Lang Resources & Evaluation (2012) 46:25–36
DOI 10.1007/s10579-011-9162-z

ORIGINAL PAPER

Annotation of sentence structure

Capturing the relationship between clauses in Czech sentences

Markéta Lopatková · Petr Homola · Natalia Klyueva

Published online: 28 August 2011

Abstract The focus of this article is on the creation of a collection of sentences manually annotated with respect to their sentence structure. We show that the concept of linear segments—linguistically motivated units, which may be easily detected automatically—serves as a good basis for the identification of clauses in Czech. The segment annotation captures such relationships as subordination, coordination, apposition and parenthesis; based on segmentation charts, individual clauses forming a complex sentence are identified. The annotation of a sentence structure enriches a dependency-based framework with explicit syntactic information on relations among complex units like clauses. We have gathered a collection of 3,444 sentences from the Prague Dependency Treebank, which were annotated with respect to their sentence structure (these sentences comprise 10,746 segments forming 6,341 clauses). The main purpose of the project is to gain a development data—promising results for Czech NLP tools (as a dependency parser or a machine translation system for related languages) that adopt an idea of clause segmentation have been already reported. The collection of sentences with annotated sentence structure provides the possibility of further improvement of such tools.

Keywords Sentence and clause structure · Dependency and coordination · Annotation

M. Lopatková (✉) · P. Homola · N. Klyueva
Charles University in Prague, Faculty of Mathematics and Physics, Prague, Czech Republic
e-mail: lopatkova@ufal.mff.cuni.cz

P. Homola
e-mail: homola@ufal.mff.cuni.cz

N. Klyueva
e-mail: kljueva@ufal.mff.cuni.cz

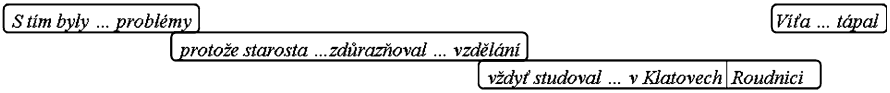
1 Motivation

Syntactic analysis of natural languages is a fundamental requirement of many applied tasks. Parsers providing automatic syntactic analysis are quite reliable for relatively short and simple sentences. However, their reliability is significantly lower for long and complex sentences, especially for languages with free word order; see e.g. Zeman (2004) for results for Czech. The identification of the overall structure of a sentence prior to its full syntactic analysis is a natural step capable of reducing the complexity of full analysis. Such methods have brought good results for typologically different languages, see e.g. Jones (1994) for English or Ohno et al. (2006) for Japanese; also first results for Czech are promising (esp. a clause segmentation in a rule-based dependency parser, see Holan and Žabokrtský 2006, or in a machine translation system for related languages, as in Homola and Kuboň 2010).

We exploit a concept of segments—linguistically motivated units, which may be easily detected automatically, as they were defined by Lopatková and Holan (2009).[1] The segment annotation captures relationships among segments, especially subordination, coordination, apposition and parenthesis. Based on the segment annotation, clauses forming complex sentences can be identified: those segments that constitute individual clauses are grouped and marked as single separate syntactic units of a higher layer, the layer of clause structures. Let us demonstrate the basic idea of segmentation on the following Czech newspaper sentence:

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

First, the sentence is split into individual segments; here the punctuation, the coordinating conjunction, and the brackets are considered as segment boundaries. Second, the mutual relations of these units are identified: apparently, local morphological analysis contains a lot of more or less reliable information that can be used (we will discuss this issue in detail in Section 3). The relations between segments can be represented by the so-called *segmentation chart* that should (1) capture a *level of embedding* for each segment, and (2) identify individual *clauses* (marked by ellipses), as in the following scheme:


S tím byly ... problémy
Víťa ... tápal
protože starosta ...zdůrazňoval ... vzdělání
vždyť studoval ... v Klatovech
Roudnici

1.1 Prague dependency treebank and segment annotation

The focus of this article is on the creation of a collection of Czech data reliably annotated with respect to sentence structure. The obvious idea is to use the *Prague*

[1] We adopt the basic idea of segments introduced and used by Kuboň (2001) and Kuboň et al. (2007). We slightly modify it for the purposes of the annotation task.

