Valency and Word Order in Czech – A Corpus Probe

Kateřina Rysová, Jiří Mírovský
Charles University in Prague, Faculty of Mathematics and Physics
Ovocný trh 5, 116 36 Praha 1, Czech Republic
Email: rysova@ufal.mff.cuni.cz, mirovsky@ufal.mff.cuni.cz

Abstract

We present a part of broader research on word order aiming at finding factors influencing word order in Czech (i.e. in an inflectional language) and their intensity. The main aim of the paper is to test a hypothesis that obligatory adverbials (in terms of the valency) follow the non-obligatory (i.e. optional) ones in the surface word order. The determined hypothesis was tested by creating a list of features for the decision trees algorithm and by searching in data of *the Prague Dependency Treebank* using the search tool *PML Tree Query*. Apart from the valency, our experiment also evaluates importance of several other features, such as argument length and deep syntactic function. Neither of the used methods has proved the given hypothesis but according to the results, there are several other features that influence word order of contextually non-bound free modifiers of a verb in Czech, namely position of the sentence in the text, form and length of the verb modifiers (the whole subtrees), and the semantic dependency relation (functor) of the modifiers.

Keywords: syntax, word order, valency, treebank, decision trees algorithm

1. Introduction and Motivation

We use a treebank and the decision trees algorithm to evaluate a hypothesis that being an obligatory member of a verbal valency frame plays an important role in the word order in Czech, as it does in some other languages. Additionally, we try to describe the strength of some other word order factors.

The knowledge of force of individual word order phenomena in a language rich in morphology can help in tasks concerning language generation, e.g. machine translation. Unsuitable word order in translated texts (especially in case of inflectional languages) is still the task to solve in machine translation (cf. e.g. Steinberger, 1994 and 1992; Engelkamp, Erbach, Uszkoreit, 1992; He, Liang, 2011). In this paper, we focus on testing valency as a word order factor in Czech by studying results of a series of corpus queries and the decision trees algorithm.

The hypothesis will be tested on adverbials of Direction (DIR1: "from where"; DIR2: "which way"; DIR3: "to where"), Locative, Manner and Extent in the role of obligatory sentence members (i.e. we deal with structures like: optional adverbial of any type – obligatory adverbial of Directional or Locative or Manner or Extent).

2. Theoretical Framework

Prague generative linguists Sgall, Hajičová and Buráňová (1980) assume that there is a general language phenomenon connected with word order called *systemic ordering*: the sequence of contextually non-bound sentence elements is not arbitrary but grammatically fixed. This sequence varies for different languages but the fact that it exists should be language independent.

There are, however, other phenomena that influence the word order. Flämig (1991) claims that valency is one of the important factors influencing word order in German: the obligatory sentence elements follow the non-obligatory ones. Valency is a universal linguistic phenomenon and its effects on systemic ordering in Czech have not yet been studied.

The aim of the paper is to test the valency-hypothesis for Czech and to find out whether the obligatoriness in verbal valency works in accordance with the systemic ordering or is one of the factors that cause changes in the order of sentence elements. By studying relevant cases in the Prague Dependency Treebank 2.0 (Hajič et al., 2006), we assemble a list of features and use a machine learning algorithm (decision trees) to evaluate the hypothesis and we complete this method with a linguistics analysis of sentences from the Prague Dependency Treebank 2.0.

3. Language Material

To test the hypothesis that the (contextually non-bound) obligatory sentence elements follow the (as well contextually non-bound) non-obligatory ones in the word order, we use language data from the Prague Dependency Treebank 2.0 (PDT).

PDT is a treebank of Czech written journalistic texts annotated manually at three layers: the morphological layer where each token is assigned a lemma and a POS tag, the so-called analytical layer, at which the surface-syntactic structure of the sentence is represented as a dependency tree, and the tectogrammatical layer, at which the linguistic meaning of the sentence is captured. Almost 50 thousand sentences have been annotated on all three layers.

