A Linguistically-Based Segmentation of Complex Sentences

Vladislav Kuboň and **Markéta Lopatková** and **Martin Plátek**
Faculty of Mathematics and Physics
Charles University in Prague
{vk,lopatkova}@ufal.mff.cuni.cz, martin.platek@mff.cuni.cz

Patrice Pognan
CENTAL INALCO, Paris
mcertal@wanadoo.fr

Abstract

The paper describes a method of dividing complex sentences into segments, easily detectable and linguistically motivated units, which may provide a basis for further processing of complex sentences. The method has been developed for Czech as a language representing languages with relatively high degree of word-order freedom. The paper introduces important terms, describes a segmentation chart, the data structure used for the description of mutual relationship between individual segments and separators. It contains a simple set of rules applied for the segmentation of a small set of Czech sentences. The issues of segment annotation based on existing corpus are also mentioned.

Introduction

There are several reasons why it seems to be a good idea to develop methods for a linguistically based segmentation of natural language sentences, especially the complex ones, which constitute the greatest challenge for every natural language processing system.

One of the areas where the usefulness of such methods is pretty obvious, is the area of syntactic analysis, a task which needs to be solved in any natural language application field. It has been shown many times in the past that parsing success depends among other things also on the length of the input sentence. Let us mention for example (Oliva 1989; Kuboň 2001a) for rule-based syntactic analyzers. It is also not true that the information allowing to divide the complex sentence into individual clauses or segments is not important and that every stochastic parser will provide it for free in the parsing process – the substantially lower results (almost 10% difference) reported for Czech compared to English for identical parsers (see (Zeman 2004; Hajič et al. 1998; McDonald et al. 2005)) support the claim that even stochastic parsers have difficulties to cope with free-word order languages. The segment analysis might be exploited for example in the context of syntactic frameworks providing support for the division of the parsing process into several steps, as e.g. in the XDG theory of D. Duchier and others, see e.g (Debusmann et al. 2005).

The area of machine (aided) translation would also profit from a linguistically-based segmentation of complex sentences, especially where the translation memories are concerned. A reliable segmentation of sentences in translation memories would allow to increase both the number of matches and their reliability with regard to the surrounding context. Last but not least is the area of document search or information retrieval, where lingusitically justified segments might increase the search precision.

This paper contains a description of a segmentation method based on properties of Czech, a western Slavic language with very strict punctuation rules, a property extremely important for segmentation.

Basic Properties of Czech

Probably the most important property of Czech which makes it different from languages as e.g. English or French is the degree of word-order freedom. Czech allows not only for a free constituent order, but also for discontinuous constituents as e.g. in a frequently cited sentence *Vánoční nadešel čas.* [Lit.: Christmas$_{Adj}$ came time.]. These constructions are relatively frequent, there might be several cases of crossing dependencies (non-projective constructions) in a single clause (Holan et al. 2000).

The freedom of word order is compensated by relatively strict grammatical rules, especially with regard to the phenomena of syntactic agreement (which is required not only at the level of subject-predicate relationship [agreement in gender and number], but also in every complex nominal group [agreement in gender, case and number]). The rich inflection (different forms used in four genders and seven cases) makes it possible to distinguish which particular word forms are in agreement with each other.

There are also very strict rules for punctuation which in principle enable sorting out where the intertwined constituents belong to. The rules for punctuation are so strict that it is even possible to use them as a very reliable source of information for grammar checking. There are, for example, only very few exceptions to a general rule saying that there must be some kind of a separator (either a comma, colon, semicolon or a conjunction) between two finite verb forms of autosemantic verbs.

The segmentation method proposed in the following sections is based on these properties of Czech. A key role is being played by the separators (punctuation marks, conjunctions etc). Apart from some very marginal cases it is almost always possible to determine whether a particular punctuation mark or a word form represents a separator of clauses or their parts. Based on the context, it is also possible to classify the separators further, to divide them into opening ones (as e.g. subordinated conjunctions which almost always open a subordinated clause), closing ones (a comma following an embedded clause) and mixed ones, those which at the same time close a previous segment of a sentence or clause and open a following one (as e.g. a coordinating conjunction).

