Segmentation Charts for Czech – Relations among Segments in Complex Sentences*

Markéta Lopatková and Tomáš Holan

Charles University in Prague, Czech Republic
lopatkova@ufal.mff.cuni.cz,
Tomas.Holan@mff.cuni.cz

Abstract. Syntactic analysis of natural languages is the fundamental requirement of many applied tasks. We propose a new module between morphological and syntactic analysis that aims at determining the overall structure of a sentence prior to its complete analysis.

We exploit a concept of segments, easily automatically detectable and linguistically motivated units. The output of the module, so-called 'segmentation chart', describes the relationship among segments, especially relations of coordination and apposition or relation of subordination.

In this text we present a framework that enables us to develop and test rules for automatic identification of segmentation charts. We describe two basic experiments – an experiment with segmentation patterns obtained from the Prague Dependency Treebank and an experiment with the segmentation rules applied to plain text. Further, we discuss the evaluation measures suitable for our task.

1 Motivation

Syntactic analysis of natural languages is the fundamental requirement of many applied tasks. The solution of this complex task is not satisfactory yet, especially for languages with free word order. Long-term efforts of many researchers brought parsers, which are quite reliable for relatively short and simple sentences. However, their reliability is significantly lower for long and complex sentences (see e.g. [1] for more citations).

A new module between morphological and syntactic analysis is a natural step capable to reduce the complexity of this task. Let us mention at least the idea of chunking [2] and cascaded parsing [3–5]. Roughly speaking, these approaches group individual tokens into more complex structures (as e.g. nominal or prepositional phrases). We propose another approach that aims at determining the overall structure of a sentence, i.e. a hierarchy of sentence segments, prior to its complete analysis. The advantage of having the estimation of sentence structure (especially for long and complex sentences) is quite obvious – it allows us to

* This paper presents the results of the project supported by the GAČR grant No. 405/08/0681 and partially also by the IS program No. 1ET100300517. The research is carried out within the project of MŠMT No. MSM0021620838.

A.H. Dediu, A.M. Ionescu, and C. Martín-Vide (Eds.): LATA 2009, LNCS 5457, pp. 542–553, 2009.

exclude inappropriate relations in syntactic trees and thus the complexity of the task is substantially reduced and the parsing process is speeded up.

We exploit a concept of segments, easily automatically detectable and linguistically motivated units. Firstly, individual segments are identified; then their mutual relationship is determined. So-called 'segmentation chart' describes the relationship among segments, especially relations of coordination and apposition or relation of subordination, i.e. the relation between governing and subordinated parts of sentence; parentheses are also identified.

We slightly modify and exploit the concept of segments which has been originally proposed in [6] and modified in [7]. The initial set of rules for segmentation of Czech sentences has also been introduced there. These rules serve for identification of (nondeterministic) segmentation charts showing the relationship of individual segments in sentences.

Let us demonstrate the basic idea of segmentation on an example of Czech sentence from the news (1). At first, the sentence is split into individual segments. We consider the punctuation marks , , the coordinating conjunction *a*, the brackets *(*, *)* and the full stop . as boundaries of segments. Then we determine mutual relations of these units – we distinguish coordination, parenthesis and subordination. Thus we obtain **segmentation chart**, which allows us to identify the overall structure of the sentence.

(1) *S tím byly trochu problémy , protože starosta v řeči rád zdůrazňoval své vzdělání (však studoval až v Klatovech a v Roudnici), a Víťa tedy občas nutně trochu tápal .*
[There was a bit problem with it , as the mayor liked to stress his education in his talk (after all he studied in Klatovy and Roudnice), and thus Víťa was occasionally a bit confused .]