At the tectogrammatical layer, the meaning of the sentence is represented as a dependency tree structure. Nodes of the tectogrammatical tree represent auto-semantic words and are labelled with a large set of attributes. Among the most important ones, there are a

tectogrammatical lemma (t-lemma) and a functor (semantic relation between the governing and depending node, e.g. Predicate (PRED), Actor (ACT), Patient (PAT), Location (LOC)). Additionally, the tectogrammatical layer includes the annotation of information structure attributes, coreference, and links to a verb valency lexicon.

Thanks to the manual annotation of the Sentence Information Structure – the sentence members (nodes) are annotated either as contextually bound or contextually non-bound. Since the context is supposed to be the highest and strongest factor influencing the word order in Czech, the presented analysis was performed only on sentence members labeled as contextually non-bound.

The annotation of contextual boundness in *the Prague Dependency Treebank* is based on the theoretical approach of Functional Generative Description (FGP), established by Prague generative linguists (Sgall, 1964, 1967, 1979; Sgall et al., 1986, 2005).

The Prague Dependency Treebank is interlinked with the valency lexicon *PDT-Vallex* (Urešová et al., 2007). Therefore, it is possible to get information about the valency obligatoriness or optionality of all individual sentence members (nodes) included in the Prague Dependency Treebank.

The distinction between obligatory and optional sentence members is based also on the theory of Functional Generative Description – especially on the valency theory carried out by Panevová (1974), see below.

3.1. Obligatoriness in the Valency in Functional Generative Description

The used valency lexicon (Urešová et al., 2007), or rather its theory, distinguishes obligatory and non-obligatory (optional) elements. As a criterion for obligatoriness, the dialogue test was introduced by Panevová (1974), see also Sgall, Hajičová, and Panevová (1986). In this context, the term obligatoriness is related to the presence of the given complementation in the deep (tectogrammatical) structure, and not to its (surface) deletability in a sentence. The dialogue test is based on the difference between questions asking about something that is supposed to be known to the speaker – because it follows from the meaning of the verb he has used, and questions about something that does not necessarily follow from its meaning. Answering a question about a semantically obligatory modification of a particular verb, the speaker – who has used the verb – cannot say: *I don't know.* Thus, for example, for the verb *přijet (= come/arrive)*, the modification answering the question *Kam? (= Where to?)* is obligatory, which can be seen from the impossibility of answering the question by saying *Nevím (= I don't know)*. The speaker used the verb *přijet*, so it would make no sense to answer the question about the goal by saying *Nevím (= I don't know)*. On the contrary, the speaker does not have to know answers to questions *Odkud? (= Where from?)* and *Proč? (= Why?)*, thus these modifications are for the given verb optional.

3.2. Free Verbal Modifiers

The Functional Generative Description distinguishes two types of verbal complementations: Inner Participants and Free Modifications. Inner Participants are (in this theory) Actor, Patient, Addressee, Effect and Origin; Free[1] Modifications are e.g. Locative, Manner, Cause, Aim, Means, and many others (see Mikulová et al., 2005).
We assume that there could be a difference between the behaviour of the Inner Participants[2] and the Free Modifications in the word order. This paper deals with the Free Modifications only.

4. Experiments

4.1. Free Verbal Modifiers Treebank Queries

To study the proposed theory, several searches were performed in the Prague Dependency Treebank 2.0, using PML-TQ[3] (PML Tree Query), a powerful client-server based query engine for treebanks (Pajas and Štěpánek, 2009), implemented as an extension for the tree editor TrEd[4], a highly customizable framework for treebank manipulation (Pajas and Štěpánek, 2008). Using the tools, we searched for sentences (utterances) that corresponded to the tested theory (in which obligatory elements followed non-obligatory ones) and sentences that contradicted the theory (in which non-obligatory elements followed obligatory ones). Both groups of sentences were linguistically analyzed and a quantitative proportion of samples was carried out. For the findings and a detailed analysis see below.

4.2. Machine Learning Experiments

During the linguistic analysis of the search results, we have assembled a list of features that seem to have an effect on the surface word order in Czech. We used them in a machine learning experiment to evaluate how much the features influence the word order in practice.
As a machine learning method, we used decision trees, namely the C5.0 algorithm[5] with 10-fold cross validation and boosting (features[6] of the C5.0 algorithm).

