dull grey, dark_black) that functionally connects several key colors[9].

As the analysis has shown, the frequency and distribution of the color black in the collection *Gift* confirms the research hypothesis that it is not tragic balladic elements that carry semantic weight, but rather a psalm-like quality with a perspective of hope – one that is further supported by the function of the color green.

Fig. 1 clearly shows that green[10] appears with a relatively high frequency. In the opening poem Gift, it symbolizes hope through the motif of green leaves and is linked to a semantic shift from despair to hope and the initiation of a new path in life. In the subsequent poems, it participates only in the motif-based depiction of nature and in imagery: *husté lesy kosodrevia* 'dense dwarf pine forest', *tmavý les* 'dark forest' (Ballad of the Great Sorrow), *jedľa* 'fir tree', *lišajník* 'lichen' (March Sunset), *hory* 'mountains', *vŕby* 'willows', *topole* 'poplars' (Highlands), *les* 'forest' (3), *hory* 'mountains', *jedlice* 'silver firs' (To Jiří Wolker), *teplý mach* 'warm moss' (Agony).

[8] In addition to explicitly expressed black color (dark_black), the attribute dark also includes color labels such as dark_grey; the characteristic of darkness is also present in separate colors like light-dark_colored, light-dark_grey.

[9] Leaf: white (Obláčik 'Little cloud'; 10) – white (biely 'white'; 14) – dark_grey (večerom 'evening'; 19) – colorful (kvety 'flowers'; 34) – red (rety 'lips'; 45) – white (Obláčik 'Little cloud'; 61) – white (biely 'white'; 65).

[10] In addition to the explicitly expressed color green, the attribute green also includes the following variants: green_lightgreen, green_darkgreen.

From these connections, it becomes clear that the green color of natural elements implies the presence of verticality – in contrast to the axis of down versus up. This subtle green element of nature serves as a kind of hint of a transcendent perspective, carrying the semantics of hope, and it is against this backdrop that lyrical experiences are situated. This may help explain why, despite the spiritual nature of poetry, the color blues appears relatively rarely in the collection. Typically associated with the vertical axis of sky and earth, blue is linked to the motif of sky (Gift, Legend). A more prominent presence of the blue sky can be observed only in the motif of clouds, which, however, tends to lean toward white/gray tones. Nevertheless, the color blue does appear in both framing poems (Gift and Legend), as well as in the poem in Part 3 (From the Sanatorium), with all three being united into a semantic triad through a leitmotif: (God's) gift – love and life.

4 CONCLUSIONS

The chosen methodological approach – using an annotated corpus to track the frequency of colors and their shades, along with graphical visualizations of color clusters and their preferences – made it possible to evaluate key semantically charged components of the composition that would not have been identified through literary analysis and interpretation alone, without the intervention of linguistic tools. This is confirmed by the comparison: while in literary research the authors point to a preference for white, red, gold, and silver, the corpus shows that silver is only marginal, and the highest frequencies are found in red, yellow with gold, white, as well as gray and green. The semantic roles of the key colors are as follows: the overarching red forms a meaning-bearing axis throughout the collection and affirms the aptness of the label „poetry of the heart"; yellow and gold emphasize the collection's polyfocal structure and the focal points of spiritualization; white becomes established as a symbolic marker of the meta-empirical dimension of reality and contrasting black (within the overarching dark) does not hold a central position. Green implies a constant presence of a transcendent perspective. From this, it follows that the initial hypothesis – that the range of colors in the debut collection *Gift* functions as a semantically charged component of lyrical structure and contributes to the processes of spiritualization – has been confirmed. The poetic symbolization of color and its compositional use in expressing spirituality are integral to the author's strategy.

Based on the presented research, future work with annotated data from poetry collections will involve annotated data from the poetry collections for overarching color categories – e.g. treating black as an independent category, unifying variants of gray, and merging of yellow and gold.

ACKNOWLEDGEMENTS

The research in this paper was supported by the Slovak National Corpus project, supported by the Slovak Academy of Sciences, Ministry of Education, Research, Development and Youth of the Slovak Republic, Ministry of Culture of the Slovak Republic and the Ľudovít Štúr Institute of Linguistics, Slovak Academy of Sciences and by the Slovak Research and Development Agency under the projects APVV-19-0244 Methodological procedures in literary research with an impact on the media environment.

References

Hajdučeková, I., and Bónová, I. (2024). O lyrickom triptychu Maše Haľamovej (život a tvorba). Košice: Univerzita Pavla Jozefa Šafárika v Košiciach, Vydavateľstvo ŠafárikPress, 286 p.

Haľamová, M. (1928). Dar. Básne Máši Haľamovej. Bratislava: Literárny odbor Umeleckej besedy slovenskej, 54 p.

Haľamová, M. (1932). Červený mak. Druhá sbierka básní Maši Haľamovej. Bratislava: Sväz slovenského študentstva, 53 p.

Haľamová, M. (1966). Smrť tvoju žijem. Bratislava: Slovenský spisovateľ, 81 p.

Sabolová, O. (2000) Z terminologických otázok výskumu kompozície umeleckej prózy. In: K. Buzássyová (ed.): Človek a jeho jazyk. 1. Jazyk ako fenomén kultúry. Bratislava: Veda, pp. 500–503.

Slovenský národný korpus – masa-2.0. (2024). Bratislava: Jazykovedný ústav Ľ. Štúra SAV. Internal Corpus.

Turner, M. (2005). Literární mysl. O původu myšlení a jazyka. Brno: Host, 278 p.

Změlík, R. (2019). Konceptualizace barev v narativní fikci na pozadí kvantitativních modelu. Olomouc: Univerzita Palackého v Olomouci, 320 p.

sciendo
DOI 10.2478/jazcas-2025-0031

SYNTACTIC COMPLEXITY AND POLITICAL IDEOLOGY: A STUDY OF CZECHOSLOVAK AND CZECH PRESIDENTIAL SPEECHES

MIROSLAV KUBÁT[1] – MICHAELA NOGOLOVÁ[2] – XINYING CHEN[3] – ŽANETA STIBORSKÁ[4]
[1]Department of Czech Language, Faculty of Arts, University of Ostrava, Ostrava, Czech Republic (ORCID: 0000-0002-3398-3125)
[2]Department of Czech Language, Faculty of Arts, University of Ostrava, Ostrava, Czech Republic (ORCID: 0000-0001-7604-9765)
[3]Department of Czech Language, Faculty of Arts, University of Ostrava, Ostrava, Czech Republic (ORCID: 0000-0002-5052-4991)
[4]Department of Czech Language, Faculty of Arts, University of Ostrava, Ostrava, Czech Republic (ORCID: 0009-0005-5137-5130)

KUBÁT, Miroslav – NOGOLOVÁ, Michaela – CHEN, Xinying – STIBORSKÁ, Žaneta: Syntactic Complexity and Political Ideology: A Study of Czechoslovak and Czech Presidential Speeches. Journal of Linguistics, 2025, Vol. 76, No 1, pp. 345 – 354.

Abstract: This study analyzes syntactic complexity in Czechoslovak and Czech presidential speeches, using a corpus of New Year's addresses spanning nearly a century. Applying quantitative stylometric methods, we measure average sentence length (ASL), average clause length (ACL), mean dependency distance (MDD), and mean hierarchical distance (MHD) to compare syntactic structures across democratic and communist regimes. The results show that democratic presidents generally use more complex sentence structures, while communist-era speeches are syntactically simpler. However, individual differences are also observed inside groups. Husák exhibits higher complexity among communist leaders, and modern democratic presidents (Klaus, Zeman, Pavel) show a trend toward simplification. These findings confirm ideological influences on presidential rhetoric and highlight a broader shift toward linguistic accessibility in contemporary political speech.

Keywords: political discourse, syntactic complexity, presidential speeches, ideology, Czechoslovakia

1 INTRODUCTION

The analysis of political speeches has a well-established tradition in stylometry, with presidential addresses receiving particular attention. As pivotal rhetorical events, these speeches not only convey national visions and political ideologies but also reflect broader linguistic and social dynamics. Research in this field has primarily examined lexical features, thematic content, and discourse structures, with U.S. presidential speeches dominating the literature (e.g. Liu 2012; Lim 2004; Savoy 2010, 2016). Beyond the United States, scholars have explored political discourse in other national contexts, such as Italy (Tuzzi, Popescu and Altmann 2010) or Russia

(Kuznetsova 2016), demonstrating how political communication adapts to historical and ideological shifts.

This study examines New Year's presidential speeches delivered by Czechoslovak and Czech presidents, a unique and continuous corpus spanning nearly a century. These annual addresses provide an opportunity for heads of state to engage with the public, summarize national progress, and outline future policy directions. The longevity and regularity of these speeches make them an ideal dataset for investigating the changes of linguistic patterns in political rhetoric over time.

Previous research on Czechoslovak and Czech presidential speeches has primarily focused on lexical and thematic analyses. Čech (2014) conducted a quantitative study analyzing the thematic concentration of these speeches, arguing that totalitarian and democratic leaders exhibited distinct linguistic patterns due to their differing ideological orientations. Additional studies, such as David et al. (2013), explored thematic structuring and ideological markers in presidential rhetoric. Čech (2011) analyzed the frequency structure of these speeches, demonstrating how certain lexical characteristics remain stable while others change in response to political shifts. A related study by Jičínský and Marek (2017) combined phonetic and textual analyses to investigate pronunciation tendencies and stylistic variations in Czech presidential speeches. Their findings indicated that ideological shifts influenced word choice and affected voice characteristics as well.

Building on these findings, Kubát, Mačutek and Čech (2021) further examined structural patterns of presidential speeches, employing multiple quantitative measures such as moving-average type-token ratio, mean verb distance, and cluster analysis of frequently used words. Their research confirmed that democratic-era speeches displayed greater lexical diversity, while communist-era addresses exhibited repetitive patterns, aligning with ideological constraints of the time. These studies underscore the importance of linguistic analysis in political speech research, illustrating how stylistic and structural patterns are shaped by ideological contexts. Despite the wealth of research on lexical and thematic aspects of presidential speeches, syntactic complexity remains an underexplored dimension, particularly within the context of Czechoslovak and Czech presidential rhetoric. This study seeks to address this gap by providing a comprehensive syntactic analysis, offering new insights into the relationship between political ideology and linguistic structuring in presidential discourse.

Syntactic complexity, broadly defined as the structural sophistication of sentences, plays a crucial role in shaping political communication. More complex syntactic structures may signal rhetorical sophistication, authority, and persuasive intent, while simpler structures may reflect an effort to communicate more directly with a broader audience. Research in political communication suggests that authoritarian regimes often favor simplistic sentence structures to facilitate ideological indoctrination, whereas democratic leaders tend to employ more complex linguistic constructions to appeal to diverse audiences (Van Dijk 2006; Fairclough 1989). Therefore, the differences between communist and democratic leaders' speeches should be reflected in their syntactic complexity, with democratic rhetoric typically employing more complex sentence structures while authoritarian communication favors simplified, repetitive patterns for ideological clarity.

2 MATERIAL

The corpus under analysis consists of 95 annual speeches delivered by all twelve Czechoslovak and Czech presidents. This tradition began in 1935 when the first Czechoslovak President, Tomáš Garrigue Masaryk, addressed the nation on the occasion of the New Year. However, Masaryk delivered only one such speech, as he resigned from office in December 1935 due to age and health concerns. Since then, these speeches have become a consistent annual tradition, providing an important window into political discourse over time.

The texts were primarily obtained from Český rozhlas, the Czech public radio broadcaster, (accessible at: https://interaktivni.rozhlas.cz/prezidentske-projevy/). This archive includes both textual transcriptions and audio recordings of the speeches. Additionally, speeches delivered by the most recent presidents, Miloš Zeman and Petr Pavel, were sourced from the official website of the Office of the President of the Czech Republic (https://www.hrad.cz/). The overview of the material can be seen in Tab. 1.

group	president	number of texts	years
Democrats	Masaryk	1	1935
Democrats	Beneš	11	1936–1938, 1941–1948
–	Hácha	7	1939–1945
Communists	Gottwald	5	1949–1953
Communists	Zápotocký	4	1954–1957
Communists	Novotný	11	1958–1968
Communists	Svoboda	6	1969–1974
Communists	Husák	15	1975–1989
Democrats	Havel	13	1990–2003
Democrats	Klaus	10	2004–2013
Democrats	Zeman	10	2014–2023
Democrats	Pavel	2	2024–2025

Tab. 1. Overview of material

3 METHODOLOGY

In this analysis, we followed a structured approach. The speeches were parsed using UDPipe 2.0 (Straka 2018) with Universal Dependencies (UD) 2.15 models (Zeman et al. 2024), a widely used framework for syntactic annotation. The resulting dependency structures were then converted into the Surface Syntactic Universal Dependencies (SUD) scheme (Gerdes et al. 2018), which provides a more syntactically oriented representation of sentence structure.

To ensure data integrity and enable valid comparisons, we included only sentences that (i) contained a predicate (i.e., a finite verb or auxiliary) as the root of the sentence and (ii) did not contain abbreviations, numerical digits, or special characters. We computed these syntactic indices – average sentence length (ASL), average clause length (ACL), mean dependency distance (MDD) and mean hierarchical distance (MHD).

ASL was computed using two complementary approaches: (i) dividing the total number of words by the number of sentences and (ii) dividing the total number of clauses by the number of sentences. The first measure reflects overall sentence length, while the second captures clause density within sentences.

ACL was determined by dividing the total number of words by the total number of clauses, offering insight into clause-level complexity.

MDD, based on Liu (2008), quantifies syntactic complexity by measuring the average dependency distance (DD) across all words in a text, excluding punctuation and root words. The DD of a word is defined as the absolute difference between its id (position of the word in the sentence) and the id of its syntactic parent. The sum of all DDs in a text was divided by the total number of dependent words (i.e., total words minus the number of sentences), as formalized in following formula:

$$\text{MDD} = \frac{\sum_{i=1}^{n-s} |DD_i|}{n-s}$$

where n is the number of words, s is the number of included sentences, and DDi is the dependency distance of the i-th word.

MHD, introduced by Jing and Liu (2015), was computed analogously to MDD but using hierarchical distances (HDs) instead of dependency distances. The HD of a word represents the number of dependency edges between the word and the root of the sentence. MHD provides a deeper structural perspective, capturing the degree of syntactic embedding in a sentence.

To assess the significance of the results, we conducted statistical comparisons across the following groups:

1. Democratic-era (41 speeches) vs. Communist-era (47 speeches) vs. Hácha's 7 speeches.
2. Individual presidents compared against one another.

Before conducting statistical tests, we evaluated the normality of each group's distribution using the Shapiro-Wilk test (Shapiro and Wilk 1965). If normality was violated in at least one group, the Mann-Whitney U test (Mann and Whitney 1947) was applied as a non-parametric alternative. If both groups followed a normal distribution, we employed the independent samples t-test to compare means.

4 RESULTS

4.1 Democratic-era vs. communist-era speeches

The findings reveal differences in syntactic complexity across the ideological contexts of democratic and communist-era presidential speeches (see Tab. 2). Democratic speeches exhibit a tendency toward higher sentence length, both in terms of words and clauses. In contrast, communist-era speeches feature shorter sentences, suggesting a more constrained syntactic style. However, only sentence length measured in number of clauses shows a statistically significant difference ($p \leq 0.05$).

Interestingly, while communist-era speeches contain shorter sentences overall, they also feature longer clauses compared to democratic-era speeches ($p \leq 0.05$), indicating a shift in the internal structuring of sentences. This suggests that instead of expanding sentences with additional clauses, communist-era speeches tend to rely on more complex clause structures.