The first segment consists of the main clause of the complex sentence (no subordinating expression appears in this segment and there is the finite verb *byly* [were] there). This segment is placed on the basic layer (layer 0) of the segmentation chart. The second segment is introduced by the subordinating conjunction *protože* [because] and it contains the finite verb *zdůrazňoval* [(he) emphasized]. This segment is identified as a segment subordinated to the first one and thus it is placed on the lower layer (layer 1) in the chart. The opening bracket follows, which is interpreted as a beginning of a parenthesis. Thus the third segment belongs to another lower layer (layer 2). The fourth segment is separated by the coordinating conjunction *a* [and], therefore it should be at the same layer as the third segment. The third segment contains a finite verb *studoval* [studied], contrary to the fourth segment – it implies with high probability that we have the case of coordination of sentence members. Embedded parenthesis ends with the closing bracket; we climb up in the segmentation chart. The last segment contains the word *tedy* [therefore] – the triplet *, a tedy* [(comma) and therefore] is considered as a characteristic of coordination. Thus the fifth segment is analyzed as a segment coordinated to either the first or the second segment.

The segmentation chart can be expressed graphically (Fig. 1 shows one of the possible charts for the sentence (1)), or as a vector of layers (e.g., two vectors reflecting two segmentation charts (01220) and (01221) for the sentence (1)).

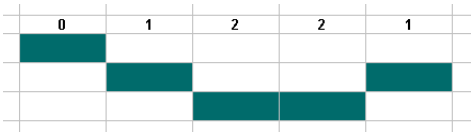

Fig. 1. Segmentation chart (01221)

Note that segmentation and analysis of segmented sentences can be formally modeled, e.g., by Parallel Communicating Grammar Systems (PCGS) and Freely Rewriting Restarting [8].

The capability to determine (reliably enough) the mutual relationship among segments and subsequently the possible structure of clauses in complex sentences prior to their full analysis would simplify the task of syntactic analysis / parsing of natural language sentences. Moreover, it appears that in a number of important application – such as information retrieval, determining the structure of documents and their main and secondary themes – there is no explicit need for full syntactic analysis. It would be of great interest to examine to which extent we can limit ourselves to the 'upper' layers of sentence structure (and ignore deeply nested segments) for such applied tasks. The achievements of similar methods for the analysis of different type of languages, e.g. [9] or [10], encourage further research in this area.

The main goal of this text is to present a framework which enables us to further develop, test and evaluate rules for automatic identification of segmentation charts. After the definition of segments and segmentation charts (Section 2), we describe two basic experiments – the experiment with obtaining segmentation patterns from tree structures stored in the Prague Dependency Treebank (Section 3.1) and the experiment with the segmentation rules applied to plain text (Section 3.2). We conclude with Section 4 where we introduce and discuss the appropriate measures for evaluating the segmentation rules. We compare segmentation charts obtained by these two sets of rules with the manually annotated sample of complex sentences and we show their limits for selected language phenomena.

2 Segment Boundaries, Segments and Segmentation Chart

An (input) sentence is understood here as a sequence of tokens $w_1 w_2 \ldots w_n$, when each token w_i represents either one word (lexical form of a given language) or one punctuation mark (comma, full stop, question mark, exclamation mark, dash, colon, semicolon, quotation marks, brackets, ...).

We do not care about dividing the text into sentences here as we dispose of tools reliable enough for sentence identification for Czech; in our experiments, we adopt it from the Prague Dependency Treebank. We also presuppose full morphological analysis of the text, i.e. we expect that each token bears its full morphological analysis.

Based on their morphological characteristics, all tokens are disjunctively divided into two groups — ordinary words and segment boundaries. After identification of boundaries, the input sentence is partitioned into individual segments.

Segment Boundaries

Boundaries are tokens and their sequences that divide a sentence into individual units referred to as segments.

In the following experiments we consider the following tokens as **elementary boundaries**:

- **punctuation marks:** comma, colon, semicolon, question mark, exclamation mark, dash (all types), opening and closing bracket (all kinds), vertical bar, quotation mark (all types), i.e. symbols *, : ; ? ! - ( )* [ ] | { } ‘ ’ “ ” , ‘ ,, “
- **punctuation ending a sentence**
- **coordinating conjunctions:** morphological tag starting with the pair J^ [11].

Several elementary boundaries may appear in a sentence following immediately one after another (as the sequence *), a* in sentence (1)). We consider a maximum sequence of such elementary boundaries as a **(compound) boundary.**

Segment S is then understood as the maximal non-empty sequence of tokens $w_1 w_2 \ldots w_s$ that does not contain any boundary.