Dependency Treebank,[2] henceforth PDT (Hajič et al. 2006), a large and elaborated corpus of Czech newspaper texts with rich syntactic annotation. The PDT adopts the dependency-based framework, thus it contains explicit information on mutual relations between individual tokens (words and punctuation marks). However, relations among more complex units, especially among clauses, are not explicitly indicated.

Rich syntactic information stored in the PDT can be used (at least to some extent) for the identification of individual clauses as well. Let us refer to the experiments described by Lopatková and Holan (2009) and by Krůza and Kuboň (2009). In both papers, the authors describe well-developed automatic procedures for identifying segments and/or clauses and their mutual relationship from the analytical layer of the PDT (i.e., a layer of surface syntax, see Hajič et al. 2004), which is based on Czech grammatical tradition, see esp. Šmilauer 1969). However, these procedures cannot be properly evaluated so far because of the lack of test data (the papers either provide evaluation on a very small sample of sentences—tens of sentences—or they focus on comparing the results of two automatic tools). Nevertheless, the preliminary evaluation shows that extracting the overall sentence structure from dependency trees of the analytical layer is not straightforward and the results are not satisfactory.[3] Further development of segmentation tools necessarily requires reliable and precisely annotated development and test data (see below).

1.2 Why to annotate sentence structure?

The development and evaluation of tools for extracting a sentence structure from PDT data is interesting from the theoretical point of view: it may reveal possible limitations of dependency-based annotation at the layer of surface syntax (especially those related to non-dependency relations like coordination, and surface ellipses).

However, the main purpose of identifying sentence structure—either manually or automatically from the PDT (using a reliable automatic tool)—is to gain development data for an automatic tool which would determine the overall structure of a (morphologically analyzed) sentence (the first results based on a small data sample are reported by Lopatková and Holan 2009).

As already mentioned, the first experiments integrating a segment and/or a clause identification have brought promising results in dependency syntactic parsing (namely a combination of several parsers, one of them exploiting the idea of segmentation, see Holan and Žabokrtský 2006) and in machine translation between related languages (namely Homola and Kuboň 2010). These results encourage us in our effort to prepare a sufficient amount of reliable data analyzed at the level of a sentence structure.

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

[3] E.g., in experiments reported by Lopatková and Holan (2009), a correct level of embedding was assigned only to approx. 75% of segments.

1.3 Related work

The proposed approach—contrary to such well known approaches as, e.g., chunking (Abney 1991) or cascaded parsing (Abney 1995; Ciravegna and Lavelli 1999), which group individual tokens into more complex structures as nominal or prepositional phrases in a 'bottom-up' direction—can be characterized as a 'top-down' method: first, a structure of sentence clauses is assessed and, second, syntactic relations within individual clauses are identified. Such an approach is quite novel; as far as we know, a similar method has been tested only for Slovene (Marinčič et al. 2010) so far.

1.4 Outline

In this article, we present a project of manual annotation of sentence structure for complex Czech sentences. The article is structured as follows: In Section 2, the basic concepts, especially boundaries, segments and segmentation charts, are introduced. The core Section 3 describes the principles of the annotation for basic linguistic phenomena. Lastly, basic statistics for the annotated sentences and first evaluation of the existing tools are presented in Section 4.

2 Boundaries, segments and segmentation charts

2.1 Segment boundaries

An (input) sentence is understood here as a sequence of tokens (word forms and punctuation marks) with their morphological tags. All tokens are automatically divided into two disjoint sets: ordinary words and segment boundaries. The following tokens are considered as *segment boundaries*:

- Punctuation marks: comma, colon, semicolon, question mark, exclamation mark, dashes (all types), brackets (all kinds), and quotation marks (all types);
- Coordinating conjunctions: tokens with morphological tag starting with the pair `J^`, see Hajič (2004) (e.g., *a* 'and', *ale* 'but', *nebo* 'or', *nebot'* 'for', *ani* 'nor').

After the identification of boundaries, the input sentence is partitioned into individual segments; a *segment* is understood as a maximal non-empty sequence of tokens that does not contain any boundary.

