Word-Order Analysis Based Upon Treebank Data

Vladislav Kuboň(✉) and Markéta Lopatková(✉)

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

Abstract. The paper describes an experiment consisting in the attempt to quantify word-order properties of three Indo-European languages (Czech, English and Farsi). The investigation 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. Unlike similar studies which concentrate either on purely linguistic or purely statistical approach, our experiment tries to combine both – the observations are verified against large samples of sentences from available treebanks, and, at the same time, we exploit the ability of our tools to analyze selected important phenomena (as, e.g., the differences of the word order of a main and a subordinate clause) more deeply.

The quantitative results of our research are collected from the syntactically annotated treebanks available for all three languages. Thanks to the HamleDT project, it is possible to search all treebanks in a uniform way by means of a universal query tool PML-TQ. This is also a secondary goal of this paper – to demonstrate the research potential provided by language resources which are to a certain extent unified.

1 Introduction

The traditional linguistics, see esp. [13], devoted considerable effort to studying various language characteristics which enabled them to classify natural languages according to their properties from various points of view. The results of these investigations have led to a generally accepted system of language types (as, e.g., the classification into four basic language types, namely isolated, agglutinative, inflectional and polysynthetic languages [12,14]).

The language phenomena enabling linguists to classify languages are numerous. Probably the most comprehensible list of features can be found in the World Atlas of Language Structures (WALS) [2], which contains 151 chapters, each describing one language phenomenon and its distribution in the languages of the world. This clearly shows that the classification of languages cannot be based upon a single phenomenon (for example the number of cases or genders or the obligatory presence of a subject in a sentence etc.), they must be characterized by a mixture of typical features.

G. Sidorov and S.N. Galicia-Haro (Eds.): MICAI 2015, Part I, LNAI 9413, pp. 47–58, 2015.
DOI: 10.1007/978-3-319-27060-9_4

In this paper we present an investigation of one particular phenomenon, word order of natural languages which we believe is very important for theoretical research as well as for practical applications. The freedom of word order to a great extent determines how difficult it is to parse 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). Its importance is also indicated by the fact that it constitutes one of the 11 major areas of WALS.

When concentrating 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 have 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 [15], have been specified on the basis of excerptions and careful examination by many linguists.

Although all these investigation have been based upon a systematic observation of linguistic material, modern computational linguistics has brought into play much larger resources providing huge volumes of language material (which can be studied by means of automatic tools), and thus it may bring a deeper linguistic insight into the language typology. Thanks to a wide range of linguistic data resources for tens of languages available nowadays, we can easily confirm (or enhance by quantitative clues) the conclusions of traditional linguists. This paper represents a step in this direction.

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 for a long time. Our previous research concentrated both on studying individual properties of languages with higher degree of word-order freedom – as, e.g., non-projective constructions (long-distance dependencies) [4] – as well as on the endeavor to find some general measures enabling more precise characterization of individual natural languages with regard to the degree of their word-order freedom [5]. Unlike similar experiments, as e.g. [3], we try to concentrate on a deeper analysis and characterization of identified patterns.

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.

When investigating syntactic properties of natural languages, it is very often the case that the discussion focuses on individual phenomena, their properties and their influence on the order of words. In this paper we concentrate upon the analysis of quantitative properties of the word order phenomenon. In order to capture the quantitative characteristic of a particular natural language, we exploit a representative sample of its syntactically annotated data and calculate the distribution of individual types of word order for the three main syntactic components – subject, predicate and object. The statistics is calculated separately for main and subordinated clauses.

2.1 HamleDT and Available Treebanks

The tools and resources we are exploiting in this paper can be found in 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 license agreements, the corpora are transformed into a common format, which enables a user – 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 project[2] (HArmonized Multi-LanguagE Dependency Treebank) [16] has already managed to transform 42 treebanks from all over the world into a common format.

