sciendo
DOI 10.2478/jazcas-2023-0037

LINEAR DEPENDENCY SEGMENTS IN FOREIGN LANGUAGE ACQUISITION: SYNTACTIC COMPLEXITY ANALYSIS IN CZECH LEARNERS' TEXTS

MICHAELA NOGOLOVÁ – MICHAELA HANUŠKOVÁ
– MIROSLAV KUBÁT – RADEK ČECH
Department of Czech Language, Faculty of Arts, University of Ostrava,
Ostrava, Czech Republic

NOGOLOVÁ, Michaela – HANUŠKOVÁ, Michaela – KUBÁT, Miroslav – ČECH, Radek: Linear Dependency Segments in Foreign Language Acquisition: Syntactic Complexity Analysis in Czech Learners' Texts. Journal of Linguistics, 2023, Vol. 74, No 1, pp. 193 – 203.

Abstract: The paper discusses a new way to measure syntactic complexity in foreign language acquisition. It is based on a recently proposed syntactic unit called linear dependency segment (LDS), the longest possible sequence of words belonging to the same clause where all linear neighbours are also syntactic neighbours. The dataset comprises 5,721 Czech texts from the CzeSL-SGT learner corpus covering five CEFR proficiency levels (A1–C1). The study covers two analyses. First, the development of the average clause length in terms of LDS and the average LDS length in the number of words across the latter language proficiency levels. Second, we consider the differences between Slavic and non-Slavic speakers. The results show an increasing tendency of the average clause length measured in LDS while the average clause length measured in words is decreasing. Results also show statistically significant differences between Slavic and non-Slavic speakers in most cases. Our results indicate that using LDS may be a useful unit of syntactic complexity measure in foreign language acquisition research.

Keywords: foreign language acquisition, dependency grammar, linear dependency segment, syntactic complexity, Czech language

1 INTRODUCTION

Syntactic complexity has long been an interest in writing in the second language acquisition domain. Over the years, the complexity of syntactic structures has become a valuable indicator of language development, both in first language acquisition and any other foreign language (FL) acquisition (see Crossley – McNamara 2014; Yang et al. 2015). However, in the last decades, the traditional syntactic complexity measures (such as average length of clause or sentence, subordinate clause per clause, or T-unit (Hunt 1965) per sentence) have been faced with critique for their lack of linguistic background, problematic use for all language proficiency levels, and vague definition of syntactic complexity itself (see e.g. Biber et al. 2020; Kuiken 2022; Ouyang et al. 2022). Recent research has focused on finding alternative ways to measure syntactic

structures, particularly those considering the dependency structure of clauses or sentences, e.g. mean dependency distances (MDD; e.g. Jiang – Ouyang 2018; Ouyang et al. 2022) and linear dependency segment (LDS; Mačutek et al. 2021). The shift towards measurements based on dependency grammar also reflects a shift towards a deeper connection between linguistics and cognitive sciences.

This research focuses on evaluating FL writing with a focus on LDS. We explore the development of average LDS length and average clause length in LDS. The language material comes from the Czech learner corpus CzeSL-SGT, a part of the Czech National Corpus (Šebesta et al. 2014). It consists of 5,721 texts on A1–C1 language proficiency levels according to The Common European Framework of Reference for Languages (CEFR).

2 LANGUAGE MATERIAL AND METHODOLOGY

2.1 Language material

The language material used in the current study comes from the Czech National Corpus. It is a collection of selected texts from the CzeSL-SGT learner corpus (Šebesta et al. 2014). The corpus comprises 8,617 texts authored by 1,965 non-native Czech speakers of all language proficiency levels defined by CEFR. The corpus contains metadata on both authors and texts. In our research, we utilise data on the learner's language proficiency level, their first language (L1), and the length of the text. To ensure the accuracy of our analysis, we excluded texts with unclear or unknown proficiency levels, as well as texts assigned to the C2 level, because only one text is tagged to this category. Furthermore, we removed texts shorter than 55 words because the standard-length requirement for passing a written exam is typically around 50–60 words. We used the L1 information to categorise learners into Slavic and non-Slavic groups. In summary, our corpus consists of 5,721 texts that cover five CEFR proficiency levels. Additionally, we compare the results with the reference corpus (REF-CZ), consisting of texts written by Czech native speakers. The data come from the SKRIPT2012 corpus (Šebesta et al. 2013). Specifically, we use texts written by fourth-grade high school students because of their comparability with the CzeSL-SGT corpus regarding authorship. For sample details, see Tab. 1.

