Studying Properties of Czech Complex Sentences from an Annotated Corpus

Vladislav Kuboň and **Markéta Lopatková**
Charles University in Prague
Czech Republic
{vk,lopatkova}@ufal.mff.cuni.cz

Abstract

The paper deals with the problem of an analysis of complex sentences in Czech on the basis of manually annotated data. The availability of a specialized corpus explicitly describing mutual relationships between segments and clauses in Czech complex sentences, together with the availability of a thoroughly syntactically annotated corpus, the Prague Dependency Treebank, provide a solid background for linguistic investigation. The paper presents quantitative, linguistic and structural observations which provide a number of clues for building an algorithm for analyzing a structure of complex sentences in the future.

1 Introduction: Boundaries and Segments

Syntactic analysis of mutual relationships between clauses in complex sentences constitutes one of possible approaches towards an improvement of results of various analyzers, regardless whether they are based upon traditional handcrafted grammars or upon stochastic or machine learning methods. Regardless of the methods used, the information about a composition of a complex sentence, about mutual relationships between clauses and about their internal composition may substantially simplify the process of syntactic analysis.

Our approach is based upon a notion of segments, naturally and unambiguously defined sequences of words. The original idea described in (Kuboň 2001) was more precisely defined in (Kuboň et al. 2007). It has been further modified in (Lopatková and Holan 2009) for the purpose of automatic as well as manual annotations. Coordinating conjunctions and punctuation marks have been defined here as **segment boundaries**; a **segment** is then understood as a maximal non-empty sequence of tokens not containing any boundary; a simple clause consists of one or more segments.

The division of a complex sentence into segments by means of the boundaries is possible thanks to a set of relatively strict rules existing in the Czech grammar for punctuation and for using coordinating (and subordinating) conjunctions; these expressions unambiguously separate individual segments.

The classification of tokens is performed by morphological analyzer. In very few cases when a particular boundary nevertheless obtains an ambiguous morphological tag in the analysis, it is possible to disambiguation it by a stochastic tagger (for example, the expression *jak* [how/yak] can either be a coordinating conjunction and therefore a boundary, or it may be as well a subordinating conjunction, pronominal adverb or noun, in which cases it is not considered as a boundary). Although the precision of best available taggers for Czech is only slightly better than 96%, they achieve over 99.3% in determining part of speech and its subpart (including the distinction between coordinating and subordinating conjunctions (Spoustová 2008)), which is sufficient for our purposes.

2 Segmentation and Syntactic Analysis

Syntactic analysis of languages with a free word-order, and Czech definitely belongs to this group, faces a wide variety of issues. Let us mention at least those which are relevant to segmentation and clause structure of complex sentences.

One of the common problems is a correct identification of **individual clauses** in complex sentences and their **mutual relationships**. This problem can be illustrated for example by a sentence *Vyskytl se i případ, kdy nájemník neplatil nájem po určitou dobu, kdy byl nezaměstnaný, a po nalezení zaměstnání dluh uhradil.* [There was also a case of a tenant who didn't pay a rent for a certain period when he had been jobless and who has settled the debt after he has found a job.] The last segment *po nalezení zaměstnání dluh uhradil* can be analyzed in two ways, either as a clause coordinated with the main clause *Vyskytl se i případ [...] a po nalezení zaměstnání dluh uhradil.*, or as a clause constituting one of two coordinated attributive clauses *kdy nájemník neplatil nájem po určitou dobu [...] a po nalezení zaměstnání dluh uhradil.*[1]

Another important issue of syntactic analysis of (not only) Czech is the determination of the scope of **embedded sentential constructions** (subordinated clauses, insertions, parentheses). Although these constructions usually have an easily recognizable beginning (subordinated conjunction, relative pronoun, pronominal adverb, etc.), to de-

[1] Both results are correct from the point of view of syntactic analysis; the preference for the second variant is determined at the level of understanding the meaning of the sentence in the context of discourse (pragmatics).