The HamleDT family of treebanks is based on the dependency framework and technology developed for the Prague Dependency Treebank (PDT),[3] i.e., large syntactically annotated corpus for the Czech language [1]. Here we focus on the so-called analytical layer, i.e., the layer describing surface sentence structure (relevant for studying word order properties). The English corpus included in HamleDT is the well known Penn Treebank[4] [7], which was automatically transformed from the original phrase-structure trees into the dependency annotation. The third corpus used in our experiments is the Persian Dependency Treebank (PerDT),[5] the collection of sentences with syntactic and morphological annotations useful for natural language processing of the Farsi language.

Figure 1 shows a sample dependency tree for an English sentence in the HamleDT format and Table 1 summarizes the basic characteristics of all corpora used in our experiment.

2.2 PML-TQ Tree Query

For searching the data, we exploit the PML-TQ search tool,[6] which has been primarily designed for processing the PDT data. PML-TQ is a query language

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

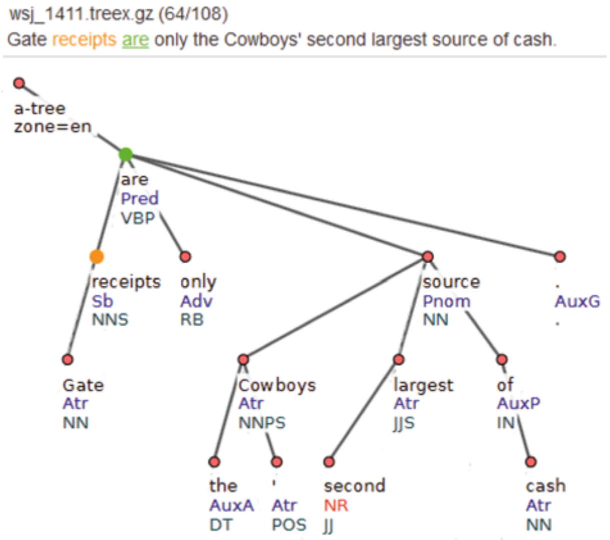
[2] http://ufal.mff.cuni.cz/hamledt.

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

[4] https://www.cis.upenn.edu/~treebank.

[5] http://dadegan.ir/en/perdt.

[6] https://lindat.mff.cuni.cz/services/pmltq/#!/treebanks.



wsj_1411.treex.gz (64/108)

Gate receipts are only the Cowboys' second largest source of cash.

Fig. 1. Sample English dependency tree in the HamleDT format

Table 1. An overview of three treebanks under scrutiny

Corpus	# Predicates	Type	Language	Genre
PDT	79,283	manual	Czech	news
Penn Treebank	51,048	manual	English	economy
PerDT	12.280	automatic	Farsi	news

and search engine designed for querying annotated linguistic data [9]; 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 in Fig. 2 gives us subtrees with an intransitive predicative verb (in a main clause), i.e. a `Pred` node with a `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.

3 Analysis of Data

Let us now look at the syntactic typology of natural languages under investigation. We are taking 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

```
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()
```

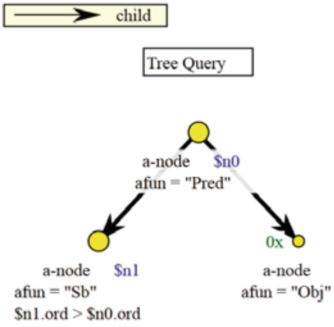


Fig. 2. Sample PML-TQ query and its visualization

analyze only sentences without too complicated or mutually interlocked phenomena. Namely we focus on sentences with the following properties:

- sentences may contain coordinated predicates (but subjects or objects common to coordinated verbs are not taken into account, due to a specific annotation of coordinated structures in the HamleDT scheme);
- we analyze only non-prepositional subjects and objects.

