The Development of Sentence and Clause Lengths in Czech L2 Texts

Michaela Nogolová https://orcid.org/0000-0001-7604-9765
Radek Čech https://orcid.org/0000-0002-4412-4588
Michaela Hanušková https://orcid.org/0000-0002-1709-0036
Miroslav Kubát https://orcid.org/0000-0002-3398-3125
Filozofická fakulta Ostravské univerzity, Ostrava, Česká republika
nogolovam@gmail.com; cechradek@gmail.com;
mi.hanuskova@gmail.com; miroslav.kubat@gmail.com

ABSTRACT: This study aims to analyse the development of sentence and clause lengths in Czech L2 texts at A1–C1 proficiency levels. The research is based on the material of the Czech National Corpus, namely the CzeSL-SGT learner corpus. The final sample contains nearly 6,000 Czech texts written by non-native speakers. Moreover, these results are compared with the reference corpus consisting of texts written by Czech native speakers. The material is syntactically annotated by the Universal Dependencies framework. The research also examines the cross-linguistic influence of learners' first language by comparing the results of native Slavic and non-Slavic speakers. The results generally show that the higher the level of language proficiency, the higher the average length of the clauses and sentences. Texts written by Slavic students reach, on average, longer sentences and clauses compared to their non-Slavic counterparts at most proficiency levels, with the gap most visible at beginner levels (A1, A2).

KEYWORDS: sentence length; clause length; Czech language; Slavic language

Průměrná délka věty a klauze v textech psaných nerodilými mluvčími češtiny

ABSTRAKT: Cílem této studie je analyzovat vývoj délky vět a klauzí textů psaných nerodilými mluvčími češtiny. Jazykový materiál je tvořen texty korpusu CzeSL-SGT, který je součástí Českého národního korpusu. Konkrétně je zde pracováno téměř s 6000 texty na jazykových úrovních A1–C1. Výsledky jsou také porovnávány s referenčním korpusem. Ten je tvořen texty psanými rodilými mluvčími češtiny. Materiál je syntakticky anotován pomocí nástroje UDPipe 2.0. Součástí analýzy je také porovnání výsledků slovanských a neslovanských mluvčích. Výsledky obecně ukazují, že s rostoucí jazykovou úrovní roste také průměrná délka věty a klauze. Texty psané slovanskými mluvčími mají v průměru delší věty i klauze téměř na všech jazykových úrovních. Největší rozdíl se však projevuje především na úrovních A1 a A2.

KLÍČOVÁ SLOVA: délka věty; délka klauze; český jazyk; slovanský jazyk

1. Introduction

Learning how to combine words to form a sentence is essential to learning any language. Numerous studies have demonstrated that as learners advance, they tend

to employ more intricate and accurate syntactic structures in their language production (Crossley – McNamara 2014; Yang et al. 2015). Hence syntax has been an essential issue in second language acquisition (SLA) research.

The term *syntactic complexity* has come to encompass research in syntax in the context of SLA over the past three decades (Lu 2017; Yoon et al. 2020; Park 2022; Khushik – Huhta 2022; Lu 2011; Kyle et al. 2021). Wolfe Quintero et al. (1998) extensively elaborated on indices used to measure syntactic complexity, such as mean lengths of sentences, clauses, and T-units, which can be measured in different units such as words or clauses. However, it is important to note that clause length alone does not always indicate syntactic intricacy. For example, coordination allows for the creation of long clauses through the application of a single rule, without necessarily introducing syntactically complex structures. In recent years, scholars have aimed to reassess existing syntactic indices and explore alternative measures to analyse SLA syntactic development, including mean dependency distances, noun or verb phrase modification (Biber et al. 2020; De Clercq – Housen 2017; Jiang – Ouyang 2018; Ouyang et al. 2022).

While SLA research has predominantly concentrated on syntactic structures in English texts (Lu – Ai 2015), there has been a recent exploration of other languages, such as Dutch (Kuiken – Vedder 2019) and German (Vyatkina et al. 2015). Slavic languages, however, have received comparatively little interest (Kisselev – Alsufieva 2017; Kisselev et al. 2021 for Russian; Trtanj – Čolić 2019 for Croatian FLA). Hence, this study aims to contribute to syntactic SLA research by investigating the Czech language.

