MATHEMATICA MONTISNIGRI
Vol XLVI (2019)

QUANTITATIVE EVALUATION OF SYNTAX SIMILARITY

E.S. KLYSHINSKY[1]*, O.V. KARPIK [2]

[1] National Research University «Higher School of Economics», Moscow, Russia
[2] Keldysh Institute of Applied Mathematics RAS, Moscow, Russia
*Corresponding author. E-mail: eklyshinsky@hse.ru

DOI: 10.20948/mathmontis-2019-46-11

Summary. Machine learning systems are facing problem of incomparability of their results in case of different languages; one of the subarea here is quantitative analysis of syntax. In this paper, we introduce a new quantitative method based on statistics of words co-occurrence in syntactically tagged corpora. The method allows quantitatively evaluate difference and similarity among languages, select most influential phenomena. Experimental setup consists materials for more than 50 languages. Our experiments demonstrate that the introduced method correctly cluster languages among language families.

1 INTRODUCTION

An advance in the area of computational linguistics and natural language processing relies on results of mathematical formulation of linguistic phenomena. One of the examples here is the distributional semantics [1] that allows introducing of vector operations in semantics using Word2Vec [2] and Glove [3] technologies. These techniques are currently improved in different non-Euclidian spaces (see [4] and [5]). Such scientific domains as quantitative linguistics [6] and theory of language complexity [7] also numerically describe language phenomena. Using these theories, we are able to describe language processes quantitatively but not qualitatively; this makes us possible numerically evaluate and compare the results achieved. One of the subareas here is the quantitative analysis of natural language syntax. In spite of many studies in this area (e.g., see [8]), the results here are still more descriptive than quantitative. It is necessary to state that the quantitative analysis uniform scheme at the moment does not exist. Therefore, there are no common algorithms of machine word processing, on which basis it would be possible to develop methods of computer modeling of language processes. There is a lack of new methods for numerical evaluation of syntactic phenomena. Such method should allow formalizing and emphasizing similarities and dissimilarities among languages and language groups. It is obvious practical usefulness in information about the syntax similarities and differences of the native and studied language during learning foreign languages. The creation of an effective formal languages description promises great prospects in automated text processing, machine translation and the creation of artificial intelligence. In addition to the applied task of syntactic structures computer modeling and text processing, syntax formalization can also help to develop updated metalanguage of linguistic research. The development of flexible and universal language describing means, which should replace the approach based on purely qualitative descriptive methods, is impossible without parameters presented in a simple numerical form.

[9] have shown that co-occurrence networks of natural language are scale-free small world graphs (for the small world graph definition see [10], another applications of modeling on small world graph represented in [11]). The same is true for syntactic representation of a text.

2010 Mathematics Subject Classification: 03E75, 68R10, 60K30.
Key words and Phrases: Statistical methods, Quantitative linguistics, Syntactic analysis.

In this paper we introduce a new method based on statistics of words co-occurrence in syntactically tagged corpora. The rest of the paper is organized as following. Section 2 briefly describes stages of natural language processing and the aim of the syntactic analysis. In section 3, we introduce a formal method for description of syntactic analysis and a new method for evaluation of the syntactic similarity of languages. Section 4 describes data used in our experiments, experimental results and evaluation of their correctness. In section 5, we are analyzing achieved results from the computational linguistics point of view. The conclusion sums up the results of this paper.

2 THEORETICAL BACKGROUND

Depending on a task statement, a natural text could be processed in several stages. The first one is tagging which divides text into sentences, word and non-word tokens. Non-word tokens are punctuation marks, numbers, IP-addresses etc. Such languages as Chinese or Japanese have their own peculiarities; the tagging accounts here for concatenating sole hieroglyphs into chunks. For example, Chinese hieroglyph 电 means *electricity*, 池 means *a tank*, while their combination 电池 means *accumulator*.

The next stage is the lexical analysis which defines a word's initial form, part of speech, and grammatical features (number, gender, grammatical case etc.). The complicity of this stage depends on a language in hand. This stage could be unnecessary for hieroglyphically languages like Chinese and Japanese; however, morphologically complicated languages need much more effort to define (disambiguate) correct initial form, part of speech, and set of grammatical features. For example, a token *text* in an English text could represent several parts of speech: verb (*to send messages*), noun (*set of characters*), and adjective (*something textual*). Thus, a correct choice depends on the context.

In this paper, we are analysing the results of the syntactic analysis, the third stage of the natural text analysis. The aim of syntactic analysis is to define connections between tokens and labelling these connections. We give a formal definition of this stage in the Section 3; here, we will just briefly discuss it in a very few words.

