sciendo
DOI 10.2478/jazcas-2021-0049

SHARING DATA THROUGH SPECIALIZED CORPUS-BASED TOOLS: THE CASE OF GramatiKat

DOMINIKA KOVÁŘÍKOVÁ
Institute of the Czech National Corpus, Faculty of Arts, Charles University, Prague, Czech Republic

KOVÁŘÍKOVÁ, Dominika: Sharing data through specialized corpus-based tools: The case of GramatiKat. Journal of Linguistics, 2021, Vol. 72, No 2, pp. 531 – 544.

Abstract: This paper presents a specialized corpus tool GramatiKat in the context of Open Science principles, namely data sharing, which offers opportunities for original research and facilitates verifiability of research and building on previous research. The tool is designed primarily for examining grammatical categories from the quantitative point of view. It offers grammatical profiles of particular lemmas (currently 14 thousand Czech nouns) and the proportion of individual grammatical categories within a part of speech, i.e., the standard behavior of a word class. The data in GramatiKat are pre-processed, statistically evaluated, and presented in charts and tables for clarity, and they are available to other linguists, especially from fields of morphology and lexicography. This article is aimed at providing inspiration and support to corpus and non-corpus linguists with utilization and enhanced use of the existing tools and with the creation of new specialized tools available to other users.

Keywords: specialized corpus tools, grammatical category, morphology, lexicography, Open Science

1 CORPUS LINGUISTICS IN THE CONTEXT OF OPEN SCIENCE

Current trends of open access to research outputs and of data sharing, which are among the principles of Open Science, are key themes of the contemporary research community. In corpus linguistics, this is not a new topic; corpora themselves, as well as corpus concordancers, are research outputs that allow both the verifiability of research conducted on corpus data and the building on previous research, while offering all users vast opportunities for original research in various fields of linguistics. These are precisely the requirements formulated by J. Chromý and V. Cvrček [1, pp. 8–11] in the article opening the monothematic issue of *Naše řeč* (1/2021), which set itself the task of opening a broad discussion on the topic of open linguistics. Contributions to this discussion range from appeals and program statements, to organizing projects aiming at data sharing, and to the actual implementation of the principles in the form of shared research articles, data, software, or other tools. I would like to contribute as well, specifically in the area of "synergy and cooperation between the researchers" [1, p. 5].

Just ten years ago, in 2011, an article was published in *Naše řeč*, that was reflected on in an editorial of the *Jazykovedný časopis* in 2019 (2/2019). The article "Možnosti a meze korpusové lingvistiky" [2] focuses, among other things, on changing trends in corpus linguistics, a discipline that adopted the principles of sharing data and tools for their analysis from the very beginning of its existence. In the first 20 years of its greatest boom since the late 1980s, corpus linguistics was devoted first to data collection and tagging, and subsequently to diverse and extensive linguistic research enabled by high-quality, large-scale data and to expanding possibilities for analysis.

In the 10 years that have passed since the 2011 article, another strong trend can be observed: development of specialized tools to process corpus data. Such tools facilitate data analysis methods, such as keyword analysis or statistical evaluation of corpus data, or they offer pre-processed data to enable research focused on particular areas (e.g., identifying metaphors or phraseology, finding n-grams, exploring translation equivalents or vocabulary of a particular text type). This trend is noticeable in corpus linguistics worldwide (see webpage providing links to various corpus tools https://corpus-analysis.com/) and the principles of Open Science have been incorporated in the Czech National Corpus project as well. Seven publicly available tools have been published in the last three years alone, offering statistical data analysis, pre-processed and organized datasets or data visualization in the form of interactive tables, graphs, and dialectological maps.[1]

As a co-author of two tools aimed at assisting other linguists with examining research areas of grammatical categories and academic vocabulary (both with Oleg Kovářík), I would like to share my experience with the development and application of such tools (specifically, I will focus on the GramatiKat tool [3]). Hopefully, this article will provide inspiration and support to corpus and non-corpus linguists in utilizing and enhancing the existing tools, sharing ideas and resources such as access to data or programming skills, and provide other researchers with new tools and pre-processed data for their original research.

2 GramatiKat: TOOL FOR RESEARCH OF GRAMATICAL CATEGORIES

The GramatiKat tool is designed primarily for researching grammatical categories in Czech. The idea of examining grammatical categories from a less traditional, quantitative point of view originated many years ago, during work on the corpus-based *Mluvnice současné češtiny* [16, pp. 205–209], and was sparked by

[1] Tools created within the Czech National Corpus project: SyD [4], Morfio [5], KWords [6], Treq [7], Pro školy [8], Slovo v kostce [9], Calc [10], Lists [11], KorpusDB [12], QuitaUp [13], Mapka [14], Akalex [15] and GramatiKat [3]. Manuals to and information on all the tools are available at https://wiki.korpus.cz/doku.php/en:manualy.

research on gradation of adjectives. This phenomenon stands between grammar and word formation, partly because it does not apply to all adjectives. In fact, we found out that comparative and superlative forms are attested only in a fraction of adjectives – in the most recent corpus of contemporary written texts SYN2020 [17], only about 10% of adjectives (with frequency 3 or more) have comparative or superlative form, i.e., less than 4 thousand adjectives. Among them, however, there are adjectives with a very high frequency, so graded forms are encountered quite often in texts, and gradation is considered a relatively common phenomenon.