The first step of our segmentation method is based on determining which tokens (word forms, punctuation marks etc.) of the input sentence may be considered separators. For punctuation marks it is relatively easy, only a comma or a full stop are to a certain extent ambiguous in this respect. A comma may serve also as a decimal separator (Czech uses commas instead of decimal points), and a full stop is used (apart from an end of sentence marker) in ordinal numerals (instead of the English *-th* in 4^{th}, there is "4." in Czech), paragraph and section numbering (e.g. Section 4.1.2), abbreviations etc. All these exceptions are relatively easy to distinguish on the basis of a local context, therefore, from the segmentation point of view, a full stop also constitutes a relatively unambiguous (closing) separator.

Although there is usually no doubt whether a comma is a separator or not, the problem is a bit more complex with regard to its syntactic role in a sentence. A comma may separate both individual members of a coordination inside a single clause as well as two clauses (or their parts). Sometimes it may also be a part of a more complex separator group, usually in connection with a subordinating conjunction. In such a case it may even be separated from the subordinate conjunction by one or more additional words, as e.g. in a sentence: *Dali gól, také ale dva dostali.* [Lit.: (they) scored goal, also but two conceded; – They scored, but they also conceded two goals.].

The Outline of the Method

Our method is based on the assumption that a thorough classification of separators accompanied by a morphemic information may provide enough information for constructing a chart describing a mutual position of individual segments in a sentence. Such a chart will probably be ambiguous due to the ambiguity issues mentioned above, but nevertheless it may provide a sound basis for subsequent processing, whether it is going to aim at an ambitious goal of creating a representation of a mutual position and a relationship of individual clauses in a complex sentence, or whether it will serve for any other kind of linguistic processing.

In this sense the method proposed in this paper falls in line with other techniques aiming at bridging the gap between results of morphological analysis (or tagging) and a full-scale rule-based syntactic analysis or stochastic parsing.

The segment analysis differs from similar approaches, out of which the chunking[1] is probably the most prominent one, by a strong stress on linguistic adequacy and a thorough classification of separators and segments.

Although the method presented in this paper had been designed primarily for Czech, it is rather useful for a whole group of related and typologically similar languages. It is still an open question whether the proposed method might not be useful even for typologically different languages. It seems to be clear that the information about individual segments of a sentence is hidden somewhere in the sentence itself, otherwise not even humans would be able to fully understand complex sentences. Finding that information might be more difficult for certain languages, but we believe that a thorough investigation of separators and segments being done for a language providing enough segmentation clues (as Czech definitely does) may help even in a more difficult task of analyzing segments in typologically different languages.

Let us now describe the method of segmentation a bit more formally (and precisely) in the following section.

Segmentation of Complex Sentences

The process of segmentation of a complex sentence needs two basic sources of data – a morphological analyzer providing a full set of morphological tags (not just a tagger guessing the most probable one) and a list of separators. The first requirement is clear, let us discuss the second one after we'll introduce a bit more formal definition of important terms.

Important Notions

In the sequel an input sentence is understood to be a sequence of lexical items $w_1w_2 \ldots w_n$. Each item w_i ($1 \leq i \leq n$) represents either a certain lexical form of a given natural language, or a punctuation mark, quotation mark, parenthesis, dash, colon, semicolon or any other special symbol which may appear in the written form of a sentence. All items are disjunctively divided into two groups – ordinary words and separators.

Let us call the words or punctuation marks which may separate two clauses (or two sentence members) **separators**. It is quite clear that there are at least three relatively easily distinguishable types of separators – opening ones, closing ones and mixed ones, those, which typically close the preceding clause or its part and open the following one. A typical opening separator is e.g a subordinating conjunction or a relative pronoun, a closing one is a full stop, question mark or exclamation mark at the end of a sentence, mixed separators are for example commas or coordinating conjunctions.

It is often the case that two clauses are separated by more than one separator (e.g. comma followed by *že* [that]), in some cases even combined with non-separators (emphasizing adverbs, prepositions, etc.). In such a case it would be more convenient to consider the whole sequence as a single item – let us call it a **separator sequence**.

Let $S = w_1w_2 \ldots w_n$ be a sentence of a natural language. A **segmentation of a sentence** S is a sequence of sections $D_0W_1D_1 \ldots W_kD_k$, where a particular section W_i ($1 \leq i \leq k$) represents so called **segment**, i.e. a (maximal)

[1] Very comprehensive explanation of this notion can be found for example at http://nltk.sourceforge.net/tutorial/chunking/

sequence of lexical items $w_j w_{j+1} \dots w_{j+m}$ not containing any separator, and section D_i ($0 \leq i \leq k$) represents a separator sequence composed of items $w_q w_{q+1} \dots w_{q+p}$. The section D_0 may be empty, all other sections D_i ($1 \leq i \leq k$) are non-empty. Each item w_i for $1 \leq i \leq n$ belongs to exactly one section D_j if it is a member of a separator sequence; in the opposite case, w_i belongs to exactly one W_j. A pair $D_{i-1}W_i$ consisting of an opening or mixed separator sequence D_{i-1} and following segment W_i is called an **extended segment**.