There are two main methods for representation of syntactic analysis results. The first one, dependencies trees [12], will be used in this paper; the second one, constituency trees [13], is not discussed here. In case of dependencies tree, words are connected by labelled edges representing a tree graph. A label represents a role which plays a word in the given context. The most common labels are a *subject* – the main actor, a *predicate* – the main action, an *object* – the object influenced by an action, etc. In order to keep uniformity of a graph representation, a fake node named *root* is introduced. The Fig. 1 represents an example of a dependencies tree for a sentence *A big black dog runs after a poor cat*. It is easy to see, that the subject here is the word *dog*, the predicate is *run*, and the object is *cat*.

A comprehensive survey of other graph representations of text could be find in [14].

Among others, there is such phenomenon as a branching. The left branching is a situation when a tail word precedes a head word; the other case is named the right branching. It is well known that different languages prefer different types of branching. For example, Germanic languages prefer left branching for adjective-noun connection, while Romance languages prefer right branching in the same situation.

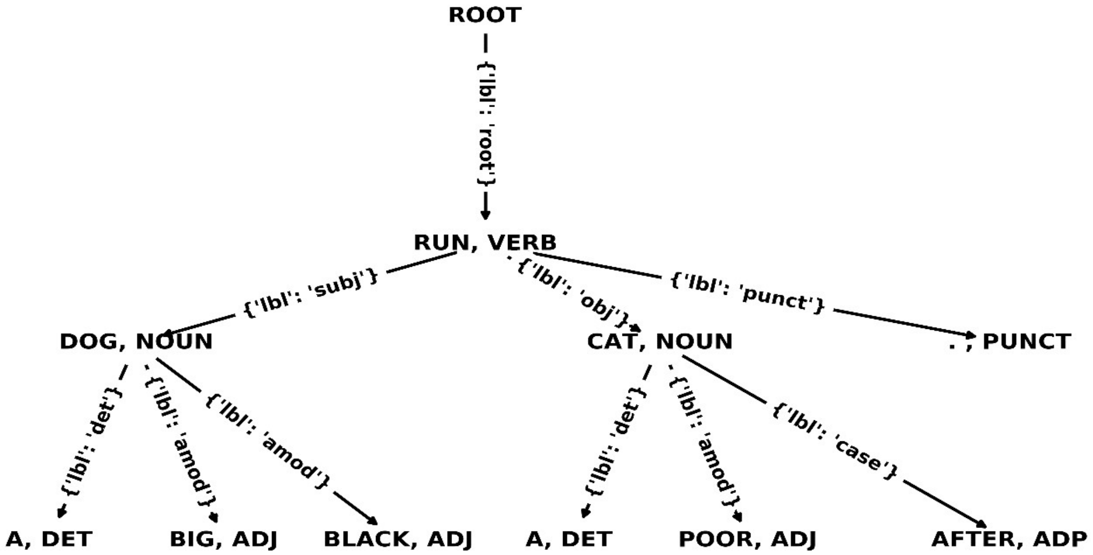

Figure 1. Dependency tree for a sentence *A big black dog runs after a poor cat*

Common sense definition of syntactical similarity of two languages means that the same ideas should be expressed in quite the same manner using quite the same parts of speech, labels, type of branching, etc. For example, Romance languages use auxiliary words for shaping tenses while Slavic languages create new word forms and new words. Thus, we can state the difference between language groups and the similarity among languages of the same group.

In this paper, we assume that similar syntactical constructions are used for expression of the similar ideas. It is not like that in common case, however we will use overall statistics that will eliminate biases. In case of serious deviance, we could state that languages use different constructions. However, such approach tends to have some disadvantages. We should guarantee that selected texts express quite the same ideas in the same style; either we should prove that language properties are hardly dependent on these features and texts could be compared with a reasonable mistake. The theoretical linguistics follows a tradition to describe a phenomenon using series of negative and positive examples. This is very convenient since a reader could draw his or her own conclusions. The quantitative linguistics gives numerical evaluation of such phenomena. In spite of numerous scientific studies discussing similarity and complexity of different phenomena in different languages (e.g., see [8]), there is a lack of overall comparison of the main world languages.

The aim of this paper is to introduce a new numerical method for a quantitative evaluation of syntactical similarity of main world languages. For this purposes, we give a formal description of dependencies trees and describe a formal method for such evaluation.