The concept of linear segments serves as a good basis for the identification of clauses in Czech. This is possible due to very strict rules for punctuation in Czech: The beginning and the end of each clause in a Czech sentence must be indicated by a boundary (contrary to, e.g., English, where there are clauses with no formal markers, as in *She said she would come.*[4]); this holds for embedded clauses as well.

[4] In Czech, the subordinated clause representing the object must be separated by a comma and introduced by a subordinating conjunction, as in *Řekla, že přijde.*

This implies that a single *clause* consists of one or more segments (Section 3); several clauses then create a *complex sentence*.

2.2 Segmentation charts and clauses

Relations between clauses, especially super- or subordination, coordination, apposition or parenthesis, are described by the so-called *segmentation charts*. The segmentation chart captures the levels of embedding for individual segments, as described below.

The principal idea of the segmentation chart is quite simple, it can be described by the following basic instructions. In the examples, segments are marked by square brackets [and]$_k$, where k is the level of embedding, the boundaries are underlined. In addition, individual clauses are marked by brackets { and }$_j$, where j is an index of a particular clause.

2.2.1 Main clauses

Segments forming all main clauses[5] of a complex sentence belong to the basic level (level of embedding 0), as in the following sentence:

{[*O studium byl velký zájem*]$_0$}$_1$, {[*v pohovorech bylo vybráno 50 uchazečů*]$_0$}$_2$.
There was a lot of interest in studying , 50 applicants were selected in interviews .

2.2.2 Dependent clauses

Segments forming clauses that depend on clauses at the k-th level obtain level of embedding $k + 1$ (i.e., the level of embedding for subordinated segments is higher than the level of segments forming their governing clause):

{[*Potom zjistíte*]$_0$}$_1$, {[*že vám nikdo nedá vstupní vízum*]$_1$}$_2$.
Then you realize that nobody gives you an entrance visa .

2.2.3 Coordination and apposition

Segments forming coordinated sentence members and coordinated clauses occupy the same level. The same holds for an apposition.

{[*Hra nám jde*]$_0$}$_1$ *a* {[*forma stoupá*]$_0$}$_2$.
We're getting on well in game and our form improves .

2.2.4 Parenthesis

Segments forming parenthesis (e.g., sequence of word forms within brackets) obtain the level of embedding $k + 1$ if the level of their neighboring segments is k:

[5] We consider main clauses to be such clauses that are *syntactically/formally* independent, see also Section 3.

$\{[\textit{Návrh mluví o dvou letech u mužů}]_0\}_1 \quad \underline{(} \; \{[\textit{zvyšuje věk z 60 na 62}]_1\}_2 \quad \underline{).}$
The proposal mentions two years for men $\underline{(}$ it raises the age from 60 to 62 $\underline{).}$

Although this basic idea of segmentation charts seems simple, it appears that—when working with 'real data' from the newspaper corpus—detailed annotation guidelines are necessary for good and consistent annotation of specific linguistic phenomena and especially for their combination, see Lopatková and Kljueva (2010). In the following section, we focus on some of the guidelines.

3 Annotation of complex sentences from the PDT

The aim of the annotation is to explicitly describe relations between clauses in complex Czech sentences. We focus on the annotation of (a part of) Czech sentences from the PDT. We primarily take advantage of morphological analysis (the m-layer of PDT provides the word form, lemma and tag for every token) and partially— in case of ambiguous sentences—also from the surface syntactic analysis stored in the PDT (the a-layer; for the segment annotation, only information on analytical functions of tokens is used). The annotation only focuses on the identification of individual clauses and their mutual relations, which are not explicitly marked in the PDT.[6]

Let us stress here that the segment annotation is based on *formally expressed* structures rather than on their semantic interpretation. For example, we do not interpret text enclosed in brackets: whether it is semantically an apposition, a sentence member or an independent sentence part, as it is discussed by Kuboň et al. (2007). We annotate such text as a parenthetical segment on a lower level compared to the neighboring segments. The interpretation is postponed to the higher layers of annotation (some phenomena are resolved at the a-layer, other phenomena pertain to the t-layer).

The annotators have been instructed to *disambiguate* annotated sentences; if more readings are possible, they should respect the reading rendered in PDT.