The section D_0 is usually empty for sentences which start with a main clause. D_0 is typically nonempty if a complex sentence starts with a subordinated clause, as e.g. in the sentence *Když jsem se probudil, zavolal jsem policii.* [When I woke up, I called the police.]. D_k represents the final punctuation mark at the end of a sentence.

The segmentation of a particular sentence can be represented by one or more **segmentation charts** that describe the mutual relationship of individual sections with regard to their coordination or subordination.

Each separator is represented by at least one node. If an opening separator represented by a node D_i has a subordinating function, a copy of the node D'_i is placed directly under a node D_i in the chart and it is connected by an arrow with the original node D_i. The closing separator may by also represented by a "raised" copy of a node D_i. Let us demonstrate example of a segmentation chart on the Czech complex sentence *Zatímco neúspěch bývá sirotkem, úspěch mívá mnoho tatínků, horlivě se hlásících, že zrovna oni byli u jeho početí.* [While failure is usually an orphan, the success tends to have many fathers, claiming eagerly that particularly they were present at its conception.], see Fig. 1.

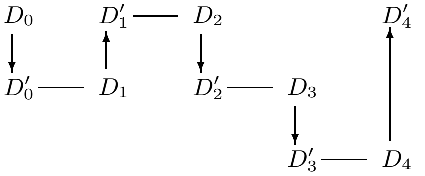

Figure 1: Example of segmentation chart

D_0 - *Zatímco* [While]
W_1 - *neúspěch bývá sirotkem* [failure is usually an orphan]
D_1 - ,
W_2 - *úspěch mívá mnoho tatínků* [the success tends to have many fathers]
D_2 - ,
W_3 - *horlivě se hlásících* [claiming eagerly]
D_3 - , *že* [that]
W_4 - *zrovna oni byli u jeho početí* [that particularly they were present at its conception]
D_4 - .

There is more than one chart in case that the segmentation of a sentence is ambiguous. It may happen if a separator is ambiguous – e.g. the Czech word form *jak*, which may be a noun [yak], a subordinating conjunction [as], a coordinating conjunction [both], or pronominal adverb [how] – or if a separator does not clearly indicate the relationship between both segments it separates, as e.g. comma.

In order to be able to present a basic set od rules for creating segmentation chart it is necessary to introduce a couple of new notions, at least informally.

A **subordination flag** is assigned to particular extended segment either if this segment contains any word form with one of the following morphological tags (for conjunctions, pronouns, and numerals, see (Hajič 2004)) or if it contains one of the listed pronominal adverbs:

- tag="J,.*" representing a subordinating conjunction;
- tag="P.*" representing a interrogative/relative pronoun, where the second position in the tag contains any of the following characters:
  – 4 (*jaký, který, čí, ...*),
  – E (*což*),
  – J (*jenž, již, ...*),
  – K (*kdo, kdož, kdožs*),
  – Q (*co, copak, cožpak*),
  – Y (*oč, nač, zač*);
- tag="C.*" representing numerals, where the second position in the tag is either
  – ? (*kolik*),
  – u (*kolikrát*) or
  – z (*kolikátý*);
- tag="D.*" for pronominal adverbs
  – adverbs (jak, kam, kde, kdy, proč)

For the sake of an easier explanation of mutual relationships of individual nodes of a segmentation chart in vertical direction we would like to introduce the notion of **chart layers**. In informal terms, a top layer of the chart (layer 1) corresponds to a main clause of the sentence and the numbers identifying layers increase in the top-down direction. The lower layers (layers with higher numbers) represent subordinated clauses. If a clause contains an embedded clause (fully embedded, that is the main clause is divided into two non-empty parts), the "tail" of the main clause is located in the same layer as its "head"; the same holds also for subordinated clauses with more deeply embedded clauses.

Building a Separator List

It is obvious that the segmentation depends very much on the quality of resources used. While the Czech morphological analysis has been tested by numerous applications including a spelling and grammar checker for Czech, the list of separators and separator sequences needs to be created. For this purpose we have decided to exploit the large scale resources available for Czech, the Prague Dependency Treebank[2] (Hajič et al. 2001) and the Czech National Corpus[3], as well as Czech grammar books, as e.g (Šmilauer 1958).

