sciendo
DOI 10.2478/jazcas-2023-0032

# COMPARATIVE LEXICAL ANALYSIS OF NOUN LEMMAS IN SLOVAK JUDICIAL DECISIONS

MIROSLAV ZUMRÍK
Slovak National Corpus, Ľ. Štúr Institute of Linguistics, Slovak Academy of Sciences, Bratislava, Slovakia

ZUMRÍK, Miroslav: Comparative Lexical Analysis of Noun Lemmas in Slovak Judicial Decisions. Journal of Linguistics, 2023, Vol. 74, No 1, pp. 140 – 149.

**Abstract:** The paper presents a comparative lexical analysis of the most frequent noun lemmas in Slovak judicial decisions. The data was taken from the large corpus of decisions provided by the Ministry of Justice of the Slovak Republic, and compared with lemmas from four other corpora. The style of administrative, legal and other highly formalized texts has been receiving attention in recent years, although research into the style of Slovak judicial decisions remains rather sparse, which could partially stem from their idiosyncratic nature. The paper's aim is thus to focus on the quantitative and qualitative characteristics of noun lemmas found specifically in the style of judicial decisions.

**Keywords:** judicial decisions, nouns, stylistics, legal linguistics

## 1 INTRODUCTION

Judicial decisions are a peculiar genre of administrative/legal texts. According to František Štícha (1985, pp. 69–71), the peculiarity of decisions in criminal cases in particular is due to the fact they regularly contain depictions of awkward or extreme life situations that are otherwise not subject to discussion in formal, official settings. Because of that, the language of judicial decisions tends to suffer from stylistic flaws. These shortcomings, as Štícha suggests, result among others from the specific circumstances in which the documents are produced (such as time pressure) and from the very nature of administrative texts. The style of judicial decisions is then characterized by a discrepancy between the elevated legal language and the often mundane content described using this language, overly complicated and prolonged sentences, the use of archaic expressions, as well as the overuse of explicitness and noun phrases. Jozef Mistrík, similarly, describes the style of administrative texts (including judicial decisions) as rather stereotypical, lexically and syntactically poor, simple in composition, yet simultaneously rich in genres, and these texts represent the least researched style in standard Slovak (1997, pp. 458–459).

General claims like the statements above are too vague to be testable, so they need to be transformed into more specific ones (Chlumská 2017, p. 53). In this paper, I will therefore examine the semantic and distributional features of the most frequent

noun lemmas in Slovak judicial decisions, as compared to lemmas in Slovak acts, specialized studies, short stories and texts from a balanced corpus. I will ask the following questions:

1. how numerous and how diverse are the most frequent noun lemmas in the decisions?
2. how do they differ from lemmas in other genres?
3. how could the different numbers and differences be explained?

## 2 THEORETICAL FRAMEWORK

Accounts on stylistic features of Slovak administrative texts, including noun lexis, have been given by Jozef Mistrík (1975; 1997) or Ján Findra (2004) and more recently by Daniela Slančová et al. (2022). Linguistic interest in legal language, termed "legal linguistics"(as the term is used in Vogel 2019 or Cvrček 2016) follows from the natural connection between language and law: "(l) laws are coded in language, and the processes of the law are mediated through language" (Gibbons 1999, p. 156). The style of legal texts has also been studied in Slovakia (Abrahámová 2012; Slančová et al. 2022, pp. 669–712), focusing mostly on the language of legal acts (Čulenová 2010; Imrichová – Turočeková 2015) or legal terminology, as discussed, for example, with respect to individual Slovak legal terms in shorter journal contributions by Rudolf Kuchár throughout the second half of the 20th century.

The main rationale for choosing the texts of judicial decisions for analysis is the fact that they are produced on a massive scale, judging by their the numbers issued by Slovak courts alone, and thus "all around us", so to say. The digital archive at the Ministry of Justice of the Slovak Republic[1] contains 4,1 M anonymized judicial decisions, with tens of new decisions added every day.

