Analysis of Coordinating Constructions in a Dependency Treebank

Vladislav Kuboň and **Markéta Lopatková** and **Jiří Mírovský**
Charles University in Prague, Faculty of Mathematics and Physics
Czech Republic
{vk,lopatkova,mirovsky}@ufal.mff.cuni.cz

Abstract

This paper summarizes results of automatic analysis of coordinating constructions and appositions in the Prague Dependency Treebank using a method of analysis by reduction. Experiments are performed on a large subset of the treebank. This subset is obtained as a result of a query providing a set of more than 4,300 suitable sentences and their tree structures containing coordinations and appositions. The automatic procedure is complemented by a manual analysis of reasons why certain sentences (trees) were not fully reduced. This analysis helps to gain a better insight into the phenomena of coordination and apposition and their formal properties.

Dependency trees have a long tradition in linguistics, especially in the description of Slavic languages of Central and Eastern Europe. Although the history of linguistics witnessed many heated discussions between the followers of the tradition of constituent trees and the linguists adoring dependency trees (both sides usually unable to persuade the opponents about the advantages of their type of trees), it seems that the dependency notation has recently been recognized as an efficient and transparent data type for the description of syntactic relations in treebanks for a number of languages. One of the dependency treebanks which became quite popular among linguists due to the thoroughness of its annotation is the Prague Dependency Treebank (PDT), see (Bejček et al. 2013), the corpus exploited in this paper.

One of the difficulties faced by the dependency notation is the necessity to express within a dependency tree not only *dependencies*, but also relations which are naturally not of a dependency nature. This is a well-known issue, a member of the Prague Linguistic Circle, Lucien Tesnière (Tesnière 1959) already distinguished between structural relations which we nowadays call as dependency ('connexion'), and between the coordinating relationships ('junction').[1]

In this paper we study primarily the relationships of *coordination and apposition*, which pose a great challenge to any dependency formalism. The analysis is performed by means of *analysis by reduction* (Lopatková, Plátek, and Kuboň 2005; Lopatková, Plátek, and Sgall 2007), a procedure which naturally defines governing and dependent words in a dependency relationship; we enrich the procedure to capture also relations of coordination and apposition. We have applied the analysis by reduction automatically to dependency trees of selected sentences from the Prague Dependency Treebank. The results obtained by this method were then manually analyzed. This approach actually brings two kinds of results – it helps to gain better insight into the problem of coordination and its annotation in the corpus, and, on top of that, it also helps to identify potential annotation inconsistencies in the corpus.

The theory we adhere to in our investigations, dependency-based Functional Generative Description (FGD), is described primarily in (Sgall, Hajičová, and Panevová 1986).

[1] However, such conception is not accepted without reservations – there are influential approaches which capture coordination as a (type of) a dependency relation, see esp. (Mel'čuk 1988).

Analysis by Reduction

The original method of analysis by reduction (AR) makes it possible to formulate the relationship between dependency and word order (Lopatková, Plátek, and Kuboň 2005). This approach is beneficial especially for modeling the syntactic structure of languages with a high degree of free word order, where the *dependency structure* and *word order* are only loosely related.

Let us now describe the ideas behind the method used for sentence analysis. Analysis by reduction is based on a stepwise simplification of an analyzed sentence. It defines possible sequences of reductions (deletions) in the sentence – each step of AR is represented by *deleting* of at least one word of the input sentence; in specific cases, deleting is accompanied by a *shift* of a word form to different word order position.

Let us stress the basic constraints imposed on the analysis by reduction, namely: (i) the obvious constraint on preserving individual word forms, their morphological characteristics and/or their surface dependency relations, and (ii) the constraint on preserving the correctness (a grammatically correct sentence must remain correct after its simplification).

The basic principles of AR can be illustrated by the following Czech sentence (1).

Example (1):

Petr se bojí o otce.
Petr - Refl - worries - about - father
(Petr worries about his father.)