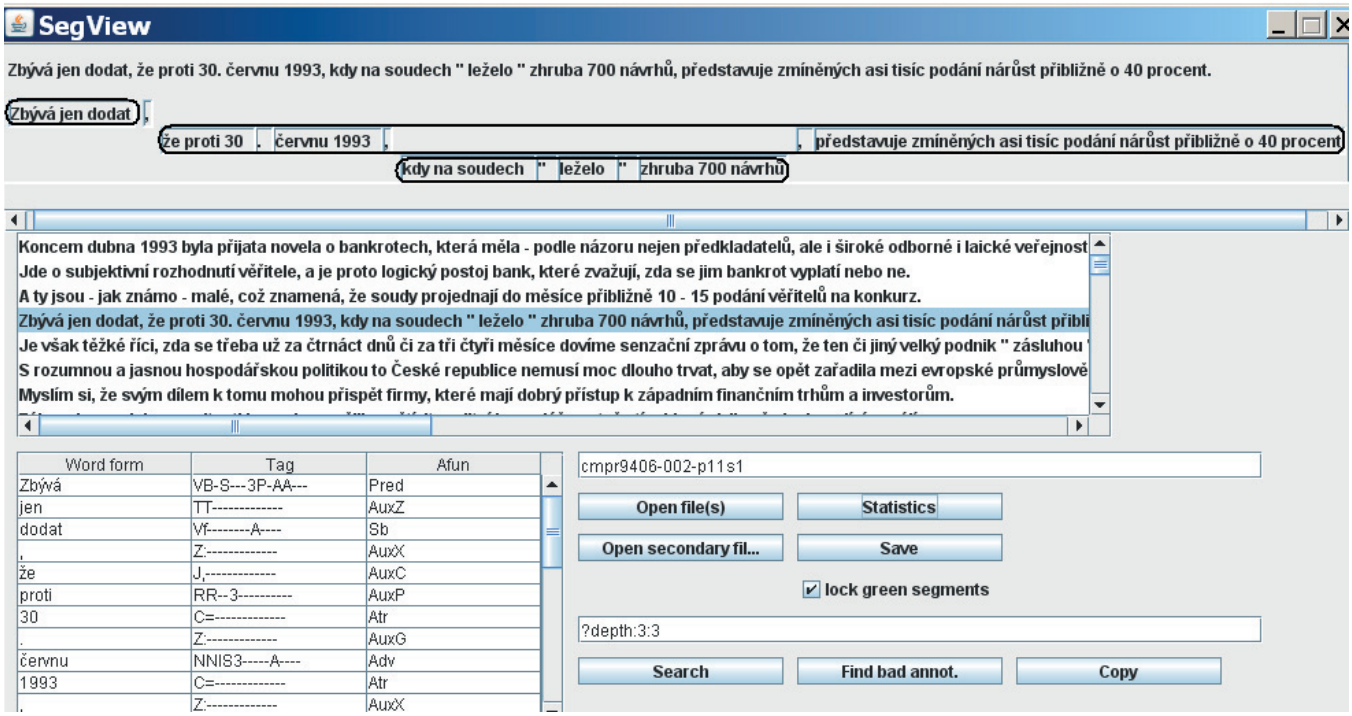

Figure 1: The SegView Editor: segmentation scheme of the sentence *Zbývá jen dodat, že proti 30. červnu 1993, kdy na soudech „leželo“ zhruba 700 návrhů, představuje zmíněných asi tisíc podání nárůst přibližně o 40 procent.* [It only remains to add that compared to June 30th, 1993, when the courts had about 700 “stalled” files, the above mentioned thousand of files constitutes approximately 40 percent increase.] (Clauses marked by ellipses, maximal level of embedding equals 2.)

termine their end constitutes far more difficult task.