level	number of texts	number of texts Slavic L1	number of texts non-Slavic L1
A1	1,854	1,364	490
A2	1,738	1,157	581
B1	1,313	833	480
B2	702	497	205
C1	114	78	36
REF-CZ	87	-	-

Tab. 1. Number of texts in each group of the analysed sample

2.2 Linear dependency segments (LDS) and data processing

Mačutek et al. (2021) defined the linear dependency segment as follows: "[...] the longest possible sequence of words (belonging to the same clause) in which all linear neighbours (i.e., words adjacent in a sentence) are also syntactic neighbours [...]". In detail, every clause is created by a predicate and other directly or indirectly dependent words. The only exception occurs when the dependent word is another predicate. In that case, the syntactic relationship between word and other predicate presents a boundary between two clauses. LDS then is created only within a clause. For illustration, clauses and LDS determination in sentence (A) are presented in Fig. 1. The circles indicate particular clauses, and the squares then individual LDS.

(A) *Petr má psa, který hodně kouše.*
'Petr has a dog that bites a lot.'

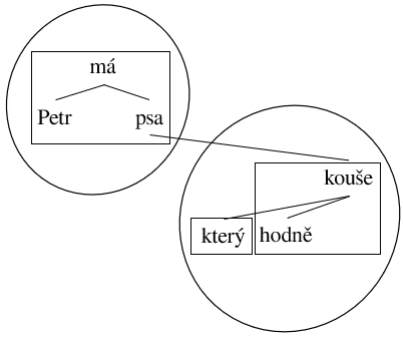

Fig. 1. Visualization of clauses and LDS determination in a sentence (A)

The sentence (A) has two clauses because two predicates – *má* ('has') and *kouše* ('bites') – are present. One LDS creates the first clause, and the second clause makes two LDS. The first word of the sentence (*Petr*) is directly dependent on the word *má.* It can be also seen that the second word (*má*) is directly connected to the word *psa*. These three words are neighbours in linear clause ordering, so creating one LDS. The third word (*psa*) is also related to the word *kouše*. However, these two words are not adjacent in the sentence, and they are not in the same clause. Therefore, they cannot be in the same LDS. The fourth word (*který*) is directly connected to the word *kouše.* These two words are not next to each other in the clause word order, so the word *který* creates one LDS. The last LDS of the second clause is created by the words *hodně* and *kouše* because they are directly connected as adjacent in the clause. As such, LDS captures a clause's linear sentence ordering and dependency structure.

The current study aims to analyse syntactic structures development of FL writing focusing on LDS. We use two indices:

(i) Average clause length measured in the number of LDS (ACL),
(ii) Average LDS length measured in the number of words (ALDSL).

As the LDS reflects both the dependency structure and word order in clause, the value of ACL and ALDSL can be in accordance with the clause complexity. For illustration, Fig. 2 and 3 show the dependency relationships of sentences (B) and (C). The squares represent individual LDS.

(B) *Petr má psa.*
'Petr has a dog.'

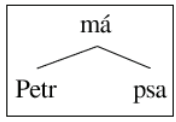

Fig. 2. Visualization of dependency relationships and LDS in a sentence (B)

(C) *Můj dobrý kamarád Petr má doma velkého psa.*
'My good friend Petr has a big dog at home.'

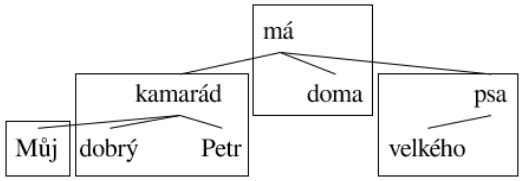

Fig. 3. Visualization of dependency relationships and LDS in a sentence (C)

Both sentences (B) and (C) contain only one clause. However, when one compares these two sentences, more complex syntactic structures are present in the longer one. The distance between two dependent words is also bigger. As the words that create an LDS must be directly connected as adjacent in the sentence word order, more complex syntactic structures will lead to more LDS within the clause and a shorter average length of LDS. The sentence (B) has an average clause length of 1 (1 LDS / 1 clause = 1), with an average LDS length of 3 (3 words / 1 LDS = 3). In contrast, sentence (C) has an average clause length of 4 (4 LDS / 1 clause = 4) and an average LDS length of 2 (1+3+2+2 = 8 words; 8 words / 4 LDS = 2). Therefore, we assume that development

towards a higher level of language proficiency will correspond with an increase in the number of LDS in a clause and a decrease in LDS length.