When determining the individual segments we presuppose that every sentence begins and ends with a boundary (if there is no boundary at the beginning or at the end of the sentence, we add the empty boundary there).

Let us point out that the boundaries specified on the basis of morphological analysis are not necessarily unambiguous. Punctuation marks are not ambiguous but this is not true for coordinating conjunctions (e.g. the wordform *ale* is either coordinating conjunctions [but] or it is a wordform belonging to three substantive lemmas *ala*). Thus we admit ambiguous segmentation of the sentence in general. However, there are highly reliable taggers for Czech (i.e., automated tools that are able to select exactly one morphological tag per token; the highest published accuracy for the first two positions of morphological tag is 99.36% [12]). Therefore, we disregard possible ambiguity of morphological analysis and we presuppose a unique morphological tag for each token (in the experiments described below, we take over the morphological tags from the Prague Dependency Treebank). It implies that boundaries and individual segments are defined unambiguously.

Segment Flags

Morphological analysis of the text contains a lot of more or less reliable information that can be used for identification of relationship among individual segments. This information is stored in a form of specific **flags** that are assigned to individual segments. In our experiments, we use only subordination flag, other flags as coordination flag or flag for finite verb are foreseen [6].

Subordination flag (SF). A subordination flag is assigned to a particular segment either if this segment contains any wordform with the morphological tag that begins with the following pair (for conjunctions, pronouns, and numerals [11]), or if this segment contains one of the listed pronominal adverbs:

- **subordinating conjunction:** `J,`
- **interrogative / relative pronoun:** `P4`, `PE`, `PJ`, `PK`, `PQ`, `PY`
- **numeral:** `C?`, `Cu`, `Cz`
- **pronominal adverb:** *jak, kam, kde, kdy, proč, kudy*

Segmentation Chart

The segmentation of a particular sentence can be represented by one or more **segmentation charts** that describe the mutual relationship of individual segments with regard to their coordination or subordination. A segmentation chart captures the **layer of embedding** for individual segments. The basic idea of the segmentation chart is very simple:

- Segments forming all main clauses of a complex sentence belong to the basic layer (layer 0);
- Segments forming clauses that depend on the clauses at the k-th layer obtain layer of embedding $k+1$ (i.e., layer of embedding for subordinated segments is higher than layer of segments forming their governing clause);
- Segments forming coordinated segments or segments in apposition have the same layer;
- Segments forming parentheses (e.g., sequence of wordforms within brackets) obtain layer $k+1$ compared to the layer k of their adjacent segments

3 Experiments with Automatic Identification of Segmentation Charts

3.1 How to Obtain Segments from Syntactic Tree?

This chapter explains the possible algorithm producing segmentation charts for individual sentences from their analytical trees in the Prague Dependency Treebank[1] (PDT [13]). Analytical layer of PDT captures the surface syntax. In principle, it contains the same information that may be directly used for the identification of segment layers.

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

A sentence at the analytical layer is represented as a dependency-based tree, i.e., a connected acyclic directed graph in which no more than one edge leads from a node. The nodes – labeled with complex symbols (sets of attributes) – represent individual tokens (wordforms or punctuation marks); one token of the sentence is represented by exactly one node of the tree. The edges represent syntactic relations in the sentence (the dependency relation and the relation of coordination and apposition being the basic ones). The actual type of the relation is given as a function label of the edge, so-called analytical function. In addition, linear ordering of the nodes corresponds to the sentence word order. In particular, there are no nonterminal nodes in PDT representing more complex sentence units – such units are expressed as (dependency) subtrees.

In order to be able to present a basic set of rules, it is necessary to introduce the concept of a path between the segments and the concept of a group of segments. For the sentence W, there is **an edge from the segment S_i to the segment S_j** ($S_i, S_j \subset W$) iff there exists a pair of words $u \in S_i$ and $v \in S_j$ such that there exists a path from u to v in the dependency tree T of the sentence W.

