DOI: https://doi.org/10.31392/NPU-nc.series9.2018.18.04
UDC: 81'0

Liudmyla H. Halii
National Aviation University,
Kyiv, Ukraine

TYPOLOGICAL STRUCTURE AS A MODEL TO STUDY LANGUAGES IN CONTRAST

Bibliographic Description:
Halii, L. H. (2018). Typological Structure as a Model to Study Languages in Contrast. Scientific Journal of National Pedagogical Dragomanov University. Series 9. Current Trends in Language Development, 18. 46–60. DOI: https://doi.org/10.31392 /NPU-nc.series9.2018.18.04

Abstract
The article studies the typological system of languages belonging to Indo-European language family and the influence of native language interference on creative written productions of English language learners. It has been defined that the components of linguistic complexity in writing follow different developmental trajectories related to different levels of language proficiency, interference being observed not only throughout closely related languages. Based on this assumption linguistic complexity has been studied in the aspect of the native language transfer and the languages have been classified according to the typological similarity of language patterns but not according to language family relation. It has been proved that typologically similar languages belonging to the same language group or family cause the same mistakes in the process of ESL studying, specifically while producing complex speech structures. The last stage of the research involves the analysis of the native language influence on English creative written productions depending on the proficiency level of the producer. Finally, languages have been classified into clusters which have the same characteristics (morphological and syntactical) in their influence on ESL studying and a new model to study language interference in contrast has been proposed.

Keywords: *interference, creative productions, language typology, hierarchical clustering.*

1. Introduction.

The importance of learner corpus-based research has been always strongly emphasized in terms of second language (L2) learning and teaching. The learners' written production data and its analysis can help to reveal the patterns that are shared among different learners, demonstrate the first language (L1) influence on the acquisition process, detect the most complicated stages in the second language mastering for individual learners and learners with the common language background. Such knowledge is a crucial part of the effective teaching methodologies, tools for second language learning, and resources development.

The analyses of such data can shed the light on how and when specific L2 structures are being acquired and how the first language shapes such developmental curves. However,

it should be highlighted that such large-scale learner corpora are still scarce (Granger et al., 2007). The number of errors produced while studying foreign language and the difficulty experienced in the process were first studied by the school of Contrastive Analysis, with its memorandum stated by Stockwell (Stockwell et al., 1965). The founding principle of the researchers at the time was to enable machine, algorithm or didactic studies to predict negative transfer by means of comparing the linguistic systems of two languages. Thus, the Contrastive Analysis group stated that the main source of errors produced and learning difficulties experienced is L1 interference.

In the 2nd part of the 20th century the interest towards the Contrastive Analysis gradually declined as it proved unable to answer to questions stated due to the lack of theoretical background and empirical studies and experiments that could predict errors produced and difficulties experienced in the course of L2 studying. Another question stated by researchers of that period was the similar mistakes observed in learners with different L1 background. This problem was addressed by Peck (Peck, 1978), Schumann (Schumann, 1979), Odlin (Odlin, 1989), Klee and Ocampo (Klee & Ocampo, 1995). Hyltenstam (Hyltenstam, 1977) supposed that better results and more profound explanations could be gained at studying one interconnected language space instead of several unrelated ones, marking the beginning of language-pairs studying dominating over attempts to build all-inclusive language structure.

2. Literature Review.

Most recent illustrations in the field discuss the degree of L1 influence on L2 acquisition regardless of the proficiency level of the learner or within one level of proficiency (e.g. elementary learners, advanced learners, etc). Having in mind that manual analysis of such data is almost unfeasible, nowadays researcher make use of Natural Language Processing techniques such as POS-tagging, lemmatization, parsing, discussing the developmental trajectories of English grammatical morphemes (Lee, 2015; Murakami, 2014), relative clauses (Alexopoulou et al., 2015) and the developmental paths of the English article accuracy (Murakami & Alexopoulou, 2016), which again leads us to the idea of wholly comprehensive research instead of describing separate language phenomena, however complex they might be.

The implementation of the aforementioned studies and other works in the same field the efficiency of linguistic complexity measures while creating readability classification of texts in specific language (Hancke et al., 2012) with accuracy of almost 90% or classifying different age groups (Vajala & Meurers, 2014) with accuracy of 95,9%. Age groups and linguistic complexity are also under study in some methodological researches and interlinguistic studies (Paradis, et al., 2017). In addition, linguistic complexity features are used in the systems that are designed for written production scoring: e-rater (electronic essay rater) system integration (Attali & Burnstein, 2006); integration in proficiency level assessment of the texts produced by learners of specific L2 (Vajala and Loo, 2013); Native Language Identification task integration (Bykh & Meurers, 2016; Bykh, Vajala, et al., 2013); partially used in related fields – as a part of studies devoted to neurodegenerative disorders (Pakhomov et al., 2011).

In this paper we address not only the question of second language development but also L1 transfer. The principle of Transfer to Somewhere (Andersen, 1983) predisposes that both L1 and L2 have impact on transfer. This idea was further developed in the study by Klee and Ocampo (1985) and can be used in other multilanguage researches exemplifying the thesis that language learners tend to adapt structures of L2 to make it more similar to L1 which can have both positive and negative impacts.

The idea that typologically similar languages (languages that belong to one family or group) can be the key factor of producing the similar mistakes by the learners as they are alike in their influence on English as second language in terms of complexity is also the subject of study in our research.

It is predicted that according to their impact on the second language acquisition native languages may be clustered not necessarily according to the distance of their language family relations, but rather according to the inner laws and patterns which coincide or don't coincide with the patterns of English.

Another hypothesis is that level of learners' proficiency may be more important factor considering the degree of language transfer than relation between L1 and L2, which means the more fluent the learner is in L2 the less likely he or she is to produce errors or use deviations in his speech production.