Further structural differences are evident in mean dependency distance (MDD) and mean hierarchical distance (MHD). Democratic speeches display greater syntactic depth and larger dependency distances, indicating a more complex structure. However, only MHD demonstrates a statistically significant difference ($p \leq 0.05$), reinforcing the notion that democratic speeches tend to be syntactically more sophisticated.

The case of Emil Hácha, whose speeches do not fit into the democratic or communist categories offers additional insights. Statistically, Hácha's speeches show a significantly higher average sentence length in clauses compared to communist-era presidents ($p \leq 0.05$), while his average clause length in words is significantly lower than that of communist leaders ($p \leq 0.05$). These findings suggest that his speeches share more similarities with democratic discourse. However, in terms of MHD, Hácha's speeches stand out as significantly lower than both democratic and communist presidents ($p \leq 0.05$), positioning his rhetorical style as distinct from both groups.

index	Democrats	sd	Communists	sd	Hácha	sd
ASL (words)	19.243	5.880	18.596	2.462	17.157	2.394
ASL (clauses)	2.429	0.573	1.888	0.344	2.154	0.363
ACL	7.895	1.184	10.010	1.456	8.032	0.688
MDD	2.306	0.188	2.174	0.111	2.199	0.207
MHD	4.339	0.578	4.226	0.440	3.747	0.413

Tab. 2. Average values (ASL, ACL, MDD, MHD) of democrats, communists and Hácha and their standard deviations (sd)

4.2 Individual presidents compared against one another

As can be seen in Tab. 3 and 4, syntactic complexity varies across individual presidents, reflecting differences in historical context, political ideology, and rhetorical strategy.

Among all presidents, Beneš consistently exhibits the highest syntactic complexity across multiple indices. His speeches feature significantly longer sentences (ASL), more clauses per sentence, and greater hierarchical depth (MHD) than those of most other presidents ($p < 0.001$ in comparisons with Klaus, Zeman, Pavel, among others). This complex rhetorical style reflects his diplomatic background and the high formal demands of early democratic leadership.

Havel also ranks among the most syntactically complex speakers, particularly in clause density (ASL in clauses) and sentence depth (MHD). His background as a playwright and philosopher likely contributed to his use of layered, introspective sentence structures. Havel's values are statistically higher than those of most communist-era and modern presidents ($p < 0.01$ in several cases), further distinguishing his rhetorical profile within the post-1989 democratic period.

In contrast, more recent presidents (Klaus, Zeman, and Pavel) exhibit a reduction in syntactic complexity. Klaus and Zeman both use shorter sentences than Beneš and Havel ($p < 0.001$ in multiple comparisons), indicating a shift toward more accessible political communication. Pavel, the most recent president, has the shortest ASL and lowest MDD in the corpus, with statistically significant differences from nearly all earlier presidents. These findings reflect a broader movement toward direct, media-adapted rhetoric in the contemporary political sphere.

Masaryk delivered only one New Year's address before resigning due to age and poor health, limiting statistical comparison. Although his background as a philosopher suggests a capacity for complex rhetoric, the speech ranks only moderately in complexity, possibly reflecting his condition at the time. Caution is therefore needed in interpreting this as representative of his typical style.

Among communist-era presidents, Husák stands out for using longer clauses and more complex sentence structures. His average clause length (ACL) is the highest of all presidents in the corpus. Unlike earlier communist leaders who used shorter, slogan-like sentences, Husák's speeches are more formal and technocratic. It could be explained by his academic background and the context of the normalization era of the 1970s and 1980s. His speeches often include bureaucratic phrasing and abstract discussions of economic planning.

Early communist presidents show varied rhetorical profiles, likely shaped by their backgrounds and ideological roles. Gottwald, with limited formal education and a background in manual labor, uses short, slogan-like sentences reflecting a simplified, mobilizing style. Zápotocký, in contrast, shows higher complexity, possibly due to his experience as a writer, suggesting a somewhat more elaborative approach, though still within ideological constraints.

Novotný and Svoboda also differ in their syntactic patterns. Novotný's relatively high values for sentence length and structural depth may reflect a more technocratic style, consistent with his engineering background. Svoboda, a military commander, ranks among the simplest speakers, favoring short, directive sentences. This may stem from a preference for clarity shaped by his military experience.

Hácha (1939–1945), whose presidency took place under foreign occupation, consistently ranks in the middle range of syntactic complexity, indicating a rhetorical style shaped by political ambiguity and limited agency.

	ASL (words)	**sd**	**ASL (clauses)**	**sd**	**ACL**	**sd**
Masaryk	16.000	0.000	1.833	0.000	8.727	0.000
Beneš	24.092	7.886	2.632	0.808	9.149	1.092
Hácha	17.157	2.394	2.154	0.363	8.032	0.688
Gottwald	17.997	1.188	1.888	0.111	9.530	0.170
Zápotocký	18.880	1.229	1.839	0.219	10.348	1.133
Novotný	19.882	1.763	2.310	0.330	8.693	0.910
Svoboda	15.970	4.539	1.652	0.227	9.586	1.722
Husák	18.828	1.597	1.684	0.133	11.215	1.008
Havel	22.225	3.437	2.712	0.382	8.207	0.675
Klaus	14.316	0.977	1.929	0.158	7.465	0.788
Zeman	16.424	2.270	2.501	0.434	6.622	0.551
Pavel	13.551	0.052	1.921	0.011	7.056	0.069

Tab. 3. Average values (ASL, ACL) of individual presidents and their standard deviations (sd)

	MDD	sd	MHD	sd
Masaryk	2.211	0.000	3.689	0.000
Beneš	2.475	0.174	4.598	0.420
Hácha	2.199	0.207	3.747	0.413
Gottwald	2.192	0.102	3.976	0.267
Zápotocký	2.330	0.086	4.284	0.145
Novotný	2.216	0.067	4.356	0.405
Svoboda	2.061	0.109	3.818	0.616
Husák	2.141	0.090	4.36	0.384
Havel	2.367	0.137	4.848	0.428
Klaus	2.273	0.141	3.775	0.251
Zeman	2.136	0.106	4.113	0.490
Pavel	2.041	0.036	3.871	0.090

Tab. 4. Average values (MDD, MHD) of individual presidents and their standard deviations (sd)

5 CONCLUSION

This study examined the relationship between syntactic complexity and political ideology in Czechoslovak and Czech presidential speeches, analyzing a nearly century-long corpus of New Year's addresses. By investigating multiple measures of syntactic complexity – including average sentence length (ASL), average clause length (ACL), mean dependency distance (MDD), and mean hierarchical distance (MHD) – the findings reveal distinctions between democratic and communist-era presidential discourse.

The results indicate that democratic-era presidents generally exhibit higher syntactic complexity across multiple metrics. Beneš and Havel stand out with the longest and most complex sentence structures, demonstrating a preference for rhetorically sophisticated discourse, which aligns with their intellectual and philosophical backgrounds. Conversely, communist-era leaders, particularly Svoboda and early communist presidents, exhibit lower syntactic complexity, favoring shorter sentences and reduced syntactic depth, a characteristic often associated with authoritarian discourse that prioritizes ideological clarity and mass accessibility. Hácha represents a transitional case, with syntactic complexity values that place him between democratic and communist leaders.

Over time, a trend toward simplification is observed in modern presidential rhetoric. Recent democratic presidents, such as Klaus, Zeman, and Pavel, exhibit significantly lower syntactic complexity compared to earlier democratic leaders. Their speeches contain shorter sentences, reduced clause length, and flatter syntactic structures, indicating a shift toward more direct and accessible political communication.

The results show that political ideology plays a key role in shaping syntactic choices in presidential speeches. The findings reveal that authoritarian regimes tend to favor structurally simpler and more constrained discourse, while democratic leaders historically have employed more complex syntactic constructions. However, the increasing simplification in modern democratic rhetoric suggests that other factors – such as media evolution, changing audience expectations, and the influence of digital communication – may now be driving linguistic change in political speech.

At the same time, the results reveal that these broader trends are significantly modulated by individual factors. Differences in educational background, professional training, and personal rhetorical style account for much of the variation observed within each ideological group. These findings indicate that although ideological context and historical shifts shape the general trajectory of political language, individual characteristics continue to play a crucial role in determining syntactic style.

ACKNOWLEDGEMENTS

This research is supported by Grant SGS04/FF/2025, University of Ostrava.

References

Čech, R. (2014). Language and ideology: quantitative thematic analysis of new year speeches given by Czechoslovak and Czech presidents (1949–2011). Quality & Quantity, 48(2), pp. 899–910.

David, J., Čech, R., Davidová Glogarová, J., Radková, L., and Šústková, H. (2013). Slovo a text v historickém kontextu – perspektivy historickosemantické analýzy jazyka. Brno: Host, 324 p.

Gerdes, K., Guillaume, B., Kahane, S., and Perrier, G. (2018). SUD or Surface-Syntactic Universal Dependencies: An annotation scheme near-isomorphic to UD. In Proceedings of the Second Workshop on Universal Dependencies (UDW 2018), pp. 66–74, Brussels, Belgium. Association for Computational Linguistics. Accessible at: https://aclanthology.org/W18-6008/.

Jičinský, M., and Marek, J. (2017). New year's day speeches of Czech presidents: phonetic analysis and text analysis. In: K. Saeed – W. Homenda – R. Chaki (eds.): Computer Information Systems and Industrial Management. Cham: Springer, pp. 110–121.

Jing, Y., and Liu, H. (2015). Mean Hierarchical Distance Augmenting Mean Dependency Distance. In Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015), pp. 161–170, Uppsala, Sweden. Accessible at: https://aclanthology.org/W15-2119/.

Kubát, M., Mačutek, J., and Čech, R. (2021). Communists spoke differently: An analysis of Czechoslovak and Czech annual presidential speeches. Digital Scholarship in the Humanities, 36(1), pp. 138–152.

Kuznetsova, J. (2016). Modern Russian history through the New Year addresses. In: S. Kübler – M. Dickinson (eds.): Proceedings of Computer Linguistics Fest 2016. Bloomington, IN: Indiana University, pp. 34–38.

Lim, E. T. (2004). Five trends in presidential rhetoric: an analysis of rhetoric from George Washington to Bill Clinton. Presidential Studies Quarterly, 32(2), pp, 328–348.

Liu, F. (2012). Genre analysis of American presidential inaugural speech. Theory and Practice in Language Studies, 2(11), pp. 2407–2411.

Liu, H. (2008). Dependency distance as a metric of language comprehension difficulty. Journal of Cognitive Science, 9(2), pp. 159–191.

Mann, H. B., and Whitney, D. R. (1947). On a Test of Whether One of Two Random Variables Is Stochastically Larger than the Other. Annals of Mathematical Statistics, 18, pp. 50–60.

Straka, M. (2018). UDPipe 2.0 Prototype at CoNLL 2018 UD Shared Task. In Proceedings of CoNLL 2018: The SIGNLL Conference on Computational Natural Language Learning, pp. 197–207. Association for Computational Linguistics, Stroudsburg, PA, USA. Accessible at: https://aclanthology.org/K18-2020/.

Savoy, J. (2010). Lexical analysis of US political speeches. Journal of Quantitative Linguistics, 17(2), pp. 123–141.

Savoy, J. (2016). Text representation strategies: an example with the State of the Union addresses. Journal of the Association for Information Science and Technology, 67(8), pp. 1858–1870.

Shapiro, S. S., and Wilk, M. B. (1965). An Analysis of Variance Test for Normality (Complete Samples). Biometrika, 52, pp. 591–611.

Tuzzi, A., Popescu, I.-I., and Altmann, G. (2010). Quantitative Analysis of Italian Texts. Lüdenscheid: RAM-Verlag, 161 p.

Van Dijk, T. A. (2006). Ideology and discourse analysis. Journal of Political Ideologies, 11(2), pp. 115–140.

Zeman, D. et al. (2024). Universal Dependencies 2.15, LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Prague: Charles University. Accessible at: http://hdl.handle.net/11234/1-5787.

sciendo
DOI 10.2478/jazcas-2025-0032

LONGER WORDS, EASIER-TO-PRONOUNCE PHONEMES: A PILOT STUDY

JÁN MAČUTEK[1] – RADEK ČECH[2] – MICHAELA KOŠČOVÁ[3]
[1]Mathematical Institute, Slovak Academy of Sciences, Bratislava, Slovakia
(ORCID: 0000-0003-1712-4395)
[2]Department of Czech Language, Faculty of Arts, Masaryk University, Brno, Czech Republic (ORCID: 0000-0002-4412-4588)
[3]Mathematical Institute, Slovak Academy of Sciences, Bratislava, Slovakia
(ORCID: 0000-0002-5541-1224)

MAČUTEK, Ján – ČECH, Radek – KOŠČOVÁ, Michaela: Longer Words, Easier-to-pronounce Phonemes: A Pilot Study. Journal of Linguistics, 2025, Vol. 76, No 1, pp. 355 – 365.

Abstract: The study investigates the relationship between word length and phoneme sonority in six languages across diverse language families. Building on the principle of least effort and the Menzerath-Altmann law, the research is aimed to analyze the phoneme sonority using translated New Testament texts in Bilua, Bola, Czech, Gagauz, Jamamadi, and Tongan. The findings reveal that in languages with complex syllables, the tendency of longer words to contain shorter syllables—consistent with the Menzerath-Altmann law—results in a higher proportion of vowels, thereby increasing the mean phoneme sonority. In contrast, languages with simple syllable structures exhibit either a decrease in mean phoneme sonority or no clear trend. Further, mean consonant sonority increases with word length in Bilua, Czech, and Gagauz, while no clear trend is observed in Bola, Jamamadi, and Tongan. Conversely, mean vowel sonority increases with word length in Bola, Jamamadi, and Tongan, but remains stable or decreases in Bilua, Czech, and Gagauz. Overall, the analysis reveals consistent patterns linking word length and sonority across all six languages.

Keywords: principle of least effort, phoneme, sonority, lenition, word length, syllable

1 INTRODUCTION

The principle of least effort (Zipf 1949) is one of key forces shaping structure and properties of language units. According to the principle, in language there is a tendency towards economizing that is probably the most easily noticeable in the relationship between frequency and length of language units. The well-known Zipf's law of abbreviation is usually formulated for words (the higher the frequency of a word, the shorter it tends to be, see Ferrer-i-Cancho et al. 2022), but the same holds true e.g. also for syllables (Rujević et al. 2021) and lengths of dependency distances (Chen and Gerdes 2022).

However, the validity of the principle of least effort is not restricted only to frequencies. The Menzerath-Altmann law (Menzerath 1954; Altmann 1980) states that longer units are composed of parts that are on average shorter (e.g. longer words consist of shorter syllables). It is another manifestation of the principle of least effort – if we must use longer words, we build them from simpler syllables. But the law refers to types (see Motalová et al. 2023 and Wang and Kelih 2024 for reasons why it is not valid for tokens), and thus disregards frequencies.

Similarly, the tendency of syllables to shorten is not the only possible way how to reduce effort in longer words. Already Hřebíček and Altmann (1996, p. 55) wrote that one could use "less complicated" instead of "shorter" parts in the formulation of the Menzerath-Altmann law. In this paper, we present some tendencies in the relationship between word length and phoneme sonority in six languages. Longer words tend to contain phonemes that are easier to pronounce (either in absolute terms, or relatively with respect to their neighbours).

The study is motivated by the fact that some languages allowing only simple syllable structure (i.e. only CV and V syllables, see Maddieson 2007, p. 96) display non-standard behaviour with respect to the Menzerath-Altmann law, see Mačutek et al. (2025). The tendency to use easier-to-pronounce phonemes in longer words seems to be universal regardless of syllable types allowed in particular languages.

