Lexical Diversity of Czech L2 Texts at Different Proficiency Levels

Michaela Hanušková[1*], Miroslav Kubát[1], Michaela Nogolová[1]

[1] University of Ostrava
[*] Corresponding author's email: mi.hanuskova@gmail.com

DOI: https://doi.org/10.53482/2025_59_427

ABSTRACT

The study valuates how lexical diversity differs across language proficiency levels (A1–C1 according to the CEFR). The material used in the research comes from the CzeSL-SGT learner corpus belonging to the Czech National Corpus. This dataset contains more than 8,000 Czech texts written by non-native speakers of different proficiency levels. Moving Average Type-Token Ratio (MATTR) is used to calculate lexical diversity in this study. The results indicate that lexical diversity increases with writers' proficiency. There is also a significant difference between the development of lexical diversity of Slavic and non-Slavic native speakers.

Keywords: lexical diversity, MATTR, corpus linguistics, second language acquisition

1 Introduction

Lexical diversity (LD) is a characteristic that reflects the extent of the lexical knowledge of the writer speaker. LD refers to the variety of unique words (types) used in a spoken or written text. It assumes that every person has their own vocabulary reflected in their language production. (Kubát, 2016) A specific situation occurs with foreign language learners. Their vocabulary is gradually evolving, making lexical diversity a useful tool for better understanding the development of language acquisition. Simply put, a less proficient speaker tends to use a small number of lexical units and cannot achieve significant variability. In contrast, a more proficient speaker uses more variable lexical items to accomplish a task (Webb, 2019). Hence, LD is one of the most reliable indicators of linguistic proficiency and development in second language acquisition (SLA) (cf. Nasseri & Thompson, 2021).

This study aims to analyse development of lexical diversity of texts written by non-native speakers of the Czech language across different levels of language proficiency, from beginners (A1) to advanced learners (C1)[1]. We are motivated by the fact that the lexical diversity of Czech as L2 (and other Slavic languages) has not been studied quantitatively yet.

[1] Language proficiency levels in this study refer to the Common European Framework of Reference for Languages (CEFR).

Slavic languages still share a part of the lexicon, which comes from both a common protolanguage and close contact between different Slavic languages speakers (Karlíková et al., 2017). Therefore, we will also pay attention to the cross-linguistic influence of native Slavic and non-Slavic languages on the development of lexical diversity in Czech as L2. Compared to non-Slavic native speakers, Slavic native speakers are expected to use a considerably wider vocabulary in their texts. Although this aspect might play a role in SLA research, only a few studies cover cross-linguistic influence in lexical diversity research (cf. Shatz, 2021).

Lexical diversity will be measured by the Moving Average Type-Token Ratio (MATTR) (Covington & McFall, 2010). Although this indicator has been shown to be a suitable lexical diversity indicator in various fields of linguistics (especially stylometry), its application in SLA research is still relatively rare. However, recent studies have shown that MATTR is a suitable measure in SLA research. The strength of MATTR in quantitative text analysis lies in its independence from text size as well as its simplicity and straightforward interpretation.

The analysis is based on material from Czech National Corpus, namely the corpus CzeSL-SGT (Czech as a Second Language with Spelling, Grammar and Tags) (Šebesta et al., 2014) consisting of more twihan 8,000 texts ten by about 2,000 different authors with 54 different first languages. Furthermore, the corpus covers a wide range of language proficiency levels, from beginners to advanced learners. Thus, this material can be considered a substantial corpus for SLA research.

The study aims to answer two research questions. First, how do lexical diversity values develop across Czech L2 proficiency levels? The second question is whether speakers with Slavic L1 backgrounds differ from speakers with non-Slavic L1 backgrounds regarding the evolution of their lexical diversity values. If so, what are the differences?

2 Material

The CzeSL-SGT (Czech as a Second Language with Spelling, Grammar and Tags) corpus of non-native speakers of Czech with automatic annotation (Šebesta et al., 2014) is used in this study as the language material. It is a part of the Czech National Corpus. The raw CzeSL-SGT corpus consists of 8,617 texts written by 1,965 authors with 54 different first languages and was collected from 2009 to 2013. Each of the essays was equipped with metadata about a text (e.g. topic, size limit in the assignment, text length) and student (e.g. sex, age, L1, language proficiency level) (cf. Rosen, 2015). Information on language proficiency level and mother tongue was used for this study. The number of texts in each proficiency level can be found in Table 1. A detailed description of the corpus can be found on the Czech National Corpus website.