3. Aim and Objectives.

The **aim** of the article is to define the impact of L1 on the linguistic complexity of the learner, forming cluster structure of related and unrelated language pairs. In order to achieve this aim, the following **objectives** are to be achieved:

– to define linguistic complexity as the measurable unit;

– to study the realisation of linguistic complexity at different language levels;

– the study the learners' acquisition, error production and language proficiency at different levels of studying English;

– to compare and contrast language pairs using hierarchical clustering technique.

4. Methodology.

To conduct the aforementioned analyses we use new longitudinal data source – EF-Cambridge Open Language Database (EFCamDat) that comprises 1 180 453 essays collected from 174 771 different learners from 209 countries. The essays represent different levels of proficiency. On average every student produced 7 scripts. The background information to every script includes id of the script, id of the learner, id of the topic, nationality of the learner, date of submission, received grade, level, unit number, unit title, lesson title, lesson aim.

Furthermore, 69% of the essays have an error annotation. However, in the supporting article to the database there is no information on how reliably and systematically different types of errors have been marked by correctors (Geerzen et al., 2013).

This number of submitted essays is then reduced to exclude the same essays (cases of copied texts). Then the essays where the pair nationality-first language could not be defined were also taken out of the corpus. That made 1 156 252 scripts. At this stage, our data comprised 113 languages where Portuguese, Chinese, Spanish, Russian, Arabic and German were consuming the largest part of it. Moreover, the majority of L1 in EFCamDat represented Italic and Slavic groups of Indo-European language family (71,2% of data). In order to have a solid picture that is not learner dependent, we considered instances with specific L1 that appear at least 100 times in our corpus. Ultimately, we selected 58 languages and 14 language families.

Table 1

The number of essays with different L1 background under study

Language	Number of essays submitted	Language	Number of essays submitted	Language	Number of essays submitted
Portuguese	465098	Hungarian	760	Norwegian	244
Chinese	196108	Slovak	723	Irish	224
Spanish	127899	Romanian	722	Albanian	212
Russian	70036	Greek	632	Kyrgyz	203
Arabic	58247	Finnish	610	Pashto	194
German	58046	Latvian	570	Serbian	184
Italian	44516	Azeri	539	Hebrew	179
French	42755	Belarusian	533	Danish	176
English	23442	Swedish	525	Armenian	171
Japanese	21711	Lithuanian	523	Moldovan	163
Turkish	14316	Mongolian	513	Haitian	152
Korean	5554	Czech	505	Malagasy	133
Indonesian	3021	Malay	425	Uzbek	128
Thai	2251	Farsi	392	Afrikaans	126
Ukrainian	1637	Filipino	337	Bosnian	124
Vietnamese	1570	Urdu	334	Bulgarian	110
Dutch	1476	Estonian	293	Sinhalese	106
Kazakh	1337	Georgian	275	Emakhuwa	101
Polish	1283	Slovenian	271		
Hindi	979	Croatian	252		

The overall algorithm is as follows: 1) form the dataset for investigation minding the first language of the speaker, originality of the written speech production, and number representation of essays submitted; 2) define the features for the analyses of linguistic complexity; 3) define the components of linguistic complexity and levels of their realization in speech; 4) analyze the essays according the aforementioned features via WEKA software (see table 2); 5) cluster the L1s that have alike patterns with L2; 6) study the development of linguistic complexity and the degree of L1 transfer through different levels of proficiency.

While structuring, grouping, clustering and analyzing the material in the course of our study we make use of structure identification techniques. First, we use clustering in order to group a set of data points which is closely related to unsupervised type of machine learning as the input data is not labeled (Shalev-Shwartz & Ben-David, 2014, 311; Witten et al., 2016, 81) (see Fig. 1).

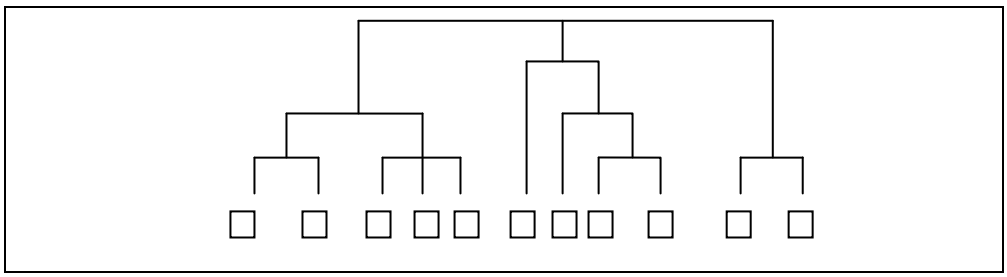

Figure 1. Representation of clusters (Witten et al., 2016)

The type of clustering which fits machine learning algorithm is hierarchical one with two prevailing forms: agglomerative (bottom-up approach) represented by Witten and divisive (top-down approach). Agglomerative type matches each unit under study to its own

cluster, after that the distance between clusters is computed and two most similar ones are merged at the bottom of the dendroid. Likewise, the higher pairs of clusters are merged at the higher levels of the hierarchy. Divisive type assigns all units under study to a single cluster and then split the cluster to two least similar clusters.

The approach represented by Mooi and Sarstedt (see Fig. 2) assumes that the type of the clustering basically depends not on the algorithm or features of the analysis, but on its direction while both types represent one and the same technique.