We emphasize that we analyze words on the phonological, and not on the phonetic level, e.g. we consider theoretical properties of phonemes in written texts, and not physical properties of sounds in actual utterances.

2 METHODOLOGY AND LANGUAGE MATERIAL

2.1 Sonority hierarchy

Phonemes in particular languages are ranked according to sonority hierarchy, see Tab. 1. We follow mostly Szigetvári (2008, p. 96); in addition, fricatives are merged into one category with plosives (Parker 2011 writes that "...the placement of affricates between stops and fricatives is a controversial issue, remaining open to disagreement. Many scales either leave affricates out entirely or group them with plosives…"). In diphthongs, both vowels are taken into account.

phonemes	sonority index
low vowels	10
mid vowels	9
high vowels and semivowels	8
rhotics	7
laterals	6
nasals	5
voiced fricatives	4
voiceless fricatives	3
voiced plosives and affricates	2
voiceless plosives and affricates	1

Tab. 1. Sonority hierarchy

Any sonority hierarchy provides only a ranking of phonemes. The ranks do not reflect actual differences (e.g. the difference between voiceless fricatives and voiced fricatives does not have to be the same as the one between rhotics and high vowels). Anyway, it can be used to characterize the mean sonority of phonemes in words.

2.2 Language material

As language material, we use translations of the New Testament (27 books) into six languages from five different language families: Bilua (from the Central Solomon language family), Bola (Austronesian), Czech (Indo-European), Gagauz (Turkic), Jamamadi (Arawan), and Tongan (Austronesian). The Bible as a source of texts has its drawbacks (e.g. there are many proper names especially of Greek, Hebrew, and Latin origin), but for many languages it is the only easily available collection of texts that are long enough to enable statistical analyses. Book titles, references to other sources etc. were deleted. Links to Bible translations can be found in Tab. 2 (if the webpage provides access only to individual chapters, the link to the first chapter of the Gospel of Matthew is given).

language	link
Bilua	https://www.bible.com/bible/2979/MAT.1.BLBNT
Bola	https://www.scriptureearth.org/data/bnp/PDF/00-PBIbnp-web.pdf
Czech	https://bible.jecool.net/wp-content/uploads/2016/03/bible-velka.pdf
Gagauz	https://www.bible.com/en-GB/bible/2554/MAT.1.GAGNTL
Jamamadi	https://www.bible.com/bible/3158/MAT.1.JAANT
Tongan	https://ebible.org/pdf/ton/ton_nt.pdf

Tab. 2. Links to texts used

Four of these languages have only simple syllables. In Bilua (Obata 2003), all monosyllables are of the CV structure. In longer words, the first syllable can be CV or V, with all other syllables being CV. Bola (van den Berg and Wiebe 2019), Jamamadi (Dixon and Vogel 2004), and Tongan (Garellek and Tabain 2020) have only CV and V syllables without positional restrictions. Words containing syllables of other types (e.g. toponyms like *Nasaret* ‘Nazareth’ in Bola) were removed.

On the other hand, Czech (Short 1993) and Gagauz (Pokrovskaja 1964) allow also more complex syllables (and, consequently, consonant clusters exist in these two languages).

As we focus on the phonological analyses of written texts, it is important to note that all these languages have shallow orthographies, i.e. the phoneme-grapheme ratios are close to one-to-one (see Coulmas 2002, pp. 101–102). Therefore, the phonological transcriptions are relatively easy to do.

3 RESULTS

In order to guarantee certain stability of the means, in the following tables and figures only those word lengths are presented for which at least ten different words occur in the text.

3.1 Menzerath-Altmann law

We first present results of the analysis of the Menzerath-Aktmann law (see Tab. 3 and Fig. 1), as they are needed to understand the development of the mean sonority in the next sections. Data for Bilua, Bola, Jamamadi, and Tongan are taken from Mačutek et al. (2025). We add the results for Czech and Gagauz. The mean syllable length decreases with the increasing word length in languages with complex syllables (Czech, Gagauz). Languages with only simple syllables (Bilua, Bola, Jamamadi, Tongan) do not display a clear Menzerathian tendency.

word length in syllables	mean syllable length in phonemes					
	Bilua	Bola	Czech	Gagauz	Jamamadi	Tongan
1	1.96	1.90	3.52	2.88	1.92	1.88
2	1.94	1.93	2.70	2.51	1.91	1.87
3	1.96	1.88	2.41	2.41	1.93	1.86
4	1.96	1.89	2.24	2.38	1.94	1.84
5	1.96	1.90	2.17	2.35	1.96	1.84
6	1.97	1.86	2.18	2.31	1.96	1.86
7	1.96		2.14	2.26	1.97	1.86
8					1.97	1.83
9					1.97	1.82
10					1.97	
11					1.97	

Tab. 3. Relationship between word length and the mean syllable length

3.2 Mean phoneme sonority

The mean phoneme sonority (Tab. 4, Fig. 2) decreases with the increasing word length in Bilua. It is a consequence of the syllable structure in this language. As only the first syllable can be V and all other must be CV, the proportion of consonants increases, and, as consonants are less sonorous then vowels, the mean sonority decreases.

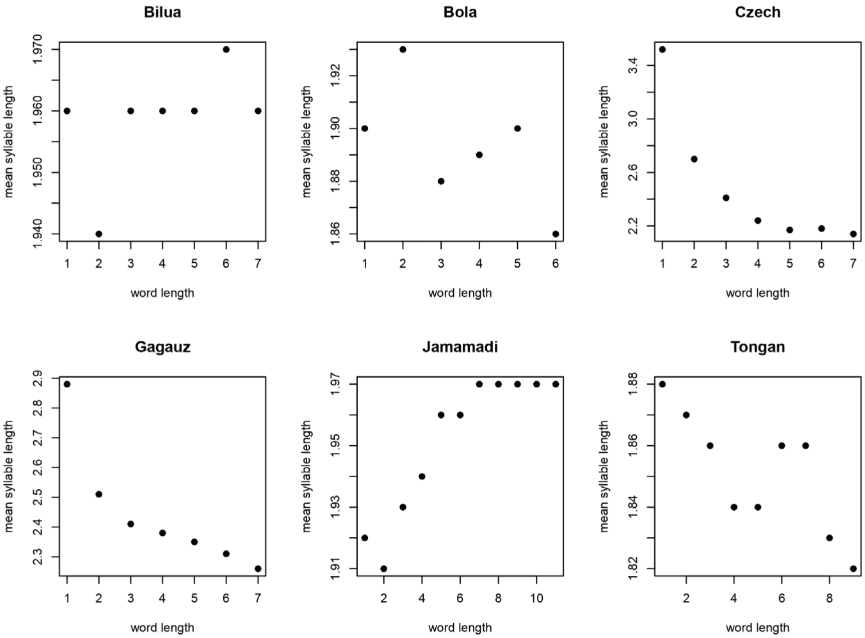


Fig. 1. Relationship between word length and the mean syllable length

In Czech and Gagauz, we observe the opposite trend, the mean phoneme sonority increases with the increasing word length. It is a consequence of the Menzerath-Altmann law – syllables get shorter, which means a higher proportion of vowels, and vowels have a higher sonority than consonants.

No clear trend is visible in Bola, Jamamadi, and Tongan.

word length in syllables	mean phoneme sonority					
	Bilua	Bola	Czech	Gagauz	Jamamadi	Tongan
1	6.88	6.47	5.22	5.45	6.67	6.19
2	6.61	6.42	5.68	5.89	6.55	6.33
3	6.60	6.66	5.92	6.14	6.48	6.28
4	6.47	6.54	6.14	6.26	6.53	6.27
5	6.44	6.52	6.26	6.37	6.57	6.27
6	6.47	6.63	6.13	6.40	6.59	6.19
7	6.48		6.20	6.44	6.62	6.21
8					6.62	6.25
9					6.64	6.31
10					6.59	
11					6.47	

Tab. 4. Relationship between word length and the mean phoneme sonority

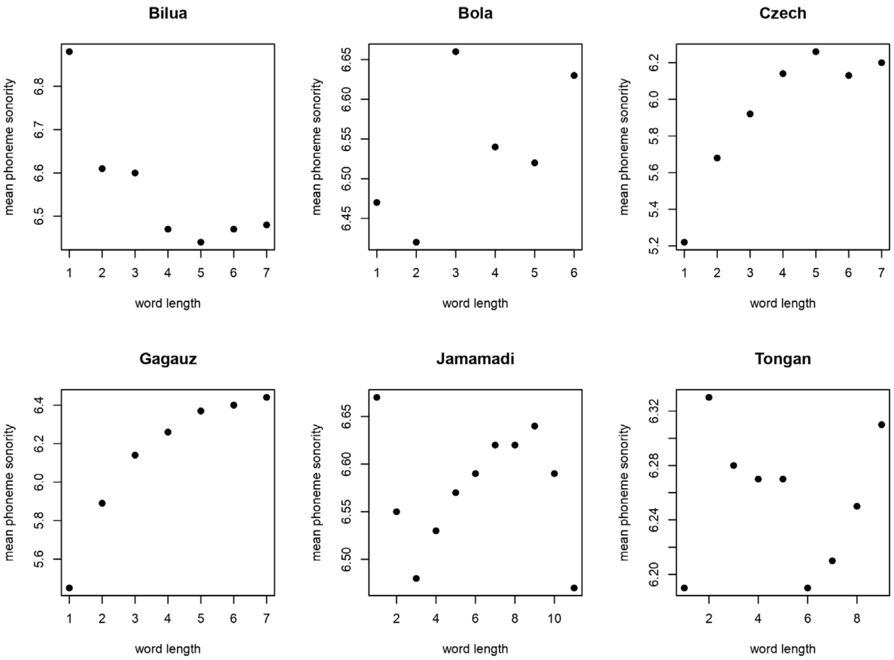


Fig. 2. Relationship between word length and the mean phoneme sonority

3.3 Mean consonant sonority

The mean consonant sonority increases in Bilua, Czech, and Gagauz (see Tab. 5 and Fig. 3). An analogy with lenition offers itself to explain this observation.

In Bilua, all consonants, possibly except for the one at the beginning of a word, are in intervocalic positions. As the mean syllable length decreases when words get longer, the probability that a consonant is in an intervocalic position in Czech and Gagauz increases.

According to Kirchner (2001, p. 138), "[i]t is fairly well established that intervocalic position is a natural lenition environment", and lenition closely correlate with increasing sonority (i.e. voicing is one of exemplifications of lenition). While we cannot apply this term literally (we observe neither a diachronic development of a language, nor preferring lenited consonants by individual speakers), we can say that Bilua, Czech, and Gagauz prefer more sonorous consonants in intervocalic positions, and these positions occur more often in longer words. These findings are in line with the paper by File-Muriel (2016) who reports an increasing lenition rates of /s/ in a Colombian variety of Spanish (though there is also an important difference – he works with tokens, not with types).

The mean consonant sonority behaves quite chaotically, or at least with a much less clear trend, in Bola, Jamamadi, and Tongan. All consonants that are not at the beginning of words are in intervocalic positions too (as in Bilua), but these languages allow also vocalic clusters (which Bilua forbids).

word length in syllables	mean consonant sonority					
	Bilua	Bola	Czech	Gagauz	Jamamadi	Tongan
1	3.22	3.46	3.63	3.08	3.79	2.67
2	3.34	3.63	3.73	3.55	3.66	3.12
3	3.45	3.83	3.77	3.86	3.51	3.02
4	3.40	3.61	3.94	4.06	3.67	2.89
5	3.45	3.48	3.99	4.17	3.79	2.85
6	3.54	3.51	3.86	4.19	3.86	2.71
7	3.53		3.99	4.27	3.89	2.66
8					3.90	2.68
9					3.93	2.71
10					3.80	
11					3.60	

Tab. 5. Relationship between word length and the mean consonant sonority

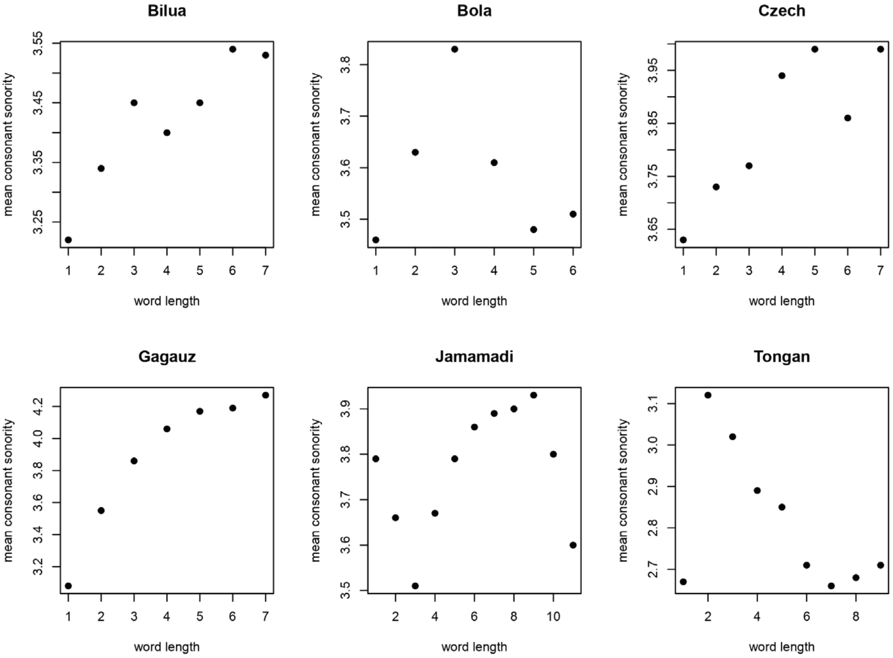


Fig. 3. Relationship between word length and the mean consonant sonority

3.4 Mean vowel sonority

The exact opposite of Section 3.3 is mean vowel sonority (see Tab. 6 and Fig. 4) – the mean sonority of vowels increases with the increasing word length in Bola, Jamamadi (admittedly, not so regularly), and Tongan, i.e. in languages in which there was no trend in the mean consonant sonority. On the other hand, Bilua, Czech, and Gagauz, languages with a systematic increase in consonant sonority, the mean vowel sonority does not increase (it rather decreases in Bilua and Gagauz, and behaves irregularly in Czech).

word length in syllables	mean vowel sonority					
	Bilua	Bola	Czech	Gagauz	Jamamadi	Tongan
1	8.97	8.92	8.86	8.95	9.19	9.00
2	9.02	8.94	8.82	8.90	9.12	9.00
3	8.98	9.08	8.85	8.88	9.21	9.01
4	8.97	9.12	8.86	8.89	9.22	9.05
5	8.96	9.21	8.90	8.90	9.21	9.09
6	8.91	9.30	8.83	8.89	9.22	9.15
7	8.94		8.83	8.86	9.26	9.22
8					9.25	9.22
9					9.28	9.24
10					9.28	
11					9.23	

Tab. 6. Relationship between word length and the mean vowel sonority

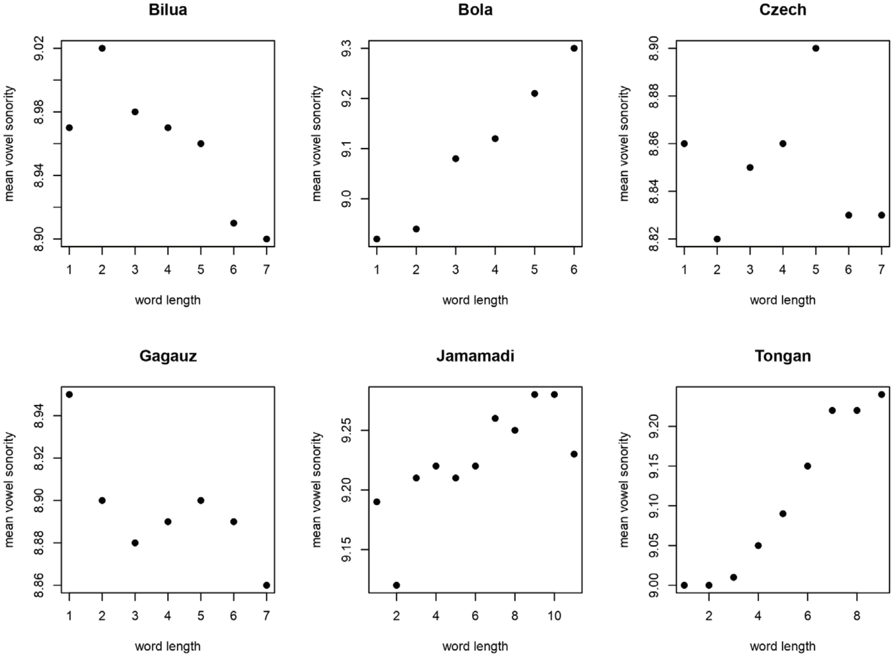


Fig. 4. Relationship between word length and the mean vowel sonority

4 CONCLUSION AND DISCUSSION

There is a systematic relationship between word length and sonority in all six languages under analysis. The nature of this relationship seems to depend on the syllable types allowed in individual languages. However, there is a connection to the principle of least effort in all six cases.