3 THE METHOD OF QUANTITATIVE EVALUATION

Following to [12] let us represent a sentence S as a sequence of words from the vocabulary W: S = $<w_1, w_2, \ldots, w_n>$, where $w_i \in W$ and $w_i = <id, l, \pi>$, where *id* is a unique identifier and position of a word in a sentence, *l* is the lemma of this word, π is its part of speech. In this paper, we are not using lexical and grammatical features of words. We also consider that all words are disambiguated, therefore, there is only one tuple for every word and it is correct. Let us also define a relation set $R = \{r_1, r_2, \ldots, r_m\}$. In this case, a dependency tree G = <V , A> for a sentence S is a labeled directed tree consisting of nodes V and arcs A: $V = \{root, w_1, w_2, \ldots, w_n\}$, $A \subseteq V \times R \times V$, if $<w_i, r, w_j> \in A$ then $<w_i, r', w_j> \notin A$ for all $r' \neq r$. 'Root' here is a fake word which is always a root of the tree G; let us take $id = 0$ for root. Let us also denote $<w_i, r, w_j>$ as $<i, r, j>$, where w_i is a parent node and w_j is a child node.

To illustrate this definition let us consider a sentence *A big black dog runs after a poor cat* (Fig.1). Here V = <root, <1, A, DET>, <2, BIG, ADJ>, <3, BLACK, ADJ>, <4, DOG, NOUN>, <5, RUN, VERB>, <6, AFTER, ADP>, <7, A, DET>, <8, POOR, ADJ>, <9, CAT, NOUN>, <10, . , PUNCT> >, A={<0, root, 5>, <5, nsubj, 4>, <4, det, 1>, <4, amod, 2>, <4, amod, 3>, <5, obl, 9>, <9, det, 7>, <9, amod, 8>, <9, case, 6>, <5, punct, 10>}. This example also demonstrates the need of introduction of unique identifiers for words. This sentence has two determiners with lemma A; if we shall write <CAT, det, A> we could not understand if we mean the word on the first or the seventh position.

Let us consider a tuple $<w_i, r, w_j>$. In case of $i < j$, we can speak about the right branching; in case of $i > j$, we can speak about the left branching.

Let us consider a tuple $<w_i, r, w_j>$. We denote here the part of speech of i-th word as π_i, and define r_{ij} as a label of an arc connecting w_i and w_j. In this case, we could calculate frequency of occurrence for every tuple $g = < \pi_i, \pi_j, r_{ij}, b_j>$ where b_j is direction of connection of j-th word: $b = 0$ if $i < j$ (right branching) and $b = 1$ otherwise (left branching).

Left and right branching are important features of a language. For example, the English language demands an adjective in preceding position before its governing noun; in Slavic languages, the preferred position is the same with some deviations; in Romance languages, preferred position is right side from the noun with some deviations. The degree of these deviations depends on the language. Calculating statistics for tuples like $<\pi_i, \pi_j, r_{ij}, b_j>$, we could find out the difference between the preferred word orders, investigate the words' aptitude for making connection with other words, and find out similarity among considered languages.

However, syntax could not be defined merely using connections between pairs of words and their order. In this paper, we are also going to investigate relations between triples of words: $h_1 = <\pi_i, \pi_j, \pi_*, r_{ij}, r_{jk}, b_j, b_k>$ and $h_2 = <\pi_i, \pi_j, \pi_*, r_{ij}, r_{ik}, b_j, b_k>$.

In this paper, we will understand a syntactic profile of a corpus as a statistics of occurrence for all pairs and triples occurred in this corpus: $Q = \{<g_i, f_i>\} \cup \{<h_j, f_j>\}$ where f_i and f_j are frequencies of occurrence for tuples g_i and h_j accordingly.

However, the syntactic label of a word could define its aptitude for making connection with other words. Thus, let us define extended tuples containing a label for a head word: $g = <\pi_i, \pi_j, r_{*i}, r_{ij}, b_j>$, $h_1 = <\pi_i, \pi_j, \pi_k, r_{*i}, r_{ij}, r_{jk}, b_j, b_k>$, and $h_2 = <\pi_i, \pi_j, \pi_k, r_{*i}, r_{ij}, r_{ik}, b_j, b_k>$. Let us also define an extended syntactic profile of a corpus Q.

In order to compare two corpora, we will use the following algorithm.