For each separator identified in the corpora we create a database entry. The entry contains the following information:

- A category of the separator – either opening, closing or mixed one;
- A regular expression describing words which directly precede the separator;

[2] http://ufal.mff.cuni.cz/pdt2.0/
[3] http://ucnk.ff.cuni.cz/

- A regular expression describing words which can be found in between a punctuation and the separator;
- A regular expression describing words which may follow the separator in a separator sequence;
- A collection of examples documenting the use of the separator.

All entries in the list are manually collected and checked on the basis of the evidence found in the data.

General Principles of Building Segmentation Charts

The process of building segmentation charts is relatively straightforward. In accordance with the principles presented above, the first step is always the morphological analysis of the input sentence. On the basis of its (typically ambiguous) results we will divide the sentence into segments, taking into account the number and position of all separators and separator sequences in the sentence.

The next step, drawing segmentation charts relevant for a given input sentence, is slightly more complicated due to the ambiguity concerning especially closing separators (mainly commas), which are generally highly ambiguous. Not only they can simply raise, lower or directly connect the following section at the same layer, they may even raise the following section several layers (in case of closing a deeply embedded subordinated clause). If there is such an ambiguous separator anywhere in the sentence, it is necessary to create more segmentation charts, each with an edge going in a different direction.

Basic Set of Rules

In order to demonstrate how the process of building the segmentation chart works, we present here a basic set of rules for Czech:

1. **Sentence start:** If the first (extended) segment does not have a subordination flag then the edge representing the first segment starts at the topmost (1^{st}) layer of the chart and continues straight to the right. Otherwise the edge for the first segment starts at the 2^{nd} layer.
2. **Comma:** If the comma is NOT followed by an item with a subordination flag, the next segment goes either straight to the right (this represents for example a comma separating two coordinated items inside a single clause) OR it jumps one or more layers (this is a highly ambiguous situation representing an end of a nested subordinated clause) upwards.
3. A **comma followed by an item with a subordination flag:** In this case the next segment moves downward.[4]
4. A **coordinating expression:** A coordinating conjunction or any other coordinating expression preserves a layer, even though it might be followed by an extended segment with subordination flag.
5. A **full stop, a question mark or an exclamation mark:** These characters represent an end of the sentence, therefore the last node of the segmentation chart always jumps to the 1^{st} layer of the chart (the layer of the main clause).
6. An **opening quotation marks:** Opening quotation marks are considered to be separators only when they are at the start of the sentence or when they are combined with other separators (comma, semicolon etc.) – in such a case the next segment drops one layer down.
7. **Closing quotation marks:** They are separators only if they follow opening quotation marks, which are considered being separators as well – in such a case the next segment jumps one or more layers up.

Segment Annotation Based on Existing Corpus

It is obvious that having a corpus annotated with information about segments would help a lot in the future development of our method. Unfortunately, there is no such corpus available for Czech at the moment, but, luckily, we can exploit the Prague Dependency Treebank[5], large and elaborated corpus with rich syntactic annotation of Czech sentences.

Unfortunately, although the annotation scheme of PDT allows for a very deep description of many kinds of syntactic relationships, there is no explicit segment annotation in the corpus. Let us check how segments are described there. From the two layers of syntactic annotation, analytical level of the corpus is the more appropriate one. It describes a surface syntactic structure and therefore constitutes much more natural information source for our purpose than the deeper, tectogrammatical level.

Formally, the structure of a sentence at the analytical layer of PDT is represented as a dependency-based tree, i.e. a connected acyclic directed graph in which no more than one edge leads into a node. The edges represent syntactic relations in the sentence (dependency relation and relation of coordination and apposition being the basic ones). The nodes – labelled with complex symbols (sets of attributes) – represent word forms and punctuation marks.

In particular, there are no nonterminal nodes in PDT that would represent more complex sentence units – such units are expressed as (dependency) subtrees, see also Figures 2, 3 and 4:

- Complex units that correspond to particular "phrases", as verb phrase, nominal phrase or prepositional phrase – such units are expressed as subtrees rooted with nodes labelled with respective governing word forms, e.g. governing verb (its "lexical" part in case of analytical verb form, or copula in verbal-nominal predicate, or modal verb), governing noun, or preposition (as a "head" of a whole prepositional phrase);

[4] There are some exceptions to this general rule, which may be handled by a set of conditions capturing those specific constructions allowing to go either right or to move the next segment upwards. Such a construction may be found for example in the sentence *Řekl, že byl, jaký byl, ŽE je, jaký je a že bude, jaký bude.* [(He) said that (he) was who (he) was, that (he) is who (he) is and the (he) is going to be who (he) is going to be.]