Czech and Gagauz have complex syllables and consonant clusters occur in them. As a consequence of the Menzerath-Altmann law, syllables are shorter in longer words. This means that consonant clusters are less probable in longer words, and syllalbes (and phonemes in them) become less difficult to be pronounced (see e.g. Stoel-Gammon 2010, p. 273). The Menzerath-Altmann law explains also why the mean phoneme sonority increases as words get longer – shorter syllables have higher proportions of vowels, and vowels are more sonorous then consonants.

In Bilua, due to the positional restrictions of its syllables types (V only at the beginning of words that have at least two syllables, CV elsewhere), the proportion of consonants increases in longer words. Therefore, there is a negative correlation between word length and the mean sonority of phonemes.

We can observe a positive correlation between word length and the mean sonority of consonants in Bilua, Czech, and Gagauz. According to Gurevich (2011), "[v]oicing, for example, has an explanation rooted in the laws of physics, specifically aerodynamics: intervocalically the vocal cords may continue to vibrate after the first vowel, through the consonant, and into the second vowel". Voicing is a typical exemplification of lenition, and voiced consonants are higher in the sonority hierarchy then their unvoiced counterparts. The increasing sonority of consonants in intervocalic positions thus weakens articulatory effort and thus compensates for increasing word length. And indeed, consonants in intervocalic positions are more probable in longer words – in Czech and Gagauz as a consequence of the Menzerath-Altmann law, and in Bilua because of more intervocalic slots available for consonants.

In Bola and Tongan (and to a slightly lesser extent also in Jamamadi) there is a positive correlation between word length and the mean vowel sonority. It means that lower vowels are preferred in longer words. But according to Jaeger (1978, p. 313), "the narrower constriction for high vowels causes air pressure in the oral cavity to be greater than that during low vowel", and (Napoli et al. 2014, p. 427) "[c] onsequently, more pulmonic effort is needed for the airflow across the glottis to overcome the resistive force of the oral cavity's higher air pressure". Thus, low vowels (with a higher sonority) are preferred because they require less effort.

Many questions appear with every (incomplete) answer. The impact of phoneme inventory size and structure must be investigated (e.g. if a language has more pairs of voiced and unvoiced consonants, it can make utterances easier by voicing consonants in intervocalic positions; if not, it can prefer lowering of vowels). We focused here on types – tokes must be studied too. And more languages must be analysed before one can reach a conclusion. But this study confirms once more that the least effort principle is (almost) ubiquitous in language.

The principle of least effort is, however, a double-edged sword. For a speaker without a hearer, it would be the most comfortable to utter only easy-to-pronounce phonemes (e.g. only low vowels). But e.g. CV syllables with a higher difference in sonority of a consonant and a vowel are easier to segment for a hearer (Yavas and Gogate 1999). Being too "lazy", a speaker would risk information loss at a hearer's side and a necessity of re-sending a message, which would require another effort. Thus, language finds itself in a state of a Zipfian equilibrium (Zipf 1935) in which the speaker's drive to economize is controlled by the hearer's feedback.

ACKNOWLEDGEMENTS

Supported by EU NextGenerationEU through the Recovery and Resilience Plan for Slovakia under the project No. 09I03-03-V04-00748 (J. Mačutek), by the European Regional Development Fund project "A lifetime with language: the nature and ontogeny of linguistic communication (LangInLife)" (reg. no.: CZ.02.01.01/00/23_025/0008726) (R. Čech), and by projects APVV-21-0216 and VEGA 2/0120/24 (M. Koščová).

References

Altmann, G. (1980). Prolegomena to Menzerath's law. In: R. Grotjahn (ed.): Glottometrika 2. Bochum: Brockmeyer, pp. 1–10.

Chen, X., and Gerdes, K. (2022). Dependency distances and their frequencies in Indo-European language. Journal of Quantitative Linguistics, 29(1), pp. 106–125.

Coulmas, F. (2002). Writing systems. An introduction to their linguistic analysis. Cambridge: Cambridge University Press. 270 p.

Dixon, R. M. W., and Vogel, A. R. (2004). The Jarawara language of Southern Amazonia. Oxford: Oxford University Press, 636 p.

Ferrer-i-Cancho, R., Bentz, C., and Seguin, C. (2022). Optimal coding and the origin of Zipfian laws. Journal of Quantitative Linguistics, 29(2), pp. 165–194.

File-Muriel, R. J. (2010). Lexical frequency as a scalar variable in explaining variation. The Canadian Journal of Linguistics, 55(1), pp. 1–25.

Garellek, M., and Tabain, M. (2020). Tongan. Journal of the International Phonetic Association, 50(3), pp. 406–413.

Gurevich, N. (2011). Lenition. In: M. van Oostendorp – C. J. Ewen – E. Hume – K. Rice (eds.): The Blackwell companion to phonology. Chichester: Wiley – Blackwell, pp. 1559–1575.

Hřebíček, L., and Altmann, G. (1996). Levels of order in language. In: P. Schmidt (ed.): Glottometrika 15. Issues in general linguistic theory and the theory of word length. Trier: WVT, pp. 38–61.

Jaeger, J. J. (1978). Speech aerodynamics and phonological universals. In: J. J. Jaeger – A. C. Woodbury – F. Ackerman – C. Chiarello – O. D. Gensler – J. Kingston – E. E. Sweetser – H. Thompson – K. W. Whistler (eds.): Proceedings of the 4th annual meeting of the Berkeley Linguistics Society. Berkeley (CA): Berkeley Linguistics Society, pp. 312–329.

Kirchner, R. (2001). An effort based approach to consonant lenition. Abingdon: Routle.g. 303 p.

Mačutek, J., Nogolová, M., Rovenchak, A., and Čech, R. (2025). What does the Menzerath-Altmann law really say? Journal of Quantitative Linguistics (accepted paper).

Maddieson, I. (2007). Issues in phonological complexity: Statistical analysis of the relationship between syllable structures, segment inventories, and tone contrasts. In: M.-J. Solé – P. Beddor Speeter – M. Ohala (eds.): Experimental approaches to phonology. Oxford: Oxford University Press, pp. 93–103.

Menzerath, P. (1954). Die Architektonik des deutschen Wortschatzes. Bonn: Dümmler. 131 p.

Motalová, T., Mačutek, J., and Čech, R. (2023). Word length in Chinese: The Menzerath-Altmann law is valid after all. Journal of Quantitative Linguistics, 30(3–4), pp. 301–321.

Napoli, D. J., Sanders, N., and Wright, R. (2014). On the linguistic effects of articulatory ease, with a focus on sign languages. Language, 90(2), pp. 424–456.

Obata, K. (2003). A grammar of Bilua: A Papuan language of the Solomon Islands. Canberra: The Australian National University. 333 p.

Parker, S. (2011). Sonority. In: M. van Oostendorp – C. J. Ewen – E. Hume – K. Rice (eds.): The Blackwell companion to phonology. Chichester: Wiley – Blackwell, pp. 1160–1184.

Pokrovskaja, L. A. (1964). Grammatika gagauzskogo jazyka (A grammar of the Gagauz language). Moskva: Nauka. 300 p.

Rujević, B., Kaplar, M., Kaplar, S., Stanković, R., Obradović, I., and Mačutek, J. (2021). Quantitative analysis of syllables in Croatian, Serbian, Russian, and Ukrainian. In: A. Pawłowski – J. Mačutek – S. Embleton – G. Mikros (eds.): Language and text: Data, Models, Information and Applications. Amsterdam, Philadelphia: Benjamins, pp. 55–67.

Short, D. (1993). Czech. In: B. Comrie – G. G. Corbett (eds.): The Slavonic languages. London: Routle.g. pp. 455–532.

Stoel-Gammon, C. (2010). The word complexity measure: Description and application to developmental phonology and disorders. Clinical Linguistics & Phonetics, 24(4–5), pp. 271–282.

Szigetvári, P. (2008). What and when? In: J. Brandão de Carvalho – T. Scheer – P. Ségéral (eds.): Lenition and fortition. Berlin, New York: de Gruyter, pp. 93–129.

van den Berg, R., and Wiebe, B. (2019). Bola grammar sketch. Ukarumpa: SIL-PNG Academic Publications, 292 p.

Wang, T., and Kelih, E. (2024). Boundary conditions for the Menzerath-Altmann law. What should be taken: Tokens, types or lemmas? Glottometrics 57, pp. 1–20.

Yavas, M., and Gogate, L. J. (1999). Phoneme awareness in children: A function of sonority. Journal of Psycholinguistic Research, 28(3), pp. 245–260.

Zipf, G. K. (1949). Human behavior and the principle of least effort. Cambridge (MA): Addison-Wesley Press, 573 p.

Zipf, G. K. (1935). The psycho-biology of language. Boston: Houghton-Mifflin, 336 p.

sciendo
DOI 10.2478/jazcas-2025-0033

QUANTITATIVE CORPUS ANALYSIS OF VLADIMIR PUTIN'S SPEECHES

PETR POŘÍZKA[1] – POLINA IVANKOVA[2]
[1]Department of Czech Studies, Faculty of Arts, Palacký University, Olomouc, Czech Republic (ORCID: 0000-0001-6980-9148)
[2]Department of Czech Studies, Faculty of Arts, Palacký University, Olomouc, Czech Republic

POŘÍZKA, Petr – IVANKOVA, Polina: Quantitative Corpus Analysis of Vladimir Putin's Speeches. Journal of Linguistics, 2025, Vol. 76, No 1, pp. 366 – 377.

Abstract: This study presents a quantitative analysis of Vladimir Putin's public speeches from 1999 to 2024, utilizing a corpus of approximately 1.75 million words sourced from the official Kremlin website. Using computational linguistic techniques, including hierarchical cluster analysis, TF*IDF keyword extraction and temporal trend analysis, the research systematically examines the evolution of Putin's rhetorical style and thematic content. Significant stylistic and thematic shifts are identified, particularly around 2012, coinciding with his return to the presidency and a notable increase in authoritarian governance. Prominent themes such as national identity, economic policy, security, and international conflicts vary significantly in prominence and shift in emphasis in response to key historical and political milestones. The findings reveal clear correlations between Putin's language patterns and major political or critical events, including the Chechen conflict, Putin's political rise and return to the presidency, the annexation of Crimea, and the recent military interventions in Ukraine. These findings demonstrate Putin's strategic rhetorical adaptability within changing geopolitical contexts.

Keywords: corpus, hierarchical clustering, keyword analysis, quantitative analysis, political discourse, Vladimir Putin, TF*IDF

1 INTRODUCTION AND PREVIOUS RESEARCH

Quantitative linguistics is concerned with measurable aspects of language, and its application to political discourse offers a systematic and potentially more objective way of analysing large amounts of textual data than standard methods based on introspection (without the use of computer-based analysis). Using quantitative analysis, it is possible to find patterns and trends that might remain hidden in a purely qualitative approach. In this paper we try to find quantitatively important patterns, tendencies or differences in Putin's language. It is also the first analysis of its kind on such a large dataset, processing Putin's speeches over the last 25 years.

Previous studies of Putin's speeches have predominantly used qualitative discursive methods. These include comparative discourse analyses of speeches by Putin and Biden (Kopik 2023), rhetorical analyses using Fairclough's framework (Shahbaz and Nawab 2024), critical discourse analyses comparing statements by Putin and Zelensky during the invasion of Ukraine (Tutar and Bağ 2023), speech act theory analyses highlighting Putin's performative language (Fafiyebi 2025), and analyses of Putin's narrative strategies during key events such as the 'special military operation' (Kadim 2023). Other studies focus on comparisons of war rhetoric (Chiluwa and Ruzaite 2024), socio-cognitive perspectives (Al-Manaseer 2025), and linguistic arguments in support of geopolitical claims, notably in Putin's 2021 essay on Ukrainian dialects (Maxwell 2025). These analyses tend to be narrow in scope, focusing on specific speeches or short periods of time.

Quantitative studies on Putin's rhetoric remain rare and tend to focus on narrowly defined topics. Only three academic papers have been identified: Wang and Zeng (2023) compare stylometric features and political themes in speeches by Putin, Medvedev, Trump, and Obama; Oleinik (2023) evaluates Cohen's d and Z-scores for term specificity in war-related political discourse; and Janda et al. (2022) apply keymorph analysis to examine changes in grammatical case usage (e.g. *Russia*, *Ukraine*, *NATO*) before and after the invasion of Ukraine.

2 CORPUS OF PUTIN'S SPEECHES

We used the speeches and texts available on the official website of the President of the Russian Federation (http://kremlin.ru/) as our data source. The total number of hours of Putin's speeches available on this website exceeds 7,500 hours; we processed and analysed approximately 4,000 hours of material. The corpus includes speeches to citizens, journalists, ambassadors, politicians, media interviews, press conference recordings, meetings with government and military officials, urgent security meetings, speeches on the occasion of holidays (Victory Day, Family Day, New Year, etc.).

The corpus covers the entire period of Putin's top political career from 1999 to 2024, the texts are in TXT format (no further annotations yet) and are bilingual: in the original Russian and in Czech translation. Because of the size of the data, the Czech translation was created automatically using the AI tool *DeepL* (https://www.deepl.com/). Given the genetic and typological similarities (Slavic languages of the inflectional type), we can assume that the translation is largely accurate. The data are segmented by years or multi-year periods (e.g. 1999–2000, 2001, 2002, ... etc. up to the 2020–2022, 2023–2024 collections). The year 2012 is divided into two sub-collections, the first up to May 2012, when Putin was still Prime Minister, and the texts after this period, which fall within his second presidential term (see below). The total size of the corpus we used for the analysis is about 1.75 million words (1,751,134 tokens).

At the moment it is a simple set of texts, but our plan is to create a standard trilingual corpus (original Russian + Czech and English translation) with linguistic annotations, which will allow a deeper and more detailed study of the linguistic aspects of Putin's speeches (including grammar).

