sciendo
DOI 10.2478/jazcas-2021-0033

# ACQUIRING WORD ORDER IN SLOVAK AS A FOREIGN LANGUAGE: COMPARISON OF SLAVIC AND NON-SLAVIC LEARNERS UTILIZING CORPUS DATA

MARTINA IVANOVÁ[1] – MIROSLAVA KYSEĽOVÁ[1,2] – ANNA GÁLISOVÁ[3]
[1] Faculty of Arts, University of Prešov, Prešov, Slovakia
[2] Faculty of Philology, Jagiellonian University, Kraków, Poland
[3] Faculty of Arts, Matej Bel University, Banská Bystrica, Slovakia

IVANOVÁ, Martina – KYSEĽOVÁ, Miroslava – GÁLISOVÁ, Anna: Acquiring word order in Slovak as a foreign language: Comparison of Slavic and Non-Slavic learners utilizing corpus data. Journal of Linguistics, 2021, Vol. 72, No 2, pp. 353 – 370.

**Abstract:** The paper deals with the acquisition of Slovak word order in written texts of students of Slovak as a foreign language. Its attention is focused on identifying the correct and incorrect placement of enclitic components, and their erroneous usage is analysed with respect to different investigated variables (types of enclitic components, types of syntactic construction, distance from lexical/syntactic anchor, and realization in pre- or post-verbal position). The paper also pays attention to the error rate regarding individual proficiency levels of students, and error distribution in two language groups, Slavic and Non-Slavic learners, is compared.

**Keywords:** word order, enclitics, error analysis, syntactic complexity, Slavic learners, Non-Slavic learners, acquisition stages, interlanguage

## 1 INTRODUCTORY REMARKS

Among morpho-syntactic phenomena, one of the most problematic challenges for students of Slovak as a foreign language is acquiring word order. The main reason is that Slovak word order is formed on the borderline of three major principles, i.e., functional sentence perspective, prosody, and grammar, which represent independent factors determining the linear order of a sentence, yet they sometimes interfere with each other. Since linearization of Slovak sentence structure is not determined by the grammatical functions of sentence components (except for attributes within noun phrases) and its major function is to express information structure (see [1]), it is characteristic of relative flexibility. On the other hand, word order flexibility is highly restricted with respect to position of attribute phrases and clitics due to grammatical and prosodic rules which govern their placement. Clitics, especially, represent one of the most specific and intricate phenomena within the word order of many languages. Even the languages with most similarities differ in

clitic placement, as pointed out by Uhlířová regarding Czech, Slovak, and Polish [2, p. 82]. This fact also opens up space for their erroneous usage in texts produced by highly proficient speakers.

The present paper is aimed at investigating acquisition of enclitics ordering by foreigners learning Slovak as L2. Based on performance analysis [3, p. 73], the aim is to map accuracy in the placement of enclitics with respect to the level of language acquisition (lower proficiency versus higher proficiency level) and the affiliation of the learners' mother tongue(s) to a language group (Slavic versus non-Slavic language family). To get a closer picture of Slovak word order acquisition, we compiled our own corpus of written texts of students learning Slovak as a foreign language at different proficiency levels, and we supplied each enclitic component present in the texts with annotation tags reflecting different variables. By measuring the error distribution and relating the statistical values with specific features of the texts (syntactic complexity), our aim was to specify how fully learners of the two language groups acquire the principles of enclitic ordering at different stages of their interlanguage.

However, determining the extent to which learners use a certain language feature accurately presupposes identification of an error and its distinction from correct use. Within the Slovak context, one particular circumstance that hinders identification of erroneous or inappropriate word orders is the absence of theoretical and practical investigation into word order, which would show preferential patterns of word order in Slovak as L1. The only work on this topic in the Slovak context was published in 1966 by J. Mistrík [4]; other works concerning Slovak word order focus mostly on syntagmatic word order (cf. [5]) and are not based on corpus data. The description of Slovak enclitics and their linear ordering within the Slavic context can be found in Frank & King [6] and Beličová & Uhlířová [7]. The situation in Slovak is largely at odds with that of Czech in which word order behaviour of clitics attracted significant attention both in investigation of Czech as L1 (cf. [8] for an overview) as well as L2 (cf. [9]).

The structure of the present paper is as follows: based on the theoretical literature available, in Section 2, we present a short definition and classification of enclitics in Slovak. In Section 3, we describe our samples and methods for annotating these enclitics. In Section 4, statistical results are presented, and Section 5 is devoted to discussion of the results and the conclusion.

## 2 CLITICS IN SLOVAK WORD ORDER

### 2.1 Definition of clitics

Prosody relates to the word order realization of phonologically non-independent elements devoid of word stress called clitics, which cannot be realized freely, i.e., in various sentence positions depending on the pragmatic and discourse function, but

their position within the sentence structure is determined phonologically. Slovak belongs to those languages which follow Wackernagel's Law and its clitic elements belong to the category of second-position clitics (2P)[1] [10], which are typical of having "dual citizenship". Within the sentence, they follow an initial element called prosodic host (anchor), a clause-initial unit, usually the first sentence constituent. However, morphologically, lexically, or syntactically they belong to the governor, most typically a verb position not conditioned prosodically within the sentence. As the prosodic host of the clitic component and its governor do not necessarily correspond, it may lead to emergence of constituent discontinuity (cf. [1]).

Phonologically, 2P clitics are enclitics, however, they may be procliticized under certain circumstances. It happens in complex sentences with a matrix clause containing a clitic item which is realized discontinuously, being disrupted from the initial sentence component by an interposed subordinated clause. If the clitic component is realized after the interposed clause (after a pause), phonologically, it is procliticized to the following sentence constituent, e.g., *Samozrejme aj to, čo je na tanieri,* ***ma*** *inšpiruje.* 'Of course, everything on the plate inspires me, too.' (Omnia Slovaca III).