In sentence (1) we can consider the subject *Petr* as dependent, therefore it is possible to remove it in the course of AR (it will become a leaf in the dependency tree). Similarly, it is also possible to reduce the prepositional group *o otce*, which is represented by a subtree in the tree (the selection of a governing node is determined by technical rules, for more details see (Lopatková, Plátek, and Kuboň 2005)). The reflexive particle (clitic) *se* is considered to be a dependent node (on the basis of the analogy principle (Sgall, Hajičová, and Panevová 1986)), because in Czech there are also verbs without a clitic, as, e.g. *odpovědět* (to answer).

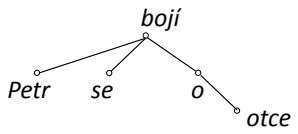

Figure 1: The dependency tree of sentence (1).

We can notice that the order of reductions reflects the dependency relations in the corresponding dependency tree, as it is described in (Lopatková, Plátek, and Sgall 2007; Plátek, Mráz, and Lopatková 2010). Informally, the words are 'cut from the bottom of the tree'; i.e., a governing node must be preserved in a simplified sentence until all its dependent words are deleted, see also (Lopatková, Plátek, and Kuboň 2005).

This basic version of the analysis by reduction, used in (Kuboň, Lopatková, and Mírovský 2013) can be naturally enriched to capture also relations of coordination and apposition – the whole (skeletal) coordination/apposition structure can be deleted in a single step of the analysis by reduction.

Data

Although the analysis by reduction captures the dependency relations in a sentence quite naturally, there is one issue which makes its use for an automatic analysis of data problematic. This issue concerns the ability to recognize correctly what can be removed (deleted) in each step of AR. Such a decision is relatively easy (apart from some ambiguous or very complicated cases) for humans who can exploit not only their knowledge of syntax, but also the understanding of the meaning of the sentence and even the knowledge of the real world. Such a level of comprehension is not achievable in an automatic process, so this kind of knowledge has to be replaced by a different type of information.

The only reliable source determining the replacement seems to be a manually syntactically annotated corpus (treebank). Treebanks to a certain extent contain also extra-syntactic knowledge inserted by human annotators as a side-effect of their endeavor to annotate syntactic relations. Human annotators must decide what is preferred in a given context also in the cases when an isolated sentence would be clearly ambiguous. This decision implicitly takes into account all important factors and thus the resulting annotation also contains more than just pure syntactic information.[2]

Prague Dependency Treebank

In our experiments we are exploiting the data from the Prague Dependency Treebank 3.0 (PDT), see (Bejček et al. 2013). The syntactic structure of sentences from the treebank – captured by dependency trees (always just one tree per sentence) – provides all information necessary for a successful application of AR.

PDT contains very detailed annotation of almost 49,500 Czech sentences. The annotation has been performed at multiple layers, out of which the most important one for our purpose is the analytical (surface syntactic) layer. It describes the (surface) syntactic structure by means of so-called analytical functions. Our experiments exploit only the training part of the treebank which contains 38,727 sentences. The remaining sentences are left aside for future testing and evaluation.

The syntactic structure of each sentence from PDT has the form of a rooted tree. The relationship between the governing and dependent word are expressed as edges between the two nodes, where the node representing the governing word is the parent and the node representing the dependent word is a child. The edge then corresponds to a syntactic relationship of dependency between the parent and its child.

As coordination and apposition represent relations which cannot be intuitively captured by dependency trees, their representation typically varies in various dependency treebanks (Štěpánek 2006). In PDT-like dependency trees, these constructions are represented as so called 'conjoining constructions' (see example (2) and dependency tree in Fig. 2). The subtrees representing coordinated/appended members are rooted in an artificial 'conjoining' node; for purely technical reasons, it is labeled by a lemma of a coordinating/appositional conjunction (conjunction *a* (and) in ex. (2)). The coordinated members (marked by `_Co` suffix) are connected to this node by non-dependency edges (in (2), see the edges *tajemníkem – a* (secretary – and) and *následníkem – a* (successor – and)); common modifiers are connected to the conjoining node as well. The whole conjoining construction is then connected to the governing node of this construction by another non-dependency edge (in (2), see the edge *a – býval* (and – used to be)).

Example (2):
Čao býval generálním tajemníkem strany a oficiálním Tengovým následníkem.
(Čao used to be the secretary general of the party and official Teng's successor.)