2.1 Historical and political context

The corpus covers different political phases in Vladimir Putin's career. He was Prime Minister from 1999 to 2000 and President during his first term from 2000 to

2008. From 2008 to 2012, he was prime minister again under President Dmitry Medvedev. Since 2012, Putin has been president during a period characterised by increasing authoritarianism, centralisation of power, assertive nationalism, growing international isolation, intensified and militarised foreign policy aggression, culminating in the annexation of Crimea in 2014 and the invasion of Ukraine, key events that have set the tone and focus of his political rhetoric in recent years.

3 CLUSTER ANALYSIS OF TEXT CORPUS

Hierarchical cluster analysis is a powerful unsupervised learning method used in stylometry to uncover latent structures within high-dimensional textual data. This approach is particularly suited to diachronic investigations, allowing researchers to trace stylistic shifts over time without making prior assumptions about text categorisation (Burrows 2002; Hoover 2003; Eder et al. 2016).

The cluster analysis (run in R and the *stylo* package with default settings) used the 100 most frequent words (MFW) in each text as input variables. The data were clustered using three different distance metrics: *cosine*, *Euclidean*, and *min-max distance*.[1] These metrics were chosen to test the stability of the clustering outcomes across different models of similarity. The results are shown in Fig.1.[2]

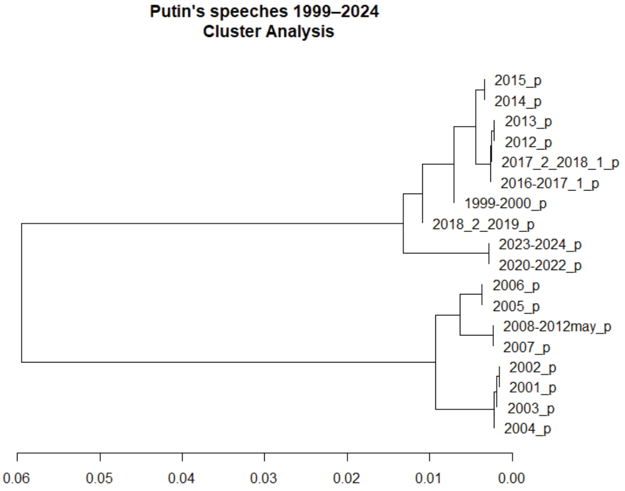


Fig. 1. Hierarchical cluster analysis of Vladimir Putin's speeches (metric: cosine; 100 most frequent words (MFW) in each text as input variables)

[1] Cosine distance measures directional similarity, highlighting relative patterns of word usage regardless of absolute frequency. Euclidean distance is sensitive to size, emphasising differences in absolute word counts. Min-max distance focuses on the range of word frequencies, making it more sensitive to outliers or rare word inflation.

[2] This is the plot with the cosine metric. All three dendrograms are available here: https://mega.nz/folder/sdtjWBbQ#vscNCLZc3kB-KT5S8jlCRw.

All three dendrograms show a coherent and converging structure, with a clear bifurcation into two primary stylistic branches, indicating a significant shift in Putin's discourse style or rhetorical strategy. Putin's political rhetoric can thus be clearly divided into pre-2012 and post-2012 stylistic periods. This chronological threshold corresponds closely with major political transitions and geopolitical events in Vladimir Putin's career. Specifically, this divide coincides with his return to the presidency in 2012, the subsequent annexation of Crimea in 2014, and the onset of increasingly centralised and confrontational governance. This shift appears to reflect a fundamental change, in line with broader political developments, including the consolidation of an increasingly autocratic regime. The high consistency of the identified clusters across different distance measures supports the reliability of the analysis.

Of interest is the position of the first speeches from 1999–2000, which appear consistently in segments of texts after 2012. In the case of the cosine and Euclidean metrics, it coincides with texts from 2012–2018, and for the min-max distance it is associated with speeches from the last years 2000–2024. Future analyses will therefore need to look more closely at possible causes. According to the dendrogram, texts from 1999–2000 can be seen as transitional or outlying observations. Such placement may reflect initial stylistic or rhetorical experimentation or variability before a consistent style is established.

4 QUANTITATIVE ANALYSIS OF KEYWORDS (TF*IDF METHOD)

The TF*IDF (Term Frequency-Inverse Document Frequency) method is a statistical measure widely used in text analysis to identify and rank significant keywords within textual data (Salton and Buckley 1988; Ramos 2003). The approach combines two aspects: term frequency (TF), which measures how often a keyword occurs in a given document or text segment, and inverse document frequency (IDF), which reflects how rarely a keyword occurs in all analysed documents or segments (Rajaraman and Ullman 2011). The resulting TF*IDF score thus assigns higher values to keywords that are characteristic of specific text periods or documents, making it particularly effective for longitudinal analyses of political discourse, such as the speeches of Vladimir Putin analysed in this study, because it reveals the main or most important expressions, motifs or themes.

In our application, keywords (hereafter KWs) were extracted and hierarchically ranked based on their TF*IDF scores across defined time slots (1999–2024), allowing for a nuanced identification of thematic shifts and continuities in political rhetoric over time.

We extracted the first 100 KW from each file using the KER tool (Libovický 2016). We then removed items that could bias the analysis, in particular proper names of persons that appeared in the interview transcript as labels preceding the dialogue line (i.e., primarily *Vladimir*, *Vladimirovich*, *Putin*, and the names of the journalists who conducted the interview). Other names of people who were part of the text (e.g. *Yanukovych*, *Bush*, *Obama*), as well as names of institutions, places, etc., remained in the lists. We also removed the repetitive formal words *dear*, *sir*, *question*, which were also related to the format of the interview (addressing the guest). Due to the large amount of data, all extracted KWs from all files are shared in the cloud[3] and we focus mainly on their interpretation.

4.1 Thematic clusters of Putin's keywords

Total KWs: 1,612

Unique KWs: 330

The average (mean) TF*IDF score across all keywords is about 0.01066, with values ranging from about 0.00586 to 0.06022.

Based on these 330 unique keywords, we have identified 8 main thematic clusters of Putin's rhetoric:

1. National identity and geopolitics
Sample KWs: *Russia, Russian, Chechnya, federation, state, citizen, national, sovereignty, territory, international, partner, relationship, region, Donbas, Crimea, Sevastopol, Dagestan, foreign*

2. Economic and financial matters
Sample KWs: *Economy, economic, financial, market, ruble, budget, growth, investment, bank, money, enterprise, business, pension, rate*

3. Security, military, and defense
Sample KWs: *Security, military, defense, weapon, terrorism, terrorist, counter-terrorism, nuclear, war, armed, threat*

4. Political institutions and governance
Sample KWs: *Government, president, Duma, law, constitution, legal, authority, organization, council, reform, process*

5. Social policies and domestic welfare
Sample KWs: *Society, social, education, housing, healthcare, population, community, future, reform* (in a social context)*, citizen* (also appears in this group)

[3] All the extracted keywords with TF*IDF values are available here: https://mega.nz/folder/sdtjWBbQ#vscNCLZc3kB-KT5S8jlCRw. For the purposes of this study, we also translated all KWs into English using the Gemini AI tool (https://gemini.google.com/). All lists are therefore bilingual.

6. International relations and conflict
Sample KWs: *NATO, sanctions, cooperation, partnership, dialogue, threat, crisis, conflict, Bush, Obama, Yanukovych, alliance, war, global*

7. Science, technology, and innovation
Sample KWs: *Technology, technological, AI, innovation, project, modern, internet*

8. Miscellaneous and context-specific terms
Sample KWs: names (e.g. *Bush, Obama, Yanukovych*), specific events or adjectives (e.g. *clear, certain, big, real*), technical descriptors and less common or specific terms such as *medium*.

For the purposes of our analyses, we have simplified and reduced these categories to the following political themes, which indicate the main trends and changes in Putin's rhetoric. The results are summarised in the following graph (Fig. 2), which shows the thematic focus over time:

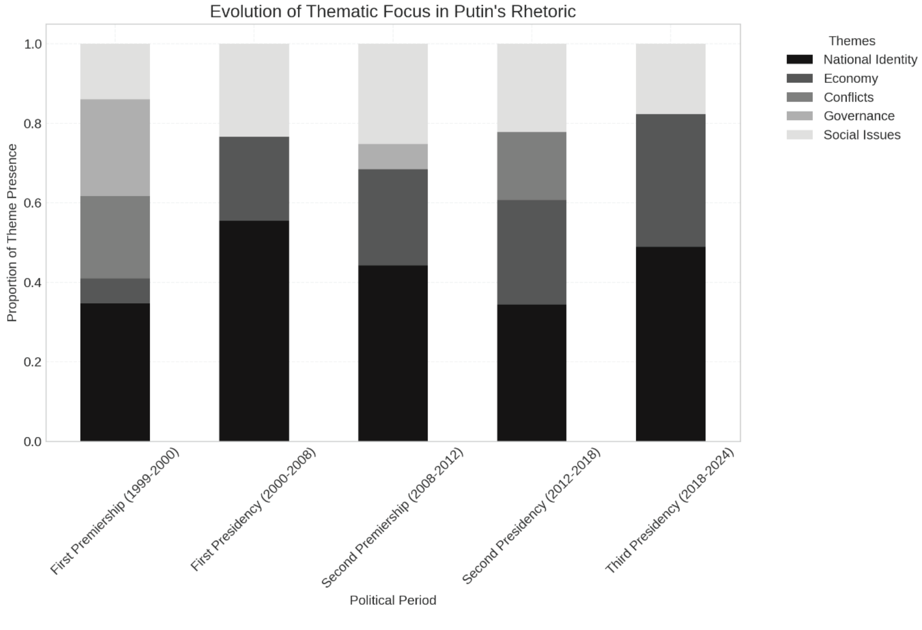


Fig. 2. Development of the thematic focus of Putin's rhetoric

National identity remains a consistently important topic (especially during the first presidential term), but its context changes – from internal identity building to Russia's international position. Conflict themes change according to the geopolitical situation: the Chechen conflict dominates the first term, while Ukraine and Crimea appear in the second term (2012–2018). Economic themes gradually become more important, especially in the later periods. Social issues (*human*, *problem*, *work*) are consistently present, but their importance increases especially in the second premiership. Governance: domestic political topics (*Duma*, *laws*) are particularly prominent in the first premiership, but later recede into the background; international themes are particularly prominent in the second premiership (*Ukraine*, *Crimea*).

4.2 Visual analysis of the top 15 KWs and keyword correlations

The most consistent keywords throughout Putin's career reveal his core rhetorical focus: '*Russia*', '*Russian*', '*country, land*', '*person, human, man*', '*problem*', '*development*', '*state*', '*economy*', '*important*', '*work*'. Russia and Russian identity remain the absolute foundation of Putin's rhetoric across all periods, followed by references to the country/state and the Russian people.

The heat-map in Fig. 3 shows how the importance of keywords has changed over the course of Putin's career. A darker colour indicates a higher ranking (lower number = more important). Blank spaces don't mean that Putin didn't use it in a particular period, but that the word was not among the top 15 keywords in that period.

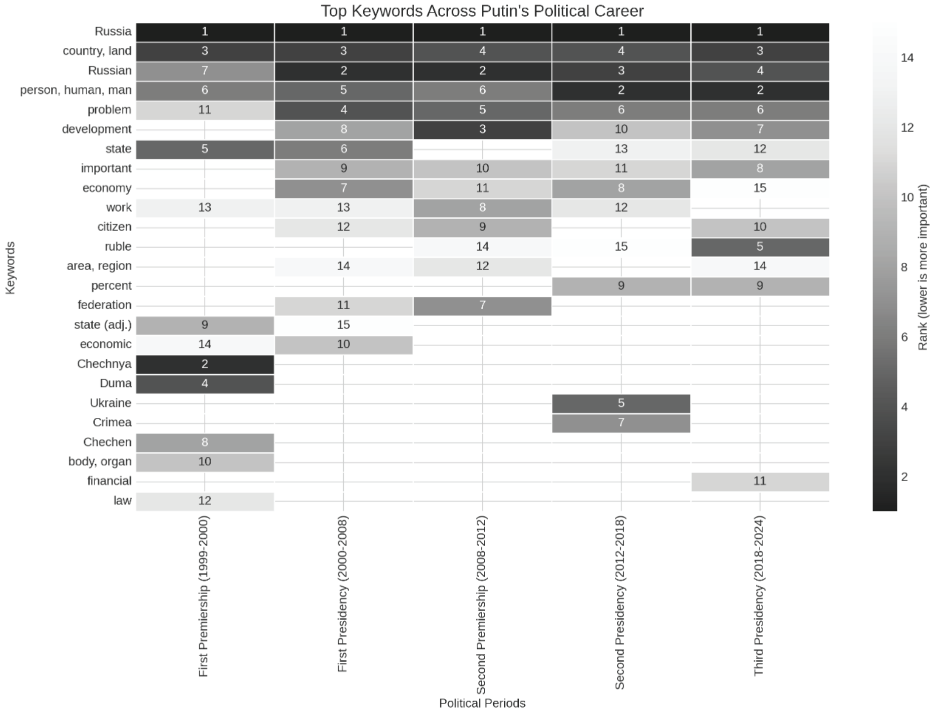


Fig. 3. Top 25 keywords from Putin's political career; blanks indicate absence from the period's top 15

The following heat-map in Fig. 4 presents a correlation map of the top 50 keywords of Vladimir Putin:

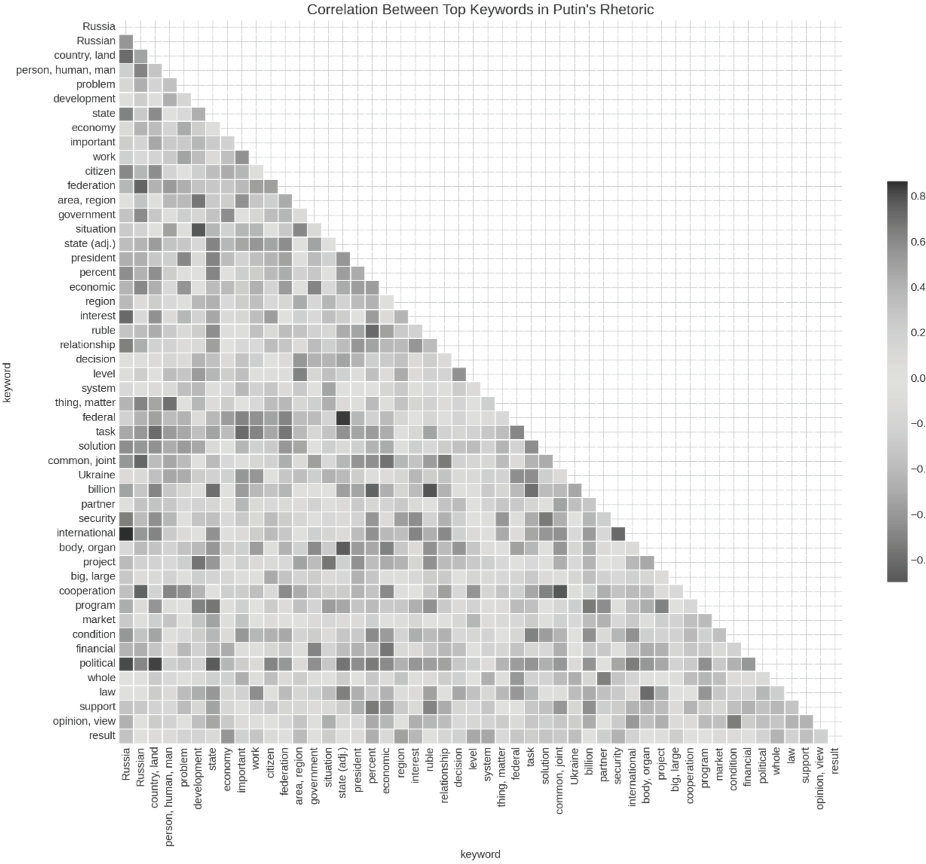


Fig. 4. Correlation between the top 50 keywords in Putin's rhetoric

This correlation heat-map reveals which keywords tend to appear together in Putin's rhetoric, helping to identify conceptual associations in his political discourse.