## 2.2 Classification of clitics in Slovak

### *2.2.1 Constant and inconstant clitics*

Prosodic deficiency (the absence of word stress) is not always considered as the defining feature of clitic components. In many theoretical works, constant and inconstant clitic components are differentiated, the former labelled enclitics tantum and the latter as volatile enclitics (cf. [4] for Slovak). The following characteristics can be stipulated for those two groups:

(i) Enclitics tantum, or pure sentential clitics (cf. [11] for the term), can be defined as prosodically deficient unstressable elements that are unstressed independently of the context in which they are realized, thus, they are unable to be focused and cannot be moved to initial position. According to Junghanns [12], they can be labelled as lexical clitics, as the clitic status represents an inherent part of their lexical "equipment".

[1] Despite the fact that Slovak clitics are defined as second position clitics, there are many deviations from that rule. Ambiguity of clitic placement holds especially true for two structure types: (i) for compound and complex sentences with certain complementizers, e.g., after *ale* ('but'), the enclitic component can either occupy the position immediately after the complementizer: *Vydala sa za nejakého Bergera, ale* ***sa*** *s ním rozviedla.* 'She married a certain Berger but she divorced him.' (Omnia Slovaca III), or after the first sentence constituent: *Cítila jeho dych za chrbtom, ale neobrátila* ***sa***. 'She felt his breath behind her back but she did not turn around.' (Omnia Slovaca III); (ii) for sentences with a multi-constituent thematic part: the enclitic component can either occupy the second position (after the complementizer): *Teraz prišiel trest za to, že* ***sa*** *kedysi vzdala svojho syna.* 'Now came the punishment for her giving up her son.' (Omnia Slovaca III), or it is realised after the first thematic item: *S. Markovič spomína, že kedysi* ***sa*** *náklad nosil hore síce ťažšie, ale oveľa romantickejšie.* 'S. Markovič recalls that carrying the load up used to be more difficult, though much more romantic.' (Omnia Slovaca III).

(ii) Volatile/inconstant enclitics or semi-clitics (cf. [11] for the term) can be defined as prosodically unstressed elements that can have phonological autonomy under certain contextual conditions. In Junghanns [12], they are labelled as phonological clitics due to the fact that their clitic status is formed "in the phonological part of the [sic] grammar usage".

However, the boundary between clitics and non-clitics is often blurred. This is especially the case of inconstant/volatile enclitics. As Hana [1, p. 74] points out, enumerating the exact set of clitics is far from trivial and probably impossible. In our approach, clitic status is ascribed to those monosyllabic auxiliary and non-auxiliary and bi-syllabic auxiliary components[2] which conform to the property specified by Hana ([1]): [1P–Cl] A word between 1P and a clitic is a clitic.[3] The dataset of investigated clitic components will be specified in the following section. Due to low frequency in students' texts, we also decided to omit clitic conjunctions and particles from our investigation.

### *2.2.2 Verbal and argument clitics*

Enclitics can be further divided into two categories depending on the possibility to independently fulfil syntactic functions within the sentence. Dependent morphological enclitics relying on their lexical governor (the verb) and functioning as exponents of grammatical categories (tense, mood, voice, person, number) can be labelled as verbal clitics. In Slovak, the following verbal enclitic components can be differentiated: (i) conditional enclitic tantum: exponent of mood (the clitic component marking conditional mood), (ii) auxiliary enclitics tantum: exponents of person and number (auxiliary components marking person and number in l-participle forms of the preterite and antepreterite *som, si, sme, ste*), (iii) reflexive enclitic tantum: exponent of voice (the reflexive clitic *sa* marking passive voice within reflexive deagentive constructions).

On the other hand, independent lexical enclitics capable of fulfilling syntactic functions within the sentence as verbal complements (with the verb as their syntactic governor) can be labelled as argument clitics. Argument clitics are represented by weak or short pronominal forms, coding both direct and indirect objects or adverbials. Among argument clitics, the following subgroups can be singled out: (i) non-prepositional personal enclitics tantum: short forms of personal pronouns which stand in opposition to long accented forms (*ma – mňa* 'me'*, ťa – teba* 'you'*, ho – jeho* 'him'*, mi – mne* 'me'*, ti – tebe* 'you'*, mu – jemu* 'him'), (ii) non-prepositional personal volatile enclitics:

[2] Due to unclear status in theoretical studies, we decided to exclude bi- and tri-syllabic prepositional-pronominal forms from our dataset, e.g., *s nami* 'with us'*, pre mňa* 'for me'*, na neho* 'on him'.

[3] Hana [1, pp. 75–76] uses two other criteria for clitic delimitation: "Clitics cannot occur in isolation, e.g., as an answer to a question.", "Clitics cannot occur sentence-finally". As inconstant clitics are also included in our dataset, we do not apply these criteria.

unparalleled forms of personal pronouns which can be used either as enclitics or as accented full forms (*ju* 'her', *nás* 'us', *vás* 'you', *ich* 'them', *jej* 'her', *nám* 'us', *vám* 'you', *im* 'them'), (iii) demonstrative volatile enclitics: forms of demonstrative pronouns which can occupy the initial position in the stressed form or under certain conditions they become prosodically dependent and behave like enclitic elements (*to* 'that', *tu* 'here', *tam* 'there', *tak* 'so', etc.), (iv) prepositional volatile enclitics: forms of personal and demonstrative pronouns (*k nám* 'to us', *k vám* 'to you', *k nej* 'to her', *s ním* 'with him', *s ňou* 'with her', *s tým* 'with that', etc.), (v) reflexive enclitic tantum: reflexive pronouns *sa, si* functioning as weak, unstressed forms of the longer forms *seba, sebe*. The clitic status of the copular *byť* 'to be' is disputable and there is disagreement as to the clitic nature of the copular 'be' in Slavic languages. The opinions on the enclitic status of *'be'* forms can be classified as follows:

– enclitic status is assigned only to the auxiliary *be*-forms (e.g., [8], [6]);
– only the auxiliary *byť* in the present tense within passive constructions is labelled as an enclitic component (e.g., [11]);
– only auxiliary forms of the past conditional and antepreterite (*byl/a* for Czech) and present forms of the non-auxiliary *být* can acquire clitic status (e.g., [13]).