Our analysis is based on 5,905 texts from the CzeSL-SGT learner corpus (Šebesta et al. 2014), a part of the Czech National Corpus. Our study focuses on fundamental measures, namely sentence and clause length, as indicators of learners' syntactic abilities. This article explores the relationship between clause/sentence length and syntactic dependency structure, shedding light on the complexity of constructing longer clauses in the Czech language. We examine how sentence and clause length evolve across different levels of language proficiency while also considering cross-linguistic influences by comparing Slavic and non-Slavic native speakers. We anticipate that learners at higher proficiency levels will employ longer sentences and clauses, and Slavic L1 learners are expected to exhibit higher values in the observed measures.

2. The relationship between clause/sentence length and syntactic dependency structure

Creating a well-formed clause or sentence in any language involves understanding and applying the syntactic rules specific to that language. These rules dictate various properties of a clause, such as word order or the grammatical forms of certain words. Moreover, they also determine some necessary conditions for the formation of grammatical clauses (e.g., verb valency) and put some limits on combining

words. As the length of a clause increases, the number of rules that are applied usually grows, too. A brief comparison of Czech clauses (1) and (2) and their structures (Figures 1 and 2), allows us to illustrate that adding three more words means the application of further syntactic and morphological rules.

(1) Petr vidí knihu.
"Peter sees the book."

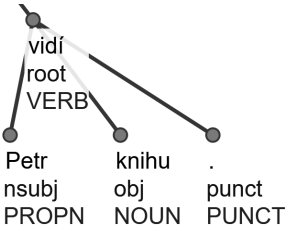

Figure 1: The syntactic tree of the sentence (1)

(2) Petr z Prahy vidí moji knihu.
"Peter from Prague sees my book."

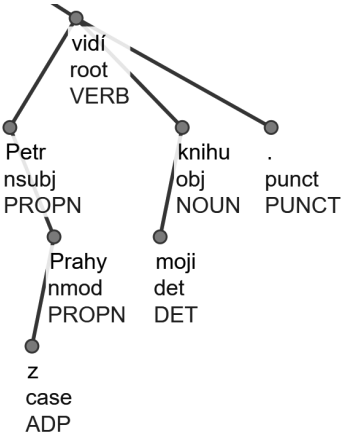

Figure 2: The syntactic tree of the sentence (2)

Specifically, the proper use of the adnominal prepositional phrase *z Prahy* "from Prague" dependent on the subject *Petr* is based on the speaker's knowledge. The phrase can occur only behind the noun, and it must have a genitive form which is expressed by a particular word form. Similarly, the proper use of *moji* "my" dependent on *knihu*

"book" reflects the knowledge of word order, the character of the syntactic relationship between the noun and dependent word (i.e., agreement in gender, case, and grammatical number), and the word form of the pronoun. However, there are some cases where the length of a clause does not correspond with its complexity. Particularly, the usage of coordination allows one to create a long clause by applying a single rule that multiplies, see clause (3) and Figure 3.

(3) Petr vidí knihu, časopis, noviny, obraz, vázu, sklenici, talíř a lžíci.
"Peter sees a book, journal, newspaper, painting, vase, glass, plate, and spoon."

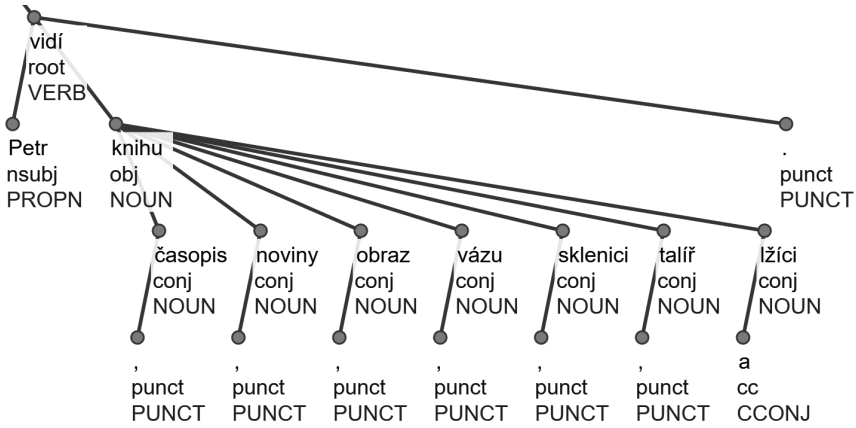

Figure 3: The syntactic tree of the sentence (3)

Sentence (3) is twice as long as sentence (2) but creating such a long sentence only requires knowledge of the rules used in clause (1) plus the rule of coordination.

To sum up, the length of the clause (or sentence) usually somewhat corresponds with its complexity and this property (i.e., length) can be used for an evaluation of the syntactic knowledge of L2 speakers. The advantage of using length for this kind of evaluation lies in its simplicity, in comparison to the other methods (Wolfe Quintero et al. 1998). On the other hand, it has some limits which must be considered (for details, see Section 3).