1. Calculate syntactic profile for both corpora and select from them top 10 most frequent pairs $T_{2,1} = \{<g_{i1}, f_{i1}>\}$, $T_{2,2} = \{<g_{i2}, f_{i2}>\}$ and top 10 most frequent triples $T_{3,1} = \{<h_{i1}, f_{i1}>\}$, $T_{3,2} = \{<h_{i2}, f_{i2}>\}$.
2. Join these sets and select frequencies for both corpora $\hat{T} = \{< \hat{d}_k, f_{k1}, f_{k2} > \}$, where $\hat{d}_k \in \{g_{k1}\} \cup \{g_{k2}\} \cup \{h_{k1}\} \cup \{h_{k2}\}$ and f_{k1}, f_{k2} are frequencies for the proper tuples of the first and the second corpora. The syntactic similarity of two corpora will be calculated as a rank correlation of these two vectors $sym = corr(<f_{k1}>, <f_{k2}>)$.

For several languages, we use the same algorithm, but we join all most frequent pairs T_2 and triples T_3 of all languages in the same list.

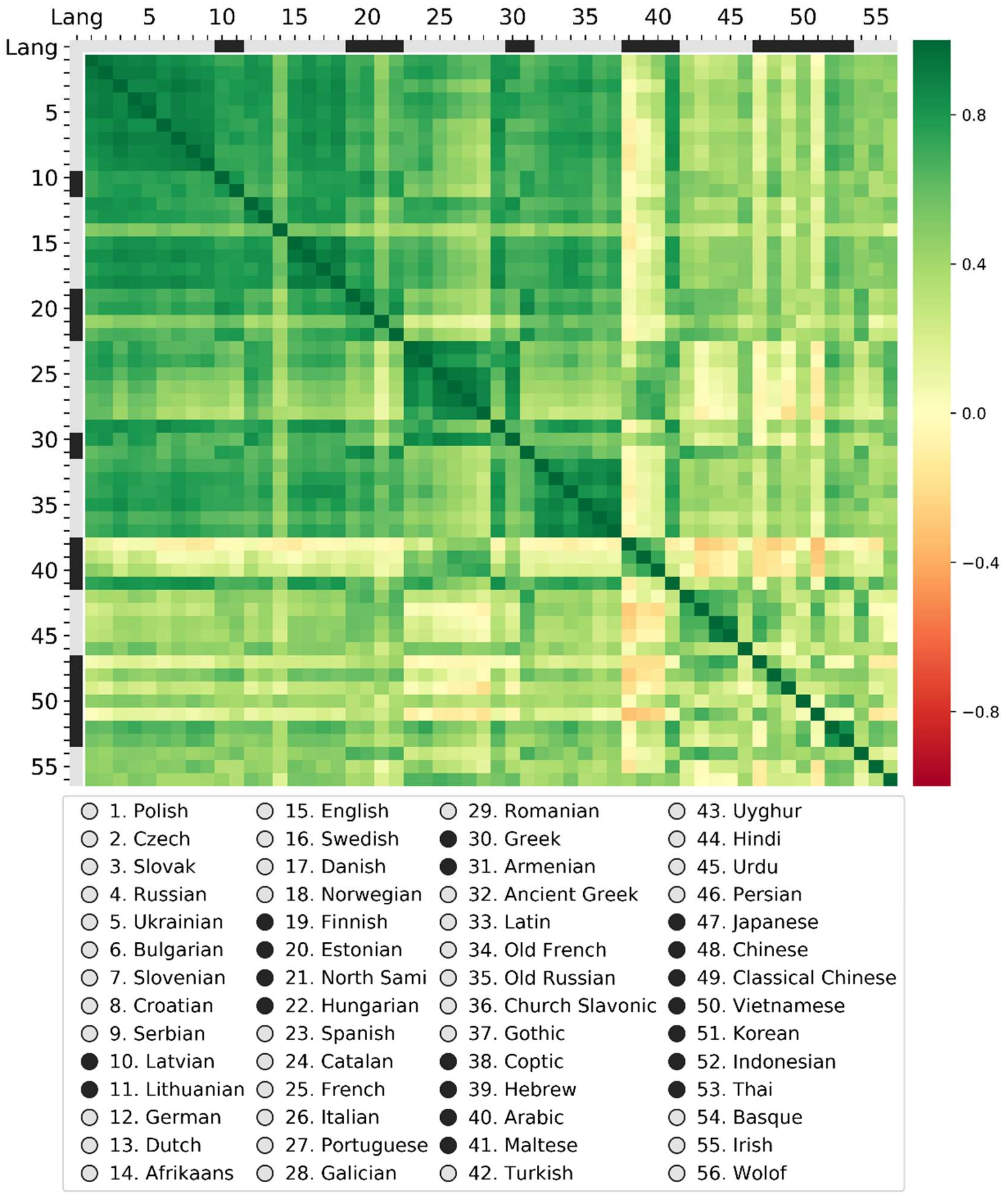

Figure 2. Correlation matrix for pairs of connected words

4 EXPERIMENTAL RESULTS AND MODEL EVALUATION

For our experiments, we used the Universal Dependencies corpus [15] ver. 2.4. (accessible at http://universaldependencies.org/). We have selected 56 languages with size of corpora varying from 22 000 of word tokens for Thai up to 3.4 mln word tokens in the German corpus. All languages were divided into language groups and families: Slavic, Baltic, Germanic, Romance, Finn-Ugric, Greek-Armenian, Semitic, Turkic, Indo-Aryan, Ancient and Old. There also was a variety of languages which did not constitute a group inside the list of selected languages: Japanese, Korean, Chinese, Thai, Vietnamese, Indonesian, Basque, Irish, and

Wolof. The Coptic language do not belong to the Ancient language group but placed here because of texts similarity. Thus, we covered the world biggest language groups excluding native languages of Americas and Central Africa. Numerical data is too big and placed in supplementary materials (http://cosyco.ru/syntax/syntax_share.zip).

We calculated syntactic profiles for all these languages and used Spearman rank correlation to evaluate the similarity of languages. Correlation matrix for extended pairs of connected words is presented in Fig. 2, for extended triples of connected words at Fig. 3.