The primary goal of GramatiKat is to expand this initial idea of quantitative research to all grammatical categories in all parts of speech, especially nouns, adjectives, and verbs. Such information is not accessible through a standard corpus concordancer search and so only someone with special resources (access to data and programming skills) is usually able to carry out such research. Through the GramatiKat tool, the data are available to all interested researchers. It would be, of course, possible to share the raw data with a presumption that experienced users can draw their own conclusions. However, we have chosen a more involved approach. The data is pre-processed, statistically evaluated, and presented in charts and tables (and of course, the raw data is also available).

The first version of the tool has been available since early 2021 and includes information about Czech nouns and their categories of number, case, and gender. For the Slovko 2021 conference, data for Slovak nouns were added (see more in section 4.4).

The information that is currently available in the GramatiKat tool includes:

- distribution of grammatical category values within a word class, e.g., distribution of all 14 cases (7 cases in singular and plural) in Czech nouns. For example, a chart (identical to figure 1 in section 4.1) shows the percentage of locative singular or dative plural;
- distribution of grammatical category values within a lemma, or, a grammatical profile [18, p. 11], e.g., what is the grammatical profile of a lemma *večer* 'evening';
- a list of words that show an unusually high frequency of a grammatical category value, e.g., individual nouns that are attested significantly more often than other nouns in a specific case;
- a list of words with a gap (or unattested form) in the paradigm, e.g., singularia tantum.

3 MATERIAL AND METHODS

We used data from the SYN2015 [19] representative corpus of contemporary written Czech with 120 million words (incl. punctuation) to create the GramatiKat tool. The corpus is balanced and consists of 1/3 fiction, 1/3 journalistic, and 1/3 non-

fiction and academic texts. In the first version of GramatiKat, we included all nouns from the SYN2015 corpus with a frequency of at least 100[2] (14 thousand noun lemmas).

For comparison of Czech and Slovak nouns, we prepared a special parallel subcorpus of InterCorp version 13, containing parallel Czech and Slovak texts. The subcorpora size is 50 million lines (incl. punctuation) in Czech, 49 million lines in Slovak. For Czech, we examined 5600 lemmas with a frequency of at least 100, for Slovak 5400 lemmas.

We examined three grammatical categories, or rather three combinations of grammatical categories: the number, the combination of case and number (i.e., 14 paradigm cells), and the combination of case and number with gender.[3]

For statistical evaluation of the data, a boxplot was used. In addition to its usual purpose, which is visualization of numerical data in quartiles, a boxplot can also serve as a guide to estimate which values are standard and which are exceptions, in other words, which values are unusually high or unusually low. Such outliers are often calculated as exceeding 1.5 times the interquartile range above the third quartile and below the first quartile (although there are other options for evaluation, e.g., using standard deviation; outliers can also be disregarded altogether).

The boxplots in GramatiKat not only show but also determine the standard and non-standard behavior of the whole part of speech. In the context of the presented research, we consider the values that do not belong to the outliers to be the standard behavior of Czech nouns in a given case, and outliers from 1.5 times above the third quartile to be exceptions – words with an unusually high representation of the given case. The lower outliers are not present in our data at all (the 1.5 times the IQR below the first quartile reach zero or negative values in all cases), and we consider the absence of a certain form in a corpus (or, a gap in the paradigm) to be non-standard behavior.[4]

In examining the quantitative properties of grammatical categories, it is necessary to be aware that the percentage of values in each grammatical category can be influenced by various factors, particularly by the size and composition of the corpus and the frequency of the lemma. A large representative and balanced corpus with a wide range of text types in balanced proportions such as SYN2015 ensures a high degree of reliability of the grammatical profiles, at least within the language

[2] We have chosen a relatively high frequency so that the probability of a given form would be high enough.

[3] In the future, we intend to include other parts of speech, primarily adjectives and verbs. In addition to traditional grammatical categories, we would like to thoroughly examine negation, which has not yet received sufficient attention in Czech grammatical or lexicographic descriptions or even corpus lemmatization (adjectives).

[4] The vocative is an exception: both singular and plural are usually unattested, so the gap in the paradigm is actually standard behavior.

variety under consideration. The researcher should always be aware of this limitation and especially of the influence a smaller or unbalanced corpus may have on the results (see sections 4.3 and 4.4).