3.1 Segments with different levels of embedding

The identification of a subordinated status of a particular segment is based on morphological properties of tokens forming this segment, i.e., on the presence of a token with 'subordinating function'. 'Subordinating tokens' are especially of the following types:

- Subordinating conjunctions (e.g., *aby* 'in order to', *dokud* 'till', *kdyby* 'if', *protože* 'because', *přestože* 'although', *že* 'that');

[6] This decision enables us to speed up the annotation as well as to avoid undesired overlapped/repeated annotation: The analytical layer of the PDT already contains the information on syntactic functions (like predicate, subject, object, nominal predicate, attribute, or adverbial); detailed semantic classification pertains to the tectogrammatical layer of the PDT.

- Relative/interrogative pronouns and some types of numerals (e.g., *kdo* ‘who’, *co* ‘what’, *jaký* ‘which’, *kolik* ‘how many’);
- Pronominal adverbs (e.g., *kde* ‘where’, *kdy* ‘when’, *jak* ‘how’, *proč* ‘why’).

3.1.1 Governing and dependent clauses

The super- and subordinated (mutually related) segments primarily capture relations between governing and dependent clauses.

A particular subordinated segment—as (a part of) a dependent clause—can precede or follow the superordinated segment(s) that create(s) its governing clause. Such a segment can also be placed in between two superordinated segments (in case of a governing clause with an embedded dependent clause, as in the following example with the embedded relative clause):

{[*Klejch*]$_0$	$\underline{,}$	{[*který dal devět ze dvanácti gólů Zlína*]$_1$}$_1$	$\underline{,}$	[*má vydatné pomocníky*]$_0$}$_2$	$\underline{.}$
Klejch		who scored nine out of twelve goals for Zlín		has good helpers	$\underline{.}$

In addition to governing and dependent clauses, there are also other constructions that should obviously be classified as subordinated segments. We will mention at least two of them, namely direct speech and parenthesis.

3.1.2 Direct speech

Segments (one or more) representing direct speech formally expressed by quotation marks[7] belong to a lower level, compared to the level of the neighboring segments:

$\underline{\text{„}}$	{[*Závod Ejpovice projevil zájem o 150 pracovníků*]$_1$}$_1$	$\underline{\text{,“}}$	{[*uvedl Ladislav Vltavský*]$_0$}$_2$	$\underline{.}$
$\underline{\text{“}}$	Ejpovice company showed interest in 150 workers	$\underline{\text{,”}}$	Ladislav Vltavský said	$\underline{.}$

On the other hand, segments representing direct speech *without quotation marks* (or other formal label(s)) are annotated as belonging to the same level as their neighboring segments. The reason is quite evident: there is no formally expressed indication of the subordination of such segments (the interpretation of these constructions is left to the higher layers of annotation, similarly as for other formally unexpressed phenomena).

{[*Přijdu později*]$_0$}$_1$	$\underline{,}$	{[*ohlásil doma Pavel*]$_0$}$_2$	$\underline{.}$
I will be late	$\underline{,}$	Pavel announced at home	$\underline{.}$

3.1.3 Parenthesis

Parenthesis marked by brackets (or other formal unambiguous label(s)), are annotated as belonging to a lower level compared to the neighboring segments. The

[7] Quotation marks marking direct speech have to be combined with another boundary in Czech, primarily with a comma. This rule serves for reliably distinguishing direct speech from the cases when quotation marks are used, e.g., for emphasizing individual words—the latter type gets the same level of embedding as its neighbors.

interpretation (whether it is an apposition, a sentence member such as, e.g., object or adverbial, or an independent sentence part) can be found at the a-layer of the PDT, see also Kuboň et al. (2007). In such cases, parenthetical expressions are captured as separate clauses even if they consist of a fragmental expression:

{[*Guido Reni*]$_0$ $\underline{(}$ {[*1575 až 1642*]$_1$}$_1$ $\underline{)}$ [*byl vynikající figuralista*]$_0$}$_2$ $\underline{.}$
Guido Reni $\underline{(}$ 1575 to 1642 $\underline{)}$ was an outstanding figural painter $\underline{.}$

In contrast, segments forming parenthesis without an explicit/unambiguous formal mark are annotated as consisting of segments on the same level as their neighboring segments (similarly as for direct speech, the annotation captures formally marked structures).