There are various approaches to processing the syntactic structure of a sentence, and these approaches influence the created annotation scheme used for sentence processing. This paper uses Surface Syntactic Universal Dependencies (SUD; Gerdes et al. 2018) for annotation. SUD is built upon the Universal Dependencies (UD; Zeman et al. 2022) which aims to provide a versatile annotation system for a wide range of languages. While UD exhibits a greater focus on semantic aspects, SUD, in contrast, adopts a more syntactically oriented approach, with auxiliaries and prepositions holding a superior position rather than being subordinate to content words, as seen in the case of UD. In conducting our analysis, we initially utilised UDPipe 2.0. (Straka 2018) to process all texts. Then, we used Grew software (Guillaume 2021) for UD to SUD conversion. The selection of SUD was motivated by our objective to conduct a syntactic analysis, making its more syntactically-oriented perspective highly suitable for our study.

The analyses were performed by the following steps. First, all texts were processed separately for each proficiency level. Second, the mean of clauses, LDS length, and standard deviations (sd) were calculated. Third, we used the Mann-Whitney test (Mann – Whitney 1947) with a significance level of $\alpha = 0.05$ to test statistical differences between pairs of proficiency levels. This statistical test was chosen due to the non-normal distribution of the data. We also performed the same test to compare groups of Slavic and non-Slavic learners at each proficiency level.

3 RESULTS

3.1 Development of ACL and ALDSL across the language proficiency levels

The results presented in Tab. 2 show the values of ACL and ALDSL for examined language proficiency levels and the values obtained from the reference corpus. We can see the increased tendency across all language proficiency levels towards the value obtained from native speakers when first focused on ACL. Fig. 4 gives a more detailed description of the obtained values. These results support our hypothesis that the ACL increases as the language proficiency level increases. ALDSL values in Tab. 2 and Fig. 5 show a descending tendency towards the value obtained from native speakers. These results also support our hypothesis that the ALDSL value decreases with increasing language proficiency level.

Additionally, the gap in syntactic abilities measured by ACL and ALDSL between learners and native speakers diminishes as language proficiency increases. Standard deviation (sd) values indicate consistent variability of the results across different proficiency levels in both ACL and ALDSL analysis.

	ACL		ALDSL	
level	mean	sd	mean	sd
A1	2.522	0.556	2.145	0.201
A2	2.572	0.522	2.118	0.192
B1	2.674	0.553	2.096	0.178
B2	2.788	0.556	2.085	0.181
C1	2.997	0.633	2.049	0.176
REF-CZ	3.216	0.573	1.935	0.097