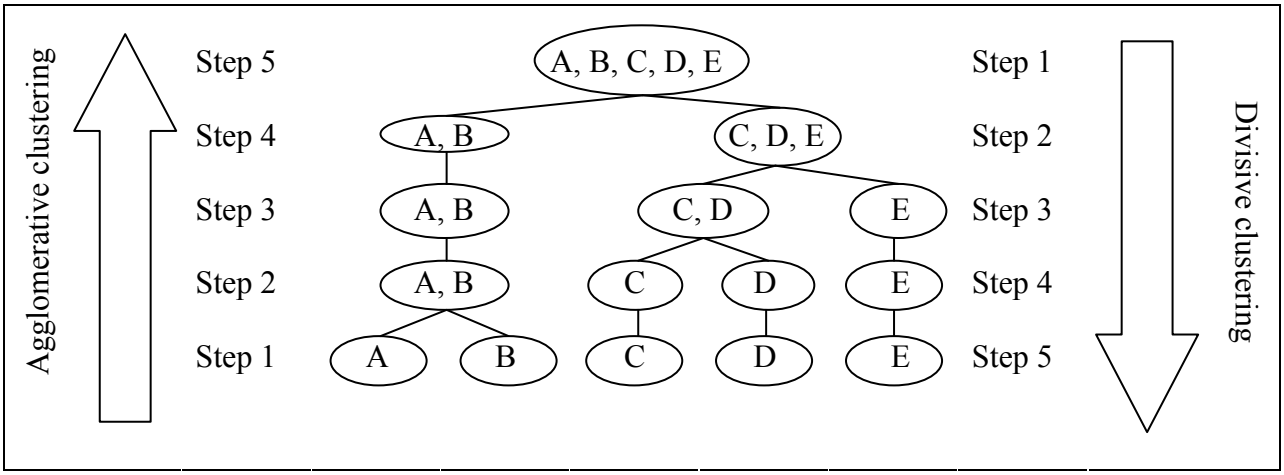

Figure 2. Agglomerative and divisive clustering (Mooi and Sarstedt, 2010, 20)

For determining the distance measure most commonly are chosen Euclidean and Manhattan equations:

Euclidean = $\sqrt{\sum}$ (ai – bi)2

Manhattan = $\sum$ |ai - bi|

However, we admit the fact that distance metrics can have different influence on clusters as, according to Pandit and Gupta (Pandit and Gupta, 2011, 29-30), some data points can be close according to one measure and far away to another, hence, the normalization of data is essential for equal data contributing. The easiest and most widely used way of doing this is z standardization where each variable has a mean 0 and standard deviation 1, however better performance is usually observed at the range 0,1 or -1,1 (Mooi and Sarstedt, 2010, 38).

5. Results.

As it can be inferred from the title of our paper, the research deals with two main concepts – language transfer and linguistic complexity. According to Jarvis and Odlin (Jarvis & Odlin, 2000) transfer is the influence resulting from similarities and differences between the target language and any other language that has been previously and perhaps imperfectly acquired.

First language transfer refers to the process of applying knowledge from the first language to the target language in speaking and writing. Transfer can have an impact on all dimensions of the target language (e.g. syntax, morphology, lexicon, etc). According to Ellis (1994, 101-120) transfer study involves the studies of negative (i.e. errors) and positive (i.e. facilitation) transfer, avoidance (underproduction) and overuse (overproduction) of the target language forms. Some studies also involve production errors and misinterpretation, however these two aspects may be studied only with the help of cognitive linguistics methods.

Let's consider the levels of transfer realization more thoroughly. Negative transfer is usually associated with the transfer of items or structures that are not alike in both languages.

According to Lado (1957, 23-35) the degree of difficulty experienced by language learners lies in a difference between first and target languages.

In our study while analysing the data via WEKA software we came across a great number of out-of-vocabulary words, i.e. tokens that NLP tool has not seen during the training (usually NPL systems are trained on newspaper corpora) – the fact that is easily explained, as learner language contains different deviations from standard language which in turn depend on different learner variables (e.g. L1, age, level of proficiency). Based on study conducted by Keeper (Keeper et al., 2016) and in order to avoid decreasing quality of POS tagging, we transferred this part of the data from misspellings to the normalized correct forms. Leaving these out-of-vocabulary units in the general amount of words analysed by WEKA software would lead to decreasing or increasing indexes of lexical sophistication level as the words would be treated not as unknown but as rarely-used, thus changing the real index of lexical proficiency. The words with improper capitalization (e.g. "i'm") were not treated as misspelling in our study as many scripts were improperly capitalized in general (e.g. fully capitalized or non-capitalized at all). All these manipulations required manual pre-processing of the error annotated essays with solving of the aforementioned issues. The final list consists of 312 tokens which gave us the opportunity to normalize 48 413 words, out of which only those who number more than 100 are represented in this study.

Table 2

The list of misspellings and the number of their occurrences

Item	*Number*	*Item*	*Number*	*Item*	*Number*	*Item*	*Number*
dont	9935	diferent	375	comming	252	tomatos	127
fourty	3144	resturant	369	apartament	236	scool	127
principal	2195	doughter	366	promiss	234	choise	125
studing	1337	peaple	357	responsability	230	shool	123
becouse	910	colum	353	rainny	223	swimm	121
coffe	828	teacher	343	beachs	221	necesary	118
nigth	809	happend	337	succesful	205	questionnair	118
swiming	800	beacause	325	goverment	197	holyday	118
departament	788	litle	320	exemple	193	thankyou	118
realy	781	sweter	320	awfull	181	tommorow	115
successfull	758	wonderfull	315	becuase	179	sity	115
isnt	757	freind	313	blu	178	therefor	110
companys	689	wether	313	dollares	177	color	109
alot	601	yong	312	foward	177	shold	108
adress	575	enviroment	305	potatos	176	vaccum	106
collegue	531	verry	286	stright	169	occured	106
yers	444	yars	280	holliday	156	immediatly	105
lifes	424	begining	276	delicius	153	atention	105
tomorow	419	beutiful	269	yelow	144	nowdays	104
recomend	416	recieve	266	programm	140	trainning	103
writting	413	tecnology	265	jewelery	138	excercise	101
finaly	411	untill	265	sucessful	137		
shoud	393	pepole	265	daugter	131		
sincerly	384	prefered	254	remeber	128		

Cases of positive transfer can be hard to identify (Ortega, 2013, 42). According to Odlin (1989) cross-linguistic similarities can lead to the following positive transfers:

– vocabulary similarities can lead to the time reduction to achieve good reading comprehension;

– vowel system similarities can make the process of vowel sounds identification easier;

– writing system similarities can make it easier to read and write in the target language;

– syntactic structures similarities can promote grammar acquisition.

