ITAT 2016 Proceedings, CEUR Workshop Proceedings Vol. 1649, pp. 11–17
http://ceur-ws.org/Vol-1649, Series ISSN 1613-0073, © 2016 V. Kuboň, M. Lopatková, T. Hercig

Searching for a Measure of Word Order Freedom

Vladislav Kuboň[1], Markéta Lopatková[1], and Tomáš Hercig[2,3]

[1] Charles University in Prague, Faculty of Mathematics and Physics,
Institute of Formal and Applied Linguistics
[2] Department of Computer Science and Engineering, Faculty of Applied Sciences,
University of West Bohemia, Univerzitní 8, 306 14 Plzeň, Czech Republic
[3] NTIS—New Technologies for the Information Society, Faculty of Applied Sciences,
University of West Bohemia, Technická 8, 306 14 Plzeň, Czech Republic
{vk,lopatkova}@ufal.mff.cuni.cz , tigi@ntis.zcu.cz

Abstract: This paper compares various means of measuring of word order freedom applied to data from syntactically annotated corpora for 23 languages. The corpora are part of the HamleDT project, the word order statistics are relative frequencies of all word order combinations of subject, predicate and object both in main and subordinated clauses. The measures include Euclidean distance, max-min distance, entropy and cosine similarity. The differences among the measures are discussed.

1 Motivation

The question of different features of natural languages has been engrossing theoretical linguists for hundred of years. They have been studying various language characteristics and classifying natural languages according to their properties, giving arise of a language typology, see esp. [1] and [2], or [3], to mention also the Czech tradition. These investigations led to a system of four basic language types, namely isolated, agglutinative, inflectional and polysynthetic languages.

Theoretical linguists have introduced an extensive list of relevant language features, a summary can be found, e.g., in the World Atlas of Language Structures (WALS) [4]. We will focus one particular phenomenon, word order of natural languages. While the classification of languages cannot be based upon a single phenomenon, the word order characteristics seems to belong among important features both for theoretical research and for practical natural language applications.

Languages are typically classified according to the degree of word order freedom to (more or less) fixed word order and free word order languages. The former type is often exemplified by English, where a word order position encodes a syntactic function (e.g., the first noun in an indicative sentence, having prototypically the function of subject, is followed by a predicative verb and a noun with the object functions); this property typically correlates with under-developed flection. The later type can be exemplified by Czech, where a syntactic function is encoded by morphological case marking [5], and word order expresses an information structure.

From the practical point of view, a freedom of word order to a great extent correlates with a parsing difficulty of a particular natural language (a language with more fixed word order is typically easier to parse than a language containing, e.g., non-projective constructions). On top of that, modern unsupervised methods of natural language processing might also profit from investigations of a similar kind as we present in this paper. If researchers would have an exact information about the properties of a language which they want to process using unsupervised methods, this knowledge might help them to choose an adequate processing method and/or to properly set its parameters.

The examination of a natural language typology have been traditionally based upon a systematic observation of linguistic material. However, linguistic research is in completely different position now: linguistic observations can be based on large amount of language data stored in corpora which have been growing not only in size but also in complexity of annotation during the last decade.

Moreover, several attempts to propose an unified annotation scheme – let us mention at least Stanford Dependencies and Stanford Universal Dependencies [6, 7, 8],[1] Google Universal Tags [9], Universal Dependencies [10],[2] – make it possible to use existing corpora for different languages.

In this paper we exploit the annotation developed in the frame of the HamleDT project (Harmonized Multi-Language Dependency Treebank [11]).[3]

We have already presented a study where we focused on word order properties of HamleDT treebanks and the languages ranking – we used a simple max-min distance based on a distribution of sentences among all variants of the word order. Here we re-calculate the results of the experiments described in [12] using standard measures like Euclidean distance, entropy, and cosine similarity.

In the remaining sections of the paper we are first going to introduce the data and tools used for the experiment, section 3 describes the setup of the experiment, section 4 presents the results and the final section discusses the conclusions and possible directions for future work.