Selection Query

For obtaining a suitable set of test sentences for AR as well as for searching the data, we use a PML-TQ search tool, which has been primarily designed for the processing

[2]See also (Mareček and Žabokrtský 2012) who use treebank data for determining possible reductions for syntactic parsing.

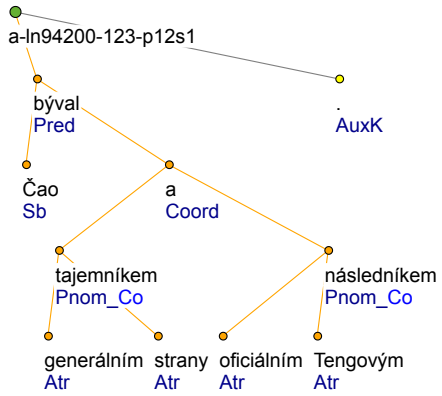

Figure 2: The dependency tree of the sentence (2), taken from PDT (in the sequel, we do not consider technical roots of the trees, containing IDs of sentences, and the final punctuation).

of PDT data. PML-TQ is a query language and search engine designed for querying annotated linguistic data (Pajas and Štěpánek 2009), based on the TrEd toolkit (Pajas and Štěpánek 2008). TrEd with the PML-TQ extension allows users to formulate complex queries on richly annotated linguistic data.

Using this tool, we extract a subset of the treebank containing sentences with the following properties:

- sentence length is between 10 and 30 words (l. 4);
- the sentence must contain at least one coordination or apposition (l. 9-10);
- the sentence does not contain numerals (l. 15-16) or parentheses (l. 11-12);
- all finite verbs are predicates (i.e., no subordinated clauses, l. 13-14);
- there is no construction *čím – tím* (the – the) (l. 5-8).

The PML-TQ query conforming to the above mentioned requirements is presented below.

```
1 t-root
2 [ file() ~ "train-[*]",
3   atree.rf a-root $r :=
4   [ file() ~ "train-[*]", [descendants()≥ 10,
descendants()≤ 30,
5    (0x descendant a-node
6     [ m/form ~ "^[Čč]ím$" ] or
7    0x descendant a-node
8     [ m/form ~ "^[Tt]ím$" ])
9    1+x descendant a-node
10     [afun in {"Coord", "Apos"}],
11   0x descendant a-node
12     [is_parenthesis_root="1"],
13   0x descendant a-node
14     [m/tag ~ "^V[Bipqt]"], afun != "Pred" ]
15   0x descendant a-node
16     [m/tag ~ "^C" ] ] ] ;
```

This PML-TQ query extracted 4,357 sentences from the training data of PDT which we used for further analysis.

AR Rules for Coordination and Apposition

Let us now summarize the rules of AR we are using in the analysis of constructions containing coordination or apposition relationships.

The rules for an automatic AR enriched by coordination and/or apposition:

0. The phenomena not participating in relations of coordination or apposition are being reduced in accordance with the original rules described in (Kuboň, Lopatková, and Mírovský 2013).

1. Common modifiers of all coordinated words are being reduced.

2. All coordinated (or appended) members are being reduced in a single step of the analysis. Unfortunately, it is not possible to limit the number of coordinated members. Although the coordinated constructions in PDT usually contain as few as 2-5 members, there is one case of coordination of as many as 57 members (TV program syntactically structured into a sentence); the richest apposition contains 15 members.

3. The conjoining expression (e.g., a coordinating conjunction or a punctuation mark) is being reduced together with the coordinated/appended expressions.

4. Also all emphasizing syntactic particles and all auxiliary words, punctuation marks, graphical symbols etc. are being reduced at the same time as their governing node (according to the rules used in previous experiments, see (Kuboň, Lopatková, and Mírovský 2013)).

5. Coordinations and appositions allow embedding (again, also in this case there is a theoretically unbounded number of levels of embedding, see (Oliva 2011); we have found 6 levels of embedding of coordinations and appositions (an overview of sport results) in the PDT data).

Similarly as in the previous experiments, we concentrate on automatic analysis of projective constructions; the non-projective ones are being analyzed manually.

Let us illustrate these principles with the following Czech sentence with coordination (2).