We are analyzing separately two types of clauses in our experiment: (i) main clauses expressing the main proposition, and (ii) subordinated clauses, i.e. clauses expressing predications embedded into the main proposition. While the former type can be easily identified in the data (main predicative verbs are labeled with the `Pred` function in the HamleDT treebanks), the latter type is more tricky: a dependency framework does not explicitly determine clauses [6] thus we approximate subordinated clauses as subtrees rooted in a finite verb (i.e., not infinitive or nominalized form of a verb) with other than `Pred` function. This approach allows us to gain deeper insight into the word order properties of the studied languages, as is documented in the following sections.

3.1 Czech

The highest quality syntactically annotated Czech data can be found in the Prague Dependency Treebank [1];[7] 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 popularizing scientific journal.

The following Table 2 summarizes the number of sentences with intransitive verbs as well as those with an omitted subject in main clauses in PDT, with respect to the word order positions of `Sb`, verbal `Pred` and `Obj` – we can see that

[7] https://lindat.mff.cuni.cz/services/pmltq/pdt30/.

the marked word order (verb preceding its subject or object preceding the verb) is quite common in Czech.

Table 3 displays the distribution of individual combinations of a subject, predicate and a single object. 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.

Table 2. Czech sentences with intransitive verbs and sentences with an omitted subject in a main clause

Word order	Number	%
SV	16,032	56.40
VS	12,395	43.60
Total	28,427	100.00
VO	8,616	75.56
OV	2,787	24.44
Total	11,403	100.00

Table 3. Czech sentences with a transitive verb in a main clause

Word order	Number	%
SVO	10,237	51.72
SOV	1,476	7.46
VSO	1,792	9.05
VOS	1,945	9.83
OVS	3,840	19.40
OSV	505	2.55
Total	19,795	100.00

Let us now turn our attention to an investigation whether the word order of subordinated clauses substantially differs from the results collected for main clauses.

As we can see, the distribution is only slightly different. The first interesting result concerns the clauses having a transitive verb on a second position, Table 5. These clauses tend to follow the unmarked word order more often than main clauses; however, the number of clauses with a subject in front (SVO) is higher by an almost equal difference as the number of clauses with the object in front (OVS) is lower. Similar correlation can be found for subordinated clauses either with an omitted subject or without object, Table 4.

The second observation concerns the subordinated clauses starting with a verb – their number is lower compared to the main clauses with the same property, regardless whether the verb is followed by a subject or an object. This drop is compensated by the increase of the cases when the verb is positioned further towards the end of the clause. Actually, the predicate positioned at the end of a main clause beginning with object is quite rare in Czech (only 2.55 % cases) and thus the increase to 7.82 % in Table 5 actually means that this special case of word order is 3 times more frequent in subordinated clauses. This is an interesting result which is not mentioned in Czech grammars.

3.2 English

The statistics concerning the distribution of word-order types for English have been calculated on the Wall Street Journal section of the Penn Treebank [7]

Table 4. Czech subordinated clauses with intransitive verbs and subordinated clauses with an omitted subject

Word order	Number	%
SV	8,625	64.66
VS	4,715	35.34
Total	13,340	100.00
VO	6,080	72.05
OV	2,358	27.95
Total	8,438	100.00

Table 5. Czech subordinated clauses with a transitive verb

Word order	Number	%
SVO	6,266	58.82
SOV	1,179	11.07
VSO	548	5.14
VOS	553	5.19
OVS	1,273	11.95
OSV	833	7.82
Total	381	100.00

whose syntactic structure has been transformed into dependency trees in the HamleDT project.[8] As was mentioned above, the transformation on the surface syntactic layer was fully automatic.

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 6 contains data for sentences with intransitive verbs and sentences with an omitted subject: the total number inadequately increases to 826 subject-less main clauses due to two reasons: first, coordinated subjects are not properly identified because of the specific annotation scheme for coordination in HamleDT; second, for analytical verb forms subject is rendered as dependent on auxiliary verbs, not as a complementation of lexical verb.