The motivation for focusing on nouns is that this part of speech, together with verbs, represents a "major part of speech" (Chlumská 2017, p. 58), capable of denoting every thinkable aspect of reality, both static, dynamic, physical or abstract (Ružička et al. 1966, p. 62). In this paper, I work with a customized set of semantic categories, which resulted from the most frequent noun lemmas found in the analysed corpora: abstract entity, action, institution, norm, object, person, physical entity, place, proper name, text-related entity, time. At the same time, I am aware that the boundary between abstract and physical objects can be blurred (Cvrček et al. 2020, p. 51). It is also to be noted that an analysis of a single stylistic feature and a singular, one-word expression can prove challenging, since stylistic layers are systematically interconnected (Mistrík 1997, p. 34), representing mutually communicating vessels.

[1] https://obcan.justice.sk/infosud/-/infosud/zoznam/rozhodnutie

## 3 METHOD

The data from the Slovak corpora was sized down to genre-specific sub-corpora. From these, random samples of 1 M tokens were extracted. The samples were described as to their composition and analysed for the relative frequency and number of lemmas. The minimal frequency threshold for a lemma was set at 5 occurrences. The frequency distribution of lemmas for each genre was visualized; the samples were tested for significant differences in variance. The lemmas were then described quantitatively and semantically. Differences in the relative frequency of lemmas found in several corpora were tested for statistical significance.

## 4 DATA

The data was taken from the following corpora, compiled at the Ľ. Štúr Institute of Linguistics, Slovak Academy of Sciences:

1. Corpus of Slovak judicial decisions *OD Justice*, version 2 (ODJ, 10.6 B tokens)
2. Corpus of Slovak legislative documents, version 1.9 (LEG, 44.9 M tokens)
3. Corpus of Slovak professional texts *prim-10.0-public-prf* (PRF, 189 M tokens)
4. Corpus of Slovak fiction *prim-10.0-public-img-sk* (IMG, 96.5 M tokens)
5. Balanced corpus of Slovak texts *prim-10.0-public-vyv* (BAL, 571.5 M tokens)

As the original corpora were different in size and document count, and because they each covered different genres, the following sub-corpora were created:

1. Sub-corpus of first instance decisions in criminal cases (CRC, 80.8 M tokens)
2. Sub-corpus of acts (ACT, 19.6 M tokens)
3. Sub-corpus of studies on marketing communication (STD, 3 M tokens)
4. Sub-corpus of Slovak short stories (COL, 27.6 M tokens)
5. Sub-corpus of the balanced corpus (BAL, 201.4 M tokens)

From these sub-corpora, random samples were created, labelled CRCS, ACTS, STDS, COLS, BALS, each containing 1 M random tokens:

| **years** | **documents** | **words** | **nouns** | **n. lemmas** | **(ipm)** |
|---|---|---|---|---|---|
| 1. CRCS | 2020 | 9,338 | 53.8 K – 145 | 251.4 K | 2,027 |
| 2. ACTS | 2002 – 2022 | 1,367 | 72 K – 198 | 274.5 K | 2,388 |
| 3. STDS | 2011 – 2020 | 42 | 123.5 K – 18.6 K | 285.3 K | 4,383 |
| 4. COLS | 1960 – 2020 | 847 | 263.2 K – 96 | 182.3 K | 5,231 |
| 5. BALS | 2010 – 2019 | 91,906 | 302.3 K – 10 | 259.7 K | 6,300 |

**Tab. 1.** Samples used in analysis

The aim behind narrowing down the corpus data was mainly to work with similarly large portions of data and comparably long texts. The texts of first instance decisions in criminal cases are usually one to several pages long, and therefore more similar in length to studies and short stories than to monographs and novels.