4.2.1 Key findings by political event

Chechen war period (1999–2000): During this early period, '*Chechnya*' was the second most prominent keyword after '*Russia*', reflecting the centrality of the Chechen conflict to Putin's early political identity. The rhetoric focused heavily on state security, terrorism, and the establishment of federal authority.

Putin's return to the presidency (2012): Keywords show a shift towards more domestic governance and economic development themes, with a continued emphasis on Russia's international position.

Annexation of Crimea (2014): After the annexation of Crimea, '*Ukraine*' emerges as a significant keyword in Putin's rhetoric.

Ukraine invasion period (2022–2024): In the final period, we see a consistent pattern of keywords with '*Russia*' remaining dominant, but with '*Ukraine*' maintaining a significant presence. The rhetoric continues to emphasise Russian identity, citizenship, and state interests.

This inspired us to investigate whether there is a correlation between keyword importance and historical events (see Fig. 5).

4.3 Trend of average keyword importance in Putin's speeches

The graph in Fig. 5 shows the temporal evolution of the average TF*IDF scores of keywords in Vladimir Putin's speeches over time.

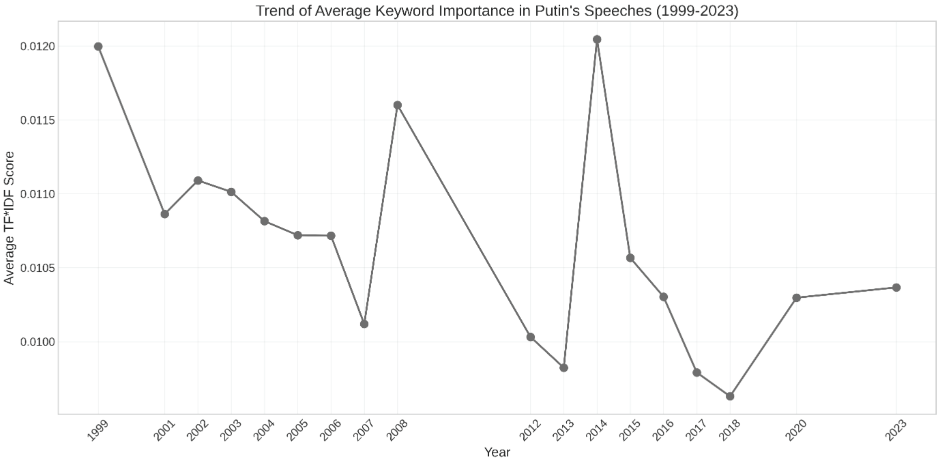


Fig. 5. Trend of average keyword importance in Putin's speeches from 1999 to 2024 according to TF*IDF score

4.3.1 Key observations

Initial peak (1999): The graph begins with a high average TF*IDF score (~0.0120) in 1999, reflecting a strong emphasis on certain keywords during Putin's early political career, likely related to his initial rise to power and the need to establish a clear and coherent rhetorical narrative when establishing nationalist themes was crucial.

Decline (2000–2006): From 2000 to 2006, there is a gradual decline in the average importance of keywords, with the score stabilising at 0.0105. This period corresponds to the first period of Putin's presidency; which may indicate a diversification of topics or a broadening of the discourse as the leadership stabilised.

Spike (2008): The sharp increase in 2008 marks a significant rhetorical shift or focus, and the average TF*IDF score peaks again. This may be related to the global financial crisis and the political transition, with Putin changing roles and emphasising a strong, unified discourse in a time of uncertainty.

Low point (2013): The lowest point in the graph occurs in 2013, with an average score below 0.0100. This period precedes the annexation of Crimea in 2014, suggesting a possible lull in focused rhetorical emphasis.

Major peak (2014): The dramatic increase in 2014 is associated with the annexation of Crimea and heightened geopolitical tensions. This suggests a concerted use of specific keywords to address the crisis and consolidate domestic and international support.

Decline and stabilization (2015–2024): After 2014, the average TF*IDF score gradually declines, suggesting a possible moderation in language intensity, reaching another low around 2018. However, there is a slight upward trend from 2020 onwards, which may reflect recent domestic or global challenges that prompt a recalibration of rhetoric, such as the COVID-19 pandemic or the invasion of Ukraine.

5 CONCLUSION

The quantitative corpus analysis of Putin's speeches over the past 25 years reveals significant stylistic and thematic shifts in his political rhetoric. Two main periods – before and after 2012 – show a clear change in rhetorical strategy and thematic focus associated with major political events, and coinciding with Putin's return to the presidency and the intensification of Russia's authoritarian governance and geopolitical assertiveness. Thematic clusters identified on the basis of 330 unique keywords confirm a continued emphasis on national identity, but shifting from internal national unity-building to internationally oriented geopolitical issues. Economic and social issues gradually gain in importance, while security and conflict issues dominate during periods marked of military crises. The temporal evolution of keyword importance, as measured by TF*IDF scores, shows notable peaks corresponding to critical moments in Russian politics, including an initial peak in 1999, a sharp increase in 2008, a dramatic increase in 2014 related to the annexation of Crimea, and subtle changes in the period 2020–2024. These findings provide a comprehensive and systematic understanding of the evolution of Putin's language, showing how his discourse adapts to changing political circumstances and strategic goals.

ACKNOWLEDGEMENTS

The research was supported by the Ministry of Education of the Czech Republic IGA_FF_2025_042 "The Czech Language and Its Worlds: Interdisciplinary Approaches to Textual Study in the 21st Century".

References

Al-Manaseer, A. (2025). Deconstructing the Ideological Frame of President Vladimir Putin's Rhetoric: A Socio-Cognitive Analysis. Wasit Journal for Human Sciences, 21(1), pp. 984–999.

Burrows, J. (2002). Delta: A Measure of Stylistic Difference and a Guide to Likely Authorship. Literary and Linguistic Computing, 17(3), pp. 267–287.

Chiluwa, I., and Ruzaite, J. (2024). Analysing the language of political conflict: a study of war rhetoric of Vladimir Putin and Volodymyr Zelensky. Critical Discourse Studies, pp. 1–17.

Eder, M., Rybicki, J., and Kestemont, M. (2016). Stylometry with R: a package for computational text analysis. R Journal, 8(1), pp. 107–121.

Fafiyebi, D. O., and Fafiyebi, O. F. (2025). A Speech Act Analysis of the Utterances of Selected Key Actors in the Russian/Ukrainian Crisis: Pragmatics. International Journal of Language and Literary Studies, 7(1), pp. 336–352.

Hidalgo-Cobo, P., López-Marcos, C., and Puebla-Martínez, B. (2024). Discourse analysis from an international relations perspective: the case study of Tucker Carlson's televised interview with Vladimir Putin. aDResearch ESIC International Journal of Communication Research, 32 (November, 2024), e285.

Hoover, D. L. (2003). Multivariate Analysis and the Study of Style Variation. Literary and Linguistic Computing, 18(4), pp. 341–360.

Janda, L., Fidler, M., Cvrček, V., and Obukhova, A. (2022). The case for case in Putin's speeches. Russian Linguistics, 47, pp. 15–40.

Kadim, E. (2023). A Critical Discourse Analysis of Vladimir Putin's Speech Announcing 'Special Military Operation' in Ukraine. International Journal of Humanities and Educational Research, 5, pp. 424–444.

Kopik, M. (2023). Comparative analysis of American and Russian political discourse: A discourse analysis study. Linguistics Beyond and Within, 9, pp. 49–59.

Leskovec, J., Rajaraman, A., and Ullman, J. D. (2014). Mining of massive datasets. Accessible at: http://www.mmds.org/#ver30.

Libovický, J. (2016). KER – Keyword extractor. [software] Accessible at: http://lindat.mff.cuni.cz/services/ker/.

Oleinik, A. (2023): A comparison of two text specificity measures analyzing a heterogenous text corpus. Glottometrics, 54, pp. 1–12.

Rajaraman, A., and Ullman, J. D. (2011). Data mining. In: J. Leskovec – A. Rajaraman – J. D. Ullman (eds.): Mining of Massive Datasets, pp. 1–19. Cambridge.

Ramos, J. (2003). Using TF-IDF to Determine Word Relevance in Document Queries. In Proceedings of the First Instructional Conference on Machine Learning, pp. 133–142. Piscataway, NJ: Rutgers University.

Salton, G., and Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. Information Processing & Management, 24(5), pp. 513–523.

Scott, M., and Tribble, Ch. (2006). Textual patterns. Key words and corpus analysis in language education. Amsterdam: Benjamins.

Shahbaz, J., and Nawab, H. (2024). Language, Politics, and Power: Unveiling Putin's Annexation Narrative through Fairclough's Model, 7(2), pp. 24–33.

stylo: Stylometric Multivariate Analyses (version 0.7.4). [software]. Accessible at: https://cran.r-project.org/web/packages/stylo/index.html.

The R Project for Statistical Computing: R (version 4.2.0) [software]. Accessible at: https://www.r-project.org/.

Tutar, H., and Bağ, S. M. (2023). Critical discourse analysis on leader statements in the Russia-Ukraine War. Etkileşim, 11, pp. 44–66.

Wang, Y., and Zeng, T. (2023): Fellow or foe? A quantitative thematic exploration into Putin's and Trump's stylometric features. Glottometrics, 54, pp. 39–57.

sciendo
DOI 10.2478/jazcas-2025-0034

CORPUS OF DRAMAS BY THE ČAPEK BROTHERS FROM A QUANTITATIVE PERSPECTIVE

PETR POŘÍZKA[1] – JÁN MAČUTEK[2]
[1]Department of Czech Studies, Faculty of Arts, Palacký University, Olomouc, Czech Republic (ORCID: 0000-0001-6980-9148)
[2]Mathematical Institute, Slovak Academy of Sciences, Bratislava, Slovakia & Department of Mathematics, Faculty of Natural Sciences and Informatics, Constantine the Philosopher University, Nitra, Slovakia (ORCID: 0000-0003-1712-4395)

POŘÍZKA, Petr – MAČUTEK, Ján: Corpus of Dramas by the Čapek Brothers From a Quantitative Perspective. Journal of Linguistics, 2025, Vol. 76, No 1, pp. 378 – 389.

Abstract: This text presents the first quantitative analysis of the plays of the Čapek brothers, exploring the linguistic and stylistic differences between their individual and collaborative works. Utilizing computational methods and quantitative approaches, it analyses a corpus of ten plays, focusing on the distribution and proportion of parts of speech in both dialogue and stage directions. The analysis reveals significant stylistic differences: Josef Čapek is characterized by a descriptive language rich in nouns with fewer words overall, while Karel Čapek uses a more dynamic approach with a predominance of verbs. Cluster analysis shows that Josef's dramas form a separate, distinct group when both dialogue and stage directions are considered, with stage directions showing particularly marked differences. Morphological coefficients, including the noun coefficient (Kn) and Busemann coefficient (B), quantitatively confirm these stylistic differences, with Josef's plays showing extreme values that indicate high descriptive saturation, especially in the stage directions. This structural analysis not only provides quantitative evidence of different authorial styles, but also lays a foundation for future research.

Keywords: computational literary studies, drama, CapekDraCor, Karel Čapek, Josef Čapek, quantitative analysis, parts of speech

1 INTRODUCTION AND STATE OF THE ART

Karel Čapek and his brother Josef Čapek are among the key figures of 20th century Czech literature. Both brothers contributed significantly to the development of drama, both in their individual works and in those they wrote together. The dramatic work of the Čapek brothers comprises ten plays (see below for details), which have been extensively discussed in the academic literature and analysed from various perspectives, including theatrical, literary, and linguistic. However, previous scholarly studies have focused on such aspects of their plays as the poetics, compositional or narrative aspects of their work (Sunbee 2011; Novák 2013; Doležel 2014), textual adaptations (e.g. Janáček's libretto in the opera *The Makropulos*

Affair, cf. Křupková 2008), the relationship between the literary version and the film adaptation (e.g. *The White Disease*), translatological and theatrological aspects, or reflections on the dramatic work in a broader social and cultural context – the ethical or philosophical aspects of the work and its influence on the literary works of other authors (Janiec-Nyitrai 2012; Drozenová 2020). In general, these are analyses and interpretations based on introspection, readerly reception and a traditional structuralist literary scholarly approach (Holý 1984, 2014; Janoušek 1989, 2018).

Rarely, there are also investigations using computer-assisted text analysis or partial quantitative analyses. However, all of them concern only selected plays by Karel Čapek, e.g. using a semi-automatic phrase recognition tool in *R.U.R.* and *The White Disease* (Kovářiková and Kopřivová 2012), quantitative analysis of proper nouns in the plays (Pořízka 2023b) or, more often, deal with other genres of his work. The works and contribution of his brother Josef have been largely overlooked, especially in terms of quantitative analysis.

The Karel Čapek Dictionary (*Slovník Karla Čapka*, Čermák 2007b), published by the Institute of the Czech National Corpus (CNC) in 2007 and based on the *capek* corpus (see below), includes the chapter *Statistical Aspects of Karel Čapek's Language, Especially His Lexicon* (Cvrček et al. 2007). It presents statistical data on Čapek's lexicon, parts-of-speech ratio (POS), and lexical richness. However, the authors themselves admit inaccuracies in the quantitative indices used, which are distorted by the length of the text (Cvrček et al. 2007, p. 675).

Previous quantitative analyses of Karel Čapek's works have focused on exploring thematic text concentration, lexical compactness across genres, and the use of selected lexical-statistical indices, such as average token length, verb distance, and vocabulary richness (Davidová et al. 2013; Čech 2015; Kubát 2016; Mačutek et al. 2016). However, these papers consistently exclude the plays of the Čapek brothers, leaving the multilayered textual structure of the plays largely unexplored.

2 CORPUS AND DATA PROCESSING

In terms of the textual structure, plays are multi-layered. This structure includes primarily the character dialogues in a form similar to spoken dialogue, with character labels (proper nouns) preceding each line of text, then structuring words (act, scene, drop-scene), comments (stage, authorial, on the characters' actions), and possibly other sections such as the author's introductory metatextual notes (foreword) and a list of characters (cast list).

There is currently a *capek* corpus (Čermák et al. 2007a) in the Czech National Corpus (CNC), which includes plays by Karel Čapek, but no those by his brother Josef. In addition, this corpus has limitations due to the way the source texts are processed, making it unsuitable for quantitative analyses. The *capek* corpus does not reflect the multi-layered structure of the plays, as the aforementioned textual and metatextual parts are not separated. This lack of segmentation makes it impossible to analyse the subparts of the plays and can significantly affect subsequent quantitative analyses and their results, as we have shown in a previous study (Pořízka 2023b).

For these reasons, and for the purpose of different types of quantitative analyses, we have recently created a new corpus in two versions, which contains all the plays by the Čapek brothers and reflects the annotation of different (meta)text layers. The first version is made as a standard corpus including also linguistic annotation (lemmatization, morphological tagging) and is available in the *SketchEngine* tool (cf. *Czech Drama Corpus* in *DraCor Drama Corpora*: https://www.sketchengine.eu/dracor-drama-corpora/).

The second version of this database called *CapekDraCor* (soon to be publicly available) which we used for this quantitative analysis focusing on the comparison of character dialogues and metatextual comments, was created specifically for the international *DraCor* project (https://dracor.org/) and its tools.