A **path from the segment S_i to the segment S_j** of the sentence W ($S_i, S_j \subset W$) exists iff there exists a sequence of segments $S_i = S_{p_1}, \ldots, S_{p_m} = S_j$, $S_{p_k} \subset W (k = 1 \ldots m)$ such that for every $k = 1 \ldots m - 1$ there is an edge from the segment S_{p_k} to the segment $S_{p_{k+1}}$.

A set of segments of the sentence W is said to be a **group of segments** G iff for each pair of segments $S_i, S_j \in G$ holds that there is a path from S_i to S_j (symmetrically, also a path from the S_j to the S_i must exist).

We use the following algorithm for obtaining the segmentation chart for individual sentences of PDT.

Determination of segments: The first step for obtaining the segmentation chart consists in the determination of boundaries; based on the boundaries, individual segments are identified.

Groups of segments: Groups of segments are identified.

Zero Layer: The segments which are connected by some path (either direct, i.e. edge, or via nodes representing elementary boundaries only) with the root node of the dependency tree T are identified; these segments as well as all segments belonging to the same groups are assigned layer 0.

Coordination and apposition: If there is a segment S_i with already assigned layer k and its adjacent segment S_j has unknown layer and, moreover, the boundary between these two segments consists of some coordinating expression or expression introducing an apposition (e.g., the node representing the elementary boundary has an analytical function `Coord` or `Apos`), then the segment S_j gets the same layer as the segment S_i has.

Deeper embedded segments: All segments with unknown layer connected by some path (either direct, i.e. edge, or via nodes representing elementary boundaries only) with segments of the layer k are assigned the layer $k + 1$; the same holds for all segments belonging to the same group of segments.

Coordination and apposition: Again all segments adjacent to the segments with already known layers are checked (see above).

This process is repeated until all segments get their layer.

The proposed algorithm assigns exactly one segmentation chart (not necessarily the correct one) to any input sentence represented by the analytical tree. Let us demonstrate it on a sentence (2); the analytical tree is in Fig. 2.

(2) _ *Po rozhovorech s majiteli našich soukromých firem a nakonec i představiteli firem zahraničních mám dojem , že v této republice nejsou schopní lidé .*
[_ After the discussions with the owners of our private companies and after all even with the representatives of foreign companies I have an idea , that there aren't clever people in this republic .]

Sentence (2) consists of four segments (the boundaries are underlined in the sentence whereas they are separated by vertical lines in Fig. 2). The first and the third segment form a group (there is an edge from the node *po* [after] to the node *mám* [(I) have] and at the same time a path leads from the node *představiteli* [representatives] to the node *s* [with], see the arrows). These two segments obtain the zero layer as there is the edge from the node *mám* [(I) have] to the root of the tree. The second segment also gets the zero layer as its boundary with the first segment is the coordination conjunction *a* [and]. The fourth segment obtains layer 1 since there is an edge leading from this segment to the third segment with already known zero layer. Therefore, the segmentation chart assigned to the sentence is (0001) (the correct segmentation chart in this case).

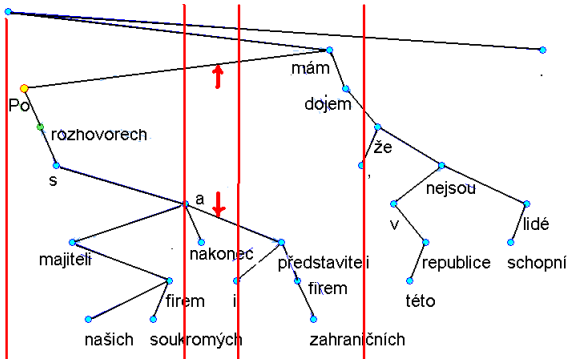

Fig. 2. Analytical tree of the sentence (2) with highlighted segments

3.2 How to Obtain Segments from Plain Text?

The basic set of (heuristic) segmentation rules for plain text was published in [7]. We have specified these rules more precisely and implemented them. That allows

us to compare the results of these rules with the results of segmentation based on the analytical trees from PDT.