Table 6. English sentences with intransitive verbs and sentences with an omitted subject in a main clause

Word order	Number	%
SV	7,633	96.18
VS	303	3.82
Total	7,936	100.00
VO	815	98.67
OV	11	1.33
Total	826	100.00

Table 7. English sentences with a transitive verb in a main clause

Word order	Number	%
SVO	7,749	84.13
SOV	1	0.01
VSO	27	0.29
VOS	3	0.03
OVS	627	6.81
OSV	804	8.73
Total	9,211	100.00

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***.

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

Out of the 303 sentences with the reversed word order, as many as 94 contained the predicate *to say*, 88 *to be*, 38 *to do*, 12 *will*, 11 *to come* and 10 *to have*. Out of all other verbs involved in these constructions, only 7 were represented more than once. A deeper analysis reveals that there are four typical phenomena that cause the marked word order in these cases (in accordance with English grammar books):

- **quotative inversion** (see above);
- **stylistic inversion**, as e.g. ***Not only can they block Wellington*** *from raising money in Japan, ... but they might be able to ... , too.*;
- **locative inversion**, as e.g. ***Here are price trends*** *on the world 's major stock markets, ...*;
- **question**, as e.g. ***And why should holders expect*** *to realize that presumed "worth"?.*

The results for sentences containing transitive verbs, Table 7, also confirm the fact that the SVO order is the prevailing order in standard sentences. The remaining types of word order represent only 15.87 % sentences in the corpus. Some of those cases, especially those with a very low number of occurrences, namely SOV nad VOS, actually represent annotation errors (esp. auxiliary verbs which have been quite often incorrectly annotated as Objects). The queries also revealed an interesting fact concerning the OVS and OSV types of sentences. Again, a vast majority of verbs appearing in these main clauses can be classified as verbs of communication (verba dicendi) – with 566 and 658 appearances of the verb *to say* within OVS and OSV patterns, respectively.

Let us now look at the results collected for English subordinated clauses (Tables 8 and 9). As we can see, for the clauses not containing either a subject[9] or an object, the distribution even more strictly follows the prevailing pattern of main clauses. In these cases, the marked word order VS is again characteristic for (i) verbs of communication, as in *As a result,* ***says Mr. Geiger,*** *lawyers think twice before appealing a judge 's ruling ...*, and also for (ii) verbs of moving, as e.g. *At 2:43 p.m. EDT,* ***came the sickening news****: ...*.

The same is actually true also for transitive verbs, where the number of marked subordinated clauses starting with an object substantially decreases in favor of the prototypical word order (SVO). However, in these cases, verbs appearing in marked word order patterns are diverse and cannot be easily characterised with respect to their semantic classes.

3.3 Farsi

We have extracted 12,280 Farsi sentences with the same requirements as for Czech and English.[10] The sentences contain a verbal predicate, no coordination

[9] The number of subject-less subordinated clauses is inadequately high due to the same reasons as for main clauses: annotation scheme for coordination and analytical verb forms.

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

Table 8. English subordinated clauses with intransitive verbs and subordinated clauses with an omitted subject

Word order	Number	%
SV	5,785	98.07
VS	114	1.93
Total	5,899	100.00
VO	3,544	99.49
OV	18	0.51
Total	3,562	100.00

Table 9. English subordinated clauses with a transitive verb

Word order	Number	%
SVO	5,118	96.38
SOV	1	0.02
VSO	1	0.02
VOS	0	0.00
OVS	4	0.08
OSV	186	3.50
Total	5,310	100.00

of dependent words and only non-prepositional objects and subjects. Table 10 contains data for main clauses with intransitive verbs, showing a total dominance of the SV word order; more interesting are the results for main clauses not containing a subject: the position of the object seems to be relatively equally distributed with a slight preference to an object located to the right of the predicate. Quite surprising ale also the results presented in Table 11, namely the low number of sentences containing both a subject and an object. They account only for about 10 % of the entire corpus.