3. Language Material and Methodology

The language material consists of selected texts from the CzeSL-SGT learner corpus (Šebesta et al. 2014). The corpus contains 8,617 texts written by 1,965 non-native Czech speakers at A1–C2 levels of language proficiency according to the Common European Framework of Reference for Languages (CEFR). The corpus contains detailed metadata of both authors and texts. In this study, information about the level

of a learner's language proficiency, first language (L1), and text length is used. Texts with unclear or unknown proficiency levels were excluded from the sample, as well as the C2 level because only one text is assigned to this category. Texts shorter than 55 words and texts with fewer than five sentences were also removed from the sample. We chose those criteria because the usual length limit for passing the written exam is around 50–60 words, and because texts with fewer than five sentences do not provide sufficient data for reliable analysis. The information about L1 was used for a determination of two groups of learners: 1) Slavic and 2) non-Slavic. To sum up, our corpus consists of 5,905 written texts that cover five CEFR proficiency levels. Since our aim is also to observe the differences between native and non-native speakers, we compare the results with the reference corpus (REF-CZ) consisting of texts written by Czech native speakers. These data come from the corpus SKRIPT2012 (Šebesta et al. 2013) containing texts written by fourth-grade high school students. For the sample details, see Table 1.

Table 1: Number of texts in each group of the analysed sample

level	number of texts	number of texts from learners with Slavic L1	number of texts from learners with non-Slavic L1
A1	1983	1447	536
A2	1778	1177	601
B1	1335	853	482
B2	697	499	198
C1	111	78	33
REF-CZ	87	–	–

All texts were processed with UDPipe 2.0 (Straka 2018). This tool is used for parsing and morphological tagging, which is necessary for the determination of clauses and their lengths. The clause is defined as a syntactic structure that contains a finite verb. For illustration, both the parsing and the determination of clauses in the sentence (4) are presented in Figure 4.

(4) Přišla jsem domů po dlouhém dni, který jsem musela strávit v práci, a šla jsem běhat.
"I came home after a long day that I had to spend at work and went for a run."

The sentence (4) has 3 clauses because there are 3 finite verbs. Every clause consists of a head (represented by a finite verb) and its directly or indirectly dependent words (if any). The boundaries of specific clauses are delimited by the presence of the syntactic relationship with the following clause. Specifically, there is a syntactic relationship between the words *dni* "day" and *musela* "had to" in the tree. However, *musela* "had to" is a predicate of the next clause, therefore, the line between them represents the border between the two clauses. Analogously, we can see the same situation between the words *přišla* "came" and *šla* "went".

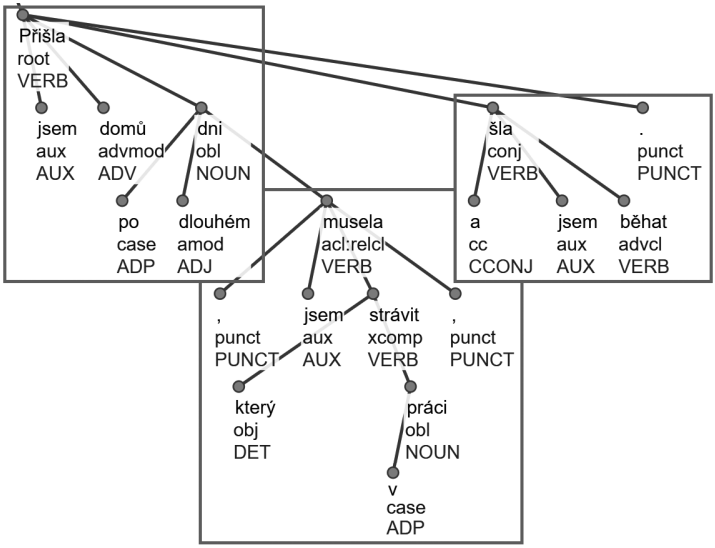

Figure 4: The syntactic tree of the sentence (4) parsed by the UDPipe 2.0. The rectangles indicate specific clauses

In this study, we use three indices to measure the syntactic development of L2 learners:

1) average sentence length measured in words (ASL),
2) average clause length measured in words (ACL),
3) average sentence length measured in clauses (CS).

We processed all texts at every chosen language level separately. Next, both the mean and standard deviation were calculated for each level. Then, the differences between each pair of proficiency levels were statistically tested. For testing, we used the Mann-Whitney test (at the significance level $\alpha = 0.05$), because the data are not normally distributed[1]. Analogously, we tested the difference between groups of Slavic and non-Slavic learners at each proficiency level.