The case of avoidance according to Schachter (1974, 193-213) can be exemplified by the cases when learners tend to avoid using specific structures different from their L1 background.

According to Ellis (1994) and Odlin (1989) overuse can be a result of avoidance, that is when learners avoid using structure uncommon for their L1 and instead of it overuse the structure that is typical for both L1 and L2. To detect overuse Ellis (1994, 761) suggests involving groups of learners with different L1s, which is just the case with our study.

The concept of linguistic complexity is used as a basic indicator of the second language development. Researchers define complexity as component, and the most difficult one, of Complexity-Accuracy-Fluency triad that assesses learners' language development and proficiency (Ebrahimi, 2015, 118; Housen & Kuiken, 2009, 461; Kuznik & Ollala-Soler, 2018, 20; Lintunen & Makila, 2014, 378; Palotti, 2009, 590-592; Shekan, 2009, 510-511; Timpe-Laughlin, 2016; Vyatkina et al., 2015).

If L2 complexity is considered in terms of language performance it can be represented by cognitive complexity and linguistic complexity. According to Housen and Kuiken (2009, 464), cognitive complexity is learner-centered and refers to the difficulty of the language features such as processing and acquisition, whereas linguistic complexity (which is of mostly of interest in our study) refers mainly to the system of the language being studied. Palotti (Palotti, 2015) bases his studying of the linguistic complexity on different linguistic structures used by learners of L1 and L2, whereas Eliot (Eliot, 2009, 474-476) defines complexity as the capacity to use more advanced language.

Linguistic complexity is multidimensional by its nature and traditionally three main components are defined – lexical, syntactic and morphological (Bulte & Housen, 2015, 46; Palotti, 2015, 121-123). Considering the fact that the native language background targets all L2 dimensions, we trace L1 influence on lexical, syntactic and morphological levels. According to Bulte and Housen (Bulte & Housen, 2015, 53), only the following concepts of linguistic complexity are represented in L2 linguistic complexity and mirror the rise in complexity level:

– more phonemes, inflectional forms, derivation, etc.;

– longer linguistic units (clauses, sentences, etc.);

– more deeply embedded units (recursion, subordination, etc.);

– more varied or diverse lexical items;

– more marked, infrequent, sophisticated, cognitively difficult or later acquired features.

The studies aimed to uncover the relationship between linguistic complexity and second language acquisition, development and proficiency were conducted by Xiaofei Lu (Lu, 2011; Lu 2012) and a number of other theorists (Gleitman L.R., et al., 2019, 9), who split it into two dimensions – syntactic and lexical. The theory further developed (Ai & Lu, 2013; Lu & Ai, 2015) provides us with a theoretical background in our study, indicating that significantly different measures observed at different levels of L2 development can predict the quality of learner's production and errors produced.

In our study the idea of linguistic complexity is a part of contrastive analysis idea implemented by computational linguistics approach as complexity is mainly used for proficiency evaluation, performance assessment and developmental level benchmarking (Ortega, 2012).

The basic measures of assessing linguistic complexity components are lexical richness, syntactic complexity and morphological complexity. As the theory goes, the notion of lexical richness is concerned with how many different words and what types of words are produced

in spoken and written production. This multidimensional feature of the language use is composed of the following interrelated components: lexical density, lexical variation (also called lexical diversity or lexical range) and lexical sophistication (also called rareness) (Read, 2000, 188-221).

Previous research in this field has demonstrated that spoken texts are disposed to have a lower lexical density comparing to the written ones having value of 40% or even higher (Halliday, 1985; Ure, 1971). Furthermore, it was reported that language learning materials on the Internet have a higher lexical density that in traditional textbooks (Kong, 2009, 38-47). It has been suggested that the reason for this is the website longing to diminish difficulty of possessing online texts, as a result, the number of sentences in each paragraph is reduced and the processing difficulty is unintentionally increased, due to the fact that more content is packed into a single sentence (Kong, 2009, 48-51).

The notion of lexical variation refers to the range of different lexical items used in specific texts. Basically, the written or spoken production has a high index of lexical variation, if the writer or speaker uses many different words and the percentage of word repetition is very low.

Lexical sophistication can be referred to the number of advanced or sophisticated words used in a text. The notions of advanced and sophisticated items strongly depend on the frequency list or lists of basic words used in the study.

The notion of syntactic complexity refers to the range and degree of sophistication of syntactic structures produced. Syntactic complexity itself takes a significant place in second language research, as the growth of syntactic repertoire and its appropriate usage is an integral part of a learner's development in second language (Ortega, 2003).

Lu (2011) grouped syntactic complexity measures into five categories:

– length of production;

– sentence complexity (clauses to sentences ratio);

– subordination: clauses per T-unit, complex T-units per T-unit, dependent clauses per clause, dependent clauses per T-unit;

– coordination: coordinate phrases per clause, coordinate phrases per T-unit, T-units per sentence;

– particular structures: complex nominals per clause, complex nominals per T-unit, verb phrases per T-unit.

De Clercq and Housen (2016) claim that native speaker level in morphological complexity can be approached by learners of English quickly enough. Morphological features can be split into: general features, segmentation features, stem allomorphy features, derivational transformation features, morphological analysis status, noun-verb-affix compound features in all their various levels of realization.