4 RESULTS

4.1 Distribution of a grammatical category of case in Czech nouns

The main information that the user can get from the GramatiKat tool is the overview of a given grammatical category within a certain part of speech. Figure 1 gives an overview of the case (in combination with number[5]) distribution in Czech nouns in the SYN2015 corpus. It shows the standard behavior of Czech nouns, as well as the threshold for an unusually high proportion of each of the cases. This threshold varies notably across individual cases, e.g., 2.5% is an unusually high proportion of dative plural, whereas 24.1% is an unusually high proportion of nominative plural, and the percentage is even higher (57.4%) for nominative singular. As mentioned above, the lower threshold for all cases is zero, in other words, a gap in the paradigm (vocative case is an exception). Specific values relating to the boxplots in figure 1 (median, interquartile range, and outliers) are presented in table 1.

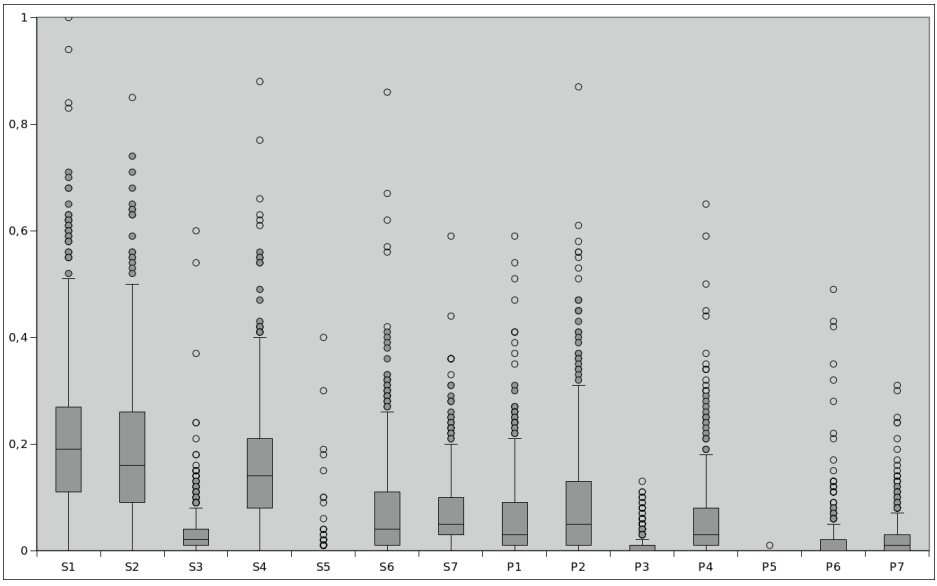

Fig. 1. Case distribution of Czech nouns in the SYN2015 corpus (nouns with a frequency of at least 100). Outliers indicate individual noun lemmas that have an unusually high percentage of the given case. Boxes, together with whiskers, represent the range of standard behavior of nouns

[5] We look at distribution in of 14 cases, i.e., 7 cases in two numbers, to capture the whole paradigm of each lemma.

	singular							plural						
	1 (nom)	2 (gen)	3 (dat)	4 (acc)	5 (voc)	6 (loc)	7 (inst)	1 (nom)	2 (gen)	3 (dat)	4 (acc)	5 (voc)	6 (loc)	7 (inst)
Unusually high	57.4	54.9	9.3	48.7	0.0	25.2	21.6	24.1	29.0	2.5	19.1	0.0	4.2	7.4
75th perc.	28.0	25.0	4.0	22.6	0.0	10.5	9.8	9.8	11.8	1.0	7.8	0.0	1.7	3.0
Median	19.3	14.8	1.8	14.5	0.0	3.9	5.5	3.5	4.2	0.2	2.8	0.0	0.4	0.9
25th perc.	12.5	7.4	0.7	7.8	0.0	0.9	2.8	0.5	0.5	0.0	0.3	0.0	0.0	0.0
Unattested	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Tab. 1. Supplementary table to Fig. 1. The value in the first line indicates the threshold for outliers or the threshold of an unusually high proportion of the given case (in %). Values between the 25th and 75th percentile form the box in the boxplot for each case. The values are calculated based on nouns that occurred with a frequency of at least 100 in the SYN2015 corpus

The overview of standard behavior of nouns within the grammatical category of case is not completely new information, it has been in shorter form presented in the books *Statistiky češtiny* [20, p. 134] and *Mluvnice současné češtiny* [16, p. 141]. Also, it is not difficult to extract this information directly from a corpus concordancer such as KonText. But this basic information is merely a gateway to grammatical profiles of all 14,000 lemmas examined, as well as to the groups of lemmas belonging to outliers (see section 4.2).