{[*Před smrtí*]$_0$ $\underline{,}$ {[*neznámo proč*]$_0$}$_1$ $\underline{,}$ [*si koupil tramvajenku*]$_0$}$_2$ $\underline{.}$
Before dying $\underline{,}$ nobody knows why $\underline{,}$ he bought a tram pass $\underline{.}$

3.2 Segments on the same level and identification of clauses

We can identify three main groups of structures in which segments are mutually related and share the same level of embedding: segments forming a clause with an embedded dependent clause, coordinated segments, and segments forming an apposition.

3.2.1 Segments forming a clause with an embedded dependent clause

Segments on the same level—unlike super/subordinated ones—can form a single clause, as in the following example of the attributive dependent clause splitting the main clause (the span of a sentence with an embedded clause being the most interesting case):

{[*V případě*]$_0$ $\underline{,}$ {[*že se nedovoláte*]$_1$}$_1$ $\underline{,}$ [*vytočte číslo ve večerních hodinách znovu*]$_0$}$_2$$\underline{.}$
In case that you do not succeed $\underline{,}$ redial the number again in the evening $\underline{.}$

For the annotators, the most important task is to *identify individual clauses*. They group those segments that constitute individual clauses of a complex sentence and thus mark them as separate syntactic units at the layer of clause structures.

3.2.2 Coordination of sentence members and coordination of clauses

The relation of coordination may occur between two (or more) sentence members (like subjects, objects, predicates, adverbials, etc.) or between two (or more) clauses, be they main clauses or dependent ones. The coordinated units are characterized by the same syntactic relations to other parts of a (complex) sentence, that is, the particular syntactic position is 'multiplied'. The annotators have to identify segments containing coordinated sentence members and put them together into a single clause. In contrast to this, coordinated clauses are marked as separate clauses sharing the same level of embedding,[8] as in the following sentence:

[8] In the PDT, a coordination of sentence members and a coordination of clauses are not distinguished (at the analytical layer).

{[*Český prezident apeloval na Čechy*]$_0$ *a* [*na Němce*]$_0$}$_1$,
The Czech president appealed to Czechs and to Germans
{[*aby odpovědně zacházeli s minulostí*]$_1$}$_2$ *a*
that they should treat their history responsibly and
{[*aby posouvali vpřed dialog*]$_1$ *a* [*spolupráci.*]$_1$}$_3$.
improve their mutual dialogue and cooperation .

This complex sentence consists of five segments (marked by [and]), which form three clauses (marked by { and }), namely one main clause (on the zero level) and two coordinated dependent clauses (on the first embedded level).

The segmentation is *purely linear* (one segment follows another). After the identification of segments, they are grouped into clauses—as we could see, a single clause prototypically consists of one or more segments. This is fully true for semantically and syntactically complete sentences, i.e., sentences without ellipses of different kinds. However, we can mention one construction where clauses identified by the annotators (i.e., clauses based on segments) do not conform with the linguistic intuition, namely the case of coordinated clauses sharing one (or more) sentence member(s) or a syntactic particle. We interpret such cases as ellipses, i.e., a shared sentence member or a particle is supposed to belong to only one of the clauses (and to be elided in the other clause); thus the shared sentence member or particle is annotated as a part of one clause only:

{[*Neopravuje* ***se***]$_0$}$_1$ *a* {[*neinvestuje*]$_0$}$_2$, {[*peníze na to nestačí*]$_0$}$_1$.
They do not renovate nor invest , there is not enough money .

The reflexive particle *se* (printed in bold) belongs to both verbs *opravovat* ‘to renovate’ and *investovat* ‘to invest’ (reflexive passive forms of the verbs); in the segmentation chart, it is marked as a part of the first clause *Neopravuje se* and elided in the second clause *neinvestuje*.

There is one exception to this basic instruction: if the shared sentence member is a predicate then the particular segments are joined together in a single clause (providing that no other formal labels as, e.g., brackets, indicate more levels), as in the following example:

{[*Petr přišel včera*]$_0$ *a* [*babička dneska*]$_0$}$_1$.
Petr came yesterday and my grandma today .