The majority of the text samples cover mostly the period of the last 20 years. The relatively low count of documents for studies and short stories results from the fact that these texts are published collectively in proceedings and collections. The low bottom range of word counts could be due to the nature of the texts in question (brief acts and stories). Interestingly, while the relative frequency of noun lemmas in the samples remains approximately the same (with the exception of short stories), the number of noun lemmas in respective samples rises, as if the diversity of the lemmas was increasing. This might also indicate that the respective genres in the given order seem to be less and less formalized: from legally regulated communication during the criminal proceedings, through various walks of life regulated by legal norms, further through the socio-economic reality of marketing communication, to artistic depictions in short stories.

## 5 ANALYSIS

### 5.1 Visualization

The relative frequencies of the 50 most frequent lemmas for each sample were first visualized in the graph below. The x-axis represents the rank of lemmas, the y-axis their relative frequencies. The lemmas are only visualized as columns:

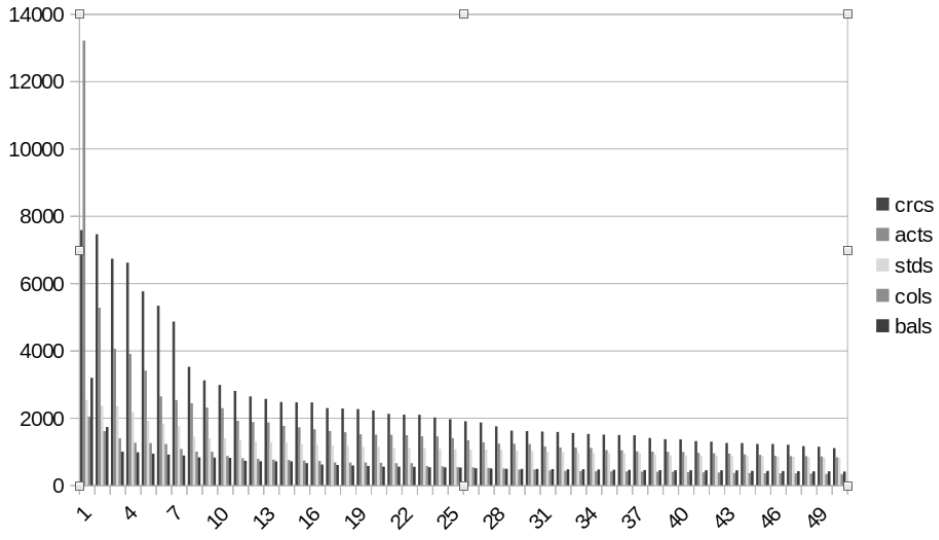



**Fig. 1.** Relative frequencies of the 50 most frequent lemmas

The red outlier column on the very left, standing for the lemma *zákon* 'act' in the sample of acts, exceeds the relative frequencies not only of the other samples, but also of the other lemmas in the act sample. This is understandable, given the nature of legislation that resembles a continuously modified corpus of legislative documents. These have a strong intertextual dimension, as new amendments refer to other acts and documents. The lemma 'act', however, also has the second most prominent frequency in the sample of decisions, so that the intertextual dimension is present here as well. This could be due to the fact that every decision has to refer to relevant articles of the Criminal Code, Criminal Procedure Code or other laws. Compared with the acts, nevertheless, the lemmas in the decisions sample have higher relative frequencies, although together shaped as a shallower slope, with several lemmas that are similarly frequent. After the bulk of these 7 lemmas, the slope drops down. This could mean that there is no such absolutely prominent lemma in the decisions, compared to acts. The relative frequencies of decisions, acts and studies are quite high, however, but the slopes later converge. The slopes for lemmas found in the short stories and the balanced corpus are lower, almost resembling a flat line. This might indicate a variety of frequent lemmas that are distributed more evenly.

### 5.2 Statistical testing

The samples, represented by 20 randomly generated lemmas from each sample, were then tested for significant differences in variance, using the online statistic calculator STATISTICS ONLINE.