4.2 Grammatical profiles of individual lemmas

In GramatiKat, it is possible to display the grammatical profile of a particular lemma against the background of standard behavior of the whole part of speech. In figure 2, we can see the case distribution of the lemma *sekera* ‘ax’ in the form of grey dots, figure 3 shows the data for the word *uvozovka* ‘quotation mark’. In both figures, there is an evident deviation from the standard. Figure 2 shows an unusually high frequency of instrumental singular (high percentage of instrumental is characteristic of other tools as well, such as *lopata* ‘spade’, *kladivo* ‘hammer’, *nůž* ‘knife’, or *hrábě* ‘rake’). In figure 3, we can see that the lemma is overall more common in the plural, and we can observe an extremely high frequency of locative plural (*v uvozovkách* ‘in quotation marks’).

Finding words that have an unusually high percentage of a certain case is also possible. For example, under dative singular, we can find 1169 lemmas where this case accounts for at least 9.3% (the threshold for outliers, see table 1). The lemmas that occur almost exclusively in this case include *mání* ‘having’, *dostání* ‘getting’, *nepoznání* ‘not recognizing’, *zahození* ‘discarding’, or **snědek* ‘eating’. All of these lemmas are components of multiword units (mostly with the verb *být* ‘to be’ and preposition *k* ‘to’: *ne/být k mání* ‘not/to be had’, *ne/být k dostání* ‘not/to be gotten’, *být k nepoznání* ‘to be unrecognizable’, *něco k snědku* ‘something to be eaten’) and their classification as nouns is entirely formal. This is especially evident in the reconstructed

nominative singular **snědek*. Among other lemmas with unusually high but not exclusive dative singular (around 20%) are *jubileum* ‘anniversary’, *politování* ‘regret’, *zlepšení* ‘improvement’, *usmrcení* ‘killing’, and *obezřetnost* ‘prudence’.

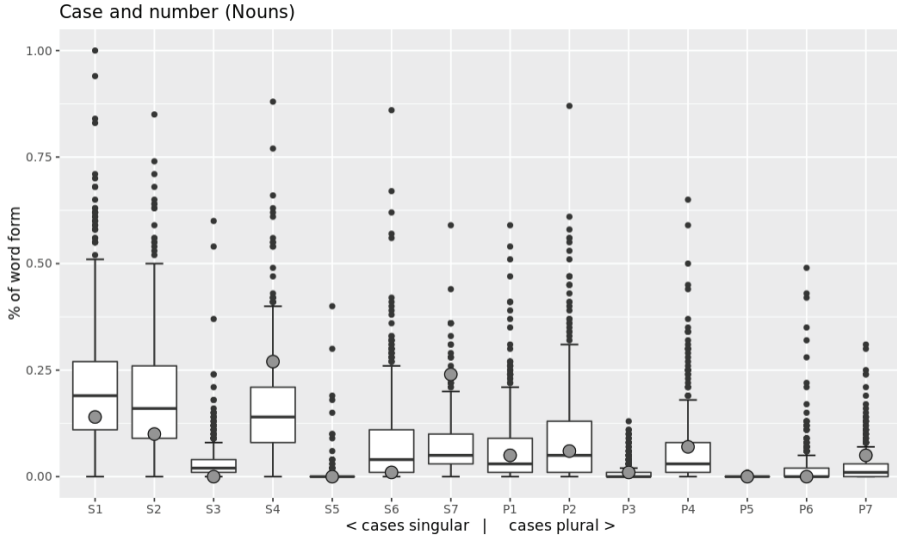

Fig. 2. The grey dots show the percentage of individual cases within the lemma *sekera* ‘ax’, the background boxplots show standard behavior as well as outliers of Czech nouns

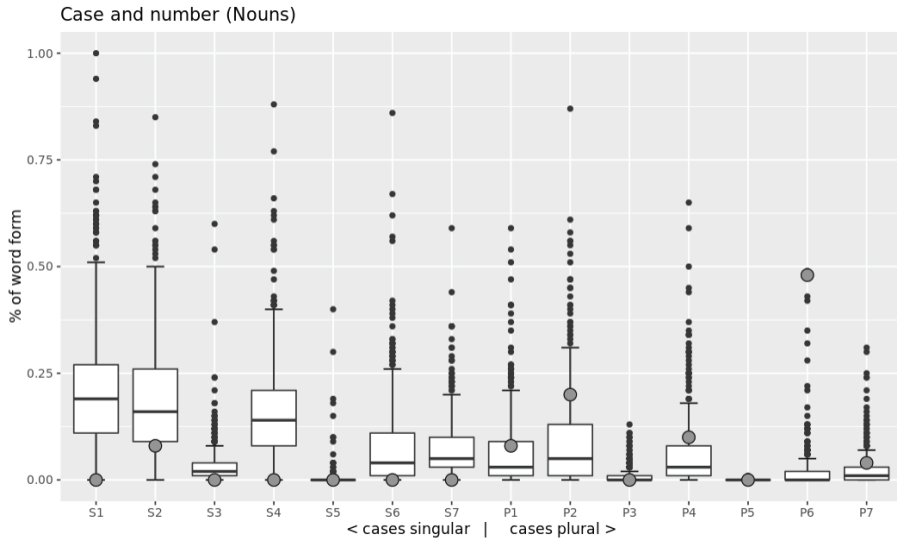

Fig. 3. The grey dots show the percentage of individual cases within the lemma *uvozovka* ‘quotation mark’, the background boxplots show standard behavior and outliers of Czech nouns