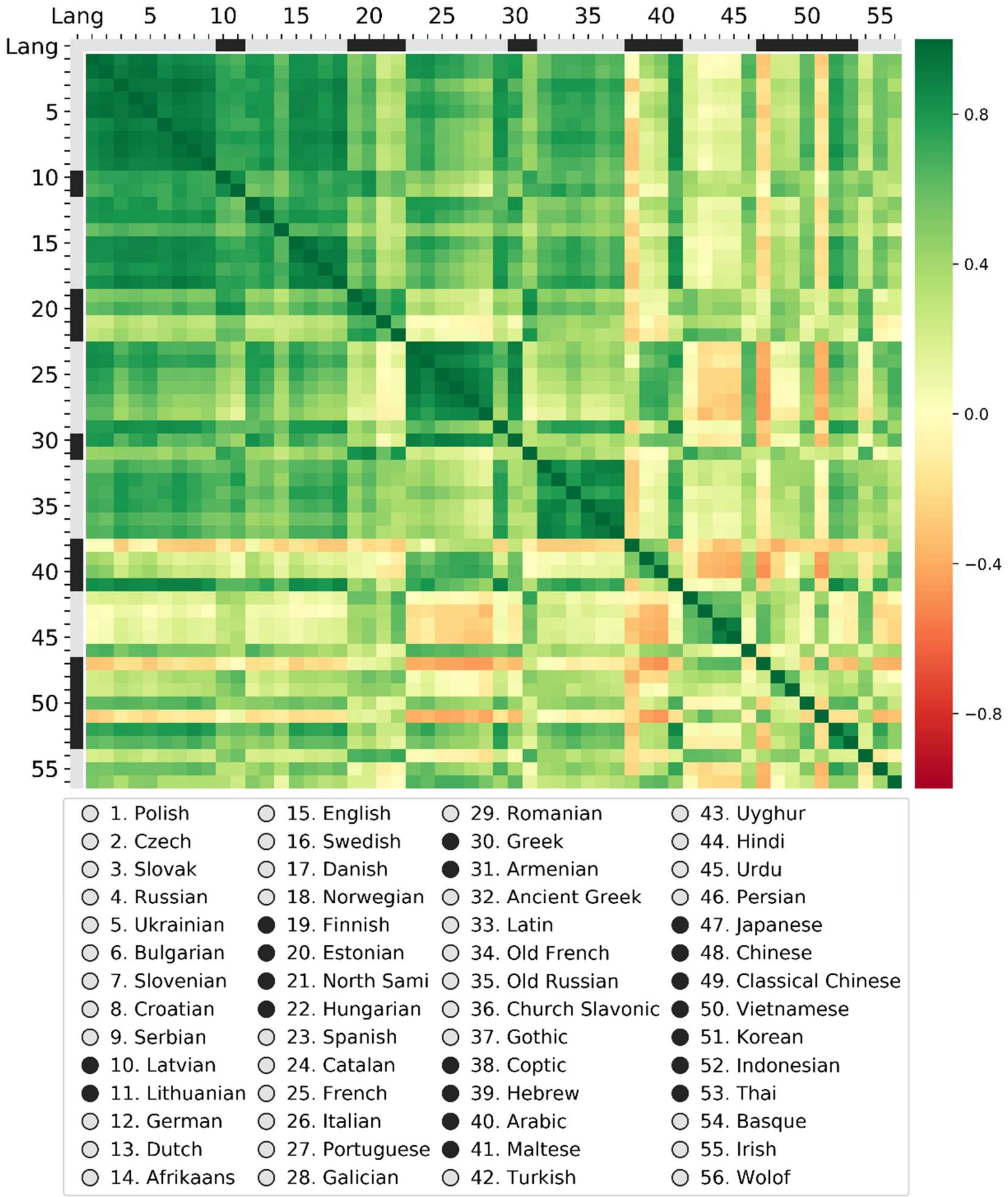

Figure 3. Correlation matrix for triples of connected words

It is easy to see that languages belonging to the same group constitute clusters on presented figures. For pairs of connected words, the most correlated language pairs are Spanish and Catalan, Polish and Czech, and Russian and Ukrainian. The most correlated language is Russian – the `average correlation is 0.64 while the average correlation for all the languages is 0.48. Coptic is the most uncorrelated language – its correlation with Arabic and Hebrew is 0.72 and 0.69 while the average correlation is equal to 0.01. Average correlations for Japanese and Korean languages are 0.03 and 0.04. The most correlated language group is Slavic – the average correlation in the group is 0.88; the second correlated group is Ancient languages (Ancient Greek, Latin, Old French, Old Russian, Church Slavonic, Gothic) – 0.86; the third group is Romance languages – 0.79 including Romanian and 0.86 excluding it.

For word triples, the situation is almost the same. One of the most correlated pairs of languages is Slovenian and Croatian. The value of correlations changes, however the qualitative image is the same. The most changed language here is Japanese; its average correlation reduces from 0.17 for word pairs to -0.11 for word triples. However, correlation keeps around zero.

The average correlation for Hindi, Urdu and Uygur reduces from 0.36, 0.34 and 0.33 for word pairs to 0.13, 0.12 and 0.11 respectively for word triples. The most tragic changes are observed for the Hungarian languages where average correlation falls from border 0.55 to neglectable 0.32.However, the average correlation among languages in the Slavic group slightly raises up to 0.91, for Ancient languages keep for 0.85, and for Romance languages keeps for 0.79 (slightly raised up to 0.89 without the Romanian language). Thus, word triples contain more language-specific information than word pairs.