Since at least the normality of distribution, one of three assumptions for the more robust, parametric ANOVA test, was debatable (the Kolmogorov-Smirnov test indicated normality, Lilliefors test did not), the Bartlett test of equal variance was dropped and the Kruskal-Wallis test, the non-parametric version of ANOVA, was performed:

The Kruskal-Wallis H test indicated that there is a significant difference in the *dependent variable* between the different *groups*, $\chi 2(4) = 55.22$, $p < .001$, with a mean rank score of 79.65 for CRCS, 66.58 for ACTS, 54.13 for STDS, 23.93 for COLS,28.23 for BALS.

The Post-Hoc Dunn's test using a Bonferroni corrected alpha of 0.005 indicated that the mean ranks of the following pairs are significantly different:

1. CRCS-COLS, CRCS-BALS
2. ACTS-COLS, ACTS-BALS
3. STDS-COLS, STDS-BALS.

Both decisions, acts and studies are significantly different from stories and the balanced corpus, while the significance of mutual differences between decisions, acts and studies has not been confirmed.

### 5.3 Semantic analysis

5.3.1 The lemmas were subsequently ranked according to *ipm* (in decreasing order) and described with the use of semantic categories (ordered alphabetically). Structures erroneously annotated as noun lemmas were excluded from the list. Number of lemmas for each semantic category and sample is given below:

| CATEGORIES | CRCS | ACTS | STDS | COLS | BALS |
|---|---|---|---|---|---|
| *ABSTRACT* | 10 | 16 | 30 | 7 | 21 |
| *ACTION* | 15 | 9 | 2 | 1 | 2 |
| *INSTITUTION* | 1 | 7 | 2 | 0 | 2 |
| *NORM* | 2 | 2 | 0 | 0 | 2 |
| *OBJECT* | 2 | 2 | 6 | 4 | 2 |
| *PERSON* | 7 | 2 | 4 | 12 | 5 |
| *PHYSICAL* | 0 | 0 | 0 | 10 | 3 |
| *PLACE* | 2 | 2 | 2 | 9 | 7 |
| *PROPER NAME* | 0 | 0 | 1 | 0 | 2 |
| *TEXT-RELATED* | 4 | 7 | 0 | 1 | 1 |
| *TIME* | 7 | 3 | 3 | 7 | 4 |

**Tab. 2.** Distribution of noun lemmas according to semantic categories

5.3.2 The 10 **abstract** nouns identified in the decisions sample denote basic terms of criminal law (e.g. *trest* ‘sentence’, *škoda* ‘damage’, *náhrada* ‘compensation’), compared to 16 lemmas in the acts denoting components of legislation (*znenie* ‘wording’, *zmena* ‘change’, *podmienka* ‘condition’). A portion of the 30 lemmas in the studies can be classified as general academic vocabulary (Kováříková 2017, pp. 26–27) (*médium* ‘medium’, *informácia* ‘information’, *cieľ* ‘goal’). The 7 lemmas in short stories represent universal existential concepts (*pravda* ‘truth’, *srdce* ‘heart’, *láska* ‘love’) and the 21 lemmas in the balanced corpus is a mixture of categories (*práca* ‘work’, *spoločnosť* ‘society’, *prípad* ‘case’).

The **action** nouns are mostly present in the decision sample (15, such as *odpor* ‘protest’, *rozkaz* ‘order’, *výrok* ‘verdict’), 9 were found in acts (*činnosť* ‘activity’, *konanie* ‘proceedings’, *rozhodnutie* ‘decision’), and only 1 – 2 in studies (*komunikácia* ‘communication’, *výskum* ‘research’), stories (*cesta* ‘road’, ‘journey’) and in the balanced corpus (*zápas* ‘struggle’).