As can be seen, the deciding criterion for assessment of enclitic status to a component is associated with the degree of grammaticalization (auxiliaries as the most grammaticalized elements). However, according to Palková [14], it is a common process that monosyllabic elements often lose stress and became part of the prosodic tact of the neighbouring word. She states that it is a matter of rhythm, not grammatical status. If Palková's assumption is right, then the nature of *be*-forms is not determined by its grammatical status (auxiliary vs. copular vs. full lexical), but by contextual distribution. In that sense, the copular *byť* can also be described as a volatile enclitic element. In our data, enclitic status is assigned to: (i) monosyllabic forms of *byť* (auxiliary, copular, full lexical) realized in the second position: **Teraz Kabula je šťastná a žije v Poľsku. – Teraz je Kabula šťastná a žije v Poľsku.* 'Kabula is now happy and lives in Poland.' = copular *byť* (Polish, A2), *Všetko v tom meste jej pripomínalo Marka, ktorý je teraz niekde nad ňou* 'Everything in the town reminded her of Mark who is now somewhere above her.' = full lexical *byť* (Serbian, B2), (ii) disyllabic forms of *byť* (auxiliary elements in past conditional, antepreterite and periphrastic passive) realized in the second position: **Víťaz vyberal cieľ charitatívny, na ktorý dávaná bola cena. – Víťaz vyberal cieľ charitatívny, na ktorý bola dávaná cena.* 'The winner would select a charitable cause for the prize to be awarded to.' (Polish, A2).

Enclitic components can aggregate into clitic clusters comprising (hypothetically) 2 – 7 components. The internal organization of clitic clusters in Slovak can be described as follows: BY > AUX > REFL > NON-ARG. DAT > ARG. DAT > ACC > GEN > CONJ. However, in real communication, such extensive clusters are rather rare. In our data, clitic clusters usually comprise 2 to 3 members.

Mono-syllabic forms of the verb *byť* 'to be' (independent of their lexical status) can occur as part of clitic clusters between 1P constituent and another clitic, e.g., *Skoro je mi ťa ľúto.* 'I almost feel sorrow for you.' (Omnia Slovaca III). Bi-syllabic forms do not show similar behaviour, compare: **Skoro bolo mi ťa ľúto., Skoro mi ťa bolo ľúto.* That's why monosyllabic forms of the verb *byť* are treated as inconstant enclitics and form part of our dataset.

## 3 METHODOLOGICAL ASPECTS OF RESEARCH

### 3.1 Data description

To conduct our investigation of word order errors in texts written by foreigners learning Slovak, we compiled our own corpus of written texts. The data come from a pre-pilot version of the Corpus of Texts of Students Learning Slovak as a Foreign Language (errkorp-0.1) [15] which is under development. In its current state, the corpus comprises 12 733 tokens and 10 428 words. As the volume of the given corpus with respect to the amount of word order errors was not sufficient, we completed it with our own texts.[4] All sentences with enclitic components were transcribed into Excel and were assigned annotation tags reflecting the investigated variables (see 4.2). Overall, we analysed 81 texts, of which 43 texts come from students of Slovak with a Slavic mother tongue and 38 texts were produced by students with a non-Slavic mother tongue.[5] The texts were divided by proficiency level into two categories: 54 texts at the lower proficiency levels A1 – B1 and 27 texts at the higher proficiency levels B2 – C1 (according to CEFR). The aim was to obtain approximately 50 errors of enclitic placement in both language categories of texts at both investigated levels (A1 – B1 and B2 – C1), which is around 200 errors in total, which we considered to be the minimum amount for our analysis purposes. Overall, 1305 sentences with clitic components were analysed out of which 217 contained errors in enclitic usage and in 1089 enclitics were used correctly. The entire database is published online at zenodo.org (cf. [16]).

### 3.2 Error annotation

In errkorp-0.1, word order errors are divided into two categories from a predefined error taxonomy: the error tag ORDER is used for errors concerning enclitic components and attributive phrases and the error tag THEME is used for errors concerning functional sentence perspective.

[4] The non-corpus texts were obtained from lecturers of Slovak as a foreign language and were collected from non-native speakers of Slovak attending university language courses abroad. The texts were produced during different types of situations, i.e., in class, as homework and in examinations, and were handed in either in electronic form or as manuscripts.

[5] The data comprise: (i) texts of students from all three language groups within Slavic languages (West Slavic – Polish, South Slavic – Serbian and East Slavic – Ukrainian), (ii) texts of students with a non-Slavic Indo-European mother tongue, with a majority of Germanic (mostly English, German) and Romance (mostly Italian) languages, (iii) texts of students with non-Indo-European mother tongues, with a majority of Finno-Ugric (Hungarian) and Sino-Tibetan (Chinese) mother tongues.

As we need to analyse highly specialized language phenomena (word order of enclitic components), we decided to annotate corpus data manually with respect to the additional variables under investigation. We did not use any commonly used programmes for the purpose of compiling a corpus as the common options that those programmes offer (like tokenization, tagging, parsing, etc.) are not relevant to our investigation.

The texts in our database were annotated using two annotators independently, the annotations were later compared to eliminate subjective evaluation of errors. Agreement in annotation solutions achieved by the annotators was evaluated using the metric κ (kappa, cf. [17]) which is used as a standard measure instrument for inter-annotation agreement. It is calculated as:

$$\kappa = \frac{P(A) - P(E)}{1 - P(E)}$$

where P(A) is observed agreement between the annotators, and P(E) is the expected agreement, i.e., the probability that the annotators agree by chance. The calculated interval oscillates between (0.1) where $\kappa = 1$ means perfect agreement and $\kappa = 0$ agreement equal to chance (cf. [18]). As a result, the reliability of the annotation has been proved as the calculations showed that = 0.92 which can be interpreted as nearly perfect agreement.[6]

### 3.3 Annotation parametres

To investigate acquisition of clitic ordering, we recorded the presence of all enclitic components in the 81 analysed texts, both correctly and incorrectly used. During text annotation, we took following parameters into account.

| Analysed parameter | Types |
|---|---|
| **Type of component** | **Reflexive enclitics**<br>R = reflexive component<br>**Verbal enclitics**<br>G = auxiliary *byť* (separate grammatical morphemes coding person and number in preterite and antepreterite forms)<br>K = conditional morpheme<br>**Argument enclitics**<br>P = short form of personal pronoun<br>PP = prepositional pronoun<br>D = monosyllabic demonstrative pronoun<br>**Be enclitics**<br>S = non-auxiliary *byť*<br>**Combination of enclitics in a row**<br>KT = clitic cluster |