This decision is based on the assumption that a single finite lexical verb form indicates a single clause, i.e., a verb constitutes (a core of) a sentence.[9,10]

3.2.3 Apposition

Apposition is a construction where the same ‘idea’ is rendered in different ways (the latter being an explanatory equivalent of the former), both having the same syntactic

[9] The reason for this decision lies in the verb-centric character of dependency syntax traditionally used for Czech.

[10] At the a-layer, the ellipsis of a predicate is marked by a special analytical function; at the t-layer, ellipsis is restored (as a node of a tree).

relation to other sentence members (e.g., a name and a position of a particular person, as in the following sentence):

{[*Oznámil to Václav Havel*]$_0$ $\underline{,}$ [*president České republiky*]$_0$}$_1$ $\underline{.}$
It was announced by Václav Havel $\underline{,}$ President of the Czech Republic $\underline{.}$

Following the PDT, apposition is treated in the same way as coordination and the members of an apposition are considered to share (multiple) syntactic position in a sentence (like in the case of coordination).

4 Basic statistics and evaluation

We have gathered a collection of 3,444 sentences from the PDT, which were annotated with respect to their sentence structure (these sentences comprise 10,746 segments forming 6,341 clauses).[11] Two graduate students with very good linguistic backgrounds serve as annotators.

4.1 Inter annotator agreement (IAA)

In order to get the idea of how difficult the annotation task is (and how good our annotation instructions are) we have measured an inter-annotator agreement (IAA) for our two annotators. As a baseline, all segments got the most frequent level, i.e., basic level of embedding (level 0); clauses were not identified.

The agreement was calculated as follows, see Table 1: (1) Both annotators got the same set of segments and they assigned a level of embedding for each segment; they agree on this segment if they assign the same level. (2) The annotators identify particular clauses; they agree on a particular clause if they identify the same span of this clause. (3) The agreement on the whole sentence means that all segments of the sentence got the same level of embedding and that the same clauses were identified by the annotators.

The annotated data obtained from the annotators were analyzed. The most frequent cause of disagreement (after the exclusion of clear annotation errors) was a different annotation of unclear syntactic constructions like sentence fragments, sport scores, or addresses and phone numbers. Based on this analysis, we have refined the instructions in the annotation manual (Lopatková and Kljueva 2010). One of the annotators then went through the sentences with disagreement and unified the annotations according to the updated manual. As a result, we got so-called *golden data* that can serve for further exploitation.

4.2 Results for the existing automatic tools

The collection of golden data makes it possible to compare and evaluate the already existing tools for automatic identification of segments and clauses. We examined

[11] We have focused on the sentences from data/full/amw/train2 portion of the PDT data, i.e., one (out of eight) directory with the PDT standard training data with the annotation both on m- and a-layers; the number of annotated sentences is approximately the same as the number of sentences in the developing data set from this portion of PDT.

Table 1 IAA (label BL stands for the baseline, labels A1 and A2 for two annotators)

	Agree on sentences	%	Agree on segments	%	Clauses A1/A2	Agree on clauses	% A1/A2
BL	–	–	7,264	67.60	–	–	–
IAA	2711	78.72	10,118	94.16	6,301/6,369	4,932	78.27/77.44

Table 2 The results of segmentation tools, measured on the golden data

	Agree on sentences	%	Agree on segments	%	Clauses	Agree on clauses	%
LH	2,654	77.06	8,028	74.71	–	–	–
KK	2,110	61.27	–	–	5,609	4,512	71.16

two segmentation tools that aim at the identification of a level of embedding for individual segments and at the identification of individual clauses, respectively, using the a-layer of PDT: (1) the rule-based tool described in Lopatková and Holan (2009) (LH in Table 2; the tool provides levels of embedding for individual segments only, clauses are not identified), and (2) the tool based on machine learning methods described in Krůza and Kuboň (2009) (KK in Table 2; the algorithm specifies clauses directly, it does not work with the concept of segments). The numbers clearly confirm that the results of existing segmentation tools are not satisfactory yet. As both these tools are based on a dependency paradigm, the comparison with a tool based on a phrase-structure paradigm will be of a great interest.