Table 1: Number of texts by proficiency level.

Proficiency Level	Number of Texts
A1	2609
A1+	315
A2	2098
A2+	570
B1	1481
B2	745
C1	123
C2	1
total	7942
unknown	675

As can be seen in Table 1, the numbers of texts at each proficiency level are unbalanced, and some texts are not even assigned to any proficiency level. Therefore, the following changes were made prior to the analysis.

- Since there is only one text at the C2 level, this level was excluded from the analysis.
- Texts with 'unknown' proficiency levels were also removed.
- Texts labelled A1+ and A2+ were excluded from the study because (a) the corpus documentation does not state on which parameters these levels are determined, and (b) these additional levels do not correspond to the CEFR framework.
- Based on Zenker and Kyle’s (2021) findings, texts shorter than 55 words were removed.

In total, 6,073 texts covering levels A1, A2, B1, B2, and C1 were analysed in this research. Since we also focus on a potential cross-linguistic mother tongue influence, texts were also divided into the Slavic or non-Slavic groups. The final adjusted structure of the corpus used in this study can be found in Table 2.

Table 2: Number of analysed texts by proficiency level.

Proficiency Level	Number of Texts		
	Slavic	non-Slavic	Mix
A1	1466	556	2022
A2	1215	625	1840
B1	879	492	1371
B2	511	211	722
C1	80	38	118
total	4151	1922	6073

3 Methodology

3.1 Lexical Diversity

Lexical diversity can be examined with multiple indices. Indices usually express the relationship between the number of different words (types) and the number of all words (tokens) in a text. It is well known that fundamental indices like type-token ratio (TTR Johnson, 1994) and its variations Root TTR (Guiard, 1960), Log TTR (Chotlos, 1944; Herdan, 1960) are sensitive to text length. The longer the text, the lower the LD score (cf. Čech et al., 2014). In an attempt to address this issue, several revised indices have been proposed, such as Maas' index (Maas, 1972), Moving-average TTR (MATTR; Covington & McFall, 2010), the hypergeometric distribution density index (HD-D; McCarthy & Jarvis, 2007), standardized type-token ratio (zTTR) based on comparing the observed TTR with the referential TTR values representing texts of identical size (Cvrček & Chlumská, 2015), or the measure of textual lexical diversity (MTLD; McCarthy & Jarvis, 2010). However, none of them has gained universal acceptance.

Previous research indicated that measures like MATTR, HD-D, or MTLD are more stable than TTR on large texts that are similar in length (McCarthy & Jarvis, 2007; McCarthy & Jarvis, 2010) but do not mention the stability in short texts. A severe concern in SLA research is the sample length that varies widely across proficiency levels, and particularly problematic are texts at the lowest CEFR levels. Their length is usually less than 100 words.

Following previous studies (Koizumi, 2012; Koizumi & In'nami, 2012), Zenker & Kyle (2021) tested the stability of the latter indices on short texts. The study demonstrates the negligible correlation between LD value and text length for MATTR, HD-D or MTLD indices. They suggest that these are suitable for working with samples as short as 50 words. They claim that, in particular, MATTR appears to be the most stable of all indices. Zenker & Kyle (2021) also focused on the relationship between LD values and proficiency levels. The analysis of MATTR values confirmed statistically significant increases through the proficiency levels. Based on these findings, we decided to use the MATTR index in our research.

In addition to its statistical robustness on short texts, MATTR has proven useful across a range of empirical contexts. It has been applied to track longitudinal lexical development in L2 learning (Lissón & Ballier, 2018), to compare lexical proficiency in academic writing across L1 and L2 writers (Nasseri & Thompson, 2021), and to document gradual gains in ESL proficiency during university study (Vidal & Jarvis). Other research has shown that MATTR can differentiate learners across proficiency bands in diverse instructional settings, including Moroccan EFL (Ait Hammou, Larouz, & Fagroud, 2021). Taken together, these applications indicate that MATTR is sensitive both to developmental change and to proficiency-related differences while remaining stable with relatively short samples. This converging evidence supports our decision to employ MATTR in the present study.