[6] 1082 cases of enclitic usage were rated as Correct by both annotators, 204 cases were rated as Errors by both annotators, so P(A) = 0.98. Annotator A rated 1088 cases as Correct, Annotator B 1095 cases; Annotator A rated 217 cases as Errors, Annotator B 210 cases, so, P(E) = 0.73. $\kappa = (0.98 - 0.73) / (1 - 0.73) = 0.92$.

| Analysed parameter | Types |
|---|---|
| **Correctness of usage** | C = correct |
| | E = error |
| **Type of syntactic construction** | JV = simple sentence (including the first main clause in a compound sentence, initial main clause in a complex sentence) |
| | PS = second main clause in a compound sentence |
| | HSH = postponed main clause of a complex sentence |
| | HSV = subordinate clause of a complex sentence |
| | IK = reflexive component in non-finite construction, mostly infinitive |
| **Correct proximity to lexical/syntactic host** | 0 = zero distance<br>1 = 1 phrase between an enclitic and the host<br>2 = 2 phrases between an enclitic and the host... |
| **Correct position in relation to lexical/syntactic host** | preV = preverbal position<br>postV = postverbal position |

**Tab. 1.** Parameters analysed with respect to the prosodic factor

## 4 RESULTS

In this section, the outcomes of error distribution measuring with respect to the stages of acquisition and the investigated variables will be presented. The aim is 1) to quantify and compare the ratio of correct and erroneous usage of enclitic components, 2) to describe the relationship between syntactic complexity and erroneous usage of enclitics, 3) to identify similarities in distribution of errors with respect to the investigated variables at early and intermediate (A1 – B1) and advanced (B2 – C1) stages of acquisition.

### 4.1 Correct and erroneous usage in the texts

The ratio of correct and erroneous enclitic placement in texts by the Slavic and the Non-Slavic groups is provided in the following figure.[7]

[7] At elementary and lower-intermediate levels, 307 sentences (26 texts) were annotated within the Slavic group and 223 sentences (28 texts) were annotated within the Non-Slavic group. At the upper-intermediate and advanced levels, 502 sentences (17 texts) were annotated within the Slavic group and 273 sentences (10 texts) were annotated within the Non-Slavic group.

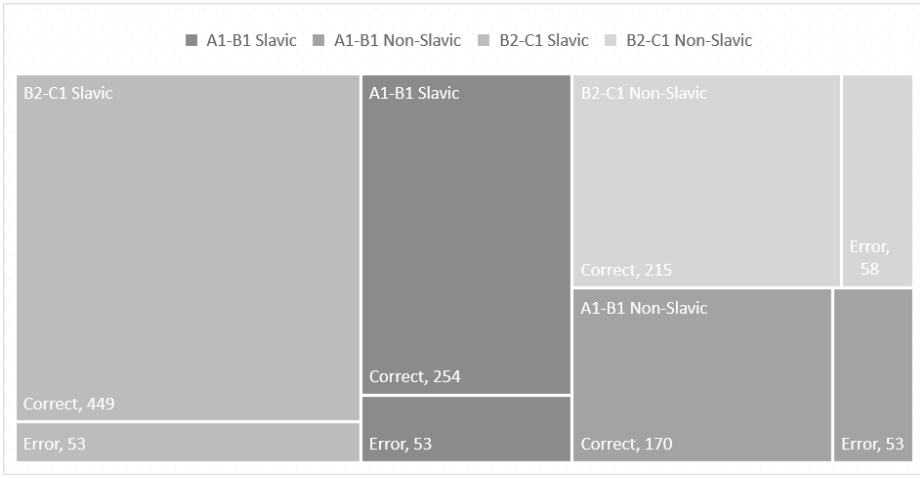



**Fig. 1.** The ratio of correct and erroneous usage in Slavic and Non-Slavic texts

The data show that, at both the lower and the higher-proficiency levels, Non-Slavic speakers produce more errors concerning enclitic elements than Slavic students: at A1-B1 levels more than every fourth enclitic component (4.2) is used erroneously in the Non-Slavic texts compared to almost every sixth erroneous component (5.8) in the Slavic texts; at B2 – C1 levels more than every fourth enclitic component (4.7) is used erroneously in the Non-Slavic texts as opposed to more than every ninth incorrectly placed component (9.5) in the Slavic texts.

### 4.2 Syntactic complexity of texts

The results presented in the previous section (4.1) should be elaborated against the background of phenomenon labelled as syntactic complexity, which is considered an indicator of overall level of L2 proficiency (cf. [19]). In numerous studies concerning second language acquisition, one aspect that syntactic complexity has been approached from was represented by subordination ratio, i.e., the ratio of embedded syntactic structures deemed to be developmentally or cognitively complex (e.g., sub-clauses) (cf. [20]).[8] Syntactic complexity in L2 is thought to expand from coordination to subordination and then to phrasal elaboration, as learners gain proficiency (cf. [27]). At beginner and low-intermediate proficiency levels, syntactic

[8] It has been claimed that, at initial levels, L2 learners start producing simple clauses organized around finite verbs (stage 1, cf. [21]), subsequent developmental stages involving the shift from simple clauses to the use of clause linking (stage 2), at advanced levels, the development involves noun and verb phrase elaboration, when information previously encoded as a clause is embedded (stage 3, cf. [22], [23]), and the highest stage is connected with sub-clausal complexification at the phrasal level, which is supposed to be characteristic of academic discourse and written prose (cf. [24], [25], [26]).

growth may show an increase in coordination (e.g., [28], [29]) and upper-intermediate levels are thought to display an increase in subordinate structures.