When processing an input sentence, we start at its beginning; we move right, identify individual boundaries and segments and determine their appropriate layers of embedding.

The following rules define the layer of embedding that is assigned to the first segment. They also determine how this layer may change when crossing the elementary boundaries. Let us point out that the rules do not always give a single unambiguous answer (e.g., comma may be considered as coordinating expression – then the layer should be preserved – or as the end of embedding segment – then it should raise the layer). Thus each segment is not assigned a single number but an **interval of** possible **layers**.

The adjacent segments may be separated by compound boundaries, i.e. by sequences of elementary boundaries. In such a case, the rules are applied to individual elementary boundaries. The segment is assigned the layer which is obtained after processing the last elementary boundary preceding this segment.

The following list introduces the rules for elementary boundaries.[2]

Beginning of the sentence: If subordination flag (SF, see Section 2) is not assigned to the first segment, then this segment gets the basic zero layer. Otherwise, it gets layer 1.

Comma: If SF is not assigned to the subsequent segment, then the lower limit of the interval of layers does not change, the upper limit is set to 0 (i.e., the case of end of any number of embedded clauses). Otherwise, the layer of embedding is increased by 1 (i.e., the beginning of embedded clause or its part).

Opening bracket (of any kind): If SF is not assigned to the subsequent segment, then the layer (or interval of possible layers) of embedding is increased by 1 (i.e., the beginning of parenthesis). Otherwise, the layer is increased by 2 (i.e., parenthesis with a deeply embedded unit).

Closing bracket (of any kind): If it is preceded by the opening bracket of the same kind, then the layer of embedding is set to the same value(s) as the segment preceding the opening bracket has. Otherwise, the layer does not change (this condition handles the cases of the list *a)... b)...*).

Coordinating conjunction: The layer remains unchanged.

Colon: If SF is not assigned to the subsequent segment, then the upper limit remains unchanged (i.e., coordination or apposition); the lower limit is increased by 1 (i.e., the beginning of (a part of) embedded clause or beginning of direct speech (together with a quotation mark)). Otherwise, the upper limit is increased by 1 and the lower limit is increased by 2 (i.e., deeper embedded (part of) clause).

Question mark, exclamation mark: The lower limit is decreased by 1, the upper limit is set to 0 (i.e., the end of any number of embedded clauses).

[2] Let us repeat that we assume the input text being already divided into sentences.

Semicolon: The lower limit of the interval of layers remains unchanged, the upper limit is set to 0.

Vertical bar, dash, quotation marks: The layer remains unchanged.[3]

These rules define a set of segmentation charts for each morphologically analyzed input sentence.

4 Evaluation and Analysis of the Results

4.1 Evaluation Data and Possible Evaluation Measures

In the previous sections, we have described the basic experiments with the automatic identification of segmentation charts from plain texts and from trees from PDT. For further development and improvement of these rules, we had to create a test set of sentences with correctly identified segmentation charts.

We chose a set of suitable sentences from development data of PDT 2.0 (the 'dtest' data, 5228 sentences) – we focused only on such sentences that contain at least five segments (707 sentences). Then we manually determined a segmentation chart for every tenth sentence from the set. Thus we received 71 relatively structurally complex sentences with attached segmentation charts.

Let us emphasize that the selection of such complex sentences (in average 6.49 segments per sentence) made the measured results significantly worse in comparison with random sample of sentences (the average number of segments per sentence in the full dtest data is 2.72).

Note also that many of the testing sentences are ambiguous, i.e. they have more (potential) syntactic trees. However, sentences in PDT are disambiguated, only one of all possible structures is stored there. When identifying the appropriate segmentation chart we consider only the structure captured in PDT. Every sentence has been assigned a single chart (e.g., sentence (1) got the only segmentation chart (01221)).

There are several possibilities how to evaluate the proposed rules. The simplest measure consists in counting the cases of correct assignment of layers to individual segments. We call this basic measure ρ.