To make a comparison with a common data, we have placed two extra columns on images (two right columns on all the images). The first one demonstrates the word order in a language; Subject-Verb-Object marked as dark green, Subject-Object-Verb marked as pale green, Verb-Subject-Object as beige, and languages without the preferred word order are marked as dark red (data taken from https://wals.info/). The second column demonstrates the order of a connected adjective and a noun: dark green marks adjective-first word order, beige marks noun-first word order, red marks languages without the preferred word order.

5 DATA ANALYSIS

We examined the data on the most frequent word pairs and triples for different languages and found out some differences among them. Table 1 demonstrates the most frequent words in connected pairs for Russian, Ukrainian, Polish, English, German, Swedish, Japanese, Chinese and Korean languages. In European languages, the most frequent connection is a noun in oblative case (prepositional phrase) entailed by a preposition labelling this case. European languages demonstrate more similar behavior – there are 21 types of connection for 6 languages, while Asian ones are more variable – 26 types of connection for 3 languages. This fact corresponds to the fact that Japanese and Korean languages have the lowest rate of average correlation with other languages. Note that Russian and Ukrainian languages have 8 identical combinations (correlation is 0.95), while Russian and Ukrainian have just 6 combinations coinciding with Polish (correlation 0.87 and 0.89 respectively). English and Swedish have 7 identical (correlation 0.93), English and German have 7 (correlation 0.82), German and Swedish have 6 (correlation 0.84).