The relationship between all indices was examined using the Kendall rank correlation coefficient. When evaluating the correlation, we determined the degree of dependence, and we tested its statistical significance ($\alpha = 0.05$).

In Section 2, we mentioned the limits of the sentence and clause average lengths' usage for analysis of the SLA development. Therefore, we examine the proportion of

[1] R software (R Core Team 2013) was used for computation.

coordination at each level. Significant differences in these proportions among certain levels would partially devalue the results. For instance, if we would find a higher proportion of coordinated nodes in syntactic trees at A2 level in comparison to A1, the potential higher average length of the clause in A2 group would not have been caused by the increasing syntactic knowledge of A2 (for reasoning, see Section 2).

We calculated the coordination rate as the proportion of coordinated words relative to the total number of words, i.e.,

$$1.\ ACN = \frac{NCorr}{NW}$$

where NCorr is the number of coordinated words[2] and NW is total number of words in the sample.

4. Results

4.1 Proportion of coordination

Since significant differences between proportions of coordination are important for interpretation of the results, we start with this metric. The results presented in Table 2 show that there are very few variations in coordination's usage at all levels of language proficiency. The differences are in the range of tenth percent. It means that the use of coordination has a minimal impact on the average length of clauses and sentences.

Table 2: The proportion of coordinated words (ACN) across the levels in the sample

level	ACN
A1	4.65 %
A2	4.16 %
B1	4.43 %
B2	4.68 %
C1	4.30 %
REF-CZ	3.73 %

4.2 Differences of ASL, CS and ACL across different language proficiency levels

In this part, our objective is to observe how average sentence and clause length develop across language proficiency levels. According to our expectation, there should be an increasing trend that reflects the growth of the learners' syntactic abilities. The results presented in Table 3 and Figures 5–7 show that there is an ascending trend through almost all indices. The only deviation from this trend occurs in the case of

[2] Coordinated words that are predicates are not counted because they represent the root of the other main clause in the sentence.

CS between levels B2 and C1. REF-CZ has the highest values for all indices, validating the assumption that native speakers should have the most developed syntactic skills. Standard deviation (SD) values show rather consistent variability of the results.

Table 3: The average sentence (ASL, CS) and clause lengths (ACL) and their standard deviations (SD)

	ASL		CS		ACL	
level	**mean**	**SD**	**mean**	**SD**	**mean**	**SD**
A1	8.570	3.116	1.542	0.441	5.537	1.170
A2	9.266	3.179	1.656	0.470	5.599	1.096
B1	10.391	3.724	1.803	0.587	5.772	1.095
B2	11.409	3.348	1.917	0.507	5.999	1.154
C1	11.675	3.240	1.863	0.387	6.298	1.361
REF-CZ	13.525	4.165	2.129	0.468	6.318	1.180