Tab. 2. The mean values of ACL, ALDSL and their sd

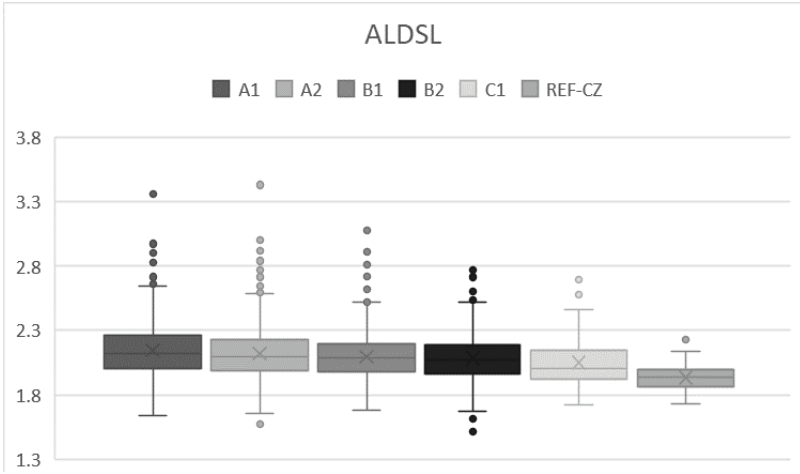

Fig. 4. ACL values from texts at A1–REF-CZ

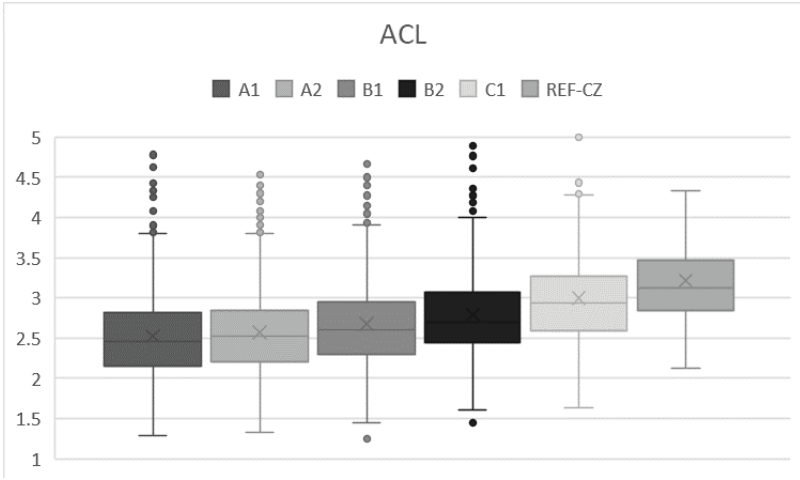

Fig. 5. ALDSL values from texts at A1–REF-CZ

The differences between pairs of levels were statistically tested. In the case of ACL, there are significant differences between all levels. These findings suggest that average clause length is relevant in FL syntactic development (for results, see Tab. 3).

ACL	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**	**0.007**	
REF-CZ	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**0.005**

Tab. 3. Statistical tests' results between each pair of the A1–REF-CZ values (ACL)

Concerning the average length of LDS, the statistically significant differences are detected between all pairs of levels except between B1 and B2 (for more details, see Tab. 4). Further research is needed to determine whether similar results from texts at B1 and B2 level represent random fluctuations or hint at some trend.

ALDSL	A1	A2	B1	B2	C1
A2	**<0.001**				
B1	**<0.001**	**<0.001**			
B2	**<0.001**	**<0.001**	0.128		
C1	**<0.001**	**<0.001**	**0.030**	**0.019**	
REF-CZ	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**0.007**

Tab. 4. Statistical tests' results between each pair of the A1–REF-CZ values (ALDSL)

3.2 Differences between Slavic and non-Slavic groups

Given that Czech is a Slavic language, with many similarities in language syntactic structures between all languages in the group, it is reasonable to assume that learners with Slavic L1 will have higher values of ACL and lower ones in the case of ALDSL. To test this hypothesis, texts at each proficiency level were divided into Slavic and non-Slavic groups, and their respective values were compared. As can be seen in Tab. 5 and Fig. 6, the data confirmed the assumption, mainly regarding ACL development. Up to the C1 level, statistically significant differences were found between these two groups, indicating that Slavic learners possess a clear advantage due to their L1 background.

	ACL_S		ACL_N		statistical tests
level	mean	sd	mean	sd	p-value
A1	2.573	0.549	2.384	0.552	**<0.001**
A2	2.632	0.519	2.456	0.508	**<0.001**
B1	2.744	0.580	2.556	0.482	**<0.001**
B2	2.851	0.553	2.637	0.536	**<0.001**
C1	3.035	0.591	2.914	0.718	0.345

Tab. 5. The ACL means and sd for Slavic and non-Slavic groups and results of statistical tests