The most problematic phenomenon, which makes syntactic analysis very difficult, concerns **coordinations**, eventually **appositions**. From the point of view of a pure syntactic analysis it is not so much about distinguishing between these two phenomena, as about the distinction between **intraclausal coordination** (coordination of words in a single clause) and **extraclausal coordination** (the coordination of clauses). This distinction determines the type of a mutual relationship between individual clauses.

The above mentioned constructions call directly for exploiting a certain mechanism making the best of the results of morphological analysis and providing information about clauses and their mutual relationship for the proper syntactic analysis. This mechanism would find certain well-defined and linguistically motivated units (segments), which would be joined later into clauses by means of a set of rules. Taking into account that the starting segments of clauses usually define the role of the clause in the complex sentence (due to various subordination markers as, e.g., subordinated conjunctions, relative pronouns, etc. which are usually located at the beginning of a first segment of the clause), these rules might help to establish mutual relationships between clauses in complex sentences. This would substantially simplify subsequent steps of the process of syntactic analysis (inserted subordinated clauses can be parsed separately), because all parsing methods are sensitive to the length of the input sentence, as it was analyzed for Czech for example in (Zeman 2004).

2.1 Available Data and Tools

In order to investigate the behavior of segments, it was necessary to create corresponding data. The available syntactically annotated corpora (for Czech it is especially the Prague Dependency Treebank (PDT), see (Hajič et al. 2006)) concentrate on relationships between pairs of words, they lack an explicit annotation of relationships between bigger sentential elements. It was therefore necessary to transform data from PDT into a form more suitable for our experiments. We have used both an automatic method described in detail in the paper (Krůza and Kuboň 2009), as well as manual a annotation described in (Lopatková, Klyueva, and Homola 2009). In the subsequent sections of this paper we are studying the manually annotated set of 3 444 sentences from PDT. The annotation describes a structure of complex sentences as it was perceived by human annotators. They had at their disposal automatically determined segments (the algorithm is very straightforward and unambiguous, the results are reliable) and they concentrated on determining mutual relationships among segments (which segments constitute individual clauses, whether they are in a relationship of coordination or subordination, whether they constitute a parenthesis, etc.) and among all clauses in complex sentences.

The annotators used SegView, a tool designed especially for this purpose. A screenshot of this tool is displayed in Fig. 1. SegView is not only an editor, it also allows a user to

No. of segments	No. of sentences	No. of clauses								
		1	2	3	4	5	6	7	8	9
1	942	942								
2	804	396	408							
3	583	165	236	182						
4	400	100	124	107	69					
5	275	48	81	7	4	29				
6	171	26	30	45	35	25	10			
7	85	10	22	24	14	7	6	2		
8	61	12	7	13	13	9	5	1	1	
9	40	7	8	7	6	3	4	1	2	2
10	26	3	1	6	2	3	3	3	4	1
> 10	57	12	8	5	6	4	7	8	4	3

Table 1: Number of clauses and number of segments in the corpus

search for interesting examples – apart from trivial queries on forms, lemmas or tags, SegView also allows the search for interesting structures, as, e.g., complex sentences with a particular number of clauses, complex sentences with a particular depth, or complex sentences containing a maximal 'jump' between individual segments.[2] The basic statistics expressing the frequency of particular linguistic phenomena playing a role in establishing mutual relationship of Czech clauses are being introduced in the subsequent section.

3 Analysis of Selected Phenomena

3.1 A Quantitative Analysis

The first type of data analysis enabled by the SegView tool is a quantitative analysis. It helped us to investigate certain properties of Czech texts, which might be important for designing an algorithm for identification of individual clauses and establishing their mutual relationship. The results of quantitative analysis are displayed in Table 1. We have identified 10 746 segments and 6 341 clauses in the 3 444 sentences of golden (hand-annotated and double checked) data.