1 The terminus “free” does not correspond to the valency characteristics of the modification, it does not mean “optional” in the valency point of view. “Free Modification” is terminus technicus. Free Modifications can be both obligatory and optional. However, not all of them can be obligatory.
2 They are much more often obligatory and there is not enough data to verify the position of optional Inner Participants in the word order.
3 http://ufal.mff.cuni.cz/pmltq/
4 http://ufal.mff.cuni.cz/tred/
5 http://www.rulequest.com/see5-info.html
6 The arguments of the C5.0 command were: C5.0 -X 10 -b -I; -X 10 means the 10-fold cross validation, -b means boosting (multiple decision trees are created that co-decide on the class of the examples), -I means that the division of the data for the cross validation is not random – it stays the

4.3. Data

For the 10-fold cross validation in the decision trees experiments, we used full data available in PDT. The training and test cases were extracted from the data, each case consisted of (up to) 46 features. All the cases were selected from declarative sentences and had a form of a verb node governing exactly two non-generated contextually non-bound nodes representing free modifications (and possibly other contextually bound nodes).

8 of the 46 features were features of the governing verb or the whole sentence, 19 features were features of each of the two governed nodes (38 in total). They can be divided according to their nature into following groups (2 for the governing verb or sentence, 6 for each of the governed nodes):

- attributes of the governing verb (e.g. its functor, voice, aspect);
- attributes of the sentence (e.g. position in the document, position in the paragraph);
- grammatical aspects of one of the governed nodes (e.g. aspect, negation);
- semantic aspects of one of the governed nodes (e.g. semantic POS, proper name);
- form and length of one of the governed nodes and its subtree (e.g. length in words, length in characters, verbal modality, dependent clause);
- contextual properties of one of the governed nodes and its subtree (e.g. presence of coreference in the subtree, number of nodes in topic/focus);
- functor of one of the governed nodes;
- obligatoriness of one of the governed nodes in the valency frame of the governing verb.

In the experiments, we first trained and evaluated the whole set of features and then removed individual feature sets in turn one at a time to determine how much the given feature set contributed to the performance of the system.

5. Results of the Machine Learning Experiment

Table 1 demonstrates the results of the experiment. In each row, error rates performed by the classifier for the given feature set are given. Statistically significant differences from the full feature set (with 95 % one sided confidence interval) are marked by "*".

same in all experiments.

feature set	**error rate (%)**
full	18.8
without attributes of the governing verb	19.9
without attributes of the sentence	* 20.2
without all governing verb/sentence related features	* 20.4
without grammatical aspects of the governed nodes	19.2
without semantic aspects of the governed nodes	19.5
without form and length of the governed nodes and their subtrees	* 22.1
without contextual properties of the governed nodes and their subtrees	18.7
without functor of the governed nodes	* 21.8
without the feature of obligatoriness of the governed nodes	19.3

Table 1: Error rate of decision trees trained and tested on various sets of features

6. Linguistic Analysis

The determined hypothesis (the obligatory adverbials follow the optional ones in word order in Czech) was tested also by searching in data of the *Prague Dependency Treebank* using the search tool *PML Tree Query.*

First, we found number of sentences in which the (contextually non-bound) **obligatory** adverbials (adverbials of Directional, Locative, Manner and Extent)[7] followed the (contextually non-bound) optional adverbials of any type (adverbials like e.g. Temporal, Cause or Condition), see an illustrative Example (1)[8].

Second, we found the number of sentences in which the (contextually non-bound) optional adverbials of any type (adverbials like e.g. Temporal, Cause or Condition) followed the (contextually non-bound) obligatory adverbials (adverbials of Directional, Locative, Manner and Extent), see an illustrative Example (2)[9].