When looking at the results of experiments, we have found out that in many cases the wrong assignment of a layer for one segment has resulted in incorrectly identified layers of other segments. However, the relationship among individual segments may be recognized correctly. For example, the sentence (3) has a correct segmentation chart (2233110). The algorithm for PDT yields the chart (1122000); although almost all relations between segments are identified correctly, there is only one correctly assigned layer and $\rho = 1/7$.

(3) *„ Když to odečtete od výplaty spolu se ztrátou při výměně slovenských korun za české a za pojištění , které se musí platit tam i u nás , nezbude manželovi z výplaty ani polovina ,“ zlobí se paní Krajčová .*

[3] Quotation marks are used in Czech either for direct speech – then they are accompanied with other boundary as comma or colon (which ensures the lower layer) or they are use for emphasizing, where the layer should stay unchanged.

[,, When you deduct this from your earnings together with the losses when exchanging Slovak crowns for Czech crowns and for insurance , which must be paid there as well as here , less then half of the sum will remain from my husband's salary ," says Mrs. Krajčová with angry .]

This drawback of the basic measure may be eliminated if we allow 'shifting' of the whole resulting segmentation chart. E.g., if we shift the vector for the sentence (3) by +1 we get (2233111) – the layers of six segments (out of seven) are identified correctly. The measure with optimal shifting will be called σ (thus $\sigma = 6/7$ for the sentence (3)).

As we are primarily interested in the relationship among segments we consider also the measure evaluating the correctness of the proposed relationship of two adjacent segments. E.g., charts (101) and (211) have the same relationship between the first and second segment (difference -1), but different relationship between the second and third segment). We call this measure δ.

4.2 Evaluation of Rules for Syntactic Trees

The proposed set of segmentation rules from PDT identifies exactly one segmentation chart for each input sentence. When evaluating these rules, we adopt only *accuracy* measure (standard *recall* and *precision* measures are equal). The results are summarized in Table 1.

Table 1. The evaluation of the proposed set of rules for segmentation charts from the trees from PDT

accuracy:	basic measure		measure with 'shifting'	
# of segments	# correct	ρ	# correct	σ
461	264	0,57	335	0,73

When evaluating the proposed relationship of two adjacent segments, the rules are reaching $\delta = 0.70$ (274 of 390 relations among segments have been proposed correctly).

Let us mention three main problems that decrease the success of the proposed rules for determining segmentation charts from the analytical trees.

1. The sentence member forming a separate segment is assigned a higher layer (by 1) than the segments with its governing member. E.g., the sentence _ *Včera , kdy tak pršelo , přišli .* [_ Yesterday , when it rained so much , they came.] with the correct segmentation chart (010) gets incorrect segmentation chart (120).
2. We postponed special (but relatively frequent) Czech construction with two subordinating expressions (underlined) appearing in one segment just one after another, as e.g. *Nevěděl, že když jsem se probral, zavolal jsem policii.* [He didn't know that when I woke up, I called the police.]
3. Coordination (and apposition) are another widespread phenomena which deserve a special treatment, especially those of more than two members.

4.3 Evaluation of Rules for Plain Text

The evaluation of rules for assigning segmentation chart to plain text consists in testing whether the resulting interval for individual segments contains the correct layer of embedding of this segment (thus we measure only *recall*), see Table 2.

Table 2. The evaluation of the proposed set of rules for charts from plain text

recall:	basic measure		measure with 'shifting'	
# of segments	# correct	ρ	# correct	σ
461	302	0,66	354	0,77

The average number of segmentation charts per sentence from our testing data is 2.17 whereas the average number of ambiguity for the entire dtest data is 1.32.

Let us mention here at least two phenomena that the proposed set of rules does not solve adequately. These phenomena have to be a subject of more precise specifications (which would require detailed linguistic examination).

1. We have not specified the segmentation rules for direct and semidirect speech. E.g., the layers of the first four segments of sentence (3) are not deep enough in all assigned segmentation charts (these segments get (1122) instead of the correct (2233) segmentation vector).
2. The case of several coordinated clauses with repeated subordinating expressions is not treated correctly yet. E.g., *Jak účelně větrat, jak nepřetápět, jak spotřebu měnit a podle toho účtovat.* [How to ventilate effectively, how to not overheat, how to change the consumption and pay according to it.] This sentence obtains the wrong chart (0122) instead of the correct one (0000).)