The *CapekDraCor* corpus used in the analysis consists of the following texts:

- plays by Karel Čapek: *Loupežník* ('The Outlaw', 1920); *R.U.R.* (1920); *Věc Makropulos* ('The Makropulos Affair', 1922); *Bílá nemoc* ('The White Disease', 1937); *Matka* ('The Mother', 1938);
- plays by Josef Čapek: *Země mnoha jmen* ('The Land of Many Names', 1923);
- plays written together by the Čapek brothers: *Lásky hra osudná* ('The Fateful Game of Love', 1910); *Ze života hmyzu* ('The Insect Play', 1921); *Adam Stvořitel* ('Adam the Creator', 1927).

The data are processed in a standardized format based on XML and general TEI guidelines for processing drama, with a defined basic drama tagset. A more detailed description of the text processing, information about the TEI-XML format and other technical aspects including illustrative examples can be found in (Pořízka 2023a).

3 QUANTITATIVE ANALYSIS AND INTERPRETATION

3.1 Basic word ratios

The two brothers are known for their different approaches to the language of their plays: while Karel Čapek used contemporary language and a more colloquial style, Josef Čapek used a bookish, even archaic style. Because of these differences, we have divided their dramas into three groups: (1) dramas by Karel Čapek, (2) dramas by Josef Čapek, and (3) dramas written by the two brothers together in order to compare these collections in terms of the composition and structure of the plays (authorial style). Using the TEI/XML data format, each play was also divided into two subgroups: (1) character dialogues (the drama itself) and (1) stage directions.

For the purposes of this quantitative analysis, additional metatextual sections were excluded, i.e. structuring words (act, scene, drop-scene), cast list, preface, and character labels (proper names) preceding individual lines of dialogue.

drama	author(s)	word proportions in the stage directions
Bílá nemoc	Karel	0.094
Loupežník	Karel	0.107
Matka	Karel	0.098
R.U.R.	Karel	0.120
Věc Makropulos	Karel	0.089
Adam stvořitel	co-authored	0.098
Lásky hra osudná	co-authored	0.093
Ze života hmyzu	co-authored	0.112
Gassirova loutna	Josef	0.046
Země mnoha jmen	Josef	0.061

Tab. 1. Word ratios in the stage directions of the Čapek brothers' plays

The basic word ratios or proportions (see Tab. 1) in the stage directions divide the ten plays into two groups. Two of Josef Čapek's plays (*Gassirova loutna*, *Země mnoha jmen*) clearly have the lowest proportions, while the three co-authored dramas do not differ from Karel Čapek's dramas in this respect.

3.2 Parts of speech proportions

The texts were linguistically annotated (lemmatization and morphological tagging via the *MorphoDiTa* tool (Straková et al. 2014)), using the new *LexaMorf* tool (Pořízka 2025), and the frequency distributions of word classes (parts of speech, hereafter POS) were calculated for both the stage directions and the characters' dialogues of each play. The individual POS categories correspond to the standard classification in Czech, e.g. according to the so-called *Academic Grammar of Czech* (*Mluvnice češtiny 2*). The results are shown in the following tables. Since the plays differ in length (as measured by the number of tokens), we relativize the proportions. Absolute frequencies and percentages of parts of speech can be found in the Tab. 2 – Tab. 5:

POS	**Loupežník**		**R.U.R.**		**Věc Makropulos**		**Bílá nemoc**		**Matka**	
	rfp	**f**	**rfp**	**f**	**rfp**	**f**	**rfp**	**f**	**rfp**	**f**
adjectives	4.21	543	6.31	927	5.45	745	6.67	935	4.78	729
adverbs	9.79	1262	8.86	1302	9.32	1274	8.83	1238	9.92	1505

conjunctions	7.29	940	6.91	1015	7.47	1021	7.76	1088	7.33	1112
interjections	1.91	246	1.59	233	1.16	159	0.39	55	0.54	82
nouns	17.29	2228	23.10	3393	20.99	2870	22.15	3105	18.34	2784
numerals	0.52	67	1.37	201	1.78	243	1.30	182	0.99	150
particles	2.72	351	2.38	350	2.81	384	3.60	505	2.51	381
prepositions	5.17	666	5.49	806	5.38	736	5.86	822	5.63	854
pronouns	23.30	3003	18.36	2697	20.86	2852	19.68	2760	23.56	3576
verbs	27.78	3580	25.64	3766	24.79	3389	23.76	3331	26.40	4006

Tab. 2. Relative frequencies in percentage (rfp) and frequency (f) of parts of speech in the dialogues of Karel Čapek's plays

POS	**Lásky hra osudná**		**Ze života hmyzu**		**Adam stvořitel**		**Země mnoha jmen**		**Gassirova loutna**	
	rfp	**f**	**rfp**	**f**	**rfp**	**f**	**rfp**	**f**	**rfp**	**f**
adjectives	6.67	453	7.08	777	5.93	980	8.23	916	6.69	280
adverbs	8.93	606	8.66	950	9.00	1487	9.88	1100	8.53	357
conjunctions	9.31	632	6.02	660	8.16	1347	8.37	932	8.96	375
interjections	0.85	58	3.12	342	1.21	200	1.25	139	0.93	39
nouns	22.55	1531	22.51	2470	17.41	2876	22.87	2545	24.74	1036
numerals	0.71	48	2.17	238	0.84	138	0.78	87	1.03	43
particles	2.28	155	2.77	304	2.71	448	2.16	240	2.34	98
prepositions	5.85	397	4.86	533	4.44	734	5.91	658	6.11	256
pronouns	18.59	1262	19.49	2138	23.04	3805	18.57	2067	15.64	655
verbs	24.26	1647	23.33	2559	27.25	4500	21.98	2446	25.03	1048

Tab. 3. Relative frequencies in percentage (rfp) and frequency (f) of parts of speech in the dialogues of plays by the Čapek brothers and Josef Čapek

POS	**Loupežník**		**R.U.R.**		**Věc Makropulos**		**Bílá nemoc**		**Matka**	
	rfp	**f**	**rfp**	**f**	**rfp**	**f**	**rfp**	**f**	**rfp**	**f**
adjectives	2.58	41	6.68	124	5.01	69	5.61	85	7.18	124
adverbs	5.16	82	5.77	107	6.03	83	4.36	66	4.87	84
conjunctions	5.03	80	3.77	70	3.78	52	3.63	55	6.2	107
interjections	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0
nouns	34.03	541	33.19	616	32.17	443	36.53	553	33.14	572
numerals	0.57	9	0.92	17	0.36	5	1.65	25	1.39	24
particles	0.31	5	0	0	0.15	2	1.19	18	0.17	3

prepositions	16.29	259	13.58	252	14.16	195	13.21	200	14.31	247
pronouns	8.74	139	9.21	171	9.08	125	8.78	133	9.97	172
verbs	27.30	434	26.89	499	29.27	403	25.03	379	22.77	393

Tab. 4. Relative frequencies in percentage (rfp) and frequency (f) of parts of speech in the stage directions of Karel Čapek's plays

POS	**Lásky hra osudná**		**Ze života hmyzu**		**Adam stvořitel**		**Země mnoha jmen**		**Gassirova loutna**	
	rfp	**f**	**rfp**	**f**	**rfp**	**f**	**rfp**	**f**	**rfp**	**f**
adjectives	4.33	28	6.23	90	7.52	141	9.49	69	8.00	16
adverbs	9.43	61	4.22	61	5.55	104	5.91	43	2.50	5
conjunctions	5.87	38	4.15	60	4.54	85	3.58	26	5.00	10
interjections	0.00	0	0.14	2	0.00	0	0.00	0	0.00	0
nouns	34.47	223	26.92	389	31.27	586	43.74	318	45.00	90
numerals	1.08	7	1.59	23	1.81	34	2.2	16	5.50	11
particles	0.15	1	0.07	1	0.16	3	0	0	0	0
prepositions	10.66	69	13.84	200	12.33	231	9.9	72	10.50	21
pronouns	10.36	67	11.56	167	11.63	218	8.25	60	3.50	7
verbs	23.65	153	31.28	452	25.19	472	16.92	123	20.00	40

Tab. 5. Relative frequencies in percentage (rfp) and frequency (f) of parts of speech in the stage directions of plays by the Čapek brothers and Josef Čapek

Even in the case of the word classes, focusing on the proportions of the parts of speech, we find the same pattern, i.e. works divided into two groups: (1) on the one hand, two plays by Josef, (2) on the other hand, other dramas (plays by brother Karel and co-authored plays). Note: There were four words for which the software did not determine the POS value; these words were not considered.

3.3 Cluster analysis

We then performed a cluster analysis to look more closely at the relationships between the POS frequency distributions. Each drama is represented by a 20-dimensional vector whose coordinates represent the percentage of parts of speech in the dialogues (the first ten coordinates) and in the stage directions (the next ten coordinates). The results of the hierarchical cluster analysis (run in R with default settings) are shown in Fig. 1.

This clear-cut and unambiguous result follows from the stage directions. If we consider only the proportions of parts of speech in the stage directions (i.e. each drama is represented by a 10-dimensional vector), we obtain the same clusters as in

Fig. 1. However, if we consider only the dialogue of the characters, we do not find any significant differences or meaningful distinction between the dramas.

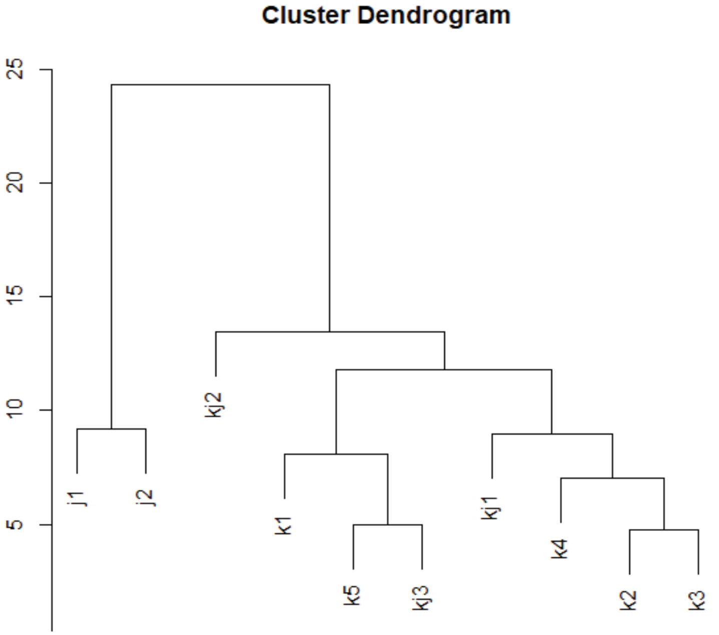


Fig. 1. Hierarchical cluster analysis of POS frequency distribution in the plays of the Čapek brothers

3.4 Morphological coefficients

Differences in the typology of the drama vocabulary can also be characterised using relatively simple statistical indicators introduced in the 1970s by the prominent Czech quantitative linguist Marie Těšitelová. She worked with nominal, verbal and neutral word-groups (Těšitelová 1974, p. 85nn). In particular, she measured the mutual proportionality of the so-called dominant components of the nominal and verbal groups and showed in her analyses that they can be used for individual characteristics of lexical styles, stylistic genres, etc. (Těšitelová 1974, p. 179). She introduced four basic indicators of morphological statistics (Těšitelová 1987, p. 89nn), which we will also use to interpret our data:

- Nominality coefficient: Kn = N / V (ratio of nouns to verbs)
- Coefficient of noun development: Krn = A / N (ratio of adjectives to nouns)
- Coefficient of verb development: Krv = D / V (ratio of adverbs to verbs)
- Busemann coefficient: B = A / V (ratio of adjectives to verbs).

Busemann's coefficient is still actively used in quantitative linguistics as an index expressing the activity vs. descriptiveness of a text (Čech et al. 2014, p. 52nn).

We calculated these coefficients for the two structural parts of all the plays under study, i.e. for the characters' dialogues and the stage directions. All the obtained data are summarized in the following heat-maps (the numbers represent the result of the given coefficients) – see Fig. 2–3.

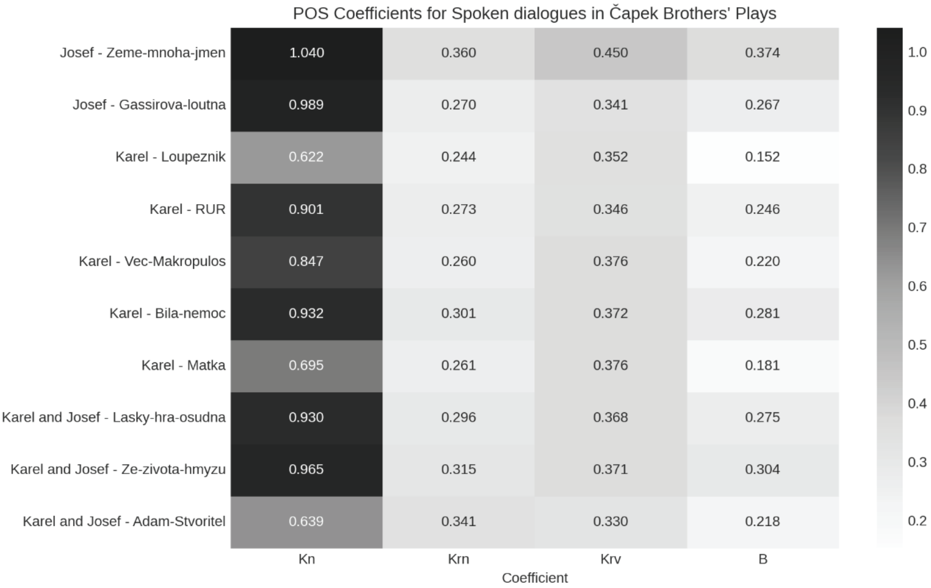


Fig. 2. Heat-map of morphological coefficients of dominant POS for the spoken dialogues in Čapek Brothers' plays

The difference between spoken dialogue and stage directions

Key findings in spoken dialogue:

- Karel Čapek's plays and collaborative works generally have consistently lower Kn values, indicating more dynamic, verb-rich and action-oriented dialogue, with relatively more verbs.
- Josef Čapek's plays have higher Kn and B values, indicating more descriptive, noun-rich dialogue, suggesting more descriptive language relative to action.
- Krn and Krv values are generally higher in spoken dialogue than in stage directions, indicating richer descriptive language of characters, more adjectives to nouns (Krn); characters qualify verbal actions more with adverbs (Krv).

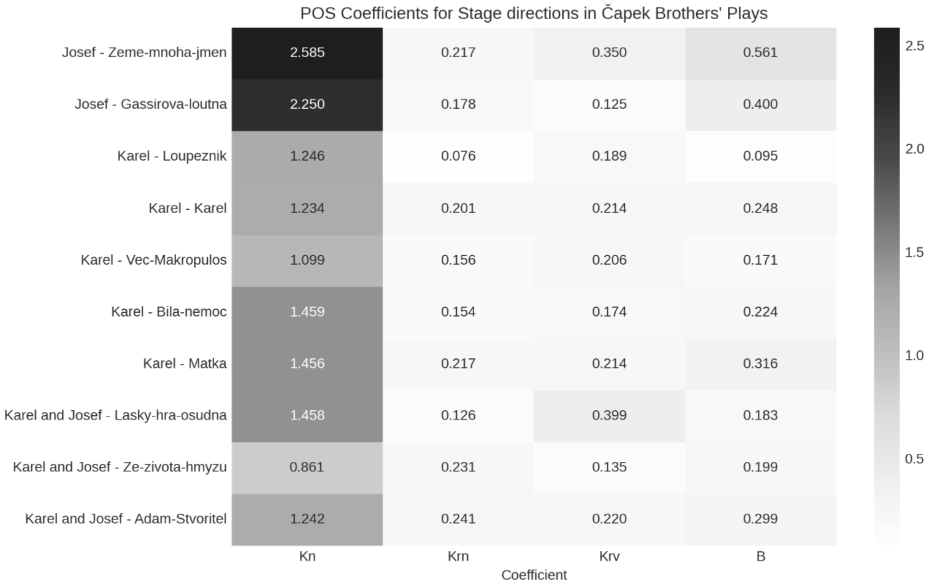


Fig. 3. Heat-map of morphological coefficients of dominant POS for the stage directions in Čapek Brothers' plays

These heat-maps clearly visualize the stylistic differences between the brothers' individual works and their collaborations, as well as the differences between writing styles of dialogue and stage directions.