To verify theoretical assumptions on L2 development reflected in growing syntactic complexity, we calculated the ratio of simple clauses, compound and complex sentences in our sample texts for both language groups (see the following table).

| | | **JV** | **PS** | **HSV + HSH** | **IS** |
|---|---|---|---|---|---|
| | Non-Slavic | 61% | 16% | 21% | 2% |
| **A1_B1** | Slavic | 49% | 18% | 30% | 3% |
| | Non-Slavic | 51% | 11% | 37% | 1% |
| **B2_C1** | Slavic | 49% | 11% | 38% | 2% |

**Tab. 2.** Syntactic complexity in the Non-Slavic and Slavic texts at different proficiency level

The data from our investigation show three major tendencies: (a) a drop in the ratio of simple clauses at higher proficiency levels in the Non-Slavic texts[9], (b) a drop in the ratio of compound sentences at higher proficiency levels in both the Slavic and Non-Slavic texts, (c) and, at the same time, a rise in the ratio of complex sentences at higher proficiency levels in both the Slavic and Non-Slavic texts.

We also calculated the distribution of compound sentences at individual lower proficiency levels, as can be seen in the following figure.

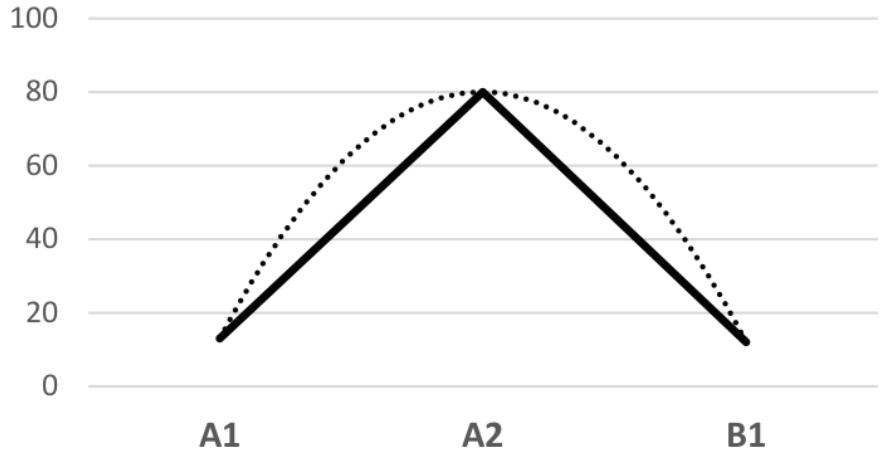


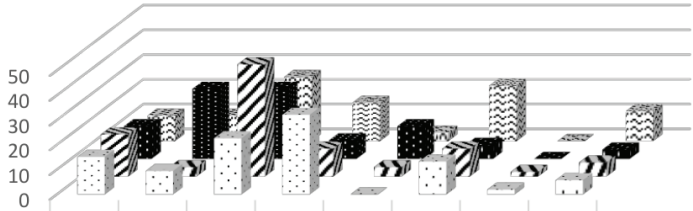

**Fig. 2.** Frequency distribution of compound sentences at individual levels

[9] Slavic texts display a relatively low number of simple clauses even at lower proficiency levels when compared with Non-Slavic texts.

The outcomes appear to confirm the developmental "omega-shaped" pattern suggested by Wolfe-Quintero et al. [28], which is to say a decrease in coordination at higher proficiency levels in favour of subordination. At the same time, our data correspond with the results of those studies which have indicated a rise in subordination rates at intermediate levels (e.g., [27]). Our data show a peak in coordination usage at A2 level and gradual decline in its usage beginning at B1 level.

### 4.3 Frequency distribution of errors with respect to enclitic type

Frequency distribution of errors with respect to enclitic type in the texts of Slavic and Non-Slavic students shows certain tendencies which can be visualized through the following figure.




| | COP | DEM | REF | PRON | PP | AUX | COND | CLUST |
|---|---|---|---|---|---|---|---|---|
| Slavic A1-B1 | 15,1 | 9,4 | 22,6 | 32,1 | 0 | 13,2 | 1,9 | 5,7 |
| Non-Slavic A1-B1 | 17 | 3,8 | 45,3 | 11,3 | 3,8 | 11,3 | 1,9 | 5,7 |
| Slavic B2-C1 | 13,2 | 28,3 | 28,3 | 7,5 | 13,2 | 5,7 | 0 | 3,8 |
| Non-Slavic B2-C1 | 10,3 | 10,3 | 25,9 | 15,5 | 3,4 | 22,4 | 0 | 12,1 |



**Fig. 3.** Error distribution in the texts of Slavic and Non-Slavic students

By comparing texts by Slavic and Non-Slavic speakers at A1 – B1 level, the most striking differences in frequency distribution can be grouped into two main categories:

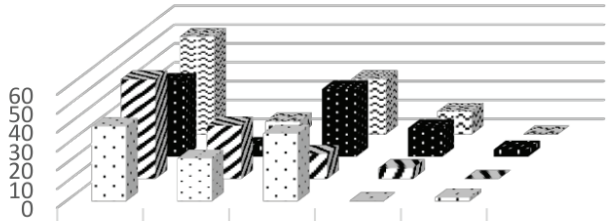
(a) argument pronominal and demonstrative clitic errors are more frequent in the texts of the Slavic students (for the Slavic A1 – B1 texts, the distribution of D, P and PP errors is 41.5%; for the Non-Slavic A1 – B1 texts it is only 18.9%), e.g., *Preto bolo* ***im*** *veľmi ľúto a mysleli, že vždy takto bude.* 'That's why they were very sorry, and they thought that it would always be that way.' (A2, Polish);

(b) verbal clitic errors (including R errors and G errors) are much more frequent in the texts of Non-Slavic students (35.8% in Slavic texts and 56.6% in the Non-Slavic texts), e.g., *A* ***sme*** *našli nejaké zaujímavé veci o čínskej a slovenskej kuchyni.* 'And we found out some interesting facts about Chinese and Slovak cuisine.' (B1, Chinese).

At higher proficiency levels, two tendencies are observed: (a) in Slavic texts, the distribution of P errors dramatically decreases in favour of an increasing number of PP and D errors, i.e., Slavic students produce more errors concerning volatile PP and D enclitics whereas the correct placement of P enclitics tantum is relatively acquired in their interlanguage (there is no striking difference between constant and volatile argument clitics in the texts of Non-Slavic students); (b) the ratio of errors concerning verbal clitics is still higher in the texts of Non-Slavic students (48.3%) when compared to Slavic students (34%), however, the difference is not so striking when compared to lower proficiency levels.