The **institutions** are represented just by one, crucial institution issuing the decisions (*súd* ‘court’), as opposed to 7 institutions in acts (*úrad* ‘office’, *orgán* ‘authority’, ‘body’, *rada* ‘council’, ‘parliament’), only 2 in studies (*firma* ‘firm’, *podnik* ‘company’) and the balanced corpus (*škola* ‘school’, *štát* ‘state’), none in stories.

The names of **norm**-related lemmas are represented with 2 lemmas in the decisions (*zákon* 'act', *poriadok* 'code'), acts (*zákon* 'act', *predpis* 'regulation') and the balanced corpus (*zákon* 'act', *právo* 'law').

The **object**-related lemmas are represented by 2 lemmas in the decisions (*euro* 'euro', *vozidlo* 'vehicle'), acts (*fond* 'fund', *liek* 'medicine') and the balanced corpus (*euro* 'euro', *kniha* 'book'), and by 4 and 6 in stories (*dvere* 'door', *stôl* 'table') and studies (*produkt* 'product', *značka* 'brand'), respectively.

There were 7 **person**-related lemmas in the decisions (*osoba* 'person', *obžalovaný* 'defendant', *prokurátor* 'prosecutor'), which indicates more individual roles and actors when compared to the 2 person-related lemmas in the acts (*osoba* 'person', *zamestnanec* 'employee'), or 4 lemmas in the studies (*človek* 'human', *zákazník* 'customer', *respondent* 'respondee') and 5 in the balanced corpus (*človek* 'human', *dieťa* 'child'). The most person-related lemmas were found in the stories (*človek* 'human', *žena* 'woman', *dieťa* 'child') as these thematize diverse human interactions, mostly within the intimate surroundings of close family members. The decisions deal with human interaction as well, but due to the anonymization process, the parties are only referred to by random letters, not being automatically annotated as individuals and proper names.

The stories sample contained the most lemmas denoting **places** (*dom* 'house', *izba* 'room'), similarly as in the balanced corpus (*mesto* 'city', *svet* 'world', *miesto* 'place'), with only 2 lemmas each in the three remaining samples (*byt* 'apartment', *republika* 'republic', *územie* 'area', *prostredie* 'environment', *svet* 'world').

The names of **physical** entities, such as body parts, were almost exclusively found in stories (*ruka* 'hand', *oko* 'eye', *hlava* 'head') but partially in the balanced corpus as well (*ruka* 'hand', *oko* 'eye', *hlava* 'head'). This might indicate that the decisions, acts and studies, despite regulating social interaction in the physical world, deal mostly with abstract procedures and concepts.

**Proper** names were almost absent in the samples, with just 1 in the studies (*Slovensko* 'Slovakia') and 2 in the balanced corpus (*Slovensko* 'Slovakia', *Bratislava)*, which can be probably related both to the anonymization of decisions, the normative universality of acts and the high representation of general concepts in the studies.

**Text**-related noun were found mostly in acts (*odsek* 'article', *slovo* 'word', *písmeno* 'letter') and decisions (*odsek* 'article', *číslo* 'number', *písmeno* 'letter'), which is understandable, given the fact that both types of documents regularly refer to other acts and decisions. In the stories and the balanced corpus, *slovo* 'word' was the most prominent.

Lastly, **time**-related lemmas were found more in decisions (*deň* 'day', *mesiac* 'month', *lehota* 'period') than in the acts (*deň* 'day', *rok* 'year', *lehota* 'period'), which is probably due to the temporal context of any given lawsuit, as compared to the normative, universal character of legislation. The presence of time lemmas is also

prominent in the balanced corpus (*rok* ‘year’, *život* ‘life’, *deň* ‘day’) and in stories, where they also can function as literary motifs (*deň* ‘day’, *rok* ‘year’, *čas* ‘time’).