3.2 Moving Average Type-Token Ratio (MATTR)

MATTR is defined as the mean of the TTR values of overlapping subtexts (the so-called windows) of the same length (L) in a text. The formula of TTR is defined as follows:

$$TTR = \frac{V}{N}$$

Where *V* is vocabulary (number of types) and *N* is text size (number of tokens).

The calculation procedure of MATTR is as follows:

1. A text is split into windows with an arbitrarily chosen size L.
2. The window moves forward one token at a time.
3. The TTR is calculated for every single window in the text.
4. Finally, the MATTR is calculated as an arithmetic mean of all the TTR values.

Let us demonstrate MATTR computation on a simple example of a sequence of 7 characters: a, b, c, d, a, a, b. Text length *N* = 7, vocabulary V = *4*. If we choose a window size of 4 tokens (*L* = 4), we obtain 4 overlapping windows:

1. a, b, c, d (TTR = 4 / 4 = 1)
2. b, c, d, a (TTR = 4 / 4 = 1)
3. c, d, a, a (TTR = 3 / 4 = 0.75)
4. d, a, a, b (TTR = 3 / 4 = 0.75)

The resulting MATTR value is calculated as the mean of the four obtained TTR values: MATTR = (1 + 1 + 0.75 + 0.75) / 4 = 0.875.

The important setting of the MATTR measurement is the window size (*L*). There is no ideal value suitable for every research. The length is usually set according to the shortest text in the corpus. The window size obviously cannot be longer than the shortest analysed text. On the other hand, the window length should be long enough to cover a sufficient text sample. Language ability, such as lexical diversity, can be barely expressed in a sequence of 5 or 10 words. It is necessary to measure such a characteristic on a longer sample. The window size can therefore vary significantly in different studies based on the analysed material. For example, stylometric analysis of novels can work with samples of hundreds of words, while investigating newspaper articles requires a window size of about 50 to 100 words. Consequently, the choice of L balances comparability and representativeness. Given that we analyse rather short texts (especially at beginner proficiency levels), we set the window size to 50 tokens, which can be considered a sufficient value for detecting lexical diversity in L2 texts.

Czech has a rich morphology in which nouns, adjectives, pronouns, numerals, and verbs are inflected to modify grammatical functions. For example, a lemma 'pes' (a dog) consists of several different word forms based on seven grammatical cases indicating their function in a sentence and two numbers (singular and plural) (see Table 3). Therefore, a lemma is a basic unit for calculating lexical diversity in this research.

Table 3: Grammatical paradigm of Lemma 'pes' (a dog).

Case / Number	singular	plural
nominative	pes	psi, psové
genitive	psa	psů
dative	psovi, psu	psům
accusative	psa	psy
vocative	pse	psi, psové
locative	psovi, psu	psech
instrumental	psem	psy

Software MATTR developed by Covington & McFall (2010) was used for the computation of lexical diversity in this research.

4 Results

In this chapter, the resulting MATTR values are presented as follows. First, the general results of different levels of language proficiency, regardless of the native language, are visualized by the graphs in Figures 1 and 2. Then the differences between the Slavic and non-Slavic groups are shown in Figures 3–5.