Similarly, it is possible to find lemmas with a missing form, for example, nominative plural. Almost 25% of the examined nouns, or 3400 lemmas, do not occur in nominative plural.[6] Such a comprehensive list of singularia tantum can lead to a better understanding and theoretical description of this phenomenon, especially the reasons for missing plural forms (usually semantic incompatibility, strong semantic preference, or limited collocability [21, p. 6]. Some of the lemmas with the plural form missing are *agresivita* 'aggressiveness', *bezpečí* 'safety', *komplexnost* 'complexity', *počasí* 'weather', or *potomstvo* 'offspring'.

4.3 Proportion of standard behavior nouns

The common presumption that most of the reasonably frequent nouns have a complete paradigm with no significant deviations is revealed as incorrect. On the contrary, the examination of the material available in GramatiKat shows that only about 25% of nouns with a frequency of at least 100 in the corpus can be considered standard concerning the distribution of cases[7] – all the cells of their paradigms are represented and there are no unusually frequent paradigm cells.[8] More specifically, approximately half of the lemmas examined show an unusually high frequency of at least one paradigm cell, and approximately 50% of the lemmas show at least one missing paradigm cell, with a significant overlap between the two groups.

However, this phenomenon is highly frequency-sensitive. The percentage of standard lemmas increases (up to a point, see figure 4) and decreases with their frequency in the corpus – the probability of attested dative plural, for example, is quite low in lemmas with a frequency lower than 100. And ultimately, a lemma with a frequency lower than 14 cannot be represented by all 14 cases.

In any case, non-standard behavior in nouns is not a marginal phenomenon but rather a frequent feature that should be monitored and described not only within the realm of grammar but also in lexicographical description (see section 5).

4.4 Comparison of Czech and Slovak nouns

The GramatiKat tool is ready to process material from languages other than Czech as well. The prerequisite is a sufficiently large morphologically tagged corpus. We have so far processed Croatian nouns (available upon request) and Slovak nouns. A major issue for comparing two languages, as well as for reliability

[6] Such lemmas do not occur or rarely occur in any of the plural forms.

[7] L. Janda and F. M. Tyers claim that "[o]nly a fraction of lexemes are encountered in all their paradigms in any corpus or even in the lifetime of any speaker" [18, p. 1]. The results presented here show that the situation is not as severe (perhaps the corpus size played a role). However, it is possible to agree that non-standard paradigms are not an exception.

[8] Again, the vocative was excluded.

of the results, is their dependency on corpus size, types of texts or, in the case of smaller corpora, even on the individual texts included. For Czech, we are satisfied with working with the representative and balanced corpus SYN2015. For other languages including Slovak, InterCorp data are large and diverse enough (even though not balanced). They offer the possibility to compare two (and even more) languages on the basis of the exact same texts, which we implemented in the GramatiKat tool for the language pair of Czech and Slovak. The comparison of Czech and Slovak is very reliable, the results for the two separate languages are less so (compare figure 1 with 5 and 6).