7 Only such types of adverbials were tested that can appear both obligatory and optional in the various sentences, i.e. in connection with various verbs (the selection of tested adverbial types was based on previous research; Rysová 2012). Most types of adverbials (as e.g. Temporal, Cause, Condition) are (in terms of valency) only optional in the theory of FGP; only a few of them can be (besides their optional function) also obligatory in some cases, e.g. *she bought some bread in a shop*.optional_adverbial_of_Locative vs. *she found herself in a strange city*.obligatory_adverbial_of_Locative – only similar adverbials like Locatives could be tested in our analysis.

8 In the Example (1), the **obligatory** adverbial of Directional follows the optional adverbial (adverbial of Temporal in this case) – i.e. Example (1) demonstrates a case corresponding to the hypothesis: obligatory adverbials follow the optional ones.

9 In the Example (2), the **obligatory** adverbial of Directional precedes the optional adverbial (adverbial of Temporal in this case) – i.e. Example (2) demonstrates a case that is contrary to the tested hypothesis.

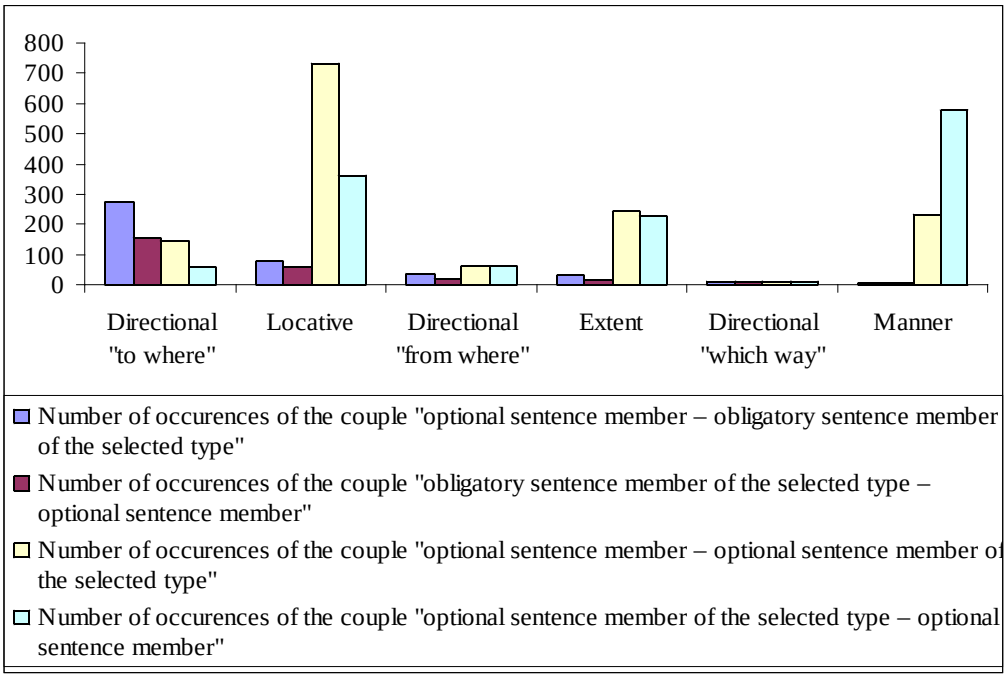

Chart 1: Number of occurrences of couples of sentence members in one and reverse word order. The obligatorinness or optionality of the sentence members is taken into account (based on data from the Prague Dependency Treebank 2.0).

The score of both searches was compared, see Chart (1) (blue and violet columns).

(1) *Přišla ráno*.optional_adverbial *domů*.obligatory_adverbial_Directional3.
Literally: (*She*) *came in the morning*.optional_adverbial *home*.obligatory_adverbial Directional3.

(2) *Přišla domů*.obligatory_adverbial Directional3 *ráno*.optional_adverbial.
Literally: (*She*) *came home*.obligatory_adverbial Directional3 *in the morning*.optional_adverbial.

For verification, we found also sentences in which the same types of contextually non-bound adverbials (Directional, Locative, Manner and Extent) appeared as **optional** sentence members. It could happen that their placement in the sentence is independent of their obligatoriness.