The numbers contained in Table 1 document a fact which is not really surprising: simple sentences and complex sentences containing only two clauses having at most two segments constitute substantial part of data. It is exactly 1 746 sentences, in other words slightly more than a half of the total number of sentences contained in the set. These sentences are of course trivial, because even those containing two clauses are relatively simple: the end of both clauses is easy to find and their mutual relationship (coordination, subordination or parenthesis) is determined by the nature of the delimiting expression (coordinating conjunction or comma, followed by subordinating conjunction or relative pronoun, etc.)

The opposite end of the table contains a couple of interesting extreme cases. One of them is a sentence having a maximal number of segments (27) in the whole corpus. This sentence at the same time consists only of a single clause, it is not even a complex sentence. It looks as follows: *Tenis Atlanta - 2. kolo: Chang - Mattar 6 : 3, 7 : 5, Martin - Dunn 6 : 3, 6 : 2, Agassi - Reneberg 4 : 6, 6 : 2, 6 : 4, Washington - Connors 6 : 4, 3 : 6, 6 : 1.* (lnd94101-082-p1s13, PDT).

Although this sentence (similarly as other extreme examples of the huge discrepancy between the numbers of segments and clauses) constitutes a very specific case, it has to be taken into account as well as other sentences because the sentences from PDT are sentences of a real written language and as such they contain also phenomena from the language periphery, with a frequency which is definitely not negligible.

A similar sentence type is represented by the following sentence, which contains 4 clauses and 20 segments: *Oslovili jsme lidi vesměs známé, zajímavé a talentované (mj. Jireš, Špáta, Vihanová, Vorel, Němec, Císařovský, Pavlásková, Svěrák, Chaun, Kačírek, Koutecký) s tím, že každý měl zároveň navrhnout „svůj objekt", hrdinu portrétu, který by rád osobně natočil.* (mf920901-025-p3s4, PDT) [We have addressed people altogether famous, interesting and talented (i.a. Jireš, Špáta, Vihanová, Vorel, Němec, Císařovský, Pavlásková, Svěrák, Chaun, Kačírek, Koutecký) with a proposal that everybody was supposed to suggest "his object", a hero of a portrait which he would like to picturize personally.].

These extreme sentences have one advantage: they are relatively easily recognizable in an ordinary text by means of simple non-linguistic rules – sport results, long 'shopping' lists, various tables, etc., contain a large number of very short segments and a small number of finite verbs. It is therefore possible to run a module for identification of frequent types of 'suspicious sentences' prior to the module of linguistically motivated analysis. Such analysis may then concentrate upon the core of the identification issue, namely upon the sentences located roughly in the middle of Table 1.

3.2 Linguistic Analysis

Apart from quantitative analysis we have also tried to analyze concrete sentences linguistically in order to obtain a set of clues making it possible to design an algorithm for connecting segments into clauses. Similar investigation has already been performed for another Slavic language, Slovenian. In the paper (Marinčič, Šef, and Gams in prass) the authors suggest an algorithm for connecting segments into clauses. They also concentrate especially on distinguishing

[2] We would like to express our thanks to the SegView author, Petr Homola, who also took care about technical support during the annotation and the evaluation of data.

between intra- and extraclausal coordinations.

Let us first summarize basic facts about Czech complex sentences, which we can extract from available set of data. The most important fact concerns the relationship between the **number of clauses** and the **number of finite verbs**: prototypically, the number of clauses is identical to the number of finite verbs. However, these two numbers may differ in some cases, for example various titles, lists, parenthesis, texts in brackets frequently don't contain any finite verb; nevertheless, we consider them to be clauses.

The question of transgressives seems to be intricate, too. Although being considered as infinite verbal forms in grammar books, in most cases they constitute their own independent segment separated from the rest of the sentence by a comma. In such cases, it is more adequate to consider a segment containing a transgressive as a separate clause. Unfortunately, this is not always the case. The evidence for this claim may even be found in such a respectable source as Šmilauer's Novočeská skladba (Šmilauer 1966) (even though most of his counterexamples come from older literature): one segment may contain both the transgressive and the main (finite) verb – *Nasytiv se chlebem usnul.* [Stuffed with bread he fell asleep] (Jirásek) or *Chlapec směje se dobře mu odpověděl.* [The boy answered him well laughing] (Němcová).