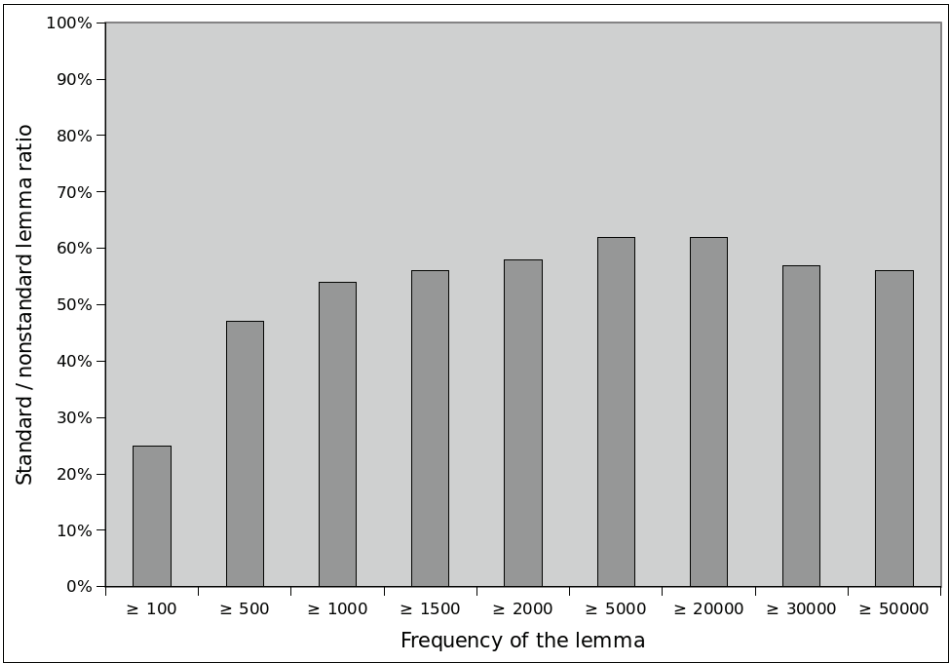

Fig. 4. Percentage of lemmas with a standard grammatical profile on different frequency levels. The figure shows that the phenomenon is frequency-dependent

A comparison of case distribution in Czech and Slovak (figure 5 for singular and figure 6 for plural, also summarized in table 2) shows that the two languages are very close in this respect. The biggest differences are between nominative singular, which is 1.1 percent more frequent in Czech, whereas accusative singular is 0.9 percent more frequent in Slovak. Whether or how these two phenomena are related to each other could only be determined through further extensive analysis.

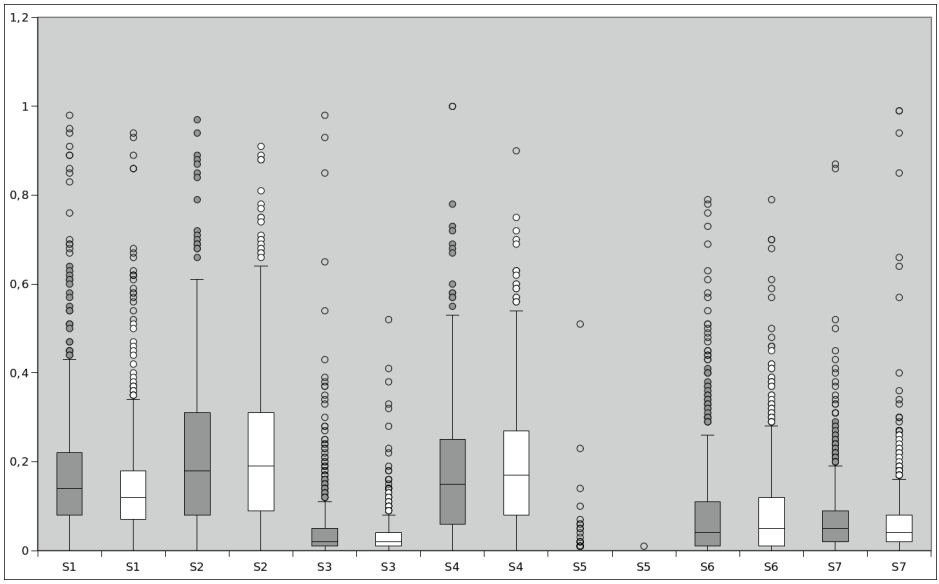

Fig. 5. Comparison of singular cases distribution in Czech (grey) and Slovak (white) nouns in InterCorp version 13, lemmas with frequency of at least 100

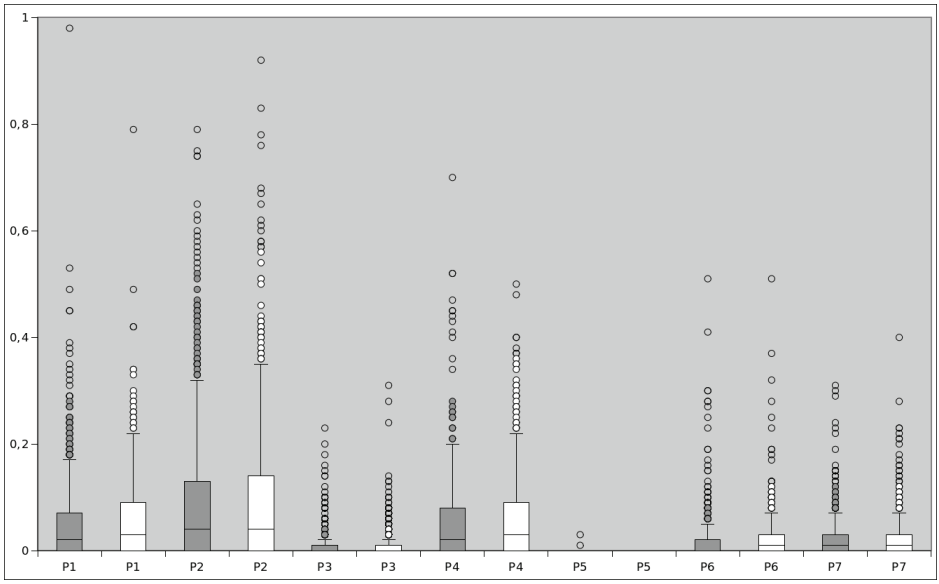

Fig. 6. Comparison of plural cases distribution in Czech (grey) and Slovak (white) nouns in InterCorp version 13, lemmas with frequency of at least 100