[1] http://nlp.stanford.edu/software/stanford-dependencies.shtml
[2] http://universaldependencies.org/
[3] https://ufal.mff.cuni.cz/hamledt

2 Available Data Resources and Tools

HamleDT (Harmonized Multi-Language Dependency Treebank, [11])[4] is a compilation of existing dependency treebanks (or dependency conversions of other treebanks), transformed so that they all conform to the same annotation style. These treebanks as well as searching tools are available through a repository for linguistic data and resources LINDAT/CLARIN.[5]

2.1 Corpora

HamleDT integrates corpora for several tens of languages. Wherever it is possible due to license agreements, the corpora are transformed into a common data and annotation format, which enables a user – after a very short period of getting acquainted with each particular treebank – to search and analyze comfortably the data of a particular language.

The HamleDT family of treebanks is based on the dependency framework and technology developed for the Prague Dependency Treebank (PDT),[6] i.e., large syntactically annotated corpus for the Czech language [15]. Here we focus on the so-called analytical layer, describing a surface sentence structure (relevant for studying word order properties). Unfortunately, due to various technical and licensing restrictions, it was not possible to use all treebanks contained in HamleDT. Thus our effort focusses on 23 treebanks with available annotation on this syntactic layer, which still represent a wide variety of languages having various word-order properties.

As an example, Figure 1 shows three dependency representations for an English sentence in the HamleDT format.[7] Tables 1 and 2 provide an overview of the languages and the size of the corpora examined in our experiment.

2.2 Querying Tool

The advantage of using a common annotation framework for multiple treebanks also has a very useful consequence – instead of developing tailor-made searching tools we can apply a common tool to all treebanks we are analyzing. In the case of HamleDT, we can use the PML-TQ [16] search tool,[8] originally developed for processing the data from PDT.

Having the treebanks in the common data format and annotation scheme, the PML-TQ framework makes it possible to analyze the data in a uniform way. A typical user interested in monolingual data can use PML-TQ in an interactive way. Such approach would, of course, not work for our set of 23 treebanks, therefore we have used a command line interface which PML-TQ also provides. This interface makes it possible to create scripts that process a specified set of treebanks automatically.

Let us now give an example of a PML-TQ query used in our analysis. It counts sentences having an SVO word order in the main clause.

```
a-node $p :=
[ depth() = "1", id ~ "prague",
  afun = "Pred", tag ~ "^V",
  1x a-node
    [ afun = "Sb" ],
  1x a-node
    [ afun = "Obj" ],
  a-node
    [ afun = "Sb", ord < $p.ord ],
  a-node
    [ afun = "Obj", ord > $p.ord ] ];
>> give count()
```

The query searches data annotated in the praguian style (`id ~ "prague"`) for sentences containing verbs (`tag ~ "^V"`) with the analytical function of a predicate (`afun = "Pred"`) at the depth of one level below the technical root of the tree (`depth() = "1"`; i.e., this query focuses on the word order in main clauses, excluding coordinated predicates and disregarding also subordinate clauses). There must be exactly one subject and one object directly depending on the predicate (for the subject: `1x a-node [afun = "Sb"]`), the subject must precede the verb (`afun = "Sb", ord < $p.ord`), and the object must follow it (`afun = "Obj", ord > $p.ord`). The result of the query is the count of such sentences (`>> give count()`). The visualization of the PML-TQ query can be found in Figure 2.

3 The Experiment

In order to avoid possible bias caused by a combination of too many language phenomena in complicated sentences, we have decided to exclude all sentences containing coordinated predicates, subjects or objects from our experiment. The phenomenon of coordination is to some extent “orthogonal” to that of word order (especially in dependency-based approaches to a language description); thus the results might have been negatively influenced if coordination of verbs or the coordination of its direct dependents would be allowed.

In this experiment, we have focused on “full” structures, i.e., sentences with core syntactic structure consisting of subject, predicate and object. We have created several queries aiming at a thorough investigation of the phenomenon of the mutual position of these syntactic units.

[4] https://ufal.mff.cuni.cz/hamledt

[5] https://lindat.mff.cuni.cz/

[6] http://ufal.mff.cuni.cz/pdt3.0

[7] Data of each treebank in HamleDT are distributed in three annotation schemes – (a) the transformation of the treebank to the praguian style (used in PDT; leftmost in Figure 1), (b) the original annotation format of the given treebank (or its dependency transformation in case of non-dependency treebanks; in the middle of Figure 1), and (c) the transformation of the treebank to the Universal Dependencies style (rigthmost in the figure).

[8] https://lindat.mff.cuni.cz/services/pmltq/

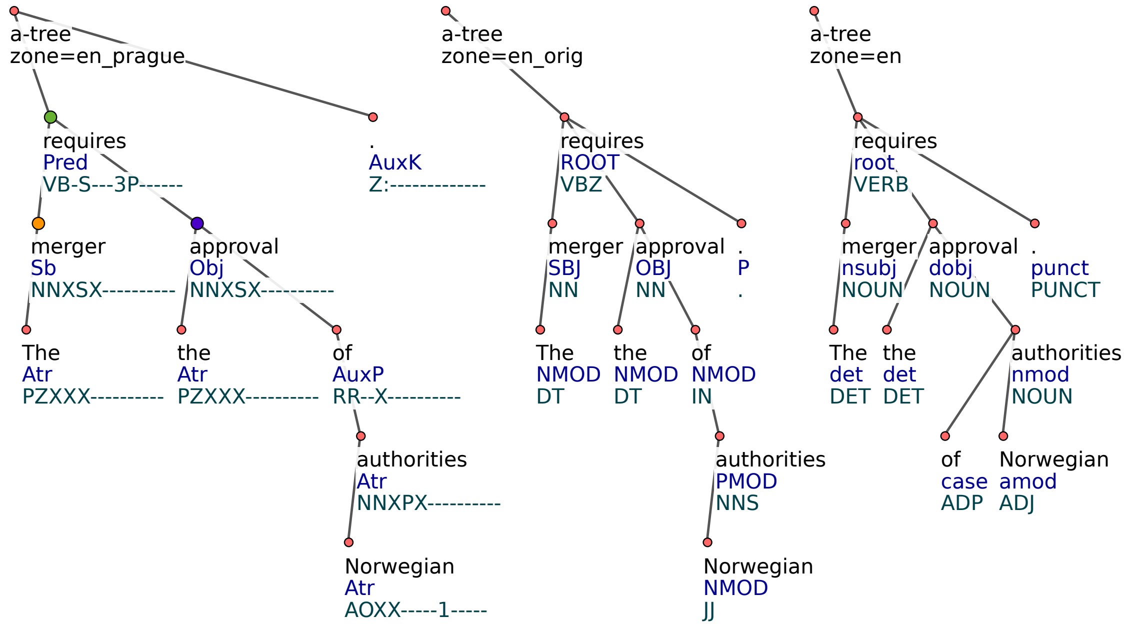

Figure 1: Three dependency representations of the sentence "The merger requires the approval of Norwegian authorities." in HamleDT 3.0. It is also one of the results of the query from Figure 2; nodes matching the query are slightly enlarged (in the left tree, nodes "requires", "merger" and "approval").

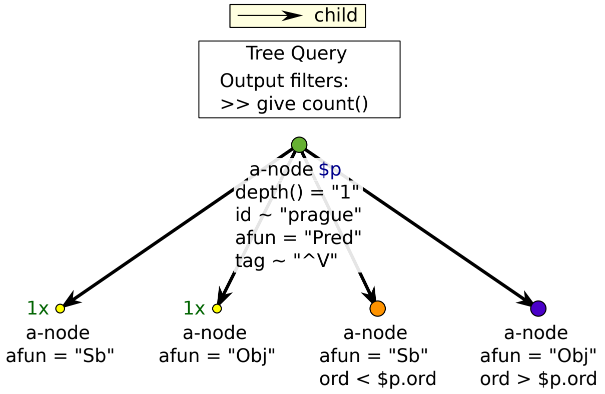

Figure 2: Visualization of the PML-TQ query

We have counted sentences for all six possible combinations (SVO, OVS, VSO, VOS, SOV, OSV), separately for main clauses and for subordinated clauses Table 1 and Table 2.

4 Comparison of Measures

The results presented in Tables 1 and 2 may serve as a basis for an estimation of a degree of word order freedom of individual languages. A typical mutual position of a subject, a predicate and an object constitutes one of the basic typological characteristic of a natural language. The problem of measuring the degree of word order freedom cannot be, of course, reduced only to this phenomenon, the freedom of word order of other sentence elements should probably be taken into account as well. Our decision to base the estimation on just these three constituents has several reasons. First of all, these constituents are present in a vast majority of sentences, they constitute a certain backbone of every sentence. Second, they are also relatively easily identifiable in all treebanks, regardless of the original annotation schemes. Although the HamleDT treebanks provide uniform annotation, the transformation of less frequent language phenomena from various languages may provide results which are not as uniform as we would like them to be. Last but not least, the three main constituents are located on top of the dependency tree, they do not require overly complex queries which might bring additional bias into the experiment.

The number we are looking for would describe how far is the distribution of individual variants of word order from the ideal absolutely free order of the main constituents. It is obvious that the languages with the highest degree of word order freedom would demonstrate the most equal distribution of sentences among all variants of the word order described in our tables, i.e., the frequency of all variants of the order of subject, verb and object will be equal to 16.66% (let us denote this "ideal vector" as Y)[9]. The difference between an actual distribution vector of each particular language from our table and this ideal vector then expresses the difference in word order freedom.

There are several measures which we can use for these

[9] The equal frequency of all variants actually means that there are probably no grammatical rules which would prefer any order of constituents over the others.

Treebank	Number of	Number of	SVO	OVS	VSO	VOS	SOV	OSV
	sentences	matches	(%)	(%)	(%)	(%)	(%)	(%)
Ancient Greek	21,173	1,648	24.6	21.1	5.1	5.2	27.2	16.8
Latin	3,473	395	25.1	6.8	8.6	4.1	41.3	14.2
Slovak	63,238	7,794	47.8	22.9	5.5	8.0	12.2	3.6
Slovenian	1,936	182	47.8	25.3	4.4	2.2	17.0	3.3
Czech	87,913	16,862	51.2	21.4	9.6	10.0	5.8	2.0
German	40,020	18,617	49.8	12.0	35.2	2.8	0.2	0.0
Tamil	600	132	0.0	6.8	0.0	0.0	59.1	34.1
Dutch	13,735	2,646	60.4	15.5	23.5	0.4	0.2	0.1
Spanish	17,709	5,569	61.3	20.4	1.0	0.5	16.5	0.4
Bengali	1,279	307	21.5	9.8	0.0	0.0	61.6	7.2
Romanian	4,042	1,132	62.1	12.9	0.4	0.8	23.6	0.2
Catalan	16,786	5,921	65.5	15.3	0.2	0.5	18.2	0.3
Polish	8,227	1,645	71.4	11.5	4.9	5.3	4.1	2.8
Telugu	1,600	254	2.4	3.9	0.0	0.0	69.7	24.0
Russian	34,895	6,194	72.2	15.2	1.6	3.5	4.2	3.4
Arabic	7,664	1,203	22.4	0.2	74.1	3.3	0.0	0.0
Turkish	5,935	802	3.2	13.3	0.6	0.1	79.2	3.5
Portuguese	9,359	2,879	80.7	9.1	1.9	5.1	3.1	0.2
Persian	12,455	2,480	15.9	0.2	0.1	0.0	81.1	2.8
English	18,791	8,585	83.1	7.3	0.3	0.0	0.0	9.2
Japanese	17,753	138	0.0	0.0	0.0	0.0	85.5	14.5
Estonian	1,315	359	85.5	4.7	7.8	1.1	0.8	0.0
Hindi	13,274	1,490	3.0	0.1	0.0	0.0	93.4	3.5

Table 1: Relative frequencies for word order variants in the main sentence in 23 studied languages.

Treebank	Number of	Number of	SVO	OVS	VSO	VOS	SOV	OSV
	sentences	matches	(%)	(%)	(%)	(%)	(%)	(%)
Ancient Greek	21,173	2,133	22.3	16.3	3.5	2.1	38.0	17.9
Latin	3,473	595	25.5	6.6	3.7	2.7	44.4	17.1
Slovak	63,238	6,354	54.0	14.8	3.1	4.2	12.2	11.7
Slovenian	1,936	137	30.7	28.5	5.1	2.2	8.8	24.8
Czech	87,913	11,849	60.2	12.2	4.8	4.9	10.4	7.6
German	40,020	9,655	14.9	0.7	8.2	0.3	70.0	6.0
Tamil	600	44	0.0	0.0	0.0	0.0	68.2	31.8
Dutch	13,735	1,155	8.7	0.4	1.8	0.1	73.4	15.5
Spanish	17,709	9,227	55.0	13.6	0.4	0.4	21.9	8.7
Bengali	1,279	54	1.9	3.7	0.0	0.0	81.5	13.0
Romanian	4,042	15	60.0	26.7	0.0	13.3	0.0	0.0
Catalan	16,786	8,612	50.6	16.6	0.1	0.7	23.3	8.7
Polish	8,227	331	71.9	6.9	2.1	2.1	10.6	6.3
Telugu	1,600	34	2.9	0.0	0.0	0.0	73.5	23.5
Russian	34,895	4,152	68.7	13.0	1.9	5.0	4.9	6.4
Arabic	7,664	1,816	48.3	0.1	17.4	34.2	0.0	0.0
Turkish	5,935	264	1.5	0.4	0.0	0.0	91.7	6.4
Portuguese	9,359	2,623	76.0	1.7	1.1	3.6	11.8	5.8
Persian	12,455	882	10.7	0.0	0.0	0.0	84.6	4.8
English	18,791	6,830	96.9	0.1	0.0	0.0	0.0	3.0
Japanese	17,753	538	0.0	0.0	0.0	0.0	70.6	29.4
Estonian	1,315	33	57.6	9.1	3.0	3.0	18.2	9.1
Hindi	13,274	1,374	0.3	0.0	0.0	0.0	95.5	4.2

Table 2: Relative frequencies for word order variants in subordinated sentences in 23 studied languages.

calculations.[10] Let us start with the simplest one, the max-min measure (marked as M_1 in the subsequent text):

$$M_1 = \max_{i \in 1,..n} x_i - \min_{i \in 1,..n} x_i$$

This measure has a value 0 for the ideal vector. The higher its value, the more fixed seems to be the word order of that particular language. The main advantage of this measure is its ability to reduce n-dimensional vectors into two dimensions only (leaving aside all four other values), thus enabling simple graphical representation. The same property also constitutes the greatest disadvantage of this measure, i.e. its insensitivity to subtle differences in distribution of values among the four variants which were actually left aside.

The second measure is the standard Euclidean distance between two vectors (marked as M_2 in the subsequent text):

$$M_2 = \|X - Y\| = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

In this formula, the symbol X represents the distribution of word order variants of a given language and Y is the “ideal vector” with equal distribution of frequencies. The Euclidean distance is more precise than M_1 because it reflects all six variants of the word order.

The third measure, very often used for measuring the similarity of two vectors in information retrieval, is the cosine similarity (marked as M_3 in the subsequent text):

$$M_3 = \frac{\sum_{i=1}^{n} (x_i \times y_i)}{\sqrt{\sum_{i=1}^{n} (x_i)^2} \times \sqrt{\sum_{i=1}^{n} (y_i)^2}}$$

Actually, because both M_2 and M_3 represent a distance between two vectors (although measured by different means and providing numerically different values), their results with regard to the estimation of word order freedom would be very similar, the main difference being the order of the numerical values of M_2 and M_3. While the values of M_2 are decreasing with the growing word order freedom, the values of M_3 are increasing.

Because M_2 and M_3 are in principle quite similar, let us therefore use one more measure which is also quite natural and widely used, namely the entropy (marked as M_4 in the subsequent text):

$$M_4 = -\sum_{i=1}^{n} P(x_i) \ln P(x_i)$$

The values $P(x_i)$ are the probabilities of individual word order variants. Because we do not know the exact probabilities, we are going to use their relative frequencies from Tables 1 and 2. The entropy is maximal for the equal distribution of relative frequencies (probabilities), minimal for an absolutely deterministic system which has only one acceptable type of the word order. In other words, the higher is the entropy for a particular language, the higher is its degree of word order freedom.

The results obtained for all four measures are presented in Tables 3 and 4. In order to enable an easier comparison of individual measures, we are presenting also the rank of all languages with regard to their degree of word order freedom for each particular measure. The ranks then show how similar the measures are. In both tables, the order of languages corresponds to their rank according to the M_1 measure applied to main sentences.

Table 3 shows the rank of individual languages with regard to the word order freedom calculated according to all measures mentioned above. It was calculated on main sentences with “full” structure, i.e. main sentences containing both subject and (exactly one) object, and although the rank according to each individual measure differs (with the exception of M_2 and M_3 which provide, not surprisingly, an absolutely identical rank), the highest rank always belongs to the two classical languages, Latin and Ancient Greek, closely followed by three Slavic languages (Slovak, Slovenian and Czech) and German. The languages with the most fixed word order are, according to all measures, English, Japanese, Estonian and Hindi.

When comparing both tables, we may notice some substantial differences in the word order freedom rank for main and subordinated clauses. We may identify two distinctive groups of languages which exhibit a relatively big rank shift. The languages with substantially higher degree of the word order freedom in subordinated clauses are Arabic, Catalan and Estonian. The languages with exactly opposite property are Bengali, German and Dutch. In case of Dutch we may recall the famous examples of phenomena exceeding the expressive power of context-free languages, namely the subordinated clauses such as *...dat Jan Piet de kinderen zag helpen zwemmen* (... that Jan saw Piet help the children swim) where the Dutch syntax requires a very strict order of words. Also in German, the word order in subordinated clauses follows much stricter rules than in the main ones. In this respect, the results obtained through our experiment correlate with the syntactic rules of the language.

5 Final Remarks and Conclusion

Although the results presented in this paper support to a relatively great extent the intuitive comprehension of the notion of word order freedom of ”big” European languages, there are at least two aspects of our experiment which are, according to our opinion, quite interesting. The first one is the fact that our experiment is based solely on data, publicly available in syntactically annotated corpora. Thanks to this fact the experiment does not require the knowledge of, or even the familiarity with all the languages under investigation. On the other hand, some of

[10] Actually, the word *measure* should not be understood as a strictly mathematical term. The cosine similarity is not a measure in a mathematical sense, it does not have all properties required by the mathematical definition of the term *measure*.