Similarly, in the sentences such as *Stejně jako další z legend, kterými hoteliér láká hosty do lokálu.* (ln95040-062-p2s9, PDT) [As well as another of the legends, by means of which the hotel owner lures guests into the pub] it is possible to find more clauses than there are finite verbs.

On the other hand, selected finite verb forms may have a function of a particle despite being verbs morphologically (esp. in 1st person sg), as in *A te si prosím najděte sedadla.* [And now please find your seats.] or *Nejdu doufám pozdě?* [Hopefully I am not late.] (Czech National Corpus). In these cases, two finite verbs are not separated by a boundary, i.e. they belong to a single segment, which is considered as a single clause. The few verbs allowing such usage can be listed.

The fact that the number of clauses roughly (with the exceptions discussed above) corresponds to the number of finite verbs can also help a lot in the identification of a coordination as intra- or extraclausal. Let us for example take a sentence *Koncem dubna 1993 byla přijata novela o bankrotech, která měla – podle názoru nejen předkladatelů, ale i široké odborné i laické veřejnosti – vyvolat dominový efekt krachu podniků, které si vzájemně neplatí.* (cmpr9406-002-p4s1, PDT) [An amendment of the bankruptcy law, which was supposed to – according to the opinion of not only the submitters, but also the broad expert and laymen public – initiate a domino effect of bankrupting companies which do not pay to each other, was passed at the end of April 1993.] This sentence contains a number of interesting phenomena. There are three finite verbs and seven segments (the combination *, ale i* is considered to constitute a single boundary between segments). Let us number them for a more easy orientation:

1. *Koncem dubna 1993 byla přijata novela o bankrotech* [An amendment of the bankruptcy law has been passed at the end of April 1993]
2. *která měla* [which was supposed to]
3. *podle názoru nejen předkladatelů* [according to the opinion of not only the submitters]
4. *široké odborné* [broad expert]
5. *laické veřejnosti* [laymen public]
6. *vyvolat dominový efekt krachu podniků* [initiate a domino effect of bankrupting companies]
7. *které si vzájemně neplatí* [which do not pay to each other]

Just the number of segments itself clearly suggests that some of them will constitute a single clause. Because the complex sentence contains a number of coordinating conjunctions, it is highly probable that it contains an intraclausal coordination. The pair of hyphens (–) also helps to determine the span of the coordination: due to the lack of presence of finite verbs between the hyphens it is possible to classify the coordinating conjunctions as intraclausal; i.e., the whole sequence between hyphens belongs to a single clause.

If we join the segments 3, 4 and 5 into a single unit as a result of these considerations, the complex sentence will still consist of five segments and three finite verbs. It is interesting that now we can find a verb in each remaining segment – three finite verbs and one infinite verb in the segment 6. This infinite form can hardly stay alone, and because the verb *měla* [was supposed to] in the segment 2 is a modal verb in Czech and as such it is related to the infinite form, it is possible to join the segments 2 and 6 into a single unit.

Only four units (candidates for clauses) will remain in the sentence now. From the point of view of syntactic analysis it then doesn't matter if the embedded group 3, 4 and 5 will be treated as a parenthesis (and analyzed separately) or whether it will be regarded as an unseparable part of the clause consisting of the original segments 2, 3, 4, 5 and 6. In this case the big difference between the number of clauses and the number of segments helps us to determine the moment when to stop joining segments: whenever the numbers are close enough.