	singular							plural						
	1 (nom)	2 (gen)	3 (dat)	4 (acc)	5 (voc)	6 (loc)	7 (inst)	1 (nom)	2 (gen)	3 (dat)	4 (acc)	5 (voc)	6 (loc)	7 (inst)
Median CZ	13.35	17.28	2.07	15.13	0.00	3.87	4.65	2.35	3.61	0.16	2.70	0.00	0.42	0.72
Median SK	12.24	17.65	1.66	16.04	0.00	4.67	4.73	2.53	3.85	0.08	2.94	0.00	0.51	0.74
Difference	**-1.11**	**0.37**	**-0.40**	**0.91**	**0.00**	**0.80**	**0.08**	**0.18**	**0.23**	**-0.08**	**0.24**	**0.00**	**0.09**	**0.01**

Tab. 2. Supplementary table to fig. 6 and 7 showing the difference between Czech and Slovak standard noun behavior. The values (in %) are calculated based on nouns that occurred with a frequency of at least 100 in the InterCorp version 13 parallel Czech-Slovak subcorpus

5 GramatiKat IN LINGUISTIC RESEARCH – SUGGESTIONS

The GramatiKat data can be utilized in various linguistic disciplines. Instant use is possible in lexicography by detecting the lemmas with non-standard behavior (gaps in paradigm, extremely frequent forms). For example, it could be helpful to supplement the entry *brva* 'eyelash' in the *Academic Dictionary of Contemporary Czech* [22] with the information that 77% occurrences of this lemma are in instrumental singular, so the lemma is overwhelmingly often a component of the idiom *ne(po)hnout (ani) brvou* 'not to bat (even) an eyelash' (the idiom itself is listed in the dictionary, without frequency information).

Similarly, the tool can be used for educational purposes, especially in teaching Czech as a second language. Adaptation of educational practices based on case distribution is discussed by Janda and Tyers [18] who suggest that "learning may be enhanced by focusing only on the word forms most likely to be encountered" [18, p. 28). For example, we can consider teaching only the genitive and accusative singular of the lemma *večer* 'evening' (nominative, genitive and accusative represent 78% of the lemma occurences), and genitive and locative singular of the lemma *zahrada* 'garden' (64% of occurrences), especially in the earlier stages of the learning process.

The obvious direction for closer examination of the pre-processed data is morphological analysis. Determination of quantitative properties of individual grammatical categories within the individual parts of speech alone can be a valuable outcome. With information on all grammatical categories completed, we can expect re-evaluation or more accurate understanding and description of morphological phenomena (as was the case of adjective gradation mentioned above). Since the anomalies in case distribution are often caused by collocational restrictions, research should be also oriented toward multi-word units which are underrated and underrepresented in current grammatical, as well as lexicographic descriptions.

As a part of the Feast and Famine project[9], research of defectivity and anomalies in grammatical profiles of Czech nouns is currently underway. The preliminary results show that GramatiKat data is very relevant to theoretical research of language potentiality and of paradigm defectivity, as well as the underlying motives (especially semantics and collocability).

6 CONCLUSION

This article is based on the plenary session of the Slovko 2021 conference. It presents an online tool for research of grammatical categories – GramatiKat. The tool reflects the current atmosphere of open access and shared data in science and humanities, as noted in Chromý and Cvrček [1]. It provides users interested in linguistic research of grammatical categories (namely in the fields of morphology and lexicography) with a large-scale, pre-processed corpus data, as well as visualizations of grammatical categories of Czech. Also available is a comparison of Czech and Slovak nouns based on a parallel corpus of the two languages.

The study gives an overview of the grammatical category of case (in combination with number) in Czech nouns – it shows the standard behavior of Czech nouns, as well as the thresholds for non-standard case distribution. On this basis, the charts in GramatiKat also show the case distribution and anomalies within paradigms of individual lemmas. The anomalies are not a peripheral phenomenon within nouns; section 4.3 shows that a significant number of lemmas exhibit non-standard case distribution – either a paradigm gap or an unusually high frequency of a certain case.

However, the main goal of this article is not to present the tool itself (although as a co-author, I am grateful for this opportunity); my ambition is to inspire others to undertake similar projects which provide other linguists with otherwise inaccessible data and facilitate a broader and deeper examination of a specific phenomenon. I demonstrated in several examples how a tool such as GramatiKat can be versatile and can serve researchers of various linguistic fields or interests. The data is relevant to morphology, as well as lexicology and lexicography, to theoretical research of language potentiality and defectivity, and can be also used for educational purposes.