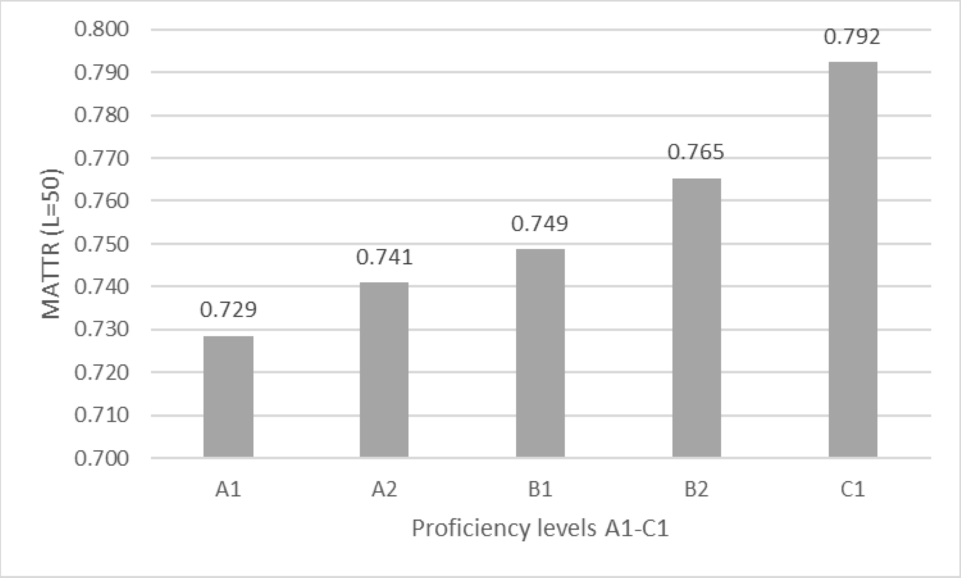

Figure 1: Average MATTR values at proficiency levels A1–C1 regardless of the native language.

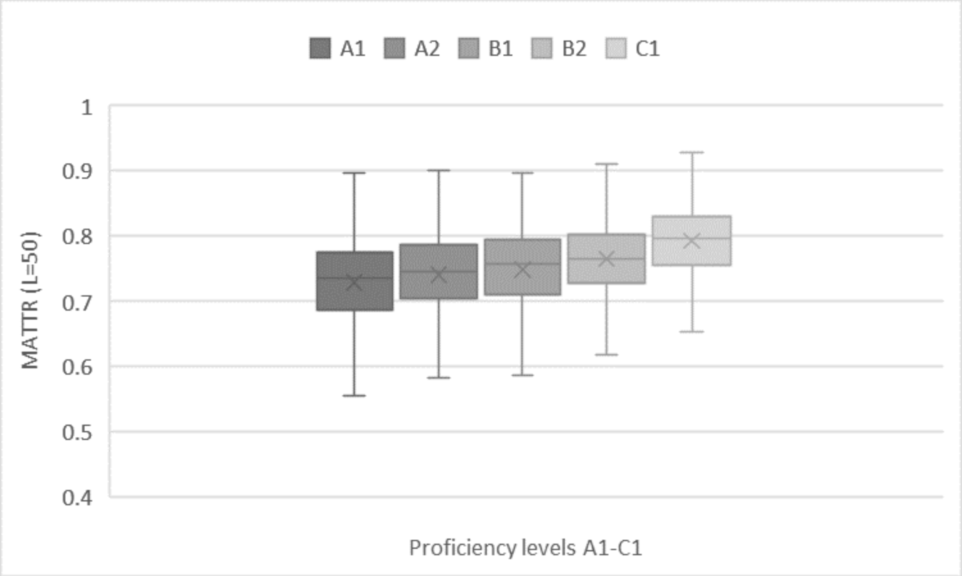

Figure 2: MATTR values at proficiency levels A1–C1 regardless of native language (boxplot without outliers).

The average MATTR values in Figure 1 show an evident increasing trend for lexical diversity in all levels of language proficiency analysed (A1–C1). Since the arithmetic means could be misleading, the dispersion of the resulting values obtained is visualized by a boxplot in Figure 2, where the tendency is also evident. All pairs of levels were statistically tested to ensure that the differences between the levels are significant. Since the obtained data are not normally distributed, we decided to apply the Wilcoxon-Mann-Whitney test, which is generally considered a non-parametric alternative to the t-test. The results show that all differences between all levels of language proficiency are statistically significant (p-value ≤ 0.05). Therefore, we can conclude that lexical diversity increases significantly with each proficiency level.

Our findings agree with previous research focused on the development of lexical diversity in other languages. Similar results can be found in Shatz (2021), who analysed lexical diversity on a large material of several thousand English texts across different CEFR levels (A1–B2).