Table 10. Farsi sentences with intransitive verbs and sentences with an omitted subject in a main clause

Word order	Number	%
SV	5,975	99.70
VS	18	0.30
Total	5,993	100.00
VO	947	56.81
OV	720	43.19
Total	1,667	100.00

Table 11. Farsi sentences with a sentences with a transitive verb in a main clause

Word order	Number	%
SVO	447	35.70
SOV	795	63.50
VSO	1	0.08
VOS	0	0.00
OVS	2	0.16
OSV	7	0.56
Total	1,252	100.00

The whole picture looks slightly different if we look at the word order in subordinated clauses. Our query searching for embedded predication with verbal predicate has found 21,649 clauses. A vast majority of them (19,868 clauses) did contain neither subject nor object. Those containing only a subject and a predicate actually confirmed the results for main clauses, i.e., a dominance of the SV word order, Table 12 (top). In case of clauses without a subject, the slight majority has turned into the other direction with almost two thirds of clauses having the OV word order, Table 12 (bottom). Also the subordinated clauses having both object and subject (though rather rare) show similar distribution as in the main clauses, Table 13.

Table 12. Farsi subordinated clauses with intransitive verbs and subordinated clauses with an omitted subject

Word order	Number	%
SV	2,252	99.73
VS	6	0.10
Total	2,258	100.00
VO	443	37.77
OV	730	62.23
Total	1,173	100.00

Table 13. Farsi subordinated clauses with a transitive verb

Word order	Number	%
SVO	101	26.51
SOV	262	68.77
VSO	0	0.00
VOS	0	0.00
OVS	1	0.26
OSV	17	4.46
Total	381	100.00

3.4 Comparison of Results

Let us summarize the statistics presented in the previous section for the main and subordinated clauses separately. The results are displayed in the charts in Fig. 3 for all three languages.

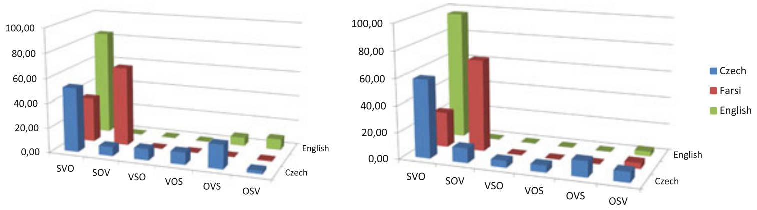


Fig. 3. Comparison of results – main clauses (left) and subordinated clauses (right)

3.5 Application of Results

The analysis of properties of the three languages described in this paper is definitely not a purely theoretical endeavor, it also has numerous practical consequences. One example of such consequence may be the creation of a test suite for a particular language. Unlike ordinary evaluation sets, which are in most cases randomly chosen data set aside from the training set, the test suites aim at more sophisticated selection of sentences or phenomena which should mirror their distribution in a given language. The guidelines for creating natural language test suites were mentioned for example in [8]. The method of corpus data analysis sketched in this paper may provide an important numerical input for particular language phenomena.

Another area where our method may be practically useful is the area of cross-lingual (or delexicalized) parsing where a parser trained on (one or more) syntactically annotated treebank(s) using non-lexical features is applied on a “similar” language with minimal available resources [17]. The notion of language similarity is crucial here: while it is often understood in terms of language relatedness (closely related languages are usually more or less similar) recent experiments show that even languages which are not related may bring useful information, see e.g. [11]. It is quite clear that in order to develop a similarity measure which would allow to determine the degree and the type of similarity, it is impossible to take into account all 151 phenomena of the WALS. One way is to focus on purely statistically-based approaches, like in [10]. Our experiment represents another, more linguistically-based approach to searching for a representative set of characteristics of natural languages.