[5] http://ufal.mff.cuni.cz/pdt2.0/

- Dependent clauses – in principle, they are rendered as subtrees rooted with a node labelled with the governing verb of dependent clause (e.g. for attributive dependent clauses), or with a node for subordinating conjunction (adverbial and content clauses);
- Coordinated structures and sentence units in apposition (whether they are sentence members or whole clauses) – they are represented by subtrees rooted with nodes that correspond to a coordinating conjunction or some formal flag for an apposition (e.g. comma, brackets).

What does it mean for our segmentation structure? It is obvious that particular segments are not represented by nonterminal nodes. With respect to the segmentation, we have intuitively supposed that subtrees might be rooted either with the opening separator or with the governing node of the segment as a natural solution. Unfortunately, it turned out that this is not the case at the analytical level of PDT.

Although a bit unfortunate for our purpose, the situation is quite understandable – there are too many syntactic phenomena for which it is extremely difficult, if not impossible, to find a general consensus about annotation. A huge number of decisions had to be made concerning the annotation of complex linguistic phenomena like coordination, verbal complexes, the proper place of prepositions etc. and if a particular phenomenon has not been taken into an account since the beginning, it is of course impossible to find it annotated in a consistent manner.

Let us demonstrate this on a very simple example of bracketing – nothing can probably be more easy to detected as a single segment than a fragment of a sentence inside brackets; unlike punctuation marks, the brackets unambiguously show the beginning and the end of a text inserted into a sentence. It is therefore quite natural to expect this easily detectable segment to be annotated in a single consistent way. The following examples of structures assigned to segments in brackets have been borrowed from (Hajič et al. 1999), see Fig. 2, 3 and 4.

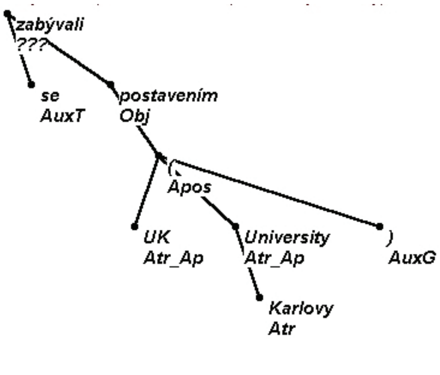

Figure 2: PDT: Segment in parenthesis as an apposition: *zabývali se postavením UK (Univerzity Karlovy)* [(they) dealt-with *refl.* a_position of_UK (University of_Charles)]

Let us point out that not only the annotation of a content of these parenthesis differs, but even the mutual position of both types of brackets in the tree is different.

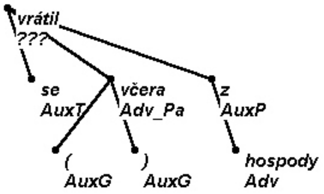

Figure 3: PDT: Segments in parenthesis as a sentence member: *vrátil se (včera) z hospody* [(he)... returned *refl* (yesterday) from a_pub]

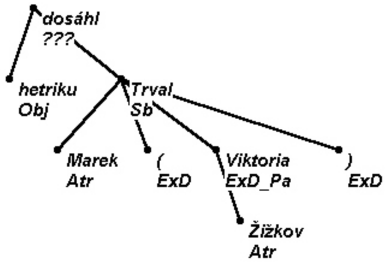

Figure 4: PDT: Segments in parenthesis as an independent sentence part: *a hetriku dosáhl Marek Trval (Viktorie Žižkov)* [and hattrick$_{Obj}$ achieved Marek Trval$_{Subj}$ (Viktorie Žižkov)]

Although seemingly inconsistent at the first sight, the trees contain an extremely valuable syntactic information which may serve as a basis for an automatic re-annotation procedure. Such a procedure might exploit the fact that in other respects, with regard to other syntactic phenomena, the corpus had been annotated with an extreme care and with a great deal of consistency. Actually, the reason for inconsistency wrt. the annotation of obvious segments is the consistency achieved wrt. a different syntactic phenomenon, this time the endeavor to annotate the apposition and coordination in a similar way, i.e. with the whole construction depending on a conjunction / appositional comma.