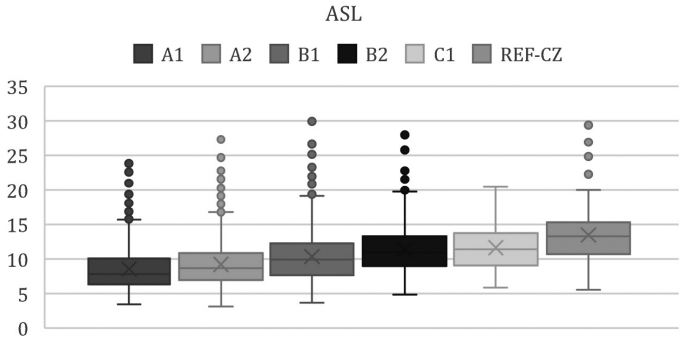

Figure 5: ASL values from texts at A1–REF-CZ

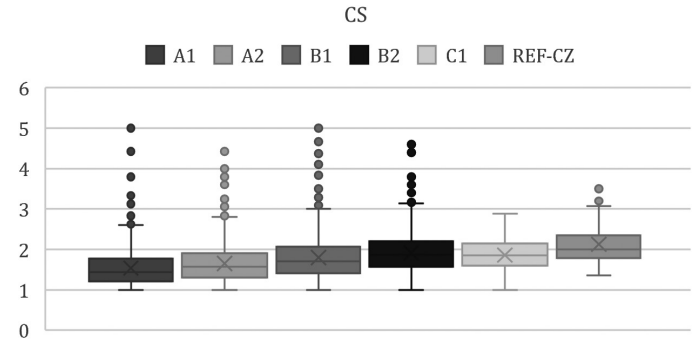

Figure 6: CS values from texts at A1–REF-CZ

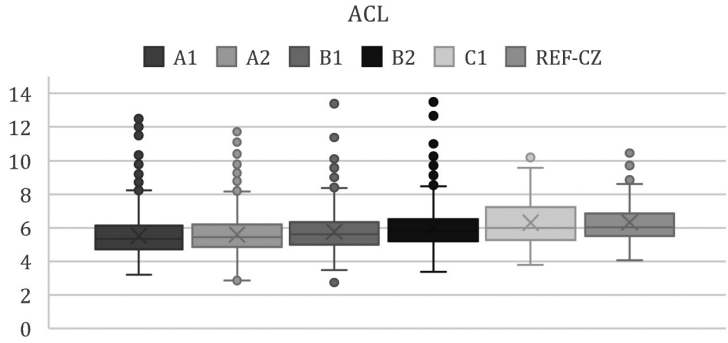

Figure 7: ACL values from texts at A1–REF-CZ

To illustrate the results of ASL and ACL, we choose one sentence[3] (5–9) at each level of language proficiency (A1–C1) from analysed sample.

(5) Myslím že to byli nejzajímavější prázdniny.
"I think it was the most interesting holidays."

(6) Mám hodně kamarádů a proto mám veselé prázdniny.
"I have got many friends therefore I have got joyful holidays."

(7) Líbily se mi ty prázdniny moc a nikdy na ně nezapomenu.
"I liked the holidays a lot and I will never forget it."

(8) Ale přibližně budeme mít dva týdny vánoční prázdniny, které začíná kolem 20.12.
"But we will have approximately two weeks of Christmas holidays that start around 20th December."

(9) Loní jsem své letní prázdniny strávila se svou kamarádkou v malem městečku vedle more.
"Last year I spent my summer holidays with my friend in a little city by the sea."

Sentences 5–9 are examples of latter mentioned increase in the complexity of syntactic structures. The more complex structures are also reflected in the ACL and ASL values that gain higher figures with every other language level. For more details see Table 4.

The differences between levels were statistically tested. There are significant differences between all levels, except B2 and C1 in ASL and CS, and C1 and REF-CZ in ACL (see Tables 5–7). These results demonstrate that the average sentence length is a relevant characteristic in the development of L2 until the level B2. Even in the case of average clause length, we can see that the p-value between B2 and C1 is slightly

[3] The analysis was performed on raw texts, therefore grammatical errors may occur.

below the chosen significance level, which can be interpreted as a weakly significant difference. However, the non-significant difference between C1 and REF-CZ in ACL is surprising. The results indicate that C1 learners' skills are nearing that of a native speaker in the case of average clause length.

Table 4: ACL and ASL results of chosen sentences

level	sentence	text id	ACL	ASL
A1	5	ttt_P2_286	6.0	3.0
A2	6	TOU_H210_152	8.0	4.0
B1	7	ttt_B6_210	11.0	5.5
B2	8	UJA_UZ_004	12.0	6.0
C1	9	UJA_KT_001	14.0	14.0

Table 5: Statistical tests' results between each pair of the A1–REF-CZ values (ASL)

ASL	A1	A2	B1	B2	C1
A2	**<0.001**				
B1	**<0.001**	**<0.001**			
B2	**<0.001**	**<0.001**	**<0.001**		
C1	**<0.001**	**<0.001**	**<0.001**	0.336	
REF-CZ	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**0.001**

Table 6: Statistical tests' results between each pair of the A1–REF-CZ values (CS)

CS	A1	A2	B1	B2	C1
A2	**<0.001**				
B1	**<0.001**	**<0.001**			
B2	**<0.001**	**<0.001**	**<0.001**		
C1	**<0.001**	**<0.001**	**<0.001**	0.647	
REF-CZ	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**0.001**

Table 7: Statistical tests' results between each pair of the A1–REF-CZ values (ACL)

ACL	A1	A2	B1	B2	C1
A2	**0.013**				
B1	**<0.001**	**<0.001**			
B2	**<0.001**	**<0.001**	**<0.001**		
C1	**<0.001**	**<0.001**	**<0.001**	**0.045**	
REF-CZ	**<0.001**	**<0.001**	**<0.001**	**0.006**	0.503

4.3. Differences of ASL, CS and ACL between Slavic and non-Slavic groups

Since Czech is a Slavic language, we expect that learners with Slavic L1 should have reached higher values of observed indices. Therefore, each proficiency level was divided into a Slavic and a non-Slavic group, and the results were compared. The obtained data confirm our expectations. The biggest difference between Slavic and non-Slavic groups is in the ACL development (see Table 8 and Figure 8). Until the C1 level, there are statistically significant differences between these two groups. Therefore, Slavic learners have an obvious advantage up to the C1 level thanks to their L1 similarities with the Czech language.