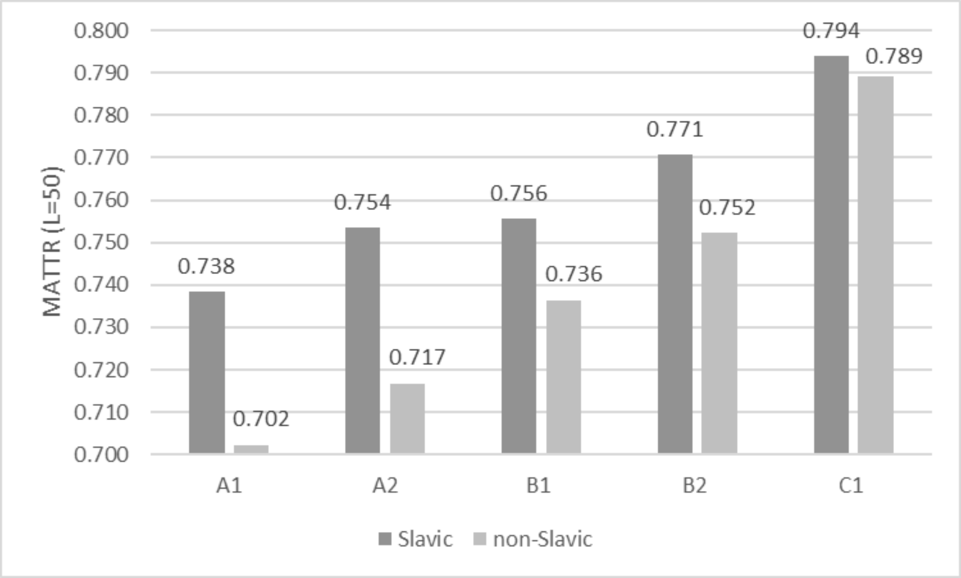

Figure 3: Comparison of average MATTR values at proficiency levels A1–C1 between Slavic and non-Slavic learners.

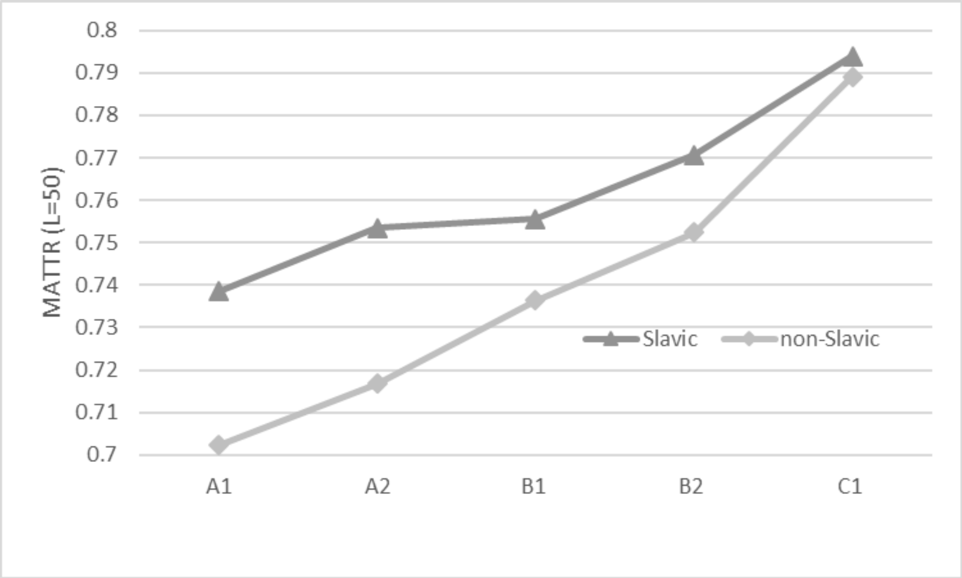

Figure 4: Comparison of average MATTR values at proficiency levels A1–C1 between Slavic and non-Slavic learners.