In order to obtain some evaluation of the proposed method, we have manually annotated a small sample of texts according to the definition of the segmentation chart. Several articles from newspapers containing political commentaries have been selected, namely from daily news Lidové noviny (LN in Table 1) and from journals MF Plus (MF) and Neviditelný pes[6] (NP).

The table below shows the degree of ambiguity of segmentation charts created automatically using the set of rules presented above, i.e. very local rules which do not presuppose understanding the sentence meaning.

Although the set of texts is relatively small, the table clearly shows that the simple rules presented above provide a very good starting point. Let us point out the high cover-

[6] http://pes.eunet.cz

	LN	MF	NP	total
sentences	33	57	15	105
tokens	553	990	334	1877
segments	78	153	57	288
1 chart	28	47	12	87
2 chart	2	3	3	8
3 chart	1	4	-	5
4 chart	1	2	-	3
5 chart	1	1	-	2
more	-	-	-	-

Table 1: Degree of ambiguity of segmentation charts

age of our method – only one correct segmentation chart has been omitted by our algorithm.

Conclusion

The method presented in this paper shows that (at least for a language displaying inflectional morphology similar to that of Czech) it is possible to draw a chart reflecting the mutual position of segments in complex sentences without applying the full-fledged syntactic parsing of the whole sentence first. The method is based on the identification of separators and their classification. The subsequent steps (which are not covered by this paper) may then decide, on the one hand, which of the charts are not valid (in case that there are several variants of charts as an output of our method), and, on the other hand, to exploit the charts for faster and more effective methods of analysis of complex sentences.

The results achieved so far encourage further research in two areas. The first area concerns the further development of more precise segmentation rules, the second one might concern the step from segmentation charts towards the chart reflecting the mutual position of clauses, not only segments.

Acknowledgments

This paper is a result of the project supported by the grants No. 1ET100300517 and MSM0021620838.

References

Debusmann, R., Duchier, D., Rossberg, A.: *Modular grammar design with typed parametric principles.* In: Proceedings of FG-MOL 2005, Edinburgh, 2005

Hajič, J.: *Disambiguation of Rich Inflection (Computational Morphology of Czech).* UK, Nakladatelství Karolinum, Praha, 2004

Hajič, J., Hajičová, E., Pajas, P., Panevová, J., Sgall, P., Vidová-Hladká, B.: *Prague Dependency Treebank 1.0 (Final Production Label).* In: CD-ROM, Linguistic Data Consortium, 2001

Hajič, J., Panevová, J., Buráňová, E., Urešová, Z. Bémová, A.: *Anotace Pražského závislostního korpusu na analytické rovině: pokyny pro anotátory.* Technical Report No. 28, ÚFAL MFF UK, Prague, Czech Republic, 1999

Hajič, J., Vidová-Hladká, B., Zeman, D.: *Core Natural Language Processing Technology Applicable to Multiple Languages.* The Workshop 98 Final Report. Center for Language and Speech Processing, Johns Hopkins University, Baltimore, 1998

Holan, T., Kuboň, V., Oliva, K., Plátek, M.: *On Complexity of Word Order.* In: Les grammaires de dépendance – Traitement automatique des langues, Vol 41, No 1, pp. 273-300, 2000

Kuboň, V.: *Problems of Robust Parsing of Czech.* Ph.D. Thesis, MFF UK, Prague, 2001a

Kuboň, V.: *A Method for Analyzing Clause Complexity.* The Prague Bulletin of Mathematical Linguistics 75, pp. 5-27, 2001

Kuboň, V., Lopatková, M., Plátek, M., Pognan P.: *Segmentation of Complex Sentences.* In: Lecture Notes in Computer Science 4188, Text, Speech and Dialogue, TSD 2006, Springer Berlin / Heidelberg, pp. 151-158, 2006

McDonald, R., Pereira, F., Ribarov, K., Hajič, J.: *Non-Projective Dependency Parsing using Spanning Tree Algorithms.* In: Proceedings of HTL/EMNLP, pp. 523-530, 2005

Oliva, K.: *A Parser for Czech Implemented in Systems Q.* In: Explizite Beschreibung der Sprache und automatische Textbearbeitung, MFF UK Praha, 1989

Šmilauer, V.: *Učebnice větného rozboru.* SPN, Praha, 1958

Zeman, D.: *Parsing with a Statistical Dependency Model.* Ph.D. Thesis. MFF UK, Prague, 2004