All the aforementioned features make up the list of items that form the grounds for analyses. For this purpose two feature selection algorithms provided by WEKA were used – *CfsSubsetEval* and *InfoGainAttributeEval*. CfsSubsetEval is a correlation based selector that assesses the predictive ability of each attribute individually and prefers sets of features that are highly correlated with the class but have low intercorrelation (Witten et al., 2016, 422). InfoGainAttributeEval measures the information gain of attributes with respect to the class (Witten et al., 2016, 393, 422).

The difference between selected features by two algorithms is basically insignificant.

Table 3

Feature selection on the whole data

CfsSubsetEval	InfoGainAttributeEval
Lexical complexity	
Conjunction density	Conjunction density
Determiner density	Determiner density
Modifier variation	Modifier variation
Modal verbs density	Modal verbs density
Past participle verb density	CTTR
RTTR	RTTR
Lexical sophistication	MTLD
Syntactic complexity	
Clauses per sentence	Clauses per sentence
Sentences	Sentences
Constituents per sentence	Constituents per sentence
Noun phrases per sentence	Noun phrases per sentence
Verb phrases per sentence	Verb phrases per sentence
Subordinating conjunctions per sentence	Subordinating conjunctions per sentence
T-units per sentence	T-units per sentence
Average parse tree height	Average parse tree height
Average sentence length	Average sentence length
Prepositional phrases per sentence	Prepositional phrases per sentence
Clauses per T-unit	Complex T-units per T-unit
Wh-phrases per sentence	Subtrees per sentence
Morphological complexity	
Words of foreign origin	Words of foreign origin
Words not found in CELEX	Words not found in CELEX
Noun-verb-affix compounds, derivations	

Consequently, for syntactic and lexical complexity we selected attributes that occurred at least three times in the sets obtained by WEKA, for morphological complexity we included measures that appeared twice – words with stem with unmarked transitivity and words with derivational transformation.

The list of features for analyzing the essays that belong to the learners with the higher rates of proficiency should be modified in order to reveal the picture of error-producing more accurately.

Table 4

Feature selection on the higher proficiency level subset

CfsSubsetEval	InfoGainAttributeEval
Lexical complexity	
Conjunction density	Conjunction density
Determiner density	Determiner density
Modifier variation	Modifier variation
Lexical sophistication	Modal verbs density
Sophisticated token ratio	
Syntactic complexity	
Subtrees per sentence	Subtrees per sentence
Clauses per sentence	Clauses per sentence
Sentences	Sentences
Constituents per sentence	Constituents per sentence
Noun phrases per sentence	Noun phrases per sentence
Subordinating conjunctions per sentence	Subordinating conjunctions per sentence
T-units per sentence	T-units per sentence
Average sentence length	Average sentence length

	Verb phrases per sentence
	Clauses per T-unit
	Wh-parses per sentence
Morphological complexity	
Words of foreign origin	Words of foreign origin
Words not found in CELEX	Words not found in CELEX
Verbal stem with unmarked transitivity	Verbal stem with unmarked transitivity
Words with derivational transformation	Words with derivational transformation

Based on the results we come to a number of conclusions. It seems that Sino-Tibetan and Altaic families have similar patterns in linguistic complexity. Speakers of Baltic language (i.e., Lithuanian and Latvian) are very close in English written production in terms of linguistic complexity. Moreover, we observe that Baltic and Slavic languages from different language groups have similar influence on the complexity. Moldovan, Romanian, and Portuguese have alike influence in linguistic complexity to Farsi, Hindi, and Sinhalese.

We observe that native speakers of Germanic languages have similar patterns in English complexity, especially learners with Dutch and German, Danish and Norwegian background. Data analysis shows that French and Italian are closely connected and result in a very similar influence on English written production.

Further insights into the database reveal that proficiency level has more profound effect on linguistic complexity than L1 background of learners and is more influential. In our study we use the representation based on the development curve. The developmental curve shows a relation between the language learner proficiency and the amount of errors produced. Theis language transfer can be represented on the developmental curve in two ways. First, it is the direct relation between the proficiency level of the learner and his/her fluency in L2. It is based on the logical assumption that the better the learner know the language and the higher his/her level is, the less mistakes are produced at each level, meaning the acquisition of language skills is getting easier and easier when it is built upon the knowledge of L2. The second way of developmental curve representation is based upon the notion "U-shaped behavioural development" first introduced by Kellerman (Kellerman, 1985). It denotes a process of L2 learning, when the learner production is error-free at the early stage, then the deviation from the target norm is observed and finally the use of the feature is correct. Such a developmental curve can be explained by the fact that at first stages learners make use of corresponding forms of their first language, i.e. L1 has a facilitative effect at the first stages of acquisition.

The number of conclusions inferred from the analysis of the database subset with regard to the proficiency level of the learners is as follows. Through all the levels Arabic and Hebrew have alike patterns, which means that the languages of Afro-Asiatic family have similar linguistic complexity patterns independently of the proficiency level. Regarding Altaic family and the level differences, we infer that the distance between languages of this group is getting smaller with every level.

The next language family under examination was Austronesian. Even though Filipino, Indonesian, Malay and Malagasy belong to one family, learners of English with these L1 have absolutely different performance in terms of language complexity.

Finnish, Estonian and Hungarian that are of Uralic family also perform independently. However, it is worth noticing that within levels A1-A2-B1-B2 of European Framework Finnish and Estonian have short distances, it means that some patterns of linguistic complexity can overlap. Unfortunately, we are not able to state whether the relation between these languages is retained at the advanced levels, because we do not have essays that correspond to learners with Estonian background.