Let us apply the same method to one more example: *Abyste mohla tento nárok s spěchem ve stanovené lhůtě uplatnit, bylo by třeba, abyste byla nejenom československou, a později českou občankou, ale měla i trvalý pobyt na zemí ČR.* (cmpr9407-005-p10s1, PDT) [In order to be able to apply this pretence successfully within a given period, it would be necessary to be not only Czechoslovak, and later Czech citizen, but to have also a permanent residence in the territory of the CR.] and let us divide it into segments.

1. *Abyste mohla tento nárok s spěchem ve stanovené lhůtě uplatnit* [In order to be able to apply this pretence successfully within a given period]
2. *bylo by třeba* [it would be necessary]
3. *abyste byla nejenom československou* [to be not only Czechoslovak]
4. *později českou občankou* [later Czech citizen]
5. *měla* [to have]
6. *trvalý pobyt na zemí ČR* [a permanent residence in the territory of the CR.]

This complex sentence contains 6 segments and 4 finite verbs in Czech. In the case of segments 3 and 4 we can easily

identify intraclausal coordination *československou a později českou občankou* [Czechoslovak and later Czech citizen] on the basis of morphological information (identical cases of an adjective and nominal group). These segments can therefore be joined. When searching for additional candidates for joining we can rely on a coordinating conjunction *i* coordinating segments 5 and 6. This conjunction isn't a standard coordinating conjunction (this role belongs to the conjunction *a* [and]), here it has a slightly contrastive role. This is, of course, an information which is not available during this stage of the analysis of complex sentences. Instead we will notice that this conjunction *i* connects a verbal form on the left hand side with a nominal group on the right hand side, where the segment containing the verbal form preceded by a conjunction *ale* [but]; it is therefore highly probable to determine that it is a contrastive connection expressed by a pair *ale i* [but also] (in a distant position). Would there be a finite verb in the segment 6, it would be on the contrary a clear coordination of clauses. This observation clearly documents that a wide variety of conjunctions has to be considered in a proper wider context and that it is important to take into account not only their morphological tag, but also their lexical value.

This claim is supported by yet another example: *Je však těžké říci, zda se třeba už za čtrnáct dnů či za tři čtyři měsíce dovíme senzační zprávu o tom, že ten či jiný velký podnik „zásluhou" příslušné banky zbankrotoval.* (cmpr9406-002-p18s1A, PDT) [It is however difficult to say whether in fourteen days or in three four months we will perhaps learn a sensational news that this or that big company went bankrupt "thanks" to a bank in question.]. Individual segments look like this:

1. *Je* [It is]
2. *těžké říci* [difficult to say]
3. *zda se třeba už za čtrnáct dnů* [whether in fourteen days]
4. *za tři čtyři měsíce dovíme senzační zprávu o tom* [in three four months we will perhaps learn a sensational news]
5. *že ten* [that this]
6. *jiný velký podnik* [that big company]
7. *zásluhou* [thanks]
8. *příslušné banky zbankrotoval* [a bank in question went bankrupt]

Three finite verbs can be found in segments 1, 4 and 8. The conjunction *či* [or] connects segments 5 and 6 into a single unit, because there is no verb between the subordinated conjunction *že* [that] and the conjunction *či* [or] which would be coordinated with the verb on the right hand side of this conjunction. It must therefore be an intraclausal coordination. A single verb on the right of the segment 5 also suggests that the quotation marks between segments 6, 7 and 8 have only emphasizing role and all segments 5, 6, 7 and 8 can be connected into a single unit (on top of that, if the quotation marks would indicate a direct speech and thus also another clause, they would be combined with punctuation). The whole set of segments is therefore a single clause (separated from the rest of the complex sentence by a subordinating conjunction on the left and by a full point closing the sentence from the right). This means that only 5 units are left. A clear candidate for joining with other segment is the segment 2 which cannot remain independent and it must be joined with the segment 1. The word *však* [however] does have a role of an adverb in this context, although in principle it could also have been a coordinating conjunction. The segments 3 and 4 can be connected on the basis of the fact that the verb *dovědět se* [to learn] is a reflexivum tantum and thus the reflexive particle *se* in the segment 3 belongs to the same clause. After this operation the number of segments equals the number of finite verbs and the joining is finished.