Key findings for stage directions:

- All stage directions have significantly higher Kn values than spoken dialogue, which is to be expected as they are more descriptive and much richer in nouns relative to verbs.
- Josef Čapek's plays have the highest Kn values in stage directions (are particularly rich in nouns) and have particularly high B values (higher ratio of adjectives to verbs), which confirms his more descriptive style, indicating an emphasis on descriptiveness and detail.

Overall stylistic differences

Karel Čapek's style:

- More action-oriented with more use of verbs than nouns.
- More direct with fewer modifiers (adjectives and adverbs).
- Maintains this style in both spoken dialogue and stage directions, though less pronounced in the latter.

Josef Čapek's style:

- More descriptive with greater use of nouns and adjectives.
- More elaborate with more modifiers.
- Very descriptive, especially in stage directions.

Collaborative works:

- Often and generally intermediate values.
- Stage directions in collaborative works are particularly noun-rich.
- Suggest a blending (or fusion) of the brothers' individual styles.

4 CONCLUSION

Karel and Josef Čapek had completely different styles of writing stage directions. Josef used considerably fewer words. He used more nouns (and adjectives) and fewer verbs than Karel. The stage directions in the co-authored plays are written in the same style (in terms of relative numbers of words and proportions of parts of speech) as those in which Karel is the sole author. The proportions of parts of speech in the spoken dialogue do not indicate one or the other of the two brothers. However, the difference in the stage directions is so clear and strongly pronounced that Josef's dramas form a separate cluster when both the stage directions and the dialogues are taken into account.

Morphological coefficients confirm these differences, and this analysis reveals clear stylistic differences between the Čapek brothers, with Josef preferring a more descriptive language and Karel favouring a more dynamic, action-oriented approach. Their collaborative works often combine these styles. The analysis also shows that there are noticeable differences between the spoken dialogue and stage directions.

In general, it can be said that this comparison shows that Josef Čapek's plays in particular are different from the others – cf. the extreme values of the coefficients Kn and B, which express the high saturation of descriptiveness (especially in the stage directions).

It should be noted that this is the first quantitative analysis focusing on structure that shows some indications of a different authorial style. In the future, we would like to explore the structure of the plays in more detail, by act or by scene, and to look at other aspects and phenomena of the plays of these authors, such as keyword analysis, methods of determining authorship, and characterization of the network of literary characters within each play.

ACKNOWLEDGEMENTS

Supported by the European Commission's Horizon 2020 program (Grant Agreement No. 101004984) through the CLS INFRA project (P. Pořízka) and by VEGA 2/0120/24 (J. Mačutek).

References

Cvrček, V., Čermák, F., and Křen, M. (2007). Statistické aspekty jazyka Karla Čapka, zvláště jeho lexikonu. In: F. Čermák (ed.): Slovník Karla Čapka. Praha: NLN, pp. 671–690.

Čech, R., Popescu, I.-I., and Altmann, G. (2014). Metody kvantitativní analýzy (nejen) básnických textů. Olomouc: Univerzita Palackého.

Čech, R. (2015). The development of thematic concentration of text in Karel Čapek's travel books. Czech and Slovak Linguistic Review, 2015(1), pp. 8–21.

Čermák, F., et al. (2007a). Capek: korpus pouze vlastních textů Karla Čapka. Praha: ÚČNK. Accessible at: http://www.korpus.cz.

Čermák, F. (ed.). (2007b). Slovník Karla Čapka. Praha: NLN.

Davidová Glogarová, J., Čech, R., and David, J. (2013). Tematické charakteristiky textu – kvantitativní analýza publicistiky Jaroslava Durycha, Ladislava Jehličky a Karla Čapka. In: J. David – R. Čech – J. Davidová Glogarová – L. Radková – H. Šústková (eds.): Slovo a text v historickém kontextu – perspektivy historickosémantické analýzy jazyka. Brno: Host, pp. 62–84.

Doležel, L. (2014). Dva moderní vypravěči: Karel Čapek a Vladislav Vančura. In L. Doležel: Narativní způsoby v české literatuře. Příbram: Pistorius & Olšanská, pp. 97–114.

Drozenová, W. (2020). Technika, autonomie a etika: ke stému výročí Čapkova dramatu R.U.R. In 100 let R. U. R. Sborník z konference na Pedagogické fakultě Masarykovy univerzity v Brně (September 11, 2019). Brno: Masarykova univerzita, pp. 9–16.

Holý, J. (1984). Funkce jmen postav v dílech Karla Čapka a Vladislava Vančury. Česká literatura, 32(5), pp. 459–476.

Holý, J. (2014). Komentář k edici Tři hry Karla Čapka. In: K. Čapek – J. Holý (eds.): Tři hry. Brno: Host, pp. 253–279.

Janiec-Nyitrai, A. (2012). Zrcadlení: Literárněvědné sondy do tvorby Karla Čapka. Nitra: Univerzita Konštantína Filozofa v Nitre, Fakulta stredoeurópskych štúdií.

Janoušek, P. (1989). Mezi problémovým dramatem a groteskou (Druhá etapa dramatické tvorby Karla Čapka). Česká literatura, 37(3), pp. 193–205.

Janoušek, P. (2018). Bratří Čapkové: Lásky hra osudná. Theatralia, 21(1), Supplementum, pp. 33–38.

Kováříková, D., and Kopřivová, M. (2012). Authorial phraseology: Karel Čapek and Bohumil Hrabal. In: V. Jesenšek – D. Dobrovoľskij (eds.): Phraseology and culture. Maribor: Europhras, pp. 522–532.

Křupková, L. (2008). Proměny Čapkova dramatického textu v Janáčkově opeře Věc Makropulos. In: J. Nowak – K. Steimetz (eds.): Leoš Janáček světový a regionální. Ostrava: Ostravská univerzita, pp. 25–42.

Kubát, M. (2016). Kvantitativní analýza žánrů. Ostrava: Ostravská univerzita.

Kořenský, J., et al. (1986). Mluvnice češtiny 2. Tvarosloví. Praha: Academia.

Mačutek, J., Koščová, M., and Čech, R. (2016). Lexical compactness across genres in works by Karel Čapek. In: Mayaffre, D., Poudat, C., Vanni, L., Magri, V., and Follette, P. (eds.): Statistical Analysis of Textual Data. Nice: University Nice Sophia Antipolis, pp. 825–832.

MorphoDiTa: Morphological Dictionary and Tagger (version 1.3) [Software]. Accessible at: http://lindat.mff.cuni.cz/services/morphodita/.

Novák, R. (2013). Nástin poetiky Karla Čapka. Ostrava: Ostravská univerzita.

The R Project for Statistical Computing: R (version 4.2.0) [Software]. Accessible at: https://www.r-project.org/.

Pořízka, P. (2023a). CapekDraCor: A New Contribution to the European Programable Drama Corpora. Jazykovedný časopis, 74(1), pp. 244–253.

Pořízka P. (2023b). The Function of Proper Nouns in Quantitative Analysis of Dramas: A Case Study of Karel Čapek's Plays. In Onomastics in Interaction With Other Branches of Science. Vol. 3: General and Applied Onomastics. Literary Onomastics. Chrematonomastics. Reports, pp. 351–379.

Pořízka, P. (2025). LexaMorf Tool [Software]. Olomouc.

SketchEngine [Software]. Accessible at: http://www.sketchengine.eu.

Straková, J., Straka, M., and Hajič, J. (2014). Open-source tools for morphology, lemmatization, POS tagging and named entity recognition. In Proceedings of the 52nd annual meeting of the Association for Computational Linguistics: System demonstrations. Baltimore (MD): Association for Computational Linguistics, pp. 13–18.

Sunbee, Y. (2011). Znaky, styl a dramata Karla Čapka [Dissertation]. Praha: Univerzita Karlova.

Těšitelová, M. (1974). Otázky lexikální statistiky. Praha: Academia.

POKYNY PRE AUTOROV

Redakcia JAZYKOVEDNÉHO ČASOPISU uverejňuje príspevky **bez poplatku** za publikovanie.
Akceptované jazyky: všetky slovanské jazyky, angličtina, nemčina. Súčasťou vedeckej štúdie a odborného príspevku je abstrakt v angličtine (100 – 200 slov) a zoznam kľúčových slov v angličtine (3 – 8 slov).
Súčasťou vedeckej štúdie a odborného príspevku v inom ako slovenskom alebo českom jazyku je zhrnutie v slovenčine (400 – 600 slov) – preklad do slovenčiny zabezpečí redakcia.
Posudzovanie príspevkov: vedecké príspevky sú posudzované anonymne dvoma posudzovateľmi, ostatné príspevky jedným posudzovateľom. Autori dostávajú znenie posudkov bez mena posudzovateľa.
Technické a formálne zásady:

- Príspevky musia byť v elektronickej podobe (textový editor Microsoft Word, font Times New Roman, veľkosť písma 12 a riadkovanie 1,5). V prípade, že sa v texte vyskytujú zvláštne znaky, tabuľky, grafy a pod., je potrebné odovzdať príspevok aj vo verzii pdf alebo vytlačený.
- Pri mene a priezvisku autora je potrebné uviesť pracovisko.
- Text príspevku má byť zarovnaný len z ľavej strany, slová na konci riadku sa nerozdeľujú, tvrdý koniec riadku sa používa len na konci odseku.
- Odseky sa začínajú zarážkou.
- Kurzíva sa spravidla používa pri názvoch prác a pri uvádzaní príkladov.
- Polotučné písmo sa spravidla používa pri podnadpisoch a kľúčových pojmoch.
- Na literatúru sa v texte odkazuje priezviskom autora, rokom vydania a číslom strany (Horecký 1956, s. 95).
- Zoznam použitej literatúry sa uvádza na konci príspevku (nie v poznámkovom aparáte) v abecednom poradí. Ak obsahuje viac položiek jedného autora, tie sa radia chronologicky.

Bibliografické odkazy:

- knižná publikácia: ONDREJOVIČ, Slavomír (2008): *Jazyk, veda o jazyku, societa*. Bratislava: Veda, vydavateľstvo SAV. 204 s.
- slovník: JAROŠOVÁ, Alexandra – BUZÁSSYOVÁ, Klára (eds.) (2011): *Slovník súčasného slovenského jazyka. H – L*. [2. zv.]. Bratislava: Veda, vydavateľstvo SAV.
- štúdia v zborníku: ĎUROVIČ, Ľubomír (2000): Jazyk mesta a spisovné jazyky Slovákov. In: S. Ondrejovič (ed.): *Sociolinguistica Slovaca 5. Mesto a jeho jazyk.* Bratislava: Veda, vydavateľstvo SAV, s. 111 – 117.
- štúdia v časopise: DOLNÍK, Juraj (2009): Reálne vz. ideálne a spisovný jazyk. In: *Jazykovedný časopis*, roč. 60, č. 1, s. 3 – 12. DOI 10.2478/v10113-009-0001-3. [cit. DD-MM-RRRR].
- internetový zdroj: Slovenský národný korpus. Verzia prim-5.0-public.all. Bratislava: Jazykovedný ústav Ľudovíta Štúra SAV. Dostupné na: https://korpus.juls.savba.sk [cit. DD-MM-RRRR].

INSTRUCTION FOR AUTHORS

JOURNAL OF LINGUISTICS publishes articles **free of publication charges**.
Accepted languages: all Slavic languages, English, German. Scientific submissions should include a 100-200 word abstract in English and a list of key words in English (3-8 words).
Scientific articles in a language other than Slovak or Czech should contain a summary in Slovak (400-600 words) – translation into Slovak will be provided by the editor.
Reviewing process: scientific articles undergo a double-blind peer-review process and are reviewed by two reviewers, other articles by one reviewer. The authors are provided with the reviews without the name of the reviewer.
Technical and formal directions:

- Articles must be submitted in an electronic form (text editor Microsoft Word, 12-point Times New Roman font, and 1.5 line spacing). If the text contains special symbols, tables, diagrams, pictures etc. it is also necessary to submit a pdf or printed version.
- Contributions should contain the full name of the author(s), as well as his/her institutional affiliation(s).
- The text of the contribution should be flush left; words at the end of a line are not hyphenated; a hard return is used only at the end of a paragraph.
- Paragraphs should be indented.
- Italics is usually used for titles of works and for linguistic examples.
- Boldface is usually used for subtitles and key terms.
- References in the text (in parentheses) contain the surname of the author, the year of publication and the number(s) of the page(s): (Horecký 1956, p. 95).
- The list of references is placed at the end of the text (not in the notes) in alphabetical order. If there are several works by the same author, they are listed chronologically.

References:

- Monograph: ONDREJOVIČ, Slavomír (2008): *Jazyk, veda o jazyku, societa*. Bratislava: Veda, vydavateľstvo SAV. 204 p.
- Dictionary: JAROŠOVÁ, Alexandra – BUZÁSSYOVÁ, Klára (eds.) (2011): *Slovník súčasného slovenského jazyka. H – L*. [2. zv.]. Bratislava: Veda, vydavateľstvo SAV.
- Article in a collection: ĎUROVIČ, Ľubomír (2000): Jazyk mesta a spisovné jazyky Slovákov. In: S. Ondrejovič (ed.): *Sociolinguistica Slovaca 5. Mesto a jeho jazyk*. Bratislava: Veda, vydavateľstvo SAV, pp. 111–117.
- Article in a journal: DOLNÍK, Juraj (2009): Reálne vz. ideálne a spisovný jazyk. In: *Jazykovedný časopis*, Vol. 60, No. 1, pp. 3–12. DOI 10.2478/v10113-009-0001-3. [cit. DD-MM-RRRR].
- Internet source: Slovenský národný korpus. Verzia prim-5.0-public.all. Bratislava: Jazykovedný ústav Ľudovíta Štúra SAV. Available at: https://korpus.juls.savba.sk [cit. DD-MM-YEAR].

ISSN 0021-5597 (tlačená verzia/print) ISSN 1338-4287 (verzia online) MIČ 49263

JAZYKOVEDNÝ ČASOPIS

VEDECKÝ ČASOPIS PRE OTÁZKY TEÓRIE JAZYKA

JOURNAL OF LINGUISTICS

SCIENTIFIC JOURNAL FOR THE THEORY OF LANGUAGE

Objednávky a predplatné prijíma/Orders and subscriptions are processed by:
SAP – Slovak Academic Press, s. r. o., Bazová 2, 821 08 Bratislava
e-mail: sap@sappress.sk

Registračné číslo 7044

Evidenčné číslo 3697/09

IČO vydavateľa 00 167 088

Ročné predplatné pre Slovensko/Annual subscription for Slovakia: 18 €, jednotlivé číslo 6 €
Časopis je v predaji v kníhkupectve Veda, Štefánikova 3, 811 06 Bratislava 1