It should be noted that across all six levels (A1-C2) Kazakh language has very short distance to East-Slavic languages – Ukrainian and Russian. This can be explained by the fact that the Russian language is one of the official languages in Kazakhstan, so that both languages (Kazakh and Russian) have impact on English written production in Kazakhstan.

East-Slavic languages (i.e., Ukrainian, Russian, Belorussian) independently of the proficiency level have extremely short distance between each other.

Germanic group of languages across A and B levels with exception of Swedish, has alike patterns of written speech production. Moreover, Dutch and German, that represent West-Germanic group, seem to have very similar impact on complexity in written production. We should admit that at the levels A1-A2 and B1-B2 the performance of Swedish learners of English is absolutely different from Germanic languages.

The learners that are native speakers of Baltic languages have similarities in complexity only at A1-A2 levels, however, it is hard to draw conclusions about C1-C2 levels, because we do not have essays submitted by learners with Lithuanian as L1.

Native speakers of Italic languages at the levels B1-B2 and C1-C2 have some similarities in their production. Notably, at the B1-B2 and C1-C2 levels not only Italian and Portuguese are connected, but also Moldovan and Spanish have fairly short distance.

According to the obtained data native speakers of Indo-Iranian languages tend to have similar patterns in written production only at the A1-A2 levels. Furthermore, it is worth noticing that Albanian and Bosnian languages that represent different groups are connected across all proficiency levels. Consequently, we can assume that these L1s have alike effect on English writings.

6. Discussion.

Considering the aims and the methodology of analyses, we explored the interrelations of L1s within six levels of proficiency (A1-C2). Our clustering experiment revealed that the production of learners with typologically similar L1s (i.e. belonging to one language family or group) has the following interrelations:

– alike patterns in English linguistic complexity regardless of the proficiency level (Afro-Asiatic family or East-Slavic group of Indo-European family);

– alike patterns in English linguistic complexity at the intermediate level (Austronesian and Uralic family);

– the distance increases with every level of proficiency (Baltic or Indo-Iranian groups of Indo-European family, emphasizing the use of forms and structures shared within one family or group);

– the distance decreases with every successive level of proficiency (Altaic family, Italic group of Indo-European family, some languages of Germanic and Slavic groups of Indo-European family, implying that at the C levels the typologically similar L1s have alike influence on linguistic complexity).

Additionally, we observed the development of linguistic complexity components with regard to the first language. In general, the learners with different L1 background have common tendencies in the development of linguistic complexity; however, at some stages of acquisition the rates of some structures are higher or lower for specific languages.

Further works in this area can be dedicated to a more detailed comparison of different L1 and L2 structures that can assist in the research on how the presence or absence of L1 structures can affect the developmental trajectory. Moreover, the knowledge about the differences in L2 linguistic complexity development of learners that have mastered one or more foreign languages or even bilingual learners can be beneficial in educational resource development and teaching.

7. Conclusions.

In this paper we investigated the influence of the native language on linguistic complexity patterns in English production using the large-scale educational resource – EFCamDat. We target three dimensions of linguistic complexity namely lexical, syntactic and morphological and consider a wide set of languages. To present the L1s interrelations and similarities in L2 written production complexity we used the software presented by WEKA and defined the features to be analysed. Additionally, developmental trajectories of lexical, syntactic, and morphological complexity components with regard to the first language have been introduced. Based on the results obtained by clustering, we proved that the level of learners' proficiency affects linguistic complexity much stronger than the first language background.

Another aim of research was to compare the developmental trajectories of linguistic complexity through all levels of proficiency. The results lead us to the conclusion that linguistic complexity is hard to be analysed as a whole unit as most of its components develop in different way. Although some degree of first language transfer is still retained, the key factor is still the level of learners' proficiency and not the first language background.

R e f e r e n c e s

Ai, H., Lu, X. (2013). A corpus-based comparison of syntactic complexity in NNS and NS university students' writing. *Automatic treatment and analysis of learner corpus data*, 249–264. doi: https://doi.org/10.1075/scl.59.15ai

Alexpoulou, T. et al. (2015). Exploring big educational learner corpora for SLA research: Perspectives on relative clauses. *International Journal of Learner Corpus Research*, 1 (1), 96–129. doi: https://doi.org/10.1075/ijlcr.1.1.04ale

Andersen, R. (1983). Transfer to Somewhere. *Language transfer in language learning*, 177–201.

Attali, Yi., Burnstein J. (2006) Automated essay scoring with e-rater V. 2. *The Journal of Technology, Learning and Assessment 4* (3). doi: https://doi.org/10.1002/j.2333-8504.2004.tb01972.x

Bulte, B., Housen, A. (2015). Evaluating short-term changes in L2 complexity development. *Circulo de linguistica aplicada a la comunicacion*, 63, 43–76. doi: https://doi.org/10.5209/rev_CLAC.2015.v63.50169

Bykh, S., Meurers, D. (2016) Advancing Linguistic Features and Insights by Label-informed Feature Grouping: An Exploration in the Comtext of Native Language Identification. Retrieved December, 2, 2018 from https://pdfs.semanticscholar.org/a837/40e9cd0b7070c38f373ce8b8147ad00d31f4.pdf?_ga=2.226138970.1737945330.1548930370-2072522481.1548930370

Bykh, S., Vajala, S., et al. (2013) Combining shallow and linguistically motivated features in native language identification. Retrieved January 2, 2019 from http://aclweb.org/anthology/W13-1726

Clercq, B. D., Housen, A. (2019). The development of morphological complexity: A cross-linguistic study of L2 French and English. *Second Language Research*, 35 (1), 71–97. doi: https://doi.org/10.1177/0267658316674506

Ebrahimi, E. (2015). The effect of dynamic assessment on complexity, accuracy, and fluency in EFL learners' oral production. *International Journal of Research Studies in Language Learning 4* (3), 107–123. doi: https://doi.org/10.5861/ijrsll.2015.982

Ellis, R. (1994). *The study of second language acquisition*. Oxford: Oxford University Press.