5 Conclusions

Segments are easily automatically detectable and linguistically motivated units that form (complex) sentences. Their mutual relationship – especially relations of coordination and apposition, relation of subordination as well as parenthesis – is captured in a form of a segmentation chart. The segmentation chart describes the overall structure of a sentence prior to its complete syntactic analysis.

We focused on the description of a framework that allows us to formulate and refine linguistically motivated rules for automatic detection of segmentation charts for given sentences. We have also introduced appropriate measures for evaluating segmentation analysis.

At this stage, two sets of rules were implemented, rules operating on analytical trees from the Prague Dependency Treebank and rules operating on plain text enriched with morphological analysis. We have compared the results reached with those rules using the manually annotated sample of sentences from PDT.

The experiments brought clear specification of segmentation charts and exact rules for manual annotation. The results show that for further research it is

necessary to work with a large set of reliably annotated data. It turns out that these data cannot be obtained without extensive (semi)manual annotation of a large set of sentences. Such data would also allow us to adopt machine learning techniques for automatic identification of segmentation charts.

References

1. Holan, T.: O složitosti Vesmíru. In: Obdržálek, D., Štanclová, J., Plátek, M. (eds.) Malý informatický seminář MIS 2007, pp. 44–47. MatFyz Press, Praha (2007)
2. Abney, S.: Parsing By Chunks. In: Berwick, R., Abney, S., Tenny, C. (eds.) Principle-Based Parsing, pp. 257–278. Kluwer Academic Publishers, Dordrecht (1991)
3. Abney, S.: Partial Parsing via Finite-State Cascades. Journal of Natural Language Engineering 2, 337–344 (1995)
4. Brants, T.: Cascaded Markov Models. In: Proceedings of EACL 1999, pp. 118–125. University of Bergen (1999)
5. Ciravegna, F., Lavelli, A.: Full Text Parsing using Cascades of Rules: An Information Extraction Procedure. In: Proceedings of EACL 1999, pp. 102–109. University of Bergen (1999)
6. Kuboň, V.: Problems of Robust Parsing of Czech. Ph.D. Thesis, MFF UK, Prague (2001)
7. Kuboň, V., Lopatková, M., Plátek, M., Pognan, P.: A Linguistically-Based Segmentation of Complex Sentences. In: Wilson, D.C., Sutcliffe, G.C.J. (eds.) Proceedings of FLAIRS Conference, pp. 368–374. AAAI Press, Menlo Park (2007)
8. Pardubská, D., Plátek, M.: On Parallel Communicating Grammar Systems and Correctness Preserving Restarting Automata. In: Dediu, A.H., Ionescu, A.M., Martín-Vide, C. (eds.) LATA 2009. LNCS, vol. 5457, pp. 1–18. Springer, Heidelberg (2009)
9. Jones, B.E.M.: Exploiting the Role of Punctuation in Parsing Natural Text. In: Proceedings of the COLING 1994, Kyoto, pp. 421–425 (1994)
10. Ohno, T., Matsubara, S., Kashioka, H., Maruyama, T., Inagaki, Y.: Dependency Parsing of Japanese Spoken Monologue Based on Clause Boundaries. In: Proceedings of COLING and ACL, pp. 169–176 (2006)
11. Hajič, J.: Disambiguation of Rich Inflection (Computational Morphology of Czech), UK, Nakladatelství Karolinum, Praha (2004)
12. Spoustová, D., Hajič, J., Votrubec, J., Krbec, P., Květoň, P.: The Best of Two Worlds: Cooperation of Statistical and Rule-Based Taggers for Czech. In: Proceedings of Balto-Slavonic NLP Workshop, pp. 67–74. ACL, Prague (2007)
13. Hajič, J., Hajičová, E., Panevová, J., Sgall, P., Pajas, P., Štěpánek, J., Havelka, J., Mikulová, M.: Prague Dependency Treebank 2.0. LDC, Philadelphia (2006)