Example (3):

Děti česaly zralé meruňky, modré blumy a také zelená jablka.

(Children harvested ripe apricots, blue plums and also green apples.)

Step 0: All words not participating in the coordination are being reduced – this will provide a 'skeleton of coordination structure' with a common modifier:

→* *česaly zralé meruňky, blumy a také jablka.*

Step 1: The common modifier *zralé* (ripe) of the coordinated members is reduced:

→ *česaly meruňky, blumy a také jablka.*

Step 2: All coordinated members – *meruňky* (apricots), *blumy* (plums), *jablka* (apples) – are being reduced.

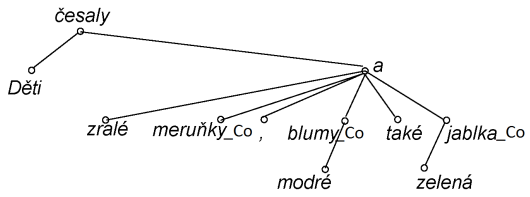

Figure 3: PDT-like dependency tree of sentence (3); the adjective *zralé* (ripe) is analyzed as a common modifier for *meruňky* (apricots), *blumy* (plums) and *jablka* (apples) – the _Co suffix distinguishes coordinated members from their common modifier.

Step 3: The coordinating node (coordinating conjunction *a* (and)) is being reduced.

Step 4: The emphasizing words and punctuation are reduced at the same time, i.e., the emphasizing *také* (also), as well as the punctuation mark (comma)
→ *česaly* ((they) harvested).

The manual analysis of automatic reductions obtained from PDT led to a further refinement of rules. It concerned especially the following phenomena:

Additional rules for the automatic AR extended towards the analysis of coordinations and appositions:

6. Expressions that function as a reference to previous context are processed (such expressions have typically a tag of a coordinating conjunction in PDT so they were excluded from the previous processing). For example: *Nemáme proto potíže se získáváním trhu pro své výrobní odpady.* (We thus do not have difficulties to acquire a market for our productional waste) →* *Nemáme proto potíže* (We thus do not have difficulties.)
(The reduction cannot proceed further due to a non-projectivity – in PDT, *proto* (thus) is marked as a coordinating conjunction and syntactically analyzed as the governing node of the verb *nemáme* ((we) do not have), which leads to a non-projective edge *nemáme – potíže*.)

7. Multiword conjunctions and syntactic particles referring to a whole coordinated clause (these are prototypically annotated in PDT as children of coordinating nodes) are being reduced in one step together with the coordinating expression. Example: *Jsou buď nedostupná, nebo nedostačující.* (They are either inaccessible, or insufficient.) → *Jsou* ((They) are) (The reduction is performed in one step: the governing node is, according to the rules of PDT, the coordinating conjunction *nebo* (or), the second part of the coordinating expression *buď* (either) must be reduced at the same time (according to this rule); coordinated members *nedostupná* (inaccessible) and *nedostačující* (insufficient) are reduced according to step 1, the punctuation comma is reduced according to step 3.)

We have also added the following rules which do not specifically concern coordinations:

8. More adequate reduction of constructions containing modal verbs (e.g., *měly tvořit*) (they were supposed to create) and verbonominal predicate (e.g., *Je učitelem* (he is a teacher)) are proposed: the inner structure of these constructions is simplified as one of the last steps of the AR.

9. Emotional and rhytmizing particles *mi*, *vám*, *si*, *to*, *ono* (to me, to you, Refl., it) etc. are being reduced anytime during the reduction, even if they serve as clitics.

Results and Findings

As we have already mentioned above, the PML-TQ query provided 4,357 sentences. The average length of the extracted sentences is 16.85 tokens. The automatic procedure does not always lead to the desired result of a complete reduction of the sentence but the average number of remaining tokens is as low as 3.69. On the one hand, almost 2 thousand sentences were completely reduced – either to a single last token (1,498) or to a skeletal coordination/apposition structure (494 cases). On the other hand, almost a third of sentences (1,403) can be reduced only to 5 or more tokens (but only a negligible number of 14 sentences result in more than 15 tokens). The complete results of the automatic procedure are presