J. Yaghob (Ed.): ITAT 2015 pp. 23–29
Charles University in Prague, Prague, 2015

Free or Fixed Word Order: What Can Treebanks Reveal?

Vladislav Kuboň and Markéta Lopatková

Charles University in Prague, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics
Malostranské nám. 25, Prague 1, 118 00, Czech Republic
{lopatkova,vk}@ufal.mff.cuni.cz

Abstract: The paper describes an ongoing experiment consisting in the attempt to quantify word-order properties of three Indo-European languages (Czech, English and German). The statistics are collected from the syntactically annotated treebanks available for all three languages. The treebanks are searched by means of a universal query tool PML-TQ. The search concentrates on the mutual order of a verb and its complements (subject, object(s)) and the statistics are calculated for all permutations of the three elements. The results for all three languages are compared and a measure expressing the degree of word order freedom is suggested in the final section of the paper.

This study constitutes a motivation for formal modeling of natural language processing methods.

1 Introduction

General linguistics, see esp. [1, 2] studies natural languages from the point of view of similarities and differences in their syntactic structure, their development and historical changes, as well as from the point of view of language functions. It studies mutual influence of particular groups of features and, on the basis of similarities of language phenomena it introduces the so called language typology [3, 4]. The freedom or, on the other hand, strictness of the word order definitely belongs among the most important phenomena. General linguistics, for example, studies whether and how a particular language handles the order of words in sentences – whether the word is determined primarily by syntactic categories (e.g., a noun or a pronoun, without any additional morphological signs, located on the first sentential position represents a subject in English), or whether syntactic categories are primarily determined by other means than by the word order (for example, in Slavic languages, the subject tends to be a noun in the nominative case, regardless of its position in the sentence).

Particular natural languages cannot be, of course, strictly characterized by a single feature (for example word order), they are typically categorized into individual language types by a mixture of characteristic features. If we concentrate on word order, we study the prevalent order of the verb and its main complements – indo-european languages are thus characterized as SVO (SVO reflecting the order Subject, Verb, Object) languages. English and other languages with a fixed word order typically follow this order of words in declarative sentences; although Czech, Russian and other Slavic languages are the so-called languages with a high degree of word order freedom, they still stick to the same order of word in a typical (unmarked) sentence. As for the VSO-type languages, their representatives can be found among semitic (Arabic, classical Hebrew) or Celtic languages, while (some) Amazonian languages belong to the OSV type. These characteristics, which are traditionally mentioned in classical textbooks of general linguistics [5], have been specified on the basis of excerptions and careful examination by many linguists.

Today, when we have at our disposal a wide range of linguistic data resources for tens of languages, we can easily confirm (or enhance by quantitative clues) their conclusions. This paper represents one of the steps in this direction.

The Institute of Formal and Applied Linguistics at the Charles University in Prague, has established a repository for linguistic data and resources LINDAT/CLARIN[1]. This repository enables experiments with syntactically annotated corpora, so called treebanks, for several tens of languages. Wherever it is possible due to the license agreements, the corpora are trasformed into a common format, which enables – after a very short period of getting acquainted with each particular treebank – a comfortable search and analysis of the data from a particular language. The HamleDT[2] (HArmonized Multi-LanguagE Dependency Treebank) project has already managed to transform more than 30 treebanks from all over the world [6] into a common format.

In this pilot study we concentrate on three Indo-european languages which substantially differ by the degree of word freedom – Czech, German and English. We investigate their typological properties on the basis of the Prague Dependency Treebank [7], the English part of the Prague Czech-English Dependency Treebank[8] and the German treebank TIGER [9] by means of the interface of PML-TQ Tree Query [10], which enables the access to the treebanks from the HamleDT.[3]

2 Setup of the Experiment

The analysis of syntactic properties of natural languages constitutes one of our long term goals. The phenomenon of word order has been in a center of our investigations

[1] https://lindat.mff.cuni.cz/cs/
[2] http://ufal.mff.cuni.cz/hamledt
[3] https://lindat.mff.cuni.cz/services/pmltq/

for a very long time. Our previous investigations concentrated both on studying individual properties of languages with higher degree of word-order freedom (as, e.g., non-projective constructions (long-distance dependencies) [11] as well as on the endeavor to find some general measures enabling to more precisely characterize concrete natural languages with regard to the degree of their word-order freedom (see, e.g. [12]).

The experiment presented in this paper continues in the same direction. It is driven by the endeavor to find an objective way how to compare natural languages from the point of view of the degree of their word-order freedom. While the previous experiments concentrated on more formal approach, this one builds upon a thorough analysis of available data resources. Let us briefly introduce them in the subsequent subsections.

When investigating syntactic properties of natural languages, it is very often the case that the discussion concentrates on individual phenomena, their properties and their influence on the order of words. The mere presence of some phenomenon (or its more detailed properties) is, of course, important and definitely influences the degree of word-order freedom but this kind of investigation cannot be complete without stating also the quantitative properties of the given phenomenon. A linguistically interesting, but marginal phenomenon does not tell us so much as a basic phenomenon occurring relatively frequently. This observation constitutes the basis of our current experiment. In order to capture the quantitative characteristic of a natural language, let us take a representative sample of its syntactically annotated data and let us calculate the distribution of individual types of word order for the three main syntactic components – subject, predicate and object. It is obvious that the more free is the word order of a given language, the more equally they are going to be distributed.

2.1 Available Treebanks

The extensive quantitative analysis of the same linguistic phenomenon for different languages would not be feasible without a common platform which makes it possible to compare various data resources from the same point of view. Thanks to the initiative HamleDT[4] (HArmonized Multi-LanguagE Dependency Treebank) it is now possible to compare the data from more than 30 languages in a uniform way [6].

The HamleDT family of treebanks is based on the dependency framework and technology developed for the Prague Dependency Treebank (PDT)[5] [7], i.e., large syntactically annotated corpus for the Czech Language. Here we focus on the so-called analytical layer, i.e., the layer describing surface sentence structure (relevant for studying word order properties). The framework and its language independence was verified within (the English part of) the Prague Czech English Dependency Treebank (PCEDT)[6] [8] – within this project, syntactically annotated Penn Treebank[7] [13] was automatically transformed from the original phrase-structure trees into the dependency annotation.[8] Based on this experience, the HamleDT initiative goes further, syntactically annotated corpora for different languages are collected and transferred into the common format. Here we make use of the TIGER corpus[9] for the German language [9], the corpus with native phrase-structure annotation enriched with the information about the head for each phrase (and thus bearing also information on dependencies). Figures 2, 6 and 7 show sample trees for Czech, English and German, respectively, and Table 1 summarizes the size of these corpora.

corpus	# preds	lang	type	genre
PDT	79,283	Czech	manual	news
PCEDT	51,048	English	automatic	economy
TIGER	36,326	German	automatic	news

Table 1: Overview of all three treebanks (# preds represents the number of predicates in the given corpora)

2.2 HamleDT and PMLTQ Tree Query

For searching the data, we exploit a PML-TQ search tool,[10] which has been primarily designed for processing the PDT data. PML-TQ is a query language and search engine designed for querying annotated linguistic data [10] – it allows users to formulate complex queries on richly annotated linguistic data.

Having the treebanks in the common data format, the PML-TQ framework makes it possible to analyse the data in a uniform way – the following sample query gives us trees with an intransitive predicate verb (in a main clause), i.e. `Pred` node with `Sb` node and no `Obj` nodes among its dependent nodes, where `Sb` follows the `Pred`; the filter on the last line (`>> for $n0.lemma give $1, count() `) outputs a table listing verb lemmas with this marked word order position and number of their occurrences in the corpus, see also Figure 1.

```
a-node $n0 :=
[ afun = "Pred",
    child a-node $n1 :=
    [ afun = "Sb", $n1.ord > $n0.ord ],
    0x child a-node
    [ afun = "Obj"]]
>> for $n0.lemma give $1, count()
```

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

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

[6] http://ufal.mff.cuni.cz/pcedt2.0/cs/index.html

[7] https://www.cis.upenn.edu/ treebank

[8] This dependency-based surface annotation then served as a basis for deep syntactic dependency-based annotation of English; however, as for Czech, only surface structure is interesting for the studied phenomenon of word order.

[9] http://www.ims.uni-stuttgart.de/forschung/ressourcen/korpora/tiger.html

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

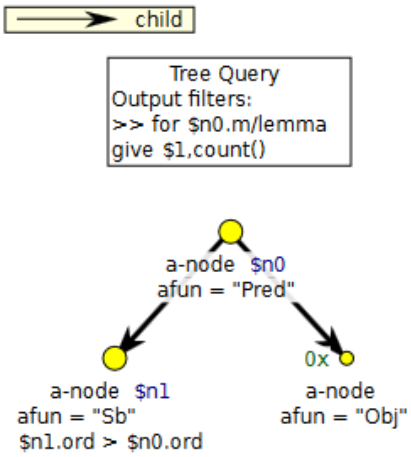

Figure 1: Visualization of the PML-TQ query

3 Analysis of Data

Let us now look at the syntactic typology of natural languages under investigation. We are going to take into account especially the mutual position of subject, predicate and direct object. After a thorough investigation of the ways how indirect objects are annotated in all three corpora, we have decided to limit ourselves – at least in this stage of our research – to basic structures and to extract and analyse only sentences without too complicated or mutually interlocked phenomena. Namely we focus on sentences with the following properties:

- A predicate under scrutiny belongs to the main clause (as e.g. in the sentence ***Jsou***$_{Pred}$ *vám nejasná některá ustanovení daňových zákonů?* 'Are$_{Pred}$ certain provisions of the tax laws unclear to you?', see the dependency tree in Fig. 2); i.e., we do not analyse word order of dependent clauses;
- We analyse only non-prepositional subjects and objects (compare e.g. with the sentence *V 2180 městech a obcích žije* ***na 2.6 milionu obyvatel***$_{Sb}$*;* 'There are (about 2.6 milion of inhabitants)$_{Sb}$ living in 2 180 towns and villages;', see Fig. 3);
- Sentences may contain coordinated predicates (as, e.g., predicates *následoval* and *opakovalo* in the corpus sentence *Vzápětí* ***následoval***$_{Pred}$ *další regulační stupeň a vše se* ***opakovalo***$_{Pred}$. 'The next level of regulation immediately followed$_{Pred}$ and everything repeated$_{Pred}$ again.', see Fig. 4);

  However, sentences with common subjects (or objects) are not taken into account (thus sentences as, e.g., ***Koupelna***$_{Sb}$ *nebo teplá* ***voda***$_{Sb}$ *nejsou trvale k dispozici.* 'A bathroom$_{Sb}$ or hot water supply$_{Sb}$ are not at the permanent disposal.', see Fig. 5 are not counted in the tables).[11]

[11] Including coordination phenomena in all their complexity would require much robust queries in any dependency framework; thus we have decided to disregard this type of sentences at all.

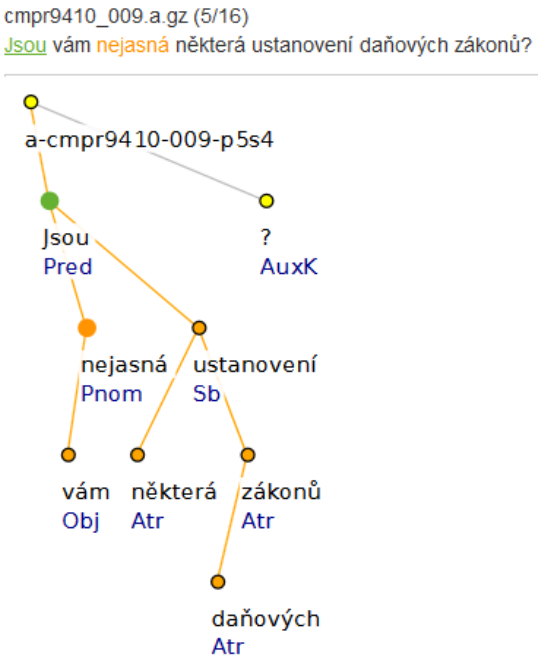

Figure 2: Sample Czech dependency tree from PDT

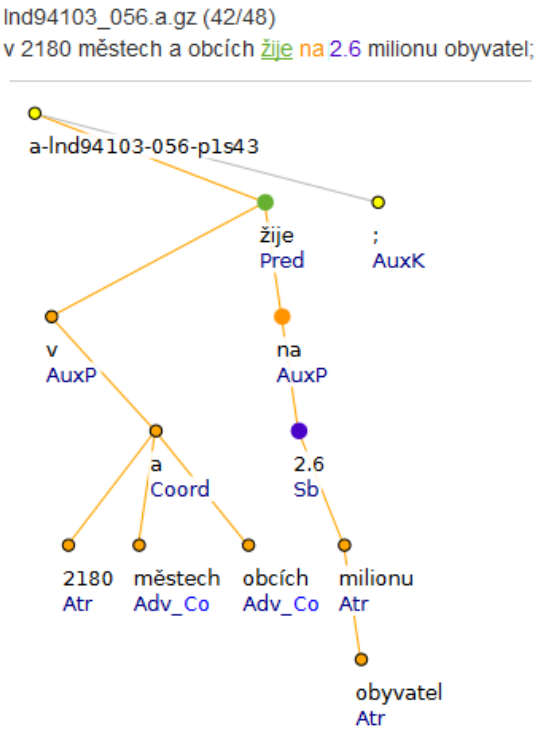

Figure 3: Sample Czech dependency tree from PDT with prepositional subject (excluded from the resulting tables)

3.1 Czech

The highest quality syntactically annotated Czech data can be found in the Prague Dependency Treebank; in fact, it is the only corpus we work with that has been manually annotated and thoroughly tested for the annotation consistency. The texts of PDT belong mostly to the journalism genre, it consists of newspaper texts and (in a limited scale) of texts from a popularizing scientific journal.

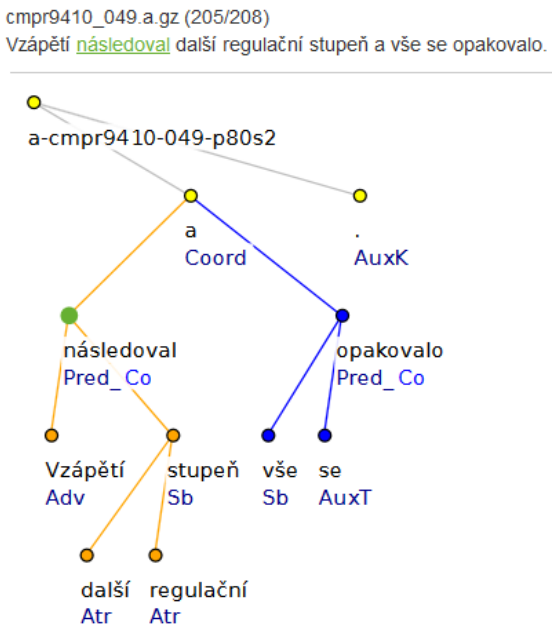

Figure 4: Sample Czech dependency tree from PDT with coordinated predicates (included in the resulting tables)

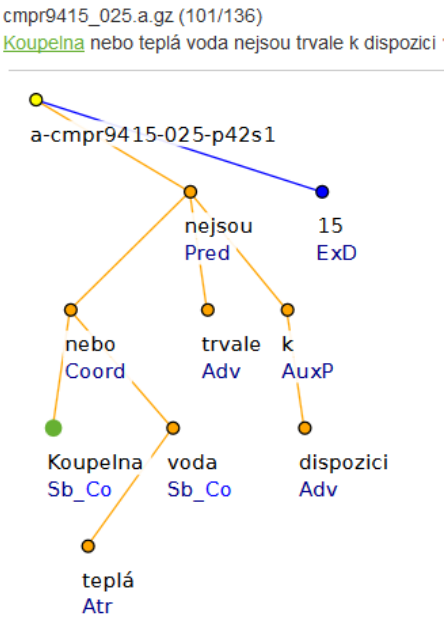

Figure 5: Sample Czech dependency tree from PDT with coordinated subject (excluded from the resulting tables)

The following Table 2 summarizes the number of sentences with intransitive verbs in main clauses in PDT with respect to the word order positions of `Sb` and `Pred` – we can see that the marked word order (verb preceding its subject) is quite common in Czech.[12]

The second table displays the distribution of individual combinations of a subject, predicate and a single object.

[12]In our settings, we do not checked the part of speech of the predicate; however, out of the 79,283 sentences conforming to the properties mentioned above, only 329 have other than verbal predicate.

Word order type	Number	%
SV	16,909	56.66
VS	12,932	44.34
Total	29,841	100.00

Table 2: Sentences with intransitive verbs

It is not surprising that the unmarked – intuitively "most natural" – word order type, SVO, accounts for only slightly more than half of cases. The relatively high degree of word order freedom is thus supported also quantitatively.

Word order type	Number	%
SVO	11,158	52.42
SOV	1,533	7.20
VSO	1,936	9.10
VOS	2,136	10.04
OVS	4,001	18.80
OSV	521	2.45
Total	21,285	100.00

Table 3: Sentences with a single object

Even more interesting (and also supporting the claim that the word order freedom of Czech is relatively high) are the results for sentences with at least two objects. They are summarized in Table 4. The distribution is even flatter than in Table 3 with all types being represented (even those starting with two objects, see the following example) and none of them exceeding 30%.

Plán mu *v úterý* ***předložil*** *velvyslanec USA v Chorvatsku Peter* ***Galbraith***.

Word order type	Number	%
SVOO	293	26.95
SOVO	223	20.52
SOOV	33	3.04
VSOO	45	4.14
VOSO	16	1.47
VOOS	27	2.48
OSVO	70	6.44
OSOV	10	0.92
OOSV	15	1.38
OOVS	124	11.41
OVSO	78	7.18
OVOS	153	14.08
Total	1,087	100.00

Table 4: Sentences with two objects

3.2 English

The statistics concerning the distribution of word-order types for English have been calculated on the English part of the Prague Czech English Dependency Treebank

(PCEDT). This corpus actually contains the same set of sentences as the Wall Street Journal section of Penn Treebank,[13] (see above for references) but unlike its predecessor, its syntactic structure has been annotated using dependency trees. As was mentioned above, the transformation on the surface syntactic layer was fully automatic, which has of course affected the quality of annotation.

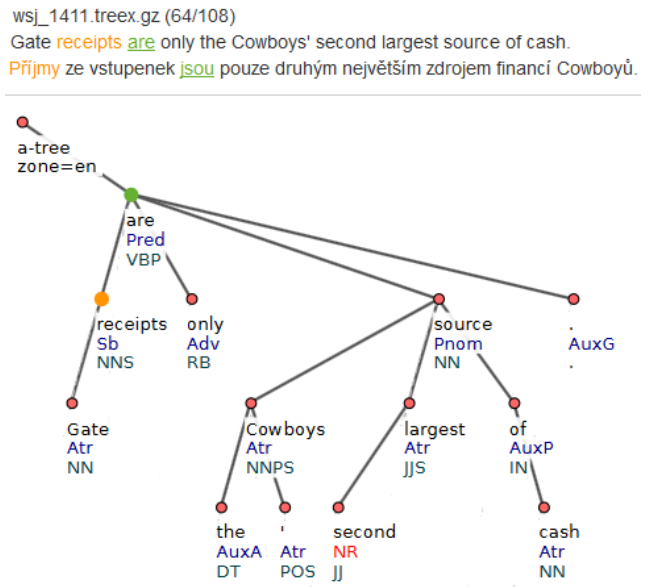

Figure 6: Sample English dependency tree from PCEDT

The statistics of different types of word order have been collected in the same manner as in the previous subsection. We have also applied identical filters as for Czech sentences from PDT. Table 5 contains data for sentences with intransitive verbs. Only as few as 40 sentences have other than verbal predicate.

Word order type	Number	%
SV	28,236	96.91
VS	900	3.09
Total	29,136	100.00

Table 5: English sentences with intransitive verbs

As we can see, the strict word order of English sentences manifests itself in a vast majority of sentences having the prototypical word order of the subject being followed by a predicate. The examples of the opposite word order include sentences containing direct speech with the following pattern:

*"It's just a matter of time before the tide turns," **says** one Midwestern **lobbyist**.*

Out of the 900 sentences with the reversed word order, as many as 630 contained the predicate *to say*, 121 *to be*. Each of all other verbs involved in these constructions were represented less than 10 times. In total, 23 verbs appear in these sentences at least twice, out of them 16 can be classified as verbs of communication (verba dicendi) (in total, it means 678 occurrences out of 822, i.e., 82,5 % of all occurrences with at least two hits in the corpus).

The results for sentences containing one object also strongly confirm the fact that the order Subject - Predicate - Object (SVO) is practically the only acceptable order in standard sentences. The remaining types of word order (representing only 1.06% sentences in the corpus) mentioned in Table 6 actually represented annotation errors in a vast majority of cases (esp. auxiliary verbs which have been quite often incorrectly annotated as Objects).

Word order type	Number	%
SVO	12,481	98.94
SOV	77	0.61
VSO	9	0.07
VOS	1	0.01
OVS	2	0.02
OSV	45	0.36
Total	12,615	100.00

Table 6: English sentences with a single object

It turns out that for English, it does not make sense to construct a similar table as Table 4 sentences with more than one object. The automatic annotation of PCEDT is, unfortunately, biased in what should be considered an Object (in the original Penn Treeank annotation, the verbal complements are labeled just as noun (or prepositional) phrases (NPs and PPs), no distinction between Objects and Adverbials.) As a consequence, adverbial constructions are very often incorrectly annotated as Objects and thus it is impossible to rely on this distinction (and the analysis shows that the numbers would be highly misleading).

3.3 German

German has more constraints on word order than Czech and less than English, therefore it constitutes a very natural candidate for our experiment. On top of that, there are also numerous high quality resources which can be exploited. We have used the German treebank conforming to the HamleDT initiative, which is located in the Lindat repository.[14]

The statistics for German were collected in the same way and with the same constraints as Czech and English ones. The statistics for German sentences with intransitive predicates are presented in Table 7.

The almost equal number of sentences with SV and VS word order types is quite surprising. The fact that SV represents the typical word order in declarative sentences, while VS in interrogative ones provides an obvious explanation. Unfortunately, this explanation does not

[13] The Czech part had been created as translation of original English sentences.

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

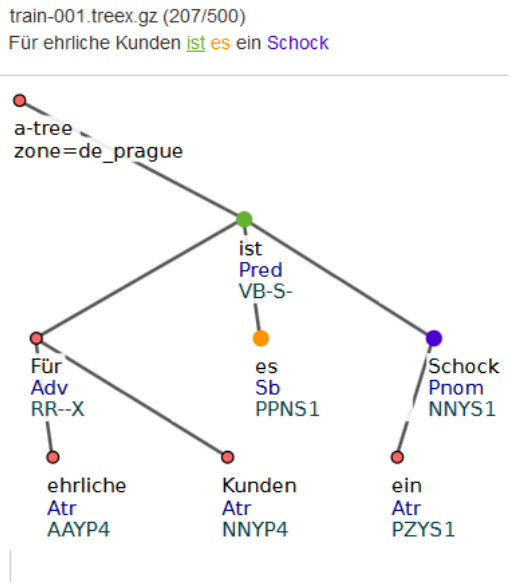

Figure 7: Sample German dependency tree from HamleDT

Word order type	Number	%
SV	6,165	56.67
VS	4,713	43.33
Total	10,878	100.00

Table 7: German sentences with intransitive verbs

cover all occurrences because the analyzed corpus (consisting mostly of newspaper texts) contains only a very small proportion of interrogative sentences. We have not investigated the reason for the surprisingly high number of VS sentences, but it definitely constitutes a very interesting topic for further research. The same is valid also for the results contained in Table 8, where we have found relatively high number of sentences having the word order of an interrogative sentence, too.

Word order type	Number	%
SVO	10,662	50.31
SOV	193	0.91
VSO	7,425	35.04
VOS	690	3.26
OVS	2,206	10.41
OSV	15	0.07
Total	21,191	100.00

Table 8: German sentences with a single object

Neither for German we have investigated the sentences with two or more objects due to annotation inconsistencies.

4 Proposed Measure of Word Order Freedom

The statistics presented in the previous section actually confirm the well known fact that Czech has the highest degree of word order freedom from all three languages investigated in our experiment. This fact is also reflected in the chart 8 comparing the results for sentences with one object for all three languages.

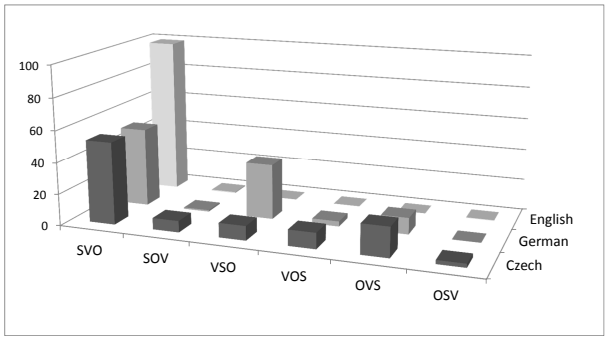

Figure 8: Comparison of results

Let us now try to suggest a formula which might allow to express the degree of word order freedom in a more precise way. Intuitively, the more free is the word order, the more equally distributed should be the results of all six word order types. The more strict the word order, the more distant are the values from the ideal (equal distribution). This leads directly to the application of a least squares method:

$$M = \frac{1}{6}\sqrt{\sum_{i=1}^{6}(V_i - Av)^2}, \qquad (1)$$

where M is the proposed measure, V_i the percentual value of the i-th word order type and Av is the average percentage for each word type (i.e., 100/6). For the three languages in our experiment we then get the following values:

- Czech: 6.82
- German: 19.20
- English: 36.79

These values seem to correspond to the intuitive feeling that the word order order of English is really strongly fixed, while German and Czech have more free word order with Czech having the highest degree of word order freedom. If we express the results in the form of percentages of the absolutely fixed word order (i.e., one of the word order types accounts for 100% and all others do not appear at all), we'll get the following results:

- Czech: 18.31%
- German: 51.52%
- English: 98.73%

5 Conclusions

The experiment described in this paper brought several interesting results which may be taken as a basis for further experiments. First of all, it shows that the endeavor to unify the annotation schemes used for various treebanks in the HamleDT project provides new opportunities for linguistic research. The treebank data can now be studied in a relation to other treebanks using the common search tool and obtaining results which are not dependent on peculiarities of individual annotation schemes.

These new opportunities have been demonstrated on a small-scale experiment involving three languages (Czech, German and English). We have managed to extract quantitative clues confirming the linguistic hypothesis about the degree of word order freedom of all three languages under consideration. The main advantage of our approach is the fact that our research is based on a large number of sentences of each language and thus it provides a representative sample of the actual language usage in a given genre. Contrary to theoretical linguistic research, our approach does not concentrate upon marginal (but definitely linguistically interesting) phenomena, but it is based upon the real language captured in the treebanks.

In the future we would like to continue the research in two directions. One will be the obvious endeavor to collect the statistics for more languages, the second one will be a more subtle treatment of linguistic phenomena appearing in treebanks, as, e.g. the investigation including also subordinated clauses or interrogative sentences.

Grant support

This paper exploits language data developed and/or distributed in the frame of the project MŠMT ČR LINDAT/CLARIN (project LM2010013).

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] Saussure, F.: Kurs obecné lingvistiky. Academia, Praha (1989) (translated by F. Čermák)

[3] 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`).

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

[5] Čermák, F.: Jazyk a jazykověda. Pražská imaginace, Praha (1994)

[6] 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

[7] Hajič, J., Panevová, J., Hajičová, E., Sgall, P., Pajas, P., Štěpánek, J., Havelka, J., Mikulová, M., Žabokrtský, Z., Ševčíková-Razímová, M.: Prague Dependency Treebank 2.0. LDC, Philadelphia, PA, USA (2006)

[8] Hajič, J., Hajičová, E., Panevová, J., Sgall, P., Bojar, O., Cinková, S., Fučíková, E., Mikulová, M., Pajas, P., Popelka, J., Semecký, J., Šindlerová, J., Štěpánek, J., Toman, J., Urešová, Z., Žabokrtský, Z.: Announcing Prague Czech-English Dependency Treebank 2.0. In: Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012), Istanbul, Turkey, ELRA, European Language Resources Association (2012), 3153–3160

[9] Brants, S., Dipper, S., Eisenberg, P., Hansen, S., König, E., Lezius, W., Rohrer, C., Smith, G., Uszkoreit, H.: TIGER: Linguistic Interpretation of a German Corpus. Journal of Language and Computation (2004), 597–620

[10] 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

[11] Holan, T., Kuboň, V., Oliva, K., Plátek, M.: On complexity of word order. Les grammaires de dépendance – Traitement automatique des langues (TAL) **41** (2000) 273–300

[12] Kuboň, V., Lopatková, M., Plátek, M.: On formalization of word order properties. In: Gelbukh, A., (ed.), Theoretical Computer Science and General Issues, Computational Linguistics and Intelligent Text Processing, CICLing 2012, volume 7181 of LNCS., Berlin / Heidelberg, Springer-Verlag (2012) 130–141

[13] Mitchell P. Marcus, Mary Ann Marcinkiewicz, B.S.: Building a large annotated corpus of English: the Penn Treebank. Computational Linguistics **19** (1993)