Table 8: The ACL means and SDs for Slavic and non-Slavic groups and results of statistical tests

	ACL_S		ACL_N		statistical tests
level	**mean**	**SD**	**mean**	**SD**	**p-value**
A1	5.621	1.178	5.310	1.118	**<0.001**
A2	5.688	1.079	5.425	1.110	**<0.001**
B1	5.856	1.128	5.623	1.019	**<0.001**
B2	6.113	1.152	5.711	1.112	**<0.001**
C1	6.300	1.312	6.295	1.494	0.864

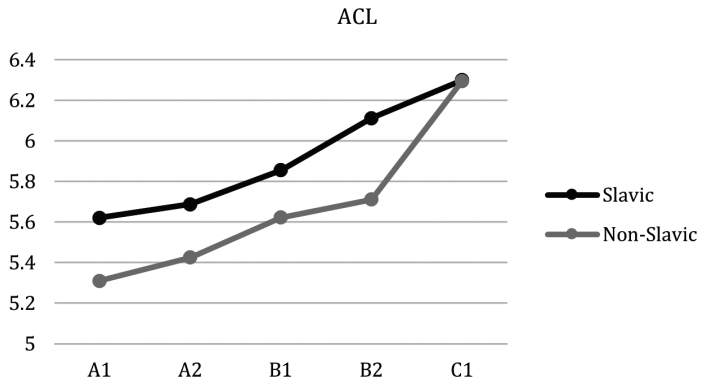

Figure 8: The ACL means for Slavic and non-Slavic groups at all language proficiency levels

While not so straightforward, a similar trend can still be seen in ASL development (see Table 9 and Figure 9). Except for the B1 and C1 levels, there are statistically significant differences between Slavic and non-Slavic learners. Further research is needed to determine whether the B1 learners' results represent random fluctuations or hint at some trend.

Table 9: The ASL means and SDs for Slavic and non-Slavic groups and results of statistical tests

	ASL_S		ASL_N		statistical tests
level	**mean**	**SD**	**mean**	**SD**	**p-value**
A1	8.835	3.055	7.853	3.169	**<0.001**
A2	9.566	3.182	8.679	3.091	**<0.001**
B1	10.255	3.299	10.633	4.368	0.681
B2	11.712	3.493	10.644	2.819	**<0.001**
C1	11.775	2.818	11.441	4.110	0.672

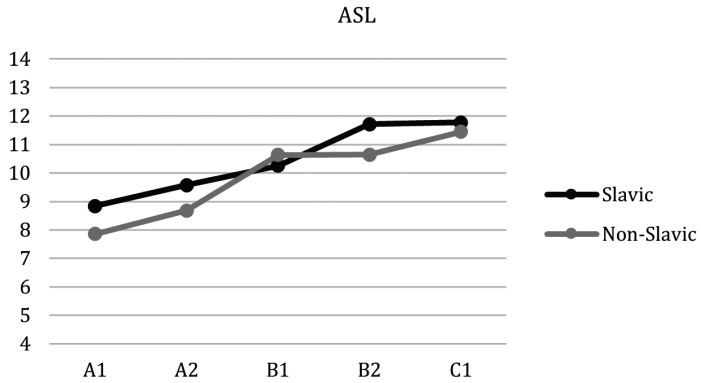

Figure 9: The ASL means for Slavic and non-Slavic groups at all language proficiency levels

In the CS results (see Table 10 and Figure 10), we can see that Slavic learners have an advantage only at lower language levels (statistically significant differences at A1 and A2). The results of the B1 level are consistent with the findings for ASL, i.e., non-Slavic learners use longer sentences on average (which can be caused by a high correlation between these two indices; see Section 4.4).