Firstly, the optional adverbials of Directional, Locative, Manner and Extent followed other contextually non-bound optional adverbials (adverbials like e.g. Temporal, Cause or Condition), see an illustrative Example (3)[10]; secondly, they preceded other contextually non-bound optional sentence members (adverbials like e.g. Temporal, Cause or Condition), see an illustrative Example (4)[11]. Again the score of both searches was compared, see Chart (1) (yellow and green-blue columns).

(3) *Dítě malovalo ráno*.optional_adverbial *na zeď*.optional_adverbial_Directional3.
Literally: *A child painted in the morning*.optional_adverbial *on a wall*.optional_adverbial_Directional3.

(4) *Dítě malovalo na zeď*.optional_adverbial_Directional3 *ráno*.optional_adverbial.
Literally: *A child painted on a wall*.optional_adverbial_Directional3 *in the morning*.optional_adverbial.

10 In the Example (3), the **optional** adverbial of Directional follows other optional adverbial (adverbial of Temporal in this case). Example (3) is a parallel to Example (1) – the adverbial of Direction stands more in right. The difference is that in Example (1), the adverbial of Direction is obligatory but in Example (3), the adverbial of Direction is optional.

11 In the Example (4), the optional adverbial of Directional precedes other optional adverbial (adverbial of Temporal in this case). Example (4) is a parallel to Example (2) – the adverbial of Direction stands more in left. The difference is that in Example (2), the adverbial of Direction is obligatory but in Example (4), the adverbial of Direction is optional.

6.1. Results of the Linguistic Research

The results of the research are captured in the Chart (1). The first (blue) column in group named as Directional "to where" in the Chart (1) expresses the number of sentences in which an **obligatory** sentence member of Directional "to where" follows an optional sentence member of an unspecified type (e.g. Temporal), see Example (1).

The second (violet) column in group named as Directional "to where" demonstrates the number of sentences in which an **obligatory** Directional "to where" precedes an optional sentence member of an unspecified type (e.g. Temporal), see Example (2).

The third (yellow) column expresses the number of sentences in which an **optional** sentence member of Directional "to where" follows an optional sentence member of an unspecified type (e.g. Temporal), see Example (3).

The fourth (green-blue) column in group named as Directional "to where" demonstrates the number of sentences in which an **optional** Directional "to where" precedes an optional sentence member of an unspecified type (e.g. Temporal), see Example (4).

The groups of columns labeled as Locative, Directional "from where", Extent, Directional "which way" and Manner express analogical types of information.

The data of the *Prague Dependency Treebank* did not fully confirm the hypothesis that obligatory adverbials (in general) follow the optional ones. E.g. the adverbials of Directional "to where" tend to follow other sentence members regardless of obligatoriness. Locatives have a similar tendency. Other analyzed types of sentence members did not occur with high frequency as obligatory members in PDT data and therefore we cannot draw more detailed conclusions about them.

7. Conclusion

The linguistic research demonstrated that the valency (obligatoriness of sentence members) is not a strong factor influencing surface word order in Czech. The adverbials of Directional "to where" and Locative tend to follow other sentence members regardless of the obligatoriness in data from the Prague Dependency Treebank (other analyzed types of obligatory adverbials occurred with low frequency in our material and it is not possible to draw more detailed conclusions about them).

The machine learning experiment shows that there are several features in our feature selection that influence word order of contextually non-bound free modifiers of a verb in Czech, namely position of the sentence in the text, form and length of the verb modifiers (the whole subtrees), and the semantic dependency relation (functor) of the modifiers. Other features may be of importance but our experiment did not prove it (with statistical significance). One of the features whose significance for word order in Czech could not be confirmed is obligatoriness, the main feature studied in this paper. In our opinion, supported also by the analysis of manual searches in the corpus, obligatoriness of verb modifications does not seem to play an important role in the word order in Czech.

8. Acknowledgements

We gratefully acknowledge support from the Grant Agency of the Czech Republic (grant P406/12/0658 *Coreference, discourse relations and information structure in a contrastive perspective*). This work has been using language resources developed and stored and distributed by the LINDAT/CLARIN project of the Ministry of Education, Youth and Sports of the Czech Republic (project LM2010013).