3.3 Structural Analysis

The available data enable also interesting observations concerning the structure of segments and the level of their embedding. They for example show that prototypically **the embedding of a segment can be only one level deeper compared to the previous segment**. This rule is violated only in 12 sentences from our dataset, out of which in 9 cases the sentences contain phenomena similar to those described in (Lešnerová-Zikánová and Oliva 2004): there are two 'subordination markers' there, as, e.g., a pair of subordinating conjunctions *že když* [that when] in the following sentence: *Zjistili jsme, že když žijeme v Čechách, měli bychom hrát muziku pro českého posluchače.* (ln95041-042-p7s6, PDT) [We have found out that when we live in Bohemia we should play the music for Czech listeners]). The remaining 3 cases concern dependent clauses in a direct speech, e.g., *Zdeněk Müller, trenér Kladna: „Jestli mám někoho pochválit, pak točníka Tona a Chlada v brance."* (ln95040-032-p2s14, PDT) [Zdeněk Müller, Kladno coach: "If I must praise somebody, then the attacker Ton and Chlad in the goal."]). Because both these cases can be determined with a high precision on the basis of morphological analysis, this observation brings very important information about an acceptable form of a segmentation scheme of the sentence.

In a similar way we can safely determine **the level of the first segment** – prototypically, the first segment not containing any subordination marker occupies the basic level. The first segment located on the level 1 was observed 190 times in the analyzed data; in these case the analysis showed that the first segment has the following characteristics:

subordination marker	105
fragment in brackets	15
direct speech (with a pair of quotation marks)	33
direct speech (only closing quotation marks)	26
semidirect speech	11

With the exception of semidirect speech, all other cases are easily recognizable and assigning a correct level is therefore relatively easy.

The first segment located on the level 2 has been observed only 4 times in the analyzed data, always as a dependent clause (with a subordination marker) in a direct speech. There was no case of a first segment at a lower level than 2, although theoretically it cannot be ruled out, see, e.g., the sentence *„A že když se bavím s osmnáctiletými kluky, připadám si jako instituce, přiznávám se," smál se trenér.* ["And that when I talk to eighteen years old lads I feel myself as an institution, I admit," laughed the coach.]): here the first segment lies at the level 3.

4 Observations Summary

The examples mentioned above make it possible to formulate a couple of observations which have to be taken into account designing a reliable algorithm for building the segmentation structure of a sentence.

It turns out that the linguistic analysis should follow **the identification of sentences with non-standard structure of clauses or segments** (lists, addresses, sport results, etc.).

The linguistic analysis showed that from the linguistic point of view there is one issue which is apparently more important than all others. The ability to distinguish for a particular coordinating conjunction whether it is in a given context **an intra- or extraclausal coordination** is crucial for the success. For this decision it is necessary to take into account especially the presence or lack of certain word forms in coordinated segments, the agreement between the segments (it is much more likely that the segments coordinated inside a single clause will agree in gender, number and case), etc.

The **lexical value** of the **conjunction** itself is very important. Already in the process of annotation it was noticed that some conjunctions (*však* [nevertheless], *proto* [hence], *či* [or], etc.) require a special treatment, it is definitely not possible to rely on the morphological tag only, the concrete lexical value must be taken into account as well and each of these conjunction must be treated individually.

Another very important set of rules describes **joining of segments** containing certain verbal forms with segments containing certain word forms complementing these verbs. A very good example of this category are separated reflexive particles, which may be connected to reflexives tantum; another group consists of words with valency slots requiring specific form, as, e.g., infinite verb (segments containing a particular verb, whose valency frame contains a slot for an infinite verb and a segment containing this infinite verb are very likely to be a part of a single clause). The rules for joining segments require the exploitation of priorities, the highest priority will be given to the rules for the intraclausal coordinations.