5 Conclusion

In this article, a project aiming at obtaining a collection of sentences annotated with respect to their sentence structure was introduced. The data collection makes it possible to search for systematic differences between the manual and the automatic sentence structure annotation and thus it provides the possibility of further improvement of NLP tools.

Acknowledgments The article presents the results of the project supported by the grant No. 405/08/0681 and partially by the grant No. P202/10/1333, Grant Agency of the Czech Republic. Also, the authors are grateful to the unknown reviewers for their valuable suggestions.

References

Abney, S. P. (1991). Parsing by chunks. In R. Berwick, S. Abney, & C. Tenny (Eds.). *Principle-based parsing* (pp. 257–278). Dordrecht: Kluwer Academic Publishers.

Abney, S. P. (1995). Partial parsing via finite-state cascades. *Journal of Natural Language Engineering* *2*(4), 337–344.

Ciravegna, F., & Lavelli, A. (1999). Full text parsing using cascades of rules: An information extraction procedure. In *Proceedings of EACL'99* (pp. 102–109). University of Bergen, Bergen.

Hajič, J. (2004). *Disambiguation of rich inflection (computational morphology of Czech).* Prague: Karolinum Press.

Hajič, J., Panevová, J., Buráňová, E., Urešová, Z., Bémová, A., Štěpánek, J., et al. (2004). Anotace na analytické rovině. Návod pro anotátory. UFAL/CKL technical report no. 2004/TR-2004-23, ÚFAL/CKL MFF UK.

Hajič, J., Hajičová, E., Panevová, J., Sgall, P., Pajas, P., Štěpánek, J., et al. (2006). *Prague dependency treebank 2.0.* Philadelphia: Linguistic Data Consortium.

Holan, T., & Žabokrtský, Z. (2006). Combining Czech dependency parsers. In *Proceedings of TSD 2006*(pp. 95–102). Springer, LNAI, Vol. 4188.

Homola, P., & Kuboň, V. (2010). Exploiting charts in the MT between related languages. *International Journal of Computational Linguistics and Applications 1*(1–2), 185–199.

Jones, B. E. M. (1994). Exploiting the role of punctuation in parsing natural text. In: *Proceedings of the COLING'94*, (pp. 421–425).

Krůza, O., & Kuboň, V. (2009). Automatic extraction of clause relationships from a treebank. In *Computational linguistics and intelligent text processing. Proceedings of CICLing 2009* (pp. 195–206). Springer, LNCS, Vol. 5449.

Kuboň, V. (2001). *Problems of robust parsing of Czech.* PhD thesis, Faculty of Mathematics and Physics, Charles University in Prague, Prague.

Kuboň, V., Lopatková, M., Plátek, M. & Pognan, P. (2007). A linguistically-based segmentation of complex sentences. In D. Wilson & G. Sutcliffe (Eds.). *Proceedings of FLAIRS conference* (pp. 368–374). Menlo Park, CA: AAAI Press.

Lopatková, M. & Holan, T. (2009). Segmentation charts for Czech—Relations among segments in complex sentences. In A. H. Dediu, A. M. Ionescu, & C. Martín-Vide (Eds.). *Proceedings of LATA 2009* (Vol. 5457, pp. 542–553). New York: Springer, LNCS.

Lopatková, M., & Kljueva, N. (2010). Anotace segmentů. (Anotanční příručka) (in manuscript).

Marinčič, D., Šef, T., & Gams, M. (2010). Intraclausal coordination and clause detection as a preprocessing step to dependency parsing. In V. Matoušek, & P. Mautner (Eds.) *Proceedings of TSD 2009* (Vol. 5729, pp. 147–153). Springer, LNAI, New York.

Ohno, T., Matsubara, S., Kashioka, H., Maruyama, T., & Inagaki, Y. (2006) Dependency parsing of Japanese spoken monologue based on clause boundaries. In *Proceedings of COLING and ACL, ACL*, (pp. 169–176).

Šmilauer, V. (1969). *Novočeská skladba (New Czech syntax).* PhD thesis, Praha: Státní pedagogické nakladatelství.

Zeman, D. (2004). *Parsing with a statistical dependency model.* PhD thesis, Prague: Charles University in Prague.