Ellis, R. (2009). The differential effects of three types of task planning on the fluency, complexity, and accuracy in L2 oral production. *Applied linguistics*, 30 (4), 474–509. doi: https://doi.org/10.1093/applin/amp042

Geertzen, J., Alexopoulou Th., Korhonen, A. (2013) Automatic linguistic annotation of large scale L2 databases: The EF-Cambridge Open Language Database (EFCAMDAT) Proceedings of the 31st Second Language Research Forum. Somerville, MA: Cascadilla Proceedings Project. Retrieved January 2, 2019 from http://people.ds.cam.ac.uk/ta259/SLRF2013Geertzenetal.pdf

Gleitman L. R., et al., (2019) The impossibility of language acquisition (and how they do it). *Annual Review of Linguistics*, 5. 1–24. doi: https://doi.org/10.1146/annurev-linguistics-011718-011640

Granger, S. et al. (2007). Integrating learner corpora and natural language processing: A crucial step towards reconciling technological sophistication and pedagogical effectiveness. *ReCALL: the Journal of EUROCALL*, 19 (3), 252–268. doi: https://doi.org/10.1017/S0958344007000237

Halliday, M. A. K. (1985). *Spoken and written language*. Deakin university.

Hancke, J., Vajjala, S., Meurers, W.D. (2012). *Readability Classification for German using Lexical, Syntactic, and Morphological Features*. COLING, 1063–1080. Retrieved October 21, 2018 from http://aclweb.org/anthology/C12-1065

Housen, A., Kuiken, F. (2009). Complexity, accuracy, and fluency in second language acquisition. *Applied linguistics*, 30 (4), 461–473. doi: https://doi.org/10.1093/applin/amp048

Hyltenstam, K. (1977). Implicational patterns in interlanguage syntax variation. *Language Learning*, 27 (2), 383–410. doi: https://doi.org/10.1111/j.1467-1770.1977.tb00129.x

Jarvis, S., Odlin, T. (2000). Morphological type, spatial reference, and language transfer. *Studies in second language acquisition*, 22 (4), 535–556. doi: https://doi.org/10.1017/S0272263100004034

Keiper, L., Hornach A., Thater S. (2016) Improving POS tagging of German Learner Language in a Reading Comprehension Scenario. *Proceedings of the Tenth International Conference on Language Resources and Evaluation*, 198–205. Retrieved December 2, 2018 from http://www.lrec-conf.org/proceedings/lrec2016 /pdf/172_Paper.pdf

Kellerman, E. (1985) If at first you do succeed. *An Input in Second Language Acquisition*, 345–353.

Klee, C., Ocampo, A. (1995). The expression of past reference in Spanish narratives of Spanish-Quechua bilingual speakers. *Spanish in four continents: Studies in language contact and bilingualism*, 52–70.

Kong, K. (2009). A comparison of the linguistic and interactional features of language learning websites and textbooks 1. *Computer Assisted Language Learning*, 22 (1), 31–55. doi: https://doi.org/10.1080 /09588220802613799

Kuznik, A., Ollala-Soler, C. Results of PACTE group's experimental research on translation competence acquisition. The acquisition of the instrumental sub-competence. *Across Languages and Cultures*, 19 (1). 19–51. doi: https://doi.org/10.1556/084.2018.19.1.2

Lado, R. (1957). *Linguistics Across Cultures: Applied Linguistics for Language Teachers*. Ann Arbor: University of Michigan Press.

Lee, B. R. (2015). The effects of word form variation and frequency on second language incidental vocabulary acquisition through reading. *Applied Linguistics Review*, 6 (4). doi: https://doi.org/10.1515/applirev-2015-0021

Lintunen, P., Makila, M. (2014). Measuring Syntactic Complexity in Spoken and Written Learner Language: Comparing the Incomparable? *Research in Language*, 12 (4), 377–399. doi: https://doi.org/10.1515/rela-2015-0005

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. doi: https://doi.org/10.5054/tq.2011.240859

Lu, X. (2012). The relationship of lexical richness to the quality of ESL learners' oral narratives. *The Modern Language Journal*, 96 (2), 190–208. doi: https://doi.org/10.1111/j.1540-4781.2011.01232_1.x

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

Mooi, E., Sarstedt, M. (2011). Chapter 9: Cluster Analysis. *A Concise Guide to Market Research*, 273–324.

Murakami, A. (2014). *Individual variation and the role of L1 in the L2 development of English grammatical morphemes: Insights from learner corpora*. PhD thesis. University of Cambridge. doi: https://doi.org/10.17863/CAM.16509

Murakami, A., Alexpoulou, T. (2016). *Longtitudinal L2 Development of the English Article in Individual Learners:* Book Chapter. Cognitive science society, 1050–1055. doi: https://doi.org/10.17863/CAM.97

Odlin, T. (1989). *Language transfer: Cross-linguistic influence in language learning.* Cambridge University Press. doi: https://doi.org/10.1017/S027226310000975X

Odlin, T., Jarvis, S. (2000). Morphological type, spatial reference, and language transfer. *Studies in second language acquisition,* 22 (4). 535–556. doi: https://doi.org/10.1017/S0272263100004034

Ortega, L. (2003). Syntactic complexity measures and their relationship to L2 proficiency: A research synthesis of college-level L2 writing. *Applied Linguistics*, 24 (4), 492–518. doi: https://doi.org/10.1093 /applin/24.4.492

Ortega, L. (2012) Interlanguage complexity. Linguistic complexity. *Second Language Acquisition, Indigenization, Contact,* 13. 127.

Ortega, L. (2013). *Understanding second language acquisition*. Routledge.