Treebank	M_1	Rank	M_2	Rank	M_3	Rank	M_4	Rank
Ancient Greek	27.18	1	0.2145	1	0.8852	1	1.6320	1
Latin	41.27	2	0.3166	2	0.7902	2	1.5134	2
Slovak	47.82	3	0.3744	3	0.7370	3	1.4273	3
Slovenian	47.80	4	0.3978	4	0.7162	4	1.3357	5
Czech	49.19	5	0.4052	5	0.7098	5	1.3714	4
German	49.76	6	0.4693	6	0.6563	6	1.0830	6
Tamil	59.09	7	0.5508	11	0.5955	11	0.8608	14
Dutch	60.24	8	0.5259	8	0.6132	8	0.9761	10
Spanish	60.89	9	0.5269	9	0.6125	9	1.0135	9
Bengali	61.56	10	0.5227	7	0.6155	7	1.0453	7
Romanian	61.93	11	0.5398	10	0.6032	10	0.9743	11
Catalan	65.53	12	0.5650	12	0.5856	12	0.9291	13
Polish	68.63	13	0.6037	13	0.5602	13	1.0231	8
Telugu	69.69	14	0.6154	14	0.5528	14	0.8101	15
Russian	72.18	15	0.6179	15	0.5512	15	0.9509	12
Arabic	74.15	16	0.6590	16	0.5267	16	0.6805	18
Turkish	79.05	17	0.6931	17	0.5075	17	0.7219	17
Portuguese	80.51	18	0.7047	18	0.5013	18	0.7352	16
Persian	81.09	19	0.7189	19	0.4938	19	0.5780	20
English	83.13	20	0.7337	20	0.4862	20	0.5857	19
Japanese	85.51	21	0.7652	22	0.4707	22	0.4138	22
Estonian	85.52	22	0.7571	21	0.4746	21	0.5673	21
Hindi	93.42	23	0.8416	23	0.4365	23	0.2935	23

Table 3: Ranks of individual languages for word order variants in the main sentence for all four measures.

Treebank	M_1	Rank	M_2	Rank	M_3	Rank	M_4	Rank
Ancient Greek	35.91	2	0.2954	2	0.8100	2	1.5034	2
Latin	41.68	3	0.3622	3	0.7479	3	1.4092	3
Slovak	50.88	6	0.4215	4	0.6956	4	1.3645	4
Slovenian	28.47	1	0.2840	1	0.8208	1	1.5149	1
Czech	55.35	9	0.4810	9	0.6470	9	1.2862	5
German	69.69	13	0.5959	12	0.5651	12	0.9586	12
Tamil	68.18	12	0.6320	14	0.5425	14	0.6254	18
Dutch	73.33	16	0.6359	15	0.5402	15	0.8314	15
Spanish	54.62	8	0.4584	6	0.6650	6	1.1902	8
Bengali	81.48	19	0.7181	19	0.4941	19	0.6276	17
Romanian	60.00	10	0.5312	10	0.6093	10	0.9276	13
Catalan	50.53	5	0.4233	5	0.6941	5	1.2345	7
Polish	69.79	14	0.6093	13	0.5566	13	0.9980	11
Telugu	73.53	17	0.6559	18	0.5284	18	0.6702	16
Russian	66.81	11	0.5760	11	0.5782	11	1.0734	9
Arabic	48.35	4	0.4629	7	0.6614	7	1.0267	10
Turkish	91.67	21	0.8234	21	0.4442	21	0.3409	21
Portuguese	74.91	18	0.6558	17	0.5284	17	0.8639	14
Persian	84.58	20	0.7498	20	0.4781	20	0.5252	20
English	96.88	23	0.8791	23	0.4211	23	0.1441	23
Japanese	70.63	15	0.6468	16	0.5336	16	0.6054	19
Estonian	54.55	7	0.4650	8	0.6597	8	1.2757	6
Hindi	95.49	22	0.8642	22	0.4271	22	0.1946	22

Table 4: Ranks of individual languages for word order variants for subordinate sentences for all four measures.

the corpora contained in the HamleDT set are too small to constitute a reliable source of information about the properties of a given language. However, this obstacle can be easily overcome in the future with the growing size and number of treebanks available under a common annotation scheme.

The second interesting aspect is the comparison of measures which give in principle very similar results and thus they support the claim that the phenomenon of word order freedom may be quantified practically by any reasonably selected measure. In other words, it is not necessary to develop any specialized measures just for this particular purpose, it is enough if we use the well known ones, such as the Euclidean distance or entropy.

Grant support

The work on this project was partially supported by the LINDAT/CLARIN project of the Ministry of Education, Youth and Sports of the Czech Republic (project LM2015071).

This work was also supported by the project LO1506 of the Czech Ministry of Education, Youth and Sports and by Grant No. SGS-2016-018 Data and Software Engineering for Advanced Applications.

This work has been using language resources and tools developed and/or stored and/or distributed by the LINDAT/CLARIN project of the Ministry of Education, Youth and Sports of the Czech Republic (project LM2015071).