5.3.3 Finally, the overlapping lemmas, that is, those found in the decisions and in one or more other samples, were identified and their relative frequencies were tested for the significance of differences. This was done by using the Czech corpus calculator CALC. Lemmas significantly different from those in the decisions are marked in bold, those not significantly different in italics.

| OVERLAPPING LEMMAS | CRCS | ACTS | STDS | COLS | BALS |
|---|---|---|---|---|---|
| *vec* ‘matter’, ‘issue’, ‘case’ | 2,257 | | | | **442** |
| *konanie* ‘proceedings’, ‘procedure’ | 2,468 | **1,760** | | | |
| *výkon* ‘fulfilment’, ‘enforcement’ | 1,858 | **1,447** | | | |
| *rozhodnutie* ‘decision’ | 1,477 | *1,607* | | | |
| *súd* ‘court’ | 7,574 | **1,031** | | | |
| *značka* ‘signature’, ‘brand’ | 1,289 | | **872** | | |
| *zákon* ‘act’ | 7,451 | **13,203** | | | **468** |
| *euro* ‘euro’ | 1,602 | | | | **540** |
| *osoba* ‘person’ | 4,359 | *3,898* | | | |
| *republika* ‘republic’ | 1,155 | **2,527** | | | |
| *odsek* ‘article’ | 2,974 | **5,270** | | | |
| *písmeno* ‘letter’ | 1,221 | *1,330* | | | |
| *deň* ‘day’ | 6,608 | **2,431** | | **1,245** | **972** |
| *lehota* ‘period’ | 2,089 | **936** | | | |
| *čas* ‘time’ | 1,547 | | **956** | | **932** |
| *rok* ‘year’ | 1,355 | **1,858** | **2,359** | **1,068** | **3,186** |

**Tab. 3.** Dispersion of overlapping lemmas throughout the corpora

A significant difference was indicated in a total of 20 pairs. The confidence intervals for the lemmas *rozhodnutie* ‘decision’, *osoba* ‘person’, *písmeno* ‘letter’ were too close to each other, despite a significant difference at the alpha level of 0.05. The majority of overlapping lemmas stem from the legal domain, which correlates with the idea of intertextuality between decisions and acts. The lemmas most dispersed throughout the corpora were the time-related lemmas, with ‘year’ found in all and ‘day’ in almost all samples.

## 6 DISCUSSION

As shown in Tab. 2, the decisions seem to contain fewer types of lemmas, when compared to other genres or samples, but the relative frequencies of lemmas in the

decisions are higher. In other words, their noun vocabulary is more limited, but used more intensively, which could corroborate J. Mistrík's claim concerning the lexical poverty of administrative texts. The statistic section indicated, however, that significant differences only exist between specialized texts (decisions, acts, studies) and short stories/mixed texts. The semantic analysis of decisions showed a notable presence of nouns denoting "action", mostly in the legal sense, as well as presence of domain-specific legal terms (*samosudca* 'single judge', *(trestný) rozkaz* '(criminal) order' or *odpor* 'protest' (a remedial measure against an order). Since the lexical differences often result from different thematic parameters (Slančová et al. 2022, pp. 445–446), a more robust analysis should also apply more measures, be it lexical density, lexical diversity or some differential measure for the prominence of one word in several contexts, such as the difference index DIN (Chlumská 2017, pp. 55–56).

## 7 CONCLUSION

From the legal point of view, judicial decisions represent the final stage in the application of law (Prusák 2001, pp. 289–293; Ivor et al. 2017, pp. 165–188). The procedural nature of the stage of decisions within a legal action is then reflected in the texts themselves. Here it materializes as the prominence of action-related nouns, denoting procedures, institutes and measures defined by law. Action-related nouns are to a lesser degree also present in the acts, because these result from a dynamic process – the legislation – as well. The "abstract" lemmas, denoting domain-specific legal terms, are found in the decisions, although they are even more present in the acts and mostly in studies on mass and marketing communication, which focus on a large spectrum of socio-economical phenomena. As complex and legally regulated speech acts, judicial decisions have a prescribed structure which is comparable, but not identical to that of legislative norms (Weinberger 2017, pp. 195–199). This is why some of the person-related lemmas, mostly denoting parts in criminal proceedings, are recurrent. On the other hand, there are several institutional actors involved in the process of legislation, which could be the reason why they appear in the acts more often. Hopefully, this paper has showed that despite its lexical idiosyncrasies – or because of them – the style of judicial decisions deserves further linguistic attention.