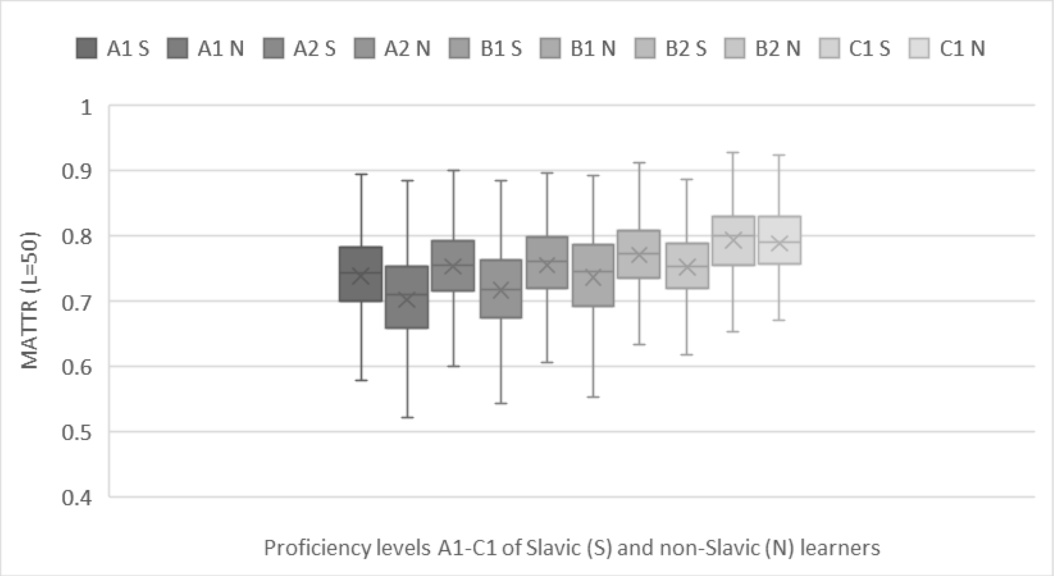

Figure 5: Comparison of MATTR values at proficiency levels A1–C1 between Slavic and non-Slavic learners (boxplot without outliers).

As can be seen in Figures 3–5, the resulting MATTR values of texts written by learners with Slavic and non-Slavic L1 show three general findings. First, the increasing tendency of lexical diversity is present in both groups (Slavic and non-Slavic). Second, students with a Slavic mother tongue reach higher average MATTR values at all levels of language proficiency. Third, the higher the level of language proficiency, the smaller the gap between Slavic and non-Slavic speakers. This shrinking gap is most visible in Figure 4, where the difference is minimal at the advanced level C1. To verify our findings statistically, we applied the Wilcoxon-Mann-Whitney test to test differences between Slavic and non-Slavic groups. The results show that the difference is always statistically significant (p-value ≤ 0.05) except for the C1 level. The statistical test, therefore, confirms the preliminary conclusions based on the graphs.

Our findings confirm our expectations that Slavic native speakers use a considerably wider vocabulary in their texts than their non-Slavic counterparts. The influence of overlapping vocabulary of one language family (Slavic languages) seems to be very strong and intuitive. Interestingly, Shatz (2021) concludes in his research on English as a second language that lexical similarity between the L1 and the L2 does not influence L2 lexical diversity, regardless of learners’ L2 proficiency. His findings are based on the lexical distance between languages measured by similarities using Swadesh lists (Swadesh, 1971), which suggests that this type of calculation may have limitations that can bias the results. At the same time, the overall influence of cross-linguistic factors on lexical diversity has been examined only to a limited extent. While our results suggest an effect related to typological proximity, further studies across different languages and contexts are necessary to assess its generalisability.

5 Conclusion

Based on the obtained data, we can answer the research questions stated in the Introduction of the study. We discovered that lexical diversity increases significantly with the writer's proficiency of Czech L2 across the whole scale of analysed levels (A1–C1). The increasing tendency was also confirmed by the statistical test, where all differences between individual levels were statistically significant. We can conclude that lexical diversity is a crucial feature in learning a second language.

Besides the overall tendency of lexical diversity development across different proficiency levels, we also focused on the differences between Slavic and non-Slavic native speakers. The results show a significant difference between the development of the lexical diversity of the two groups. According to the expectation, Slavic native speakers reached significantly higher MATTR values at all proficiency levels except for advanced level C1. We can also state that the higher the level of language proficiency, the smaller the difference between Slavic and non-Slavic speakers. The cross-linguistic influence is, therefore, most visible at beginner levels.