### 4.4 Frequency distribution of errors with respect to type of syntactic structure

The following figure features frequency distribution of errors with respect to type of syntactic construction.




| | JV | PS | HSV | HSH | IS |
|---|---|---|---|---|---|
| Slavic A1-B1 | 39,6 | 22,6 | 35,8 | 0 | 1,9 |
| Non-Slavic A1-B1 | 52,8 | 28,3 | 13,2 | 5,7 | 0 |
| Slavic B2-C1 | 39,6 | 5,7 | 35,8 | 15,1 | 3,8 |
| Non-Slavic B2-C1 | 51,7 | 6,9 | 29,3 | 12,1 | 0 |

Slavic A1-B1 Non-Slavic A1-B1 Slavic B2-C1 Non-Slavic B2-C1

**Fig. 4.** Error distribution with respect to syntactic structure

The data from Figure 4 are coherent with the results concerning syntactic complexity of texts at individual proficiency levels (i.e., the error rate in individual sentence types corresponds to the overall distribution of sentence types, cf. Table 2). At lower and intermediate proficiency levels, these findings emerged from the data:

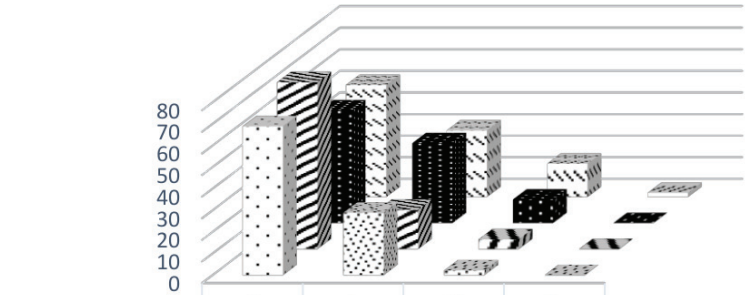
(a) the Slavic texts show the highest erroneous usage with respect to complex sentences, e.g., *Myslim si že pokoriš* ***ho***. 'I think that you will break it.' (A2, Ukrainian);

(b) in the Non-Slavic texts most errors occur in simple sentences, e.g. *Toto zaujalo detektiva, on myslel* ***si*** *že budu nasledovať ešte zločiný.* 'This interested the detective; he thought that more crimes would ensue.' (B1, German).

At upper-intermediate and advanced levels, the number of errors occurring in complex sentences in the Non-Slavic texts rises considerably (18.9% vs. 41.4%) which can be associated with the increasing syntactic complexity of the Non-Slavic texts. Both in the Slavic and Non-Slavic texts, there is a striking drop in the number of errors occurring in compound sentences, which can be explained by the "omega-shaped" pattern in the distribution of compound sentences (cf. 4.2). The number of errors occurring in simple sentences remains the same at lower and higher proficiency levels both in the Slavic and Non-Slavic texts.

### 4.5 Frequency distribution of errors with respect to distance from lexical/ syntactic host

The same comparison can be made with respect to the distance of enclitic components from their lexical/syntactic hosts as shown in Figure 5.




| | 0 | 1 | 2 | 3 |
|---|---|---|---|---|
| Slavic A1-B1 | 68,9 | 28,9 | 2,2 | 0 |
| Non-Slavic A1-B1 | 77,3 | 18,2 | 4,5 | 0 |
| Slavic B2-C1 | 52,2 | 37 | 10,9 | 0 |
| Non-Slavic B2-C1 | 51,9 | 30,8 | 15,4 | 1,9 |

**Fig. 5.** Error distribution with respect to the distance from host component

The data from Figure 5 show that in both the Slavic and Non-Slavic speakers' texts most errors relate to enclitics with 0- and 1-distance position. At elementary and lower-intermediate levels, the following tendencies can be observed:

(a) the errors occur in 0-distance in the Non-Slavic texts more often than in the Slavic texts, e.g., *Študenti a dôležití hostia sa zúčastnili a* ***sme*** *pozerali deväť krátke filmy.* 'Students and important guests took part, and we watched nine short films.' (A2, Italian);

(b) higher error distribution in 1-distance is observed in the Slavic texts, e.g., *Mne velmi* ***sa*** *páči cestovať, preto že najlepši deň pre mňa to je deň kedy začinaje cestovanije.* 'I like travelling very much, and because of this the best day for me is the day when a trip begins.' (A2, Ukrainian).

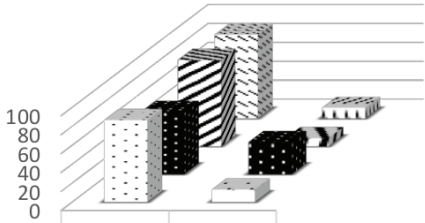
It can be linked to the fact that Slavic learners at these proficiency levels produce more complex syntactic structures (cf. 4.2) with enclitic components often put in more distant positions from their lexical/syntactic hosts which may cause a higher occurrence of 1-distance errors.

At upper-intermediate and advanced levels, there is a distinct decrease in erroneous placement of enclitics in 0-distance and an evident increase in 1-distance and 2-distance placement in both the Slavic and Non-Slavic texts. It seems that, in their interlanguage development, both Slavic and Non-Slavic learners are able to place the enclitic component in the proper position more often when it is adjacent to its lexical/syntactic host and the higher erroneous usage of enclitics relates to their distant, non-adjacent positioning.

At the same time, at higher proficiency levels, there is no striking difference between the Slavic and Non-Slavic texts, which corresponds to our findings according to which the syntactic complexity of the Non-Slavic texts is approaching the Slavic texts regarding interlanguage development, which is reflected in converging distribution of errors in the texts.