## ACKNOWLEDGEMENTS

The paper has been written within the Slovak National Corpus project supported by the Slovak Academy of Sciences, Ministry of Education, Science, Research and Sport of the Slovak Republic, Ministry of Culture of the Slovak Republic and the Ľudovít Štúr Institute of Linguistics, Slovak Academy of Sciences.

## References


Abrahámová, E. (2012). Základy právnej komunikácie. Bratislava: Heuréka, 128 p.

Corpora of the Slovak National Corpus: prim-10.0-public-prf, prim-10.0-public-img-sk, prim-10.0-public-vyv. Accessible at: https://korpus.sk/.

Corpus calculator CALC. Ústav Českého národního korpusu. Accessible at: https://www.korpus.cz/calc.

Corpus of Slovak judicial decisions OD Justice, version 2.0. Accessible at: https://www.juls.savba.sk/justicecorp.html.

Corpus of Slovak legal documents, version 1.9. Accessible at: https://www.juls.savba.sk/legalcorp.html.

Cvrček, F. (2016). Právní informatika a lingvistika. Jurisprudence, 25(6), pages 49–53.

Cvrček, V. et al. (2020). Registry v češtině. Praha: Nakladatelství Lidové noviny, 234 p.

Čulenová, E. (2010). Jazyk a štýl v slovenských zákonoch. Banská Bystrica: Univerzita Mateja Bela, 140 p.

Findra, J. (2004). Štylistika slovenčiny. Bratislava: Osveta, 232 p.

Gibbons, J. (1999). Language and the law. Annual Review of Applied Linguistics, 19, pages 156–173.

Chlumská, L. (2017). Překladová čeština a její charakteristiky. Praha: Nakladatelství Lidové noviny, 150 p.

Imrichová, M., and Turočeková, M. (2015). Lingvistická analýza právnych textov. Prešov: Filozofická fakulta Prešovskej univerzity v Prešove, 107 p.

Ivor, J., Polák, P., and Záhora, J. Trestné právo procesné II. Bratislava: Wolters Kluwer SK, 492 p.

Kováříková, D. (2017). Kvantitativní charakteristiky termínů. Praha: Nakladatelství Lidové noviny, 136 p.

Mistrík, J. (1975). Žánre vecnej literatúry. Bratislava: Slovenské pedagogické nakladateľstvo, 212 p.

Mistrík, J. (1997). Štylistika. 3. vydanie. Bratislava: Slovenské pedagogické nakladateľstvo, 598 p.

Prusák, J. (2001). Teória práva. Bratislava: Právnická fakulta Univerzity Komenského v Bratislave, 340 p.

Ružička, J. et al. (1966). Morfológia slovenského jazyka. Bratislava: Vydavateľstvo Slovenskej akadémie vied, 895 p.

Slančová, D. et al. (2022). Úvod do štúdia interaktívnej štylistiky I, II. Prešov: Prešovská univerzita v Prešove, 481 + 449 p. Accessible at: https://www.pulib.sk/web/kniznica/elpub/dokument/Slancova6.

Statistics calculator STATISTICS ONLINE. Accessible at: https://www.statskingdom.com.

Štícha, F. (1985). O jazyce soudních rozhodnutí. Naše řeč, 68(2), pages 68–77.

Vogel, F. (ed.) (2019). Legal linguistics beyond borders. Language and law in a world of media, globalisation and social conflicts. Berlin: Duncker and Humblot, 384 p.

Weinberger, O. (2017). Norma a instituce. Úvod do teorie práva. Plzeň: Aleš Čeněk, 248 p.