We can also conclude that MATTR is a suitable method for measuring the lexical diversity across CEFR levels, given the strong association between L2 proficiency level and lexical diversity found in this study and previous research (e.g. Zenker & Kyle, 2021). MATTR seems to be a reliable tool for measuring lexical diversity of texts with various text lengths, which is essential, especially in the case of very short texts typical for beginner L2 learners.

Acknowledgements

The research was supported by University of Ostrava, project No. SGS06/FF/2022.

References

Ait Hammou, B., Larouz, M., Fagroud, M. (2021). Word frequency, Range and Lexical diversity: Picking Out Changes in Lexical Proficiency among University Learners in an EFL Context. *International Journal of Linguistics and Translation Studies*, 2(2), pp. 22–38. https://doi.org/10.36892/ijlts.v2i2.131

Council of Europe. Council for Cultural Co-operation. Education Committee. Modern Languages Division. (2001). *Common European Framework of Reference for Languages Learning, teaching, assessment.* Cambridge University Press.

Čech, R., Popescu, I.-I., Altmann, G. (2014). *Metody kvantitativní analýzy (nejen) básnických textů*. Olomouc: Univerzita Palackého v Olomouci.

Cvrček, V., Chlumská, L. (2015). Simplification in translated Czech: a new approach to type-token ratio. *Russian Linguist 39*, pp. 309–325. https://doi.org/10.1007/s11185-015-9151-8

Karlíková, H., Skalka, B. & Večerka, R. (2017). SLOVANSKÉ JAZYKY. In: Karlík, P, Nekula, M, Pleskalová, J. (Eds.). *CzechEncy - Nový encyklopedický slovník češtiny*. URL: https://www.czechency.org/slovnik/SLOVANSKÉ JAZYKY

Kubát, Miroslav. *Kvantitativní analýza žánrů*. Ostrava: Ostravská univerzita, 2016.

Nasseri, M., Thompson, P. (2021). *Lexical Density and Diversity in Dissertation Abstracts: Revisiting English L1 vs. L2 text differences*. Assessing Writing, 47, 100511. https://doi.org/10.1016/j.asw.2020.100511

Rackevičienė, S., Utka, A., Bielinskienė, A., Rokas, A. (2022). Distribution of Terms across Genres in the Annotated Lithuanian Cybersecurity Corpus. *Respectus Philologicus*, 41(46), 26–42. http://dx.doi.org/10.15388/RESPECTUS.2022.41.46.105

Rosen, A. (2015). *CzeSL-SGT – a corpus of non-native speakers' Czech with automatic annotation*. https://doi.org/10.13140/RG.2.1.1906.2487

Šebesta, K. et al. (2014). *AKCES 5 (CzeSL-SGT) Release 2*, LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL) http://hdl.handle.net/11234/1-162.

Shatz, I. (2022). *The Potential Influence of Crosslinguistic Similarity on Lexical Transfer: Examining Vocabulary Use in L2 English* (Doctoral dissertation, University of Cambridge).

Zenker, F., Kyle, K. (2021). Investigating minimum text lengths for lexical diversity indices. *Assessing Writing*, 47, 100505.

Lissón, P., Ballier, N. (2018). Investigating Lexical Progression through Lexical Diversity Metrics in a Corpus of French L3. *Discours. Revue de linguistique, psycholinguistique et informatique. A Journal of Linguistics, Psycholinguistics and Computational Linguistics,* 23. https://doi.org/10.4000/discours.9950

Swadesh, M. (1971). *The Origin and Diversification of Language*. Ed. post mortem by Joel Sherzer. Chicago: Aldine. Contains final 100-word list on p. 283.

Treffers-Daller, J. (2013). Measuring lexical diversity among L2 learners of French: an exploration of the validity of D, MTLD and HD-D as measures of language ability. In: Jarvis, S., Daller, M. (Eds.), *Vocabulary Knowledge: Human Ratings and Automated Measures*, pp. 79–103. John Benjamins Publishing Company.

Vidal, K., Jarvis, S. (2020). Effects of English-medium instruction on Spanish students' proficiency and lexical diversity in English. *Language Teaching Research*, 24(5), 568–587. https://doi.org/10.1177/1362168818817945

Webb, S. (Ed.). (2020). *The Routledge Handbook of Vocabulary Studies (Vol. 2)*. London: Routledge.