4 Conclusions

The experiment described in this paper confirmed our initial hypothesis that a quantitative analysis of important linguistic phenomena based upon large scale syntactically annotated resources may bring interesting theoretical and practical conclusions. The ability to exploit a common annotation format of treebanks for queries analyzing individual linguistic phenomena across multiple languages brings observations which cannot be based upon a simple introspective analysis. Some examples of such observations are presented in this paper – the differences between the word order of subordinated clauses in Czech and Farsi compared to the word order of main clauses cannot be discovered solely on the basis of manual analysis of data: a slight shift in the frequency of constructions which are otherwise absolutely syntactically correct can be discovered only on the basis of quantitative data measured on a representative sample of a language.

The investigations described in this paper also represent a first step towards practical applications. In the future, we might be able to discover linguistic phenomena which are decisive for measuring the similarity of natural languages.

The future research will concentrate on two main directions. One is pretty obvious – to apply the queries used in this paper to a larger number of languages. The second one should investigate the possibility to create more detailed queries in PML-TQ enabling even deeper analysis of individual language phenomena.

Acknowledgments. This work has been using language resources developed, stored and distributed by the LINDAT/CLARIN project of the Ministry of Education of the Czech Republic (project LM2010013).

References

1. 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 (2013)
2. Dryer, M.S., Haspelmath, M.: The World Atlas of Language Structures Online. Harcourt, Brace and company, Leipzig (2005–2013). http://wals.info, Accessed on 28 June 2015

3. 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 University, Uppsala, Sweden (2015)
4. Holan, T., Kuboň, V., Oliva, K., Plátek, M.: On complexity of word order. Les grammaires de dépendance - Traitement automatique des langues (TAL) **41**(1), 273–300 (2000)
5. Kuboň, V., Lopatková, M., Plátek, M.: On formalization of word order properties. In: Gelbukh, A. (ed.) CICLing 2012, Part I. LNCS, vol. 7181, pp. 130–141. Springer, Heidelberg (2012)
6. Lopatková, M., Homola, P., Klyueva, N.: Annotation of sentence structure: capturing the relationship between clauses in Czech sentences. Lang. Res. Eval. **46**(1), 25–36 (2012)
7. Marcus, M.P., Marcinkiewicz, M.A., Santorini, B.: Building a large annotated corpus of English: the penn treebank. Comput. Linguist. **19**, 313–330 (1993)
8. Oepen, S., Netter, K., Klein, J.: TSNLP - Test suites for natural language processing. CSLI Lecture Notes (1998)
9. Pajas, P., Štěpánek, J.: System for querying syntactically annotated corpora. In: Proceedings of the ACL-IJCNLP 2009 Software Demonstrations, pp. 33–36. Association for Computational Linguistics, Suntec, Singapore, August 2009
10. Rosa, R., Žabokrtský, Z.: KL_{cpos^3} - a Language Similarity Measure for Delexicalized Parser Transfer (2015)
11. Rosa, R., Žabokrtský, Z.: MSTParser Model interpolation for multi-source delexicalized transfer. In: Proceedings of the 14th International Conference on Parsing Technologies, pp. 71–75. Association for Computational Linguistics, Stroudsburg (2015)
12. 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
13. Saussure, F.: Course in General Linguistics. Open Court, La Salle (1983). (prepared by C. Bally and A. Sechehaye, translated by R. Harris)
14. Skalička, V.: Vývoj jazyka. Soubor statí. Státní pedagogické nakladatelství, Praha (1960)
15. Čermák, F.: Jazyk a jazykověda. Pražská imaginace, Ptraha (1994)
16. Zeman, D., Dušek, O., Mareček, D., Popel, M., Ramasamy, L., Štěpánek, J., Žabokrtský, Z., Hajič, J.: HamleDT: harmonized multi-language dependency treebank. Lang. Res. Eval. **48**(4), 601–637 (2014)
17. Zeman, D., Resnik, P.: Cross-language parser adaptation between related languages. In: IJCNLP 2008 Workshop on NLP for Less Privileged Languages, pp. 35–42. Asian Federation of Natural Language Processing, Hyderabad (2008)