Russian				**Ukrainian**				**Polish**			
head		*tail*	*N*	*head*		*tail*	*N*	*head*		*tail*	*N*
NOUN obl	←	ADP case	1	NOUN obl	←	ADP case	1	NOUN obl	←	ADP case	1
NOUN nmod	←	ADJ amod	2	NOUN nmod	←	ADJ amod	2	VERB root	→	NOUN obl	10
NOUN nmod	←	ADP case	3	NOUN nmod	←	ADP case	3	VERB root	←	NOUN nsubj	7
NOUN nmod	→	NOUN nmod	4	NOUN nmod	→	NOUN nmod	4	NOUN nmod	←	ADP case	3
NOUN obl	←	ADJ amod	5	NOUN obl	←	ADJ amod	5	VERB root	→	NOUN obj	13
NOUN obl	→	NOUN nmod	6	NOUN obl	→	NOUN nmod	6	NOUN obl	←	ADJ amod	5
VERB root	←	NOUN nsubj	7	VERB conj	←	CCONJ cc	11	NOUN obl	→	NOUN nmod	6
NOUN nsubj	→	NOUN nmod	8	NOUN conj	←	CCONJ cc	12	VERB root	←	PART advmod	14
NOUN nsub	←	ADJ amod	9	VERB root	→	NOUN obl	10	VERB root	→	VERB conj	15
VERB root	→	NOUN obl	10	VERB root	←	NOUN nsubj	7	VERB root	←	NOUN obl	16
English				**German**				**Swedish**			
NOUN obl	←	ADP case	1	NOUN obl	←	ADP case	1	NOUN obl	←	ADP case	1
NOUN nmod	←	ADP case	3	NOUN nsubj	←	DET det	20	NOUN nmod	←	ADP case	3
NOUN obl	←	DET det	17	NOUN nmod	←	ADP case	3	VERB root	→	NOUN obl	10
NOUN obj	←	DET det	18	NOUN nmod	←	DET det	21	NOUN obl	←	ADJ amod	5
VERB root	←	PRON nsubj	19	NOUN obl	←	DET det	17	VERB root	←	NOUN nsubj	7
NOUN nsubj	←	DET det	20	VERB root	←	NOUN nsubj	7	NOUN obl	←	DET det	17
NOUN nmod	←	DET det	21	NOUN obj	←	DET det	18	VERB root	→	NOUN obj	13
VERB conj	←	CCONJ cc	11	VERB root	←	NOUN obl	16	VERB root	←	PRON nsubj	19
VERB root	→	NOUN obj	13	NOUN obl	←	ADJ amod	5	NOUN conj	←	CCONJ cc	12
VERB root	→	NOUN obl	10	VERB root	→	NOUN obl	10	VERB conj	←	CCONJ cc	11
Japanese				**Chinese**				**Korean**			
NOUN nmod	→	ADP case	22 23	VERB root	←	ADV advmod	31	VERB root	←	NOUN dislocated	35
NOUN obl	→	ADP case	24	VERB root	←	PRON nsubj	19	VERB root	←	NOUN obj	36
NOUN obj	→	ADP case	25	NOUN obl	←	ADP case	1	VERB root	←	ADV advmod	31
VERB advcl	→	SCONJ mark	26	VERB root	→	NOUN obj	13	VERB root	←	NOUN nsubj	7
NOUN nmod	←	NOUN compound	27	VERB root	←	NOUN nsubj	7	VERB root	←	SCONJ ccomp	37
VERB root	←	VERB advcl	16	VERB root	←	NOUN obl	16	NOUN obj	←	VERB acl	38
VERB root	←	NOUN obl	28	VERB acl	→	PART mark	32	NOUN obj	←	NOUN nmod	39
NOUN nmod	←	NOUN nmod	29 30	VERB root	←	VERB advcl	27	VERB acl	←	ADV advcl	40
NOUN obl	←	NOUN nmod		NOUN obj	←	NOUN compound	33	VERB root	←	CCONJ cc	41
VERB advcl	←	NOUN obl		VERB root	→	VERB ccomp	34	VERB root	←	NOUN advmod	42

Table 1. Most frequent word pairs; an arrow demonstrates left or right branching

It is easy to see that different languages demonstrate different behavior at top-10 connections. Slavic languages do not use determiners while they are a very important part in Germanic languages; the Swedish language has less determiners because they could be a postfix of a word. The Japanese language uses more adpositions because there are no grammatical cases at all while the situation with the Korean language is opposite. Thus, the method correctly demonstrates similarity and dissimilarity of world languages. However, we should check several features which could influence the achieved results.

As it was mentioned above, the Universal Dependencies (UD) corpus is a collection of corpora for a long list of languages. However, there could be several subcorpora for a language, tagged by different research groups. The tagging by different groups with different defaults could lead to corpora inconsistency. That is why we checked if different corpora for a language correlate among themselves. The correlation among Czech subcorpora is higher than 0.94, for Russian and German subcorpora is higher than 0.97.

Several languages were annotated by the same group that is responsible for UD support. They used parallel texts translated into Polish, Czech, Russian, German, English, Swedish, Finnish, Spanish, French, Italian, Portuguese, Arabic, Turkish, Hindi, Japanese, Chinese, Korean, Indonesian, and Thai languages. Using this corpus we could eliminate the influence of several factors: tagging style of a team, differences among language models created by different teams of researchers, differences among styles and genres. Using this parallel corpus has not changed the situation, the difference between correlation for the full corpus and its parallel subcorpora for the European languages stays within 0.05. However, the correlation for the Arabic and Turkish languages changes dramatically for up to ±0.25. Situation for word triples changed more than for word pairs.