A substantial role is also being played by **structural constraints** which must be applied on the shape of the segmentation scheme, in other words, on a possible structure of segments and on the level of their embedding; the structure of segments then more or less determines the structure of clauses.

When it will not be possible to apply rules for joining segments, we will apply **special heuristics**, created on the basis of specific phenomena identified in section 3. They will – among others – concentrate on joining segments without finite verbs to segments containing these verbs, on solving the cases of clauses contained in brackets, etc.

5 Conclusions and Future Work

As we have stated above, the main topic of this paper is an analysis of data obtained by a manual transformation of the PDT into a shape taking into account mutual relationship of clauses in Czech sentences. This set of data is large enough for providing the background for creating reliable rules for joining segments into clauses. The initial observations performed on the annotated data indicate that it will be possible to create relatively reliable set of linguistically motivated rules or heuristics.

Making the process of formulating the rules more automatic is the main goal for future research. The investigation performed so far clearly indicates that the most important subtasks are distinguishing between intra- and extraclausal coordination wherever coordinating conjunctions are involved; the majority of complex sentences with multiple segments in the data analyzed so far contains one or more intraclausal coordinations. Preprocessing of specific cases of problematic sentence types seems to be also one of the key factors which might help to increase the precision of the main algorithm.

Acknowledgments

The research reported in this paper was carried out under the project of MŠMT ČR No. MSM0021620838. It was supported by the grant of GAČR No. 405/08/0681.

References

Hajič, J.; Hajičová, E.; Panevová, J.; Sgall, P.; Pajas, P.; Štěpánek, J.; Havelka, J.; and Mikulová, M. 2006. *Prague Dependency Treebank 2.0*. Philadelphia, PA, USA: Linguistic Data Consortium.

Krůza, O., and Kuboň, V. 2009. Obtaining Hidden Relations from a Syntactically Annotated Corpus – From Word Relationships to Clause Relationships. In *Proceedings of Flairs 2009*, 501–505. Menlo Park: AAAI Press.

Kuboň, V. 2001. *A Robust Parser for Czech*. Ph.D. Dissertation, Charles University in Prague, Prague.

Kuboň, V.; Lopatková, M.; Plátek, M.; and Pognan, P. 2007. A Linguistically-Based Segmentation of Complex Sentences. In Wilson, D. C., and Sutcliffe, G. C., eds., *Proceedings of FLAIRS 2007 Conference*, 368–374. Menlo Park: AAAI Press.

Lešnerová-Zikánová, Š., and Oliva, K. 2004. Česká vztažná souvětí s nestandardní strukturou. *Slovo a slovesnost* 64(4):241–252.

Lopatková, M., and Holan, T. 2009. Segmentation Charts for Czech – Relations among Segments in Complex Sentences. In Dediu, A. H.; Ionescu, A. M.; and Martín-Vide, C., eds., *Proceedings of LATA 2009*, *LNCS* 5457, 542–553. Springer-Verlag.

Lopatková, M.; Klyueva, N.; and Homola, P. 2009. Annotation of sentence structure; capturing the relationship among clauses in czech sentences. In *Proceedings of the LAW III Workshop*, 74–81. Suntec, Singapore: Association for Computational Linguistics.

Marinčič, D.; Šef, T.; and Gams, M. in press. Parsing with clause and intraclausal coordination detection. *Computing and Informatics*.

Spoustová, D. 2008. Combining Statistical and Rule-Based Approaches to Morphological Tagging of Czech Texts. *The Prague Bulletin of Mathematical Linguistics* 89:23–40.

Šmilauer, V. 1966. *Novočeská skladba*. Praha: SPN.

Zeman, D. 2004. *Parsing with a Statistical Dependency Model*. Ph.D. Dissertation, Charles University, Prague.