References

[1] Saussure, F.: Course in General Linguistics. Open Court, La Salle, Illinois (1983) (prepared by C. Bally and A. Sechehaye, translated by R. Harris).

[2] Sapir, E.: Language. An Introduction to the Study of Speech. Harcourt, Brace and company, New York (1921) (http://www.gutenberg.org/files/12629/12629-h/12629-h.htm).

[3] Skalička, V.: Vývoj jazyka. Soubor statí. Státní pedagogické nakladatelství, Praha (1960)

[4] Dryer, M.S., Haspelmath, M.: The World Atlas of Language Structures Online. Harcourt, Brace and company, Leipzig (2005-2013) Available online at http://wals.info, Accessed on 2015-06-28.

[5] Futrell, R., Mahowald, K., Gibson, E.: Quantifying Word Order Freedom in Dependency Corpora. In: Proceedings of the International Conference on Dependency Linguistics (Depling 2015), Uppsala, Sweden, Uppsala University (2015)

[6] de Marneffe, M.C., MacCartney, B., Manning, C.D.: Generating typed dependency parses from phrase structure parses. In: Proceedings of LREC 2006. (2006)

[7] de Marneffe, M.C., Manning, C.D.: The Stanford typed dependencies representation. In: COLING Workshop on Cross-framework and Cross-domain Parser Evaluation. (2008)

[8] de Marneffe, M.C., Dozat, T., Silveira, N., Haverinen, K., Ginter, F., Nivre, J., Manning, C.: Universal Stanford Dependencies: A cross-linguistic typology. In: Proceedings of LREC 2014. (2014)

[9] McDonald, R., Nivre, J.: Characterizing the errors of data-driven dependency parsing models. In: Proceedings of EMNLP-CoNLL 2007. (2007)

[10] Nivre, J., de Marneffe, M.C., Ginter, F., Goldberg, Y., Hajič, J., Manning, C., McDonald, R., Petrov, S., Pyysalo, S., Silveira, N., Tsarfaty, R., Zeman, D.: Universal dependencies v1: A multilingual treebank collection. In: Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016), Portorož, Slovenia, European Language Resources Association (2016)

[11] Zeman, D., Dušek, O., Mareček, D., Popel, M., Ramasamy, L., Štěpánek, J., Žabokrtský, Z., Hajič, J.: HamleDT: Harmonized multi-language dependency treebank. Language Resources and Evaluation **48** (2014) 601–637

[12] Kuboň, V., Lopatková, M., Mírovský, J.: Analysis of Word Order in Multiple Treebanks. In: Proceedings of CICLing 2016. LNCS, Berlin Heidelberg, Springer-Verlag (2016)

[13] Lopatková, M., Kuboň, V.: Free or FixedWord Order: What can Treebanks Reveal? In Yaghob, J., ed.: Information Technologies – Applications and Theory, Prague, Charles University in Prague (2015) 23–29

[14] Kuboň, V., Lopatková, M.: Word-order analysis based upon treebank data. In Sidorov, G., Galicia-Haro, S., eds.: MICAI 2015: Advances in Artificial Intelligence and Soft Computing, Part I. Volume 9413., Berlin / Heidelberg, Springer (2015) 47–58

[15] Bejček, E., Hajičová, E., Hajič, J., Jínová, P., Kettnerová, V., Kolářová, V., Mikulová, M., Mírovský, J., Nedoluzhko, A., Panevová, J., Poláková, L., Ševčíková, M., Štěpánek, J., Zikánová, Š.: Prague Dependency Treebank 3.0. Charles University in Prague, MFF, ÚFAL, Prague (2013) (http://ufal.mff.cuni.cz/pdt3.0/).

[16] Pajas, P., Štěpánek, J.: System for Querying Syntactically Annotated Corpora. In: Proceedings of the ACL-IJCNLP 2009 Software Demonstrations, Suntec, Singapore, Association for Computational Linguistics (2009) 33–36