We could suppose that the correlation depends on the list and number of investigated features. If we change the list of languages, then we change the number of considered features. This means that we could not correctly compare results for different lists of corpora and subcorpora. However, the proposed method allows drawing a correct sketch for the situation in similarity of syntax for different languages.

6 CONCLUSIONS

In this article, we introduced a new method for the quantitative evaluation of the syntax similarity. The experimental results demonstrate that the method draws a correct picture of similarity among world languages. Correlations between different corpora of the same language are extremely high – more than 0.95; languages belonging to the same language group have higher correlation inside the group. Slavic languages demonstrate the highest correlation inside the group. They are followed by Ancient languages. The third most similar group is Romance languages which are surprisingly weakly correlated with Latin – about 0.4.

Despite of the overall correct picture of language correlation, the method has several disadvantages. The basis of the method finds the most frequent syntactical connections for corpora of different languages and joins them into a list. Therefore, changing of the list of considered languages leads to changing of the list of the most frequent connections since there is a high probability for introduction of new connections. In this article, we use the Spearman rank correlation formula which is dependent to a feature set and incomparable for different sets. Thus, the results calculated for one set of languages are incomparable for another set. Consequently, we are not able to compare the results for different multilingual corpora directly, as we tried to do for PUD and GSD. We are not able to compare languages pair to pair, since such comparison will be conducted on different feature sets and, consequently, will not be also comparable. Thus, the method needs some further improvement.

In spite of the mentioned drawback, the method could be successfully applied for quantitative research in such area as comparative linguistics. Here it could allow finding new connections among languages, and track language changes. It is also applicable to quantitative stylometric for finding differences among genres and styles.

REFERENCES

[1] S. Padó and M. Lapata, “Dependency-based construction of semantic space models”, *Computational Linguistics*. **33** (2), 161–199 (2007).

[2] T. Mikolov, K. Chen, G. Corrado and J. Dean, “Efficient estimation of word representations in vector space” *arXiv preprint arXiv:1301.3781* (2013). https://arxiv.org/abs/1301.3781

[3] J. Pennington, R. Socher and C. Manning, “Glove: Global vectors for word representation” In *Proc. of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 1532-1543 (2014).

[4] M. Nickel and D. Kiela, “Poincaré embeddings for learning hierarchical representations” *Advances in Neural Information Processing Systems*, 30, 6338–6347 (2017).

[5] B. Dhingra, C.J. Shallue, M. Norouzi, A.M. Dai and G.E. Dahl, “Embedding Text in Hyperbolic Spaces” In *Proc. of the Twelfth Workshop on Graph-Based Methods for Natural Language Processing (TextGraphs-12)*, 59–69 (2018).

[6] E. Gibson and E. Fedorenko, “The need for quantitative methods in syntax and semantics research” *Language and Cognitive Processes*, **28** (1-2), 88-124 (2013).

[7] Ö. Dahl, *The Growth and Maintenance of Linguistic Complexity*, Amsterdam: John Benjamins Publishing Company (2004).

[8] R. Köhler, *Quantitative Syntax Analysis*, Berlin; Boston: De Gruyter Mouton (2012).

[9] R. Ferrer-i-Cancho and R.V. Solé, “The small world of human language” *In Proceedings of The Royal Society of London. Series B, Biological Sciences*, **268** (1482), 2261–2265 (2001).

[10] S. Fortunato, “Community detection in graphs” *Physics Reports*, **486,** 75–174 (2010).

[11] A.P. Mikhailov, A.P. Petrov and O.G. Pronicheva, “″Power-Information-Society″ Model“ *Mathematica Montisnegri*, **XLIV**, 73-83 (2019).

[12] S. Kübler, R. McDonald and J. Nivre, *Dependency Parsing*, San Rafael: Morgan & Claypool, (2009).

[13] A. Carnie, *Syntax: A generative introduction*, 3rd edition, Malden, MA: Wiley-Blackwell, (2013).

[14] V. Nastase, R. Mihalcea, D. Radev, “A survey of graphs in natural language processing” *Natural Language Engineering*, **21** (5), 665-698 (2015).

[15] J. Nivre, M.-C. de Marneffe, F. Ginter et al., “Universal Dependencies v1: A Multilingual Treebank Collection” In *Proc. of LREC-2016*, 1659-1666 (2016).

Received August 8, 2019