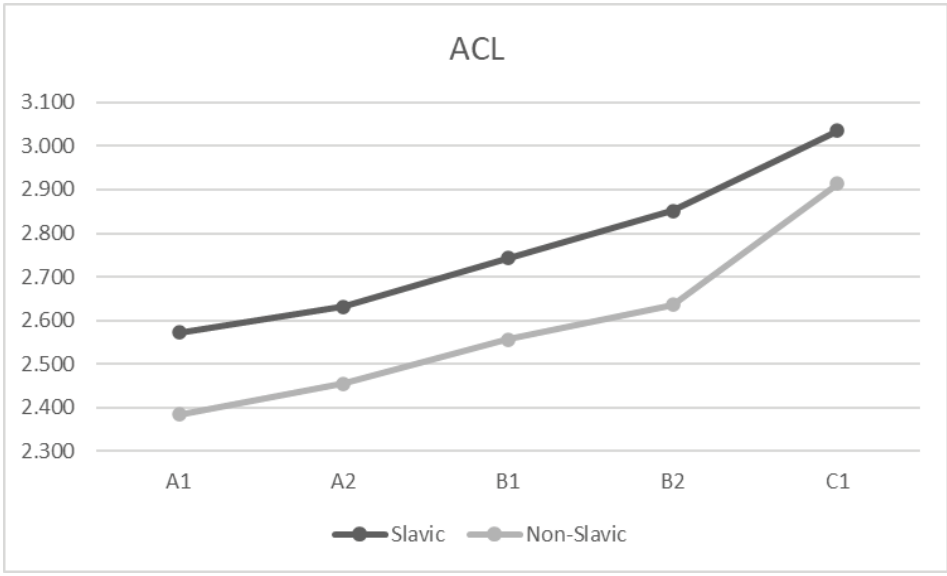

Fig. 6. The ACL means for Slavic and non-Slavic groups at all language proficiency levels

Regarding ALDSL, results in Tab. 6 and Fig. 7 demonstrate that the mean values for the Slavic groups are generally lower than those for non-Slavic counterparts across all proficiency levels. However, statistically significant differences between the two groups are observed only at levels A1 and C1.

	ALDSL_S		ALDSL_N		statistical tests
level	mean	sd	mean	sd	p-value
A1	2.131	0.184	2.182	0.237	**<0.001**
A2	2.111	0.185	2.133	0.204	0.094
B1	2.092	0.171	2.102	0.189	0.870
B2	2.081	0.178	2.092	0.187	0.425
C1	2.023	0.167	2.107	0.185	**0.013**

Tab. 6. The ALDSL means and sd for Slavic and non-Slavic groups and results of statistical tests

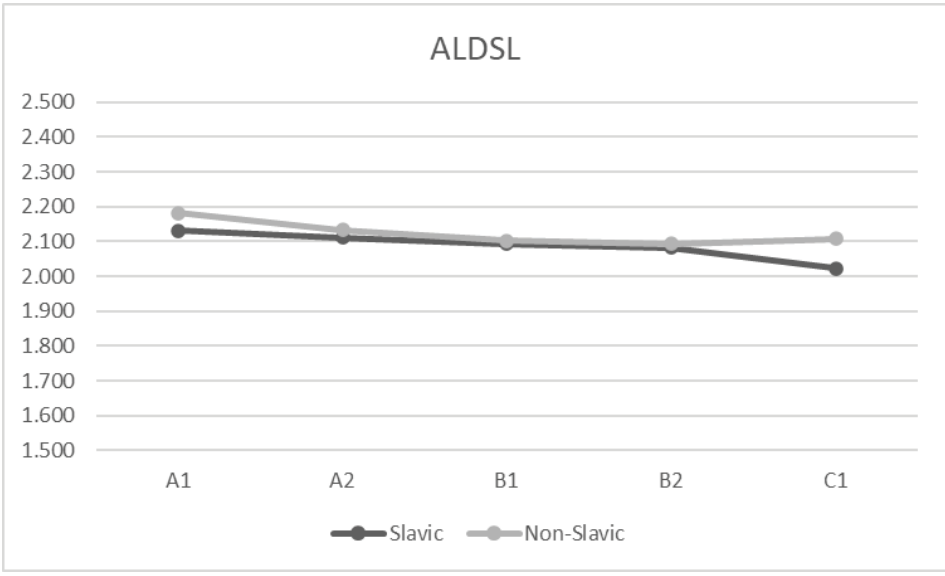

Fig. 7. The ALDSL means for Slavic and non-Slavic groups at all language proficiency levels

4 CONCLUSION

This study focused on linear dependency segments (LDS) in Czech texts written by non-native speakers. We aimed to explore the potential of LDS for measuring syntactic complexity in foreign language acquisition research. The results showed that the linear dependency segments could be useful for measuring syntactic complexity.