The benefits of a tool such as GramatiKat, offering pre-processed data, are numerous. The shared data follows the principles of Open Science, namely the verifiability of research and building on previous research. Most importantly, such tool gives all linguists, corpus and non-corpus, access to data that might otherwise be unattainable. The users can then undertake thorough research of a scale that is impossible for one person and can also use the tool in original and unexpected ways.

[9] Feast and Famine: Confronting Overabundance and Defectivity in Language is a project that takes place in several European universities and language institutes, including Sheffield University, the Faculty of Arts of Charles University and the Czech Language Institute (https://www.sheffield.ac.uk/feastandfamine).

ACKNOWLEDGEMENTS

This paper resulted from implementation of the Czech National Corpus project (LM2018137) funded by the Ministry of Education, Youth and Sports of the Czech Republic within the framework of Large Research, Development and Innovation Infrastructures. The research has been also in part funded by the United Kingdom's Arts and Humanities Research Council (AH/T002859/1) and by the European Regional Development Fund-Project "Creativity and Adaptability as Conditions of the Success of Europe in an Interrelated World" (No. CZ.02.1.01/0.0/0.0/16_019/0000734).

References

[1] Chromý, J., and Cvrček, V. (2021). Lingvistika jako otevřená a transparentní disciplína. Naše řeč, 104(1), page 514.

[2] Cvrček, V., and Kováříková, D. (2011). Možnosti a meze korpusové linvistiky. Naše řeč, 94(3), pages 113–133.

[3] Kováříková, D., and Kovářík, O. (2021). GramatiKat. Prague: ÚČNK FF UK. Praha 2021. Accessible at: http://www.korpus.cz/gramatikat.

[4] Cvrček, V., and Vondřička, P. (2011). SyD – Korpusový průzkum variant. Prague: FF UK. Accessible at: http://syd.korpus.cz.

[5] Cvrček, V., and Vondřička, P. (2013). Morfio. Prague: ÚČNK FF UK. Accessible at: http://morfio.korpus.cz.

[6] Cvrček, V., and Vondřička, P. (2013). KWords. Prague: ÚČNK FF UK. Accessible at: http://kwords.korpus.cz.

[7] Vavřín, M., and Rosen, A. (2015). Treq. Prague: ÚČNK FF UK. Accessible at: http://treq.korpus.cz.

[8] L. Lukešová (ed.). (2017). Pro školy – reportáž korpusových cvičení. Prague: ÚČNK FF UK. Accessible at: http://www.korpus.cz/protokoly.

[9] Machálek, T. (2019). Slovo v kostce – agregátor slovních profilů. Prague: ÚČNK FF UK. Accessible at: http://www.korpus.cz/slovo-v-kostce.

[10] Cvrček, V. (2019). Calc: Korpusová kalkulačka. Prague: ÚČNK FF UK. Accessible at: http://www.korpus.cz/calc.

[11] Křen, M., and Cvrček, V. (2019). Lists: Prohlížeč frekvenčních seznamů. Prague: ÚČNK FF UK. Accessible at: http://www.korpus.cz/lists.

[12] Vondřička, P. (2020). KorpusDB: Databáze slovních tvarů a lemmat doložených v korpusech ČNK. Verze 1.0. Prague: ÚČNK FF UK. Accessible at: http://db.korpus.cz/.

[13] 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/.

[14] Goláňová, H., Waclawičová, M., and Pejcha, J. (2021). Mapka: Mapová aplikace pro korpusy mluvené češtiny. Verze 1.1. Prague: ÚČNK FF UK. Accessible at: http://www.korpus.cz/mapka.

[15] Kováříková, D., and Kovářík, O. (2021). Akalex. Prague: ÚČNK FF UK. Praha 2021. Accessible at: http://www.korpus.cz/akalex.

[16] Cvrček, V. et al. (2009). Mluvnice současné češtiny I.: Jak se píše a jak se mluví. Praha: Karolinum.
[17] Křen, M. et al. (2020). SYN2020: reprezentativní korpus psané češtiny. Prague: ÚČNK FF UK. Accessible at: http://www.korpus.cz.
[18] Janda, L. A., and Tyers, F. M. (2018). Less is more: why all paradigms are defective, and why that is a good thing. Corpus linguistics and linguistic theory, 14(2). Accessible at: https://doi.org/10.1515/cllt-2018-0031.
[19] Křen, M. et al. (2015). SYN2015: reprezentativní korpus psané češtiny. Prague: ÚČNK FF UK. Accessible at: http://www.korpus.cz.
[20] Čermák, F. et al. (2009). Statistiky češtiny. Prague: NLN.
[21] Kováříková, D. et al. (2019). Lexicographer's Lacunas or How to Deal with Missing Representative Dictionary Forms on the Example of Czech. International Journal of Lexicography, 33(1), pages 90–103. Accessible at: https://doi.org/10.1093/ijl/ecz027.
[22] Akademický slovník současné češtiny (2021). Accessible at: https://slovnikcestiny.cz/uvod.php.