Pahomov S. et. al. (2011) Computerized assessment of syntactic complexity in Alzheimer's disease: a case study of Iris Murdoch's writing. *Behaviour research methods*, 43 (1), 136–144. doi: http://doi.org/10.3758/s13428-010-0037-9

Palotti, G. (2009). CAF: Defining, refining and differentiating constructs. *Applied linguistics*, 30 (4), 590–601. doi: https://doi.org/10.1093/applin/amp045

Palotti, G. (2015). A simple view of linguistic complexity. *Second Language Research*, 31 (1), 117–134. doi: https://doi.org/10.1177%2F0267658314536435

Pandit, Sh., Gupta, S. (2011) A comparative study on distance measuring approaches for clustering. *International Journal of Research in Computer Science*, 2 (1), 29-31. Retrieved January 2, 2019 from https://pdfs.semanticscholar.org/b3b4/445cb9a2a55fa5d30a47099335b3f4d85dfb.pdf

Paradis, J., et al., (2017) Children's second language acquisition of English complex syntax: the role of age, input, and cognitive factors. *Annual Review of Applied Linguistics*, 37. 148–167. doi: https://doi.org /10.1017/S0267190517000022

Peck, S. (1978). Child-child discourse in second language acquisition. *Second language acquisition: a book of readings,* 383–400.

Read, J. (2000). *Assessing vocabulary.* Cambridge: Cambridge University Press. doi: https://doi.org /10.1017/CBO9780511732942

Robinson, P. (2007). Task complexity, theory of mind, and international reasoning: effects on L2 speech production, interaction, uptake and perceptions of task difficulty. *IRAL-International Review of Applied Linguistics in Language Teaching,* 45 (3), 193–213. doi: https://doi.org/10.1515/iral.2007.009

Schachter, J. (1974). An error in error analysis. *Language learning,* 24 (2), 193–213. doi: https://doi.org/10.1111/j.1467-1770.1974.tb00502.x

Schumann, J. (1979). The acquisition of English negation by speakers of Spanish: a review of literature. *The acquisition and use of Spanish and English as first and second languages,* 3–32.

Shalev-Schwartz, Sh., Ben-David, Sh. (2014) *Understanding machine learning: from theory to algorithms*. Cambridge University Press. Retrieved December 2, 2018 from https://www.cs.huji.ac.il/~shais /UnderstandingMachineLearning/understanding-machine-learning-theory-algorithms.pdf

Skehan, P. (2009). Modelling second language performance: Integrating complexity, accuracy, fluency, and lexis. *Applied linguistics*, 30 (4), 510–532. doi: https://doi.org/10.1093/applin/amp047

Stockwell, R. P., Bowen, J. D., Martin, J. W. (1965). *The grammatical structures of English and Spanish*, 4. Chicago: University of Chicago Press.

Timpe-Laughlin, V. (2016). Adult learners acquisitional patterns in L2 pragmatics: what do we know? *Applied Linguistics Review*, 8 (1). 101–130. doi: https://doi.org/10.1515/applirev-2015-2005

Ure, J. (1971). Lexical density: A computational technique and some findings. *Talking about text*, 27–48.

Vajala, S., (2012). On improving the accuracy of readability classification using insights from second language acquisition. *Proceedings of the Seventh Workshop on Innovative Use of NLP for Building Educational Applications,* 163–173.

Vajala, S., Meurers, D. (2014). Exploring Measures of "Readability" for Spoken language: Analyzing linguistic features of subtitles to identify age-specific TV programs. *Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations,* 21-29. Retrieved December 18, 2019 from http://aclweb.org/anthology//W/W14/W14-1203.pdf

Vyatkina, N., Hirshmann, 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. doi: https://doi.org/10.1016/j.jslw.2015.06.006

Witten, I. H. et al. (2016). *Data mining: practical machine learning tools and techniques.* Morgan Kaufmann. 525. Retrieved October 21, 2018 from ftp://ftp.ingv.it/pub/manuela.sbarra/Data%20Mining% 20Practical%20Machine%20Learning%20Tools%20and%20Techniques%20-%20WEKA.pdf

Бібліографічний опис:

Галій, *Л.* Г. (2018). Типологічна структура як зіставна модель для вивчення мов. Науковий часопис Національного педагогічного університету імені М. П. Драгоманова. Серія 9. Сучасні тенденції розвитку мов, 18. 46–60. DOI: https://doi.org/10.31392/NPU-nc.series9.2018.18.04

Анотація

У статті досліджується типологічна структура мов індоєвропейської сім'ї та вплив інтерференції рідної мови на писемне мовлення творчого характеру тих людей, що вивчають англійську мову. Виявлено, що компоненти складності писемного мовлення мають різні траєкторії

розвитку, пов'язані з рівнями володіння мовою, але при цьому інтерферентність має місце не лише між близькоспорідненими мовами. Ґрунтуючись на цьому припущенні, розглянуто мовленнєву складність з огляду на вплив рідної мови та класифіковано мови на основі типологічної подібності їх структур, а не на основі спорідненості. Доведено, що типологічно близькі мови, які до того ж належать до однієї мовної родини або мовної групи, зумовлюють однакові помилки у процесі вивчення англійської мови як іноземної, зокрема на етапі творення складних конструкцій мовлення. На останньому етапі дослідження здійснено аналіз впливу рідної мови на писемне англійське мовлення творчого характеру залежно від рівня володіння цією мовою. У підсумку, класифіковано мови у кластери, що за своїм впливом на вивчення англійської мови мають однакові характеристики (морфологічні та синтаксичні), та запропоновано нову модель для дослідження інтерферентності мов при їх зіставному вивченні.

Ключові слова: *інтерференція, творче мовлення, типологія мов, ієрархічна кластеризація.*