### 4.6 Frequency distribution of errors with respect to pre- and post-verbal position

Finally, the erroneous usage of enclitics is connected to the ability of learners to shift enclitics into pre-verbal position or to place them in post-verbal position. The results are presented in Figure 6.




| | Pre-verbal | Post-verbal |
|---|---|---|
| Slavic A1-B1 | 86,7 | 13,3 |
| Non-Slavic A1-B1 | 65,9 | 34,1 |
| Slavic B2-C1 | 91,3 | 8,7 |
| Non-Slavic B2-C1 | 88,5 | 11,5 |

**Fig. 6.** Error distribution with respect to pre-verbal or post-verbal position

As the data from Figure 6 show, the erroneous usage of enclitics is connected mostly with the ability to put them into preverbal position in the Slavic texts whereas the ratio of erroneous distribution in post- and pre-verbal position is more balanced in

the texts of Non-Slavic students, however, it holds true only for lower and intermediate proficiency levels. The higher frequency of erroneous usage in post-verbal position in the Non-Slavic texts can be caused by higher frequency of compound sentences in the texts by Non-Slavic students in which the proclitic conjunction, which requires postponing enclitic elements into post-verbal position, is often employed, e.g., *Niekedy vybuchne, stále odbieha od problémov a **si** nenechá čas na ich riešenie.* 'Sometimes he loses control; he's always ignoring his problems, not leaving himself the time to solve them.' (C1, English). At upper-intermediate and advanced levels, the distribution of errors is even more attracted to pre-verbal position both in the Slavic and Non-Slavic texts, e.g., *Povedala mi, že spoznali **sa** na diskotéke.* 'She told me that they had met at a club.' (C1, Polish). It can be explained by a decrease in compound sentence structures, which usually motivates post-verbal position of enclitic components.

## 5 DISCUSSION AND CONCLUSIONS

Based on the data (see Figure 1), at first glance, it seems that Slavic speakers at higher proficiency levels show progress (producing fewer errors), and conversely, non-Slavic speakers' word order acquisition more or less stagnates, i.e., almost every fifth (4.7) enclitic is still placed inaccurately in their texts (when compared to the lower proficiency level, where erroneous distribution concerns every fourth component (4.2)). However, investigation into development of syntactic complexity sheds new light on the issue. As indicated in Section 4.2, syntactic complexity varies in our sample texts with respect to proficiency levels, achieving a higher degree at upper-intermediate and advanced levels, which holds true for both Slavic and Non-Slavic text groups. Despite more syntactically complex nature of their texts, Slavic learners produce lower – and Non-Slavic learners produce comparable – amounts of errors. In other words, increasing syntactic complexity does not result in a greater number of enclitic errors. This finding leads us to the conclusion that acquisition of word order goes hand in hand with higher L2 proficiency in both language groups.[10]

Against the background of different syntactic complexity of the Slavic and Non-Slavic texts at lower levels of proficiency, the uneven frequency distribution of errors concerning pronominal enclitics can also be explained. As shown in Section 4.3 (see Figure 3), the ratio of argument enclitic errors is higher in Slavic speakers' texts than in the Non-Slavic ones. However, not only errors, but also the overall distribution of argument enclitics is higher in the texts of Slavic students. The more frequent usage of pronominal argument enclitics may be related directly to the higher ratio of complex sentences, since object pronouns (for 3rd person) require an antecedent to which the

[10] Worth noting is the fact that the degree of syntactic complexity in the Slavic texts is higher than in the Non-Slavic texts even at elementary and lower-intermediate levels. This issue goes beyond the scope of our study but opens up space for further comparative research into acquisition of Slovak by Slavic and Non-Slavic learners at early stages of their L2 development.

form is referring, so that this type of construction often requires more than one sentence, as concepts are developed across longer strings of language. At initial stages, the use of repetition as an avoidance strategy has proved dominant (cf. [30]) which seems to be the case of the Non-Slavic students' texts in our sample. However, the issue was not explored in more detail in our study and is open to further research.

At the upper-intermediate and advanced levels there is also a slight decrease in the number of copular enclitic errors both in the Slavic and in Non-Slavic texts (see Figure 3). At the elementary and lower-intermediate level, the copular *byť* 'to be' belongs to those basic verbs which are acquired at the early stage and is typical of high dominance in terms of frequency (cf. [31]). In other words, the higher frequency of copulas in general is reflected in higher distribution of copular errors at initial stages of L2 development.[11]

Finally, the data presented in Section 6 (cf. Figure 6) also point out that, in interlanguage development, there is a cognitive barrier blocking the correct preverbal position of enclitic components. Two possible explanations are at hand with respect to this phenomenon:

(i) As to verbal enclitics: the components with grammatical function are typically realized at the right periphery of the verb which corresponds with the investigation of affix ordering in Slavic languages (cf. [32]): affixes are realized in the order: prefix-basis-suffix-thematic marker-grammatical morphemes (this can be an explanation for the fact that placement of verbal enclitics with grammatical function is attracted to the right periphery of the verb in the texts, irrespectively of prosodic conditions in syntactic structures).

(ii) As to argument enclitics: as argument clitics usually fulfil object function, their typical post-verbal position can be determined by dominant SVO order which is characteristic of the majority of European languages within the Standard Average European area (cf. [33]) and in the production of a second language it is preferred by L2 learners regardless of basic word order in their native language (cf. [34, p. 87]).

**ACKNOWLEDGEMENTS**

This work was supported by the Slovak Research and Development Agency within contract No. APVV-19-0155 Language Errors in Slovak as a Foreign Language Based on Learner Corpus.

[11] Although the ratio of any kind of enclitic error decreases at higher proficiency level, this happens in favour of increasing frequency of another kind of error. With this fact in mind, it would not be quite correct to assume that, for higher proficiency speakers, the placement of enclitic component no longer constitutes a challenge to overcome. A good example are copular errors, the frequency of which declines compared to other enclitic errors at higher acquisition level, but does not decline with respect to the total number of copulas in texts (cf. 8 erroneously placed copulas out of 79 copulas for Slavic A1 – B1 texts, and 7 erroneously placed copulas out of 65 for Slavic B2 – C1 texts; 9 errors out of 67 copulas for Non-Slavic A1 – B1 texts, and 6 errors out of 28 copulas in Non-Slavic B2 – C1 texts).

## References


[1] Hana, J. (2007). Czech Clitics in Higher Order Grammar. Columbus: The Ohio State University, 281 p.