The analysis of average clause length based on the number of LDS revealed an increasing tendency towards native speakers across all levels of language proficiency. Furthermore, the data also confirmed our expectation that Slavic L1 speakers have longer clauses on average than their non-Slavic counterparts.

According to the average length of LDS measured in words, the higher the proficiency level, the shorter the LDS. LDS lengths for texts written by native Slavic speakers and native non-Slavic speakers are generally similar, except for the A1 and C1 levels. The differences between these two groups of learners are not as apparent as we had expected.

Since this is a pioneer study examining linear dependency segments in foreign language acquisition, further research must be done to confirm our results. It is important to perform research in other languages and different contexts, such as spoken language or different writing genres. Additionally, further exploration of the relationship between linear dependency segments and other measures of syntactic complexity could provide a deeper understanding of language acquisition.

ACKNOWLEDGEMENTS

The research is supported by Grant SGS08/FF/2023, University of Ostrava.

References

Biber, D., Gray, D., Staples, S., and Egbert, J. (2020). Investigating grammatical complexity in L2 English writing research: Linguistic description versus predicative measurement. Journal of English for Academic Purposes, 46.

Crossley, A., S., and 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, pages 66–79.

Gerdes, K., Guillaume, B., Kahane, S., Perrier, G. (2018). SUD or Surface-Syntactic Universal Dependencies: An annotation scheme near-isomorphic to UD. In Universal Dependencies Workshop.

Guillaume, B. (2021). Graph Matching and Graph Rewriting: GREW tools for corpus exploration, maintenance and conversion. In Demonstrations – 16th Conference of the European Chapter of the Association for Computational Linguistics (EACL).

Hunt, K. (1965). Grammatical structures written at three grade levels. NCTE Research Report No. 3. Champaign, IL, USA: NCTE.

Jiang, J., and 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, pages 167–190.

Kuiken, F. (2022). Linguistic complexity in second language acquisition. Linguistic Vanguard.

Mačutek, J., Čech, R., and Courtin, M. (2021). The Menzerath-Altmann law in syntactic structure revisited. In Quasy, SyntaxFest 2021: Proceedings of the Second Workshop on Quantitative Syntax (March 21 – 25, 2022). Sofia: Association for Computational Linguistics, pages 65–73.

Mann, H. B., and Whitney, D. R. (1947). On a Test of Whether One of Two Random Variables Is Stochastically Larger than the Other. Annals of Mathematical Statistics, 18, pages 50–60.

Ouyang, J., Jiang, J., and Liu, H. (2022). Dependency distance measures in assessing L2 writing proficiency. Assessing Writing, 51.

Straka, M. (2018). UDPipe 2.0 Prototype at CoNLL 2018 UD Shared Task. In Proceedings of CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies (October 31 – November 1, 2018). Brussels: Association for Computational Linguistics, pages 197–207.

Šebesta, K., Bedřichová, Z., Šormová, K., Štindlová, B., Hrdlička, M., Hrdličková, T., Hana, J., Petkevič, V., Jelínek, T., Škodová, S., Poláčková, M., Janeš, P., Lundáková, K., Skoumalová, H., Sládek, Š., Pierscieniak, P., Toufarová, D., Richter, M., Straka, M., and Rosen, A. (2014). CzeSL-SGT: korpus češtiny nerodilých mluvčích s automaticky provedenou anotací, verze 2 z 28. 7. 2014. Praha. Ústav Českého národního korpusu FF UK. Accessible at: http://www.korpus.cz.

Šebesta, K., Goláňová, H., Jelínek, T., Jelínková, B., Křen, M., Letafková, J., Procházka, P., and Skoumalová, H. (2013). SKRIPT2012: akviziční korpus psané češtiny – přepisy písemných prací žáků základních a středních škol v ČR. Praha, Ústav Českého národního korpusu FF UK. Accessible at: http://www.korpus.cz.

Yang, W., Lu, X., and 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, pages 53–67.

Zeman, D., et al. (2022). Universal Dependencies 2.10, LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics, Matematicko-fyzikální fakulta, Univerzita Karlova. Accessible at: http://hdl.handle.net/11234/1-4758.