Table 10: The CS means and SDs for Slavic and non-Slavic groups and results of statistical tests

	CS_S		CS_N		statistical tests
level	**mean**	**SD**	**mean**	**SD**	**p-value**
A1	1.566	0.414	1.478	0.504	**<0.001**
A2	1.684	0.470	1.603	0.465	**<0.001**
B1	1.749	0.456	1.899	0.758	**0.012**
B2	1.931	0.527	1.881	0.454	0.531
C1	1.891	0.377	1.797	0.408	0.260

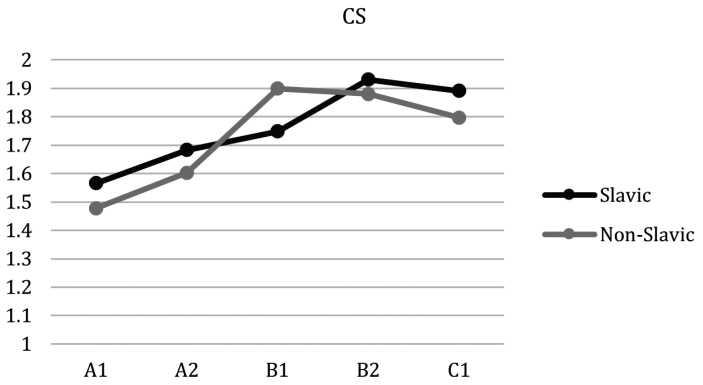

Figure 10: The CS means for Slavic and non-Slavic groups at all language proficiency levels

4.4 Correlations between indices

The last part of our analysis focused on correlations between indices (see Table 11). Not surprisingly, the highest correlation (and in all cases statistically significant) is between ASL and CS. This finding raises the question of the meaningfulness of using both indices concurrently. Both methods model development in a very similar manner, as shown in Section 4.3. On the other hand, ACL has an extremely low correlation with CS. ASL and ACL reveal similar properties despite their weaker but statistically significant differences. However, especially the differences in the results presented in Section 4.3 show a different sensitivity of ACL compared to ASL.

Table 11: Correlation results between all indices and results of statistical tests

level	ASL_CS tau	p-value	ASL_ACL tau	p-value	ACL_CS tau	p-value
A1	0.614	**<0.001**	0.475	**<0.001**	0.083	**<0.001**
A2	0.617	**<0.001**	0.394	**<0.001**	0.005	0.739
B1	0.649	**<0.001**	0.383	**<0.001**	0.027	0.147
B2	0.601	**<0.001**	0.320	**<0.001**	-0.086	**<0.001**
C1	0.486	**<0.001**	0.464	**<0.001**	-0.055	0.399

5. Conclusion

In this study, we analysed the relationship between the sentence and clause lengths and the proficiency levels of foreign Czech learners. We started with the assumption that the sentence and clause length can reflect the degree of syntactic

knowledge and, consequently, can be used for this kind of analysis. The main advantage of this approach lies in its simplicity. However, there are several ways to operationalize the length of these units. Therefore, three different approaches were applied in this study, and the results were compared.

We discovered that learners at higher proficiency levels tend to use longer sentences and clauses. In most cases, there are statistically significant differences between the proficiency levels. Furthermore, Slavic learners use longer clauses up to the C1 level. For sentence lengths, the results are not as straightforward, but we can conclude that for beginners (A1 and A2), Slavic L1 means an advantage for learning Czech language (seen from the perspective of longer clauses and sentences usage).

Finally, relationships between measurement methods were compared as well. Based on the results, we can conclude that it is reasonable to use both clause and sentence length measures because they capture different syntactic properties (especially if we compare ACL and CS).

REFERENCES

BIBER, D. – GRAY, B. – STAPLES, S. – EGBERT, J. (2020): Investigating grammatical complexity in L2 English writing research: Linguistic description versus predictive measurement. *Journal of English for Academic Purposes*, 46. https://doi.org/10.1016/j.jeap.2020.100869

CROSSLEY, A. S. – MCNAMARA, S. D. (2014): Does writing development equal writing quality? A computational investigation of syntactic complexity in L2 learners. *Journal of Second Language Writing*, 26, 66–79. https://doi.org/10.1016/j.jslw.2014.09.006

DE CLERCQ, B. – HOUSEN, A. (2017): A Cross-Linguistic Perspective on Syntactic Complexity in L2 Development: Syntactic Elaboration and Diversity. *The Modern Language Journal*, 101(2), 315–334. https://doi.org/10.1111/modl.12396

JIANG, J. – OUYANG, J. (2018): Minimization and Probability Distribution of Dependency Distance in the Process of Second Language Acquisition. In: J. Jingyang – H. Liu (eds.), *Quantitative Analysis of Dependency Structures*. De Gruyter Mouton, 167–190. https://doi.org/10.1515/9783110573565-009

KHUSHIK, G. A. – HUHTA, A. (2022): Syntactic complexity in Finnish-background EFL learners' writing at CEFR levels A1–B2. *European Journal of Applied Linguistics*, 10(1), 142–184. https://doi.org/10.1515/eujal-2021-0011