[2] Uhlířová, L. (1987). Knížka o slovosledu. Praha: Academia, 160 p.

[3] Ellis, R., and Barkhuizen, G. (2005). Analysing Learner Language. Oxford: Oxford University Press, 416 p.

[4] Mistrík, J. (1966). Slovosled a vetosled v slovenčine. Bratislava: Vydavateľstvo slovenskej akadémie vied, 276 p.

[5] Kačala, J. (2013). Syntagmatický slovosled v slovenčine. Martin: Vydavateľstvo Matice slovenskej, 220 p.

[6] Franks, S., and King, T. H. (2000). A Handbook of Slavic clitics. Oxford: Oxford University Press, 424 p.

[7] Beličová, H., and Uhlířová, L. (1996). Slovanská věta. Praha: Euroslavica, 277 p.

[8] Uhlířová, L., Kosta, P., and Veselovská, L. (2017). Klitikon. In P. Karlík, M. Nekula and J. Pleskalová (eds.), Nový encyklopedický slovník češtiny. Masarykova univerzita, Brno. Accessible at: https://www.czechency.org/slovnik/KLITIKON.

[9] Starý Kořánová, I. (2017). Příklonky a vazaly infinitivu. Studie z aplikované lingvistiky, 11 (special issue), pages 109–120.

[10] Omnia Slovaca III Maior. Accessible at: http://ske.juls.savba.sk/bonito/index.html, (Cited 17.03.2021).

[11] Avgustinova, T., and Oliva, K. (1995). On the nature of the Wackernagel position in Czech. In U. Junghanns and G. Zybatow (eds.), Formale Slavistik, pages 25–47, Vervuert, Frankfurt am Main.

[12] Junghanns, U. (2002). Clitic climbing im Tschechischen. Linguistische Arbeitsberichte, 2002(80), pages 57–90.

[13] Kosek, P. (2011). Enklitika v češtině barokní doby. Brno: Host, 360 p.

[14] Palková, Z. (1994). Fonetika a fonologie češtiny. Praha: Karolinum, 367 p.

[15] Slovenský národný korpus – errkorp-0.1. Bratislava: Jazykovedný ústav Ľ. Štúra SAV 2019. (Pre-pilot version not accesible for public.).

[16] Ivanová, M., Kyseľová, M., and Gálisová, A. (2021). Acquiring word order in Slovak as a foreign language: comparison of Slavic and Non-Slavic learners utilizing corpus data [Data set]. Zenodo. Accessible at: http://doi.org/10.5281/zenodo.4625429.

[17] Cohen, J. (1960). A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 20(1), pages 37–46.

[18] Glen, S. "Cohen's Kappa Statistic" From StatisticsHowTo.com: Elementary Statistics for the rest of us! Accessible at: https://www.statisticshowto.com/cohens-kappa-statistic/, (Cited 05.03.2021).

[19] Ortega, L. (2003). Syntactic Complexity Measures and their Relationship to L2 Proficiency: A Research Synthesis of College-level L2 Writing. Applied Linguistics, 24(4), pages 492–518.

[20] De Clercq, B., and Housen, A. (2017). A Cross-Linguistic Perspective on Syntactic Complexity in L2 Development: Syntactic Elaboration and Diversity. The Modern Language Journal, 101(2), pages 315–334.

[21] Klein, W., and Perdue, C. (1997). The Basic Variety (or: Couldn't natural languages be much simpler?). Second Language Research, 13(4), pages 301–347.

[22] Byrnes, H. (2009). Emergent L2 German writing ability in a curricular context: A longitudinal study of grammatical metaphor. Linguistics and Education, 20(1), pages 50–66.
[23] Ortega, L. (2012). Interlanguage complexity: A construct in search of theoretical renewal. In B. Kortmann and B. Szmrecsanyi (eds.), Linguistic complexity: Second language acquisition, indigenization, contact, pages 127–155, De Gruyter, Berlin.
[24] Biber, D., Nekrasova, T., and Horn, B. (2011). The Effectiveness of Feedback for L1-English and L2-Writing Development: A Meta-Analysis. ETS Research Report Series, 2011(1), pages i–99.
[25] Biber. D., and Gray, B. (2011). Grammatical change in the noun phrase: The influence of written language use. English Language and Linguistics, 15(2), pages 223–250.
[26] Norris, J. M., and Ortega, L. (2009). Towards an Organic Approach to Investigating CAF in Instructed SLA: The Case of Complexity. Applied Linguistics, 30(4), pages 555–578.
[27] Bardovi-Harlig, K., and Bofman, T. (1989). Attainment of Syntactic and Morphological Accuracy by Advanced Language Learners. Studies in Second Language Acquisition, 11(1), pages 17–34.
[28] Wolfe-Quintero, K., Inagaki, S., and Kim, H.-Y. (1998). Second Language Development in Writing: Measures of Fluency, Accuracy, and Complexity. University of Hawai'i, Second Language Teaching and Curriculum Centre, Honolulu.
[29] Vyatkina, N. (2012). The Development of Second Language Writing Complexity in Groups and Individuals: A Longitudinal Learner Corpus Study. Modern Language Journal, 96(4), pages 576–598.
[30] Amzallag, O. (2019). 3rd Person Object Pronoun Use in French Beginning Textbooks. Journal of Second and Multiple Language Acquisition – JSMULA 7(1), pages 15–47.
[31] Viberg, Å. (2002). Basic Verbs in Second Language Acquisition. Revue française de linguistique appliquée, 7(2), pages 61–79.
[32] Manova, S. (2015). Affix Order and the Structure of the Slavic Word. In S. Manova (ed.), Affix Ordering Across Languages and Frameworks, pages 205–230, Oxford University Press, New York.
[33] Haspelmath, M. (2001). The European linguistic area: Standard Average European. In M. Haspelmath (ed.), Language typology and language universals. (Handbücher zur Sprach- und Kommunikationswissenschaft), pages 1492–1510, de Gruyter, Berlin.
[34] Odlin, T. (1997; first published 1989). Language Transfer. Cross-linguistic influence in language learning. Cambridge: Cambridge University Press, 210 p.