9. References

Engelkamp J., Erbach G., Uszkoreit, H. (1992) Handling Linear Precedence Constraints by Unication. In *ACL Proceedings.* Newark, 201–208.

Flämig, W. (1991). *Grammatik des Deutschen: Einführung in Struktur-und Wirkungszusammenhänge; Erarbeitet auf der theoretischen Grundlage der, Grundzüge einer deutschen Grammatik*. Akademie Verlag.

Hajič J., Panevová J., Hajičová E., Sgall P., Pajas P., Štěpánek J., Havelka J., Mikulová M., Žabokrtský Z., Ševčíková-Razímová M. (2006). *Prague Dependency Treebank 2.0.* Software prototype, Linguistic Data Consortium, Philadelphia, PA, USA, ISBN 1-58563-370-4, www.ldc.upenn.edu

He J., Liang H. (2011). Word-reordering for Statistical Machine Translation Using Trigram Language Model. In *Proceedings of the 5th International Joint Conference on Natural Language Processing*. Chiang Mai, Thailand, 1288–1293, <http://aclweb.org/anthology/I/I11/I11-1144.pdf>

Mikulová M. et al. (2005). *Annotation on the tectogrammatical layer in the Prague Dependency Treebank. Annotation manual.* Avaiable from http://ufal.mff.cuni.cz/pdt2.0/doc/manuals/en/t-layer/pdf/t-man-en.pdf

Pajas P., Štěpánek J. (2008). Recent advances in a feature-rich framework for treebank annotation. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, Manchester, pp. 673–680

Pajas P., Štěpánek J. (2009). System for Querying Syntactically Annotated Corpora. In: *Proceedings of the ACL-IJCNLP 2009 Software Demonstrations*, Association for Computational Linguistics, Suntec, Singapore, ISBN 1-932432-61-2, pp. 33-36

Panevová J. (1974). On verbal frames in functional generative description. *Prague Bulletin of Mathematical Linguistics, 22* (3–40).

Rysová K. (2012). Možnosti jednotlivých volných slovesných doplnění být obligatorním členem věty. In Čmejrková S.; Hoffmannová J.; Klímová J. (eds.). *Čeština v pohledu synchronním a diachronním: Stoleté*

kořeny Ústavu pro jazyk český. Prague: Karolinum, 615–620.
Sgall P. (1964). Generativní systémy v lingvistice. *Slovo a slovesnost 25* (274–282).
Sgall P. (1967). *Generativní popis jazyka a česká deklinace*. Prague: Academia.
Sgall P. (1979). Towards a definition of Focus and Topic. *Prague Bulletin of Mathematical Linguistics, 31* (3–27).
Sgall P., Hajičová E. (2005). The Position of Information Structure in the Core of Language. In Carlson G. N. a Pelletier F. J. (eds.). *Referency and Quantification: The Partee Effect.* Stanford (California): CSLI Publications, 289–302.
Sgall P., Hajičová E., Buráňová E. (1980). *Aktuální členění věty v češtině.* Academia, Praha, 172 pp.
Sgall, P., Hajičová, E., Panevová, J. (1986). *The meaning of the sentence in its semantic and pragmatic aspects.* Springer.
Steinberger R. (1992). Beschreibung der Adverbstellung im deutschen und englischen Satz im Hinblick auf Maschinelle Übersetzung. EUROTRA-D Working Paper 23, Saarbrücken (IAI), 2/92, 1–49, <http://langtech.jrc.ec.europa.eu/Documents/IAI-92_Steinberger_AdvPos.pdf>
Steinberger R. (1994). Treating “Free Word Order” in Machine Translation. In *COLING 94. The 15th International Conference on Computational Linguistics Proceedings.* Kyoto, 69–75, <http://langtech.jrc.ec.europa.eu/Documents/Coling-94_Steinberger.pdf>
Urešová Z., Štěpánek J., Hajič J. (2007). *PDT Vallex for PDT 2.0*. Institute of Formal and Applied Linguistics MFF UK Prague, http://ufal.mff.cuni.cz/lindat/PDT-Vallex.html.