KISSELEV, O. – ALSUFIEVA, A. (2017): The development of syntactic complexity in the writing of Russian language learners: A longitudinal corpus study. *Russian Language Journal*, 67, 27–53. Available at: https://scholarsarchive.byu.edu/rlj/vol67/iss1/3

KISSELEV, O. – KLIMOV, A. – KOPOTEV, M. (2021): Syntactic Complexity Measures as Indices of Language Proficiency in Writing: Focus on Heritage Learners of Russian. *Heritage Language Journal*, 18(3). 1–30. http://dx.doi.org/10.1163/15507076-12340016

KUIKEN, F. – VEDDER, I. (2019): Syntactic complexity across proficiency and languages: L2 and L1 writing in Dutch, Italian and Spanish. *International Journal of Applied Linguistics*, 29, 192–210. https://doi.org/10.1111/ijal.12256

KYLE, K. – CROSSLEY, S. – VERSPOOR, M. (2021): Measuring Longitudinal Writing Development Using Indices of Syntactic Complexity and Sophistication. *Studies in Second Language Acquisition*, 43(4), 781–812. https://doi.org/10.1017/S0272263120000546

LU, X. – AI, H. (2015): Syntactic complexity in college-level English writing: Differences among writers with diverse L1 backgrounds. *Journal of Second Language Writing*, 29, 16–27. https://doi.org/10.1016/j.jslw.2015.06.003

LU, X. (2011): A corpus-based evaluation of syntactic complexity measures as indices of college-level ESL writers' language development. *TESOL Quarterly*, 45(1), 36–62. https://doi.org/10.5054/tq.2011.240859

LU, X. (2017): Automated measurement of syntactic complexity in corpus-based L2 writing research and implications for writing assessment. *Language Testing*, 34(4), 493–511. https://doi.org/10.1177/0265532217710675

OUYANG, J. – JIANG, J. – LIU, H. (2022): Dependency distance measures in assessing L2 writing proficiency. *Assessing Writing*, 51. https://doi.org/10.1016/j.asw.2021.100603

PARK, S. (2022): Syntactic complexity in a learner written corpus and L2 speaking quality: Suggestions for distinguishing L2 speaking proficiency. *Journal of Language and Linguistic Studies*, 18(1), 361–371. Doi: 10.52462/jlls.187

R CORE TEAM (2013): A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from: http://www.R-project.org/

TRTANJ, I. – ČOLIĆ, M. A. (2019): Syntactic Complexity and Subordination in Written Discourse of Speakers of Croatian as a Second and Foreign Language. *Rasprave: Časopis Instituta za hrvatski jezik i jezikoslovlje*, 45(2). https://doi.org/10.31724/rihjj.45.2.21

VYATKINA, N. – HIRSCHMANN, H. – GOLCHER, F. (2015): Syntactic modification at early stages of L2 German writing development: A longitudinal learner corpus study. *Journal of Second Language Writing*, 29, 28–50. https://doi.org/10.1016/j.jslw.2015.06.006

WOLF-QUINTERO, K. – INAGAKI, S. – KIM, H.-Y. (1998): *Second language development in writing: Measures of fluency, accuracy, & complexity*. Honolulu: University of Hawai'i, Second Language Teaching & Curriculum Center.

YANG, W. – LU, X. – WEIGLE, C. S. (2015): Different topics, different discourse: Relationships among writing topic, measures of syntactic complexity, and judgments of writing quality. *Journal of Second Language Writing*, 28, 53–67. https://doi.org/10.1016/j.jslw.2015.02.002

YOON, S-Y. – LU, X. – ZECHNER, K. (2020): Features Measuring Vocabulary and Grammar. In: K. Zechner – K. Evanini (eds.), Automated Speaking Assessment. New York: Routledge, 123–137.

CORPORA

ŠEBESTA, K. – BEDŘICHOVÁ, Z. – ŠORMOVÁ, K. et al. (2014): *CzeSL-SGT: CzeSL-SGT – a corpus of non-native speakers' Czech with automatic annotation*, version 2 from 28 Sep 2014. Praha: Ústav Českého národního korpusu FF UK.

ŠEBESTA, K. – GOLÁŇOVÁ, H. – JELÍNEK, T. et al. (2013): *SKRIPT2012: akviziční korpus psané češtiny, přepisy písemných prací žáků základních a středních škol v ČR*. Praha: Ústav Českého národního korpusu FF UK.

ACKNOWLEDGMENT

The research was supported by University of Ostrava, project No. SGS06/FF/2022, Quantitative analysis of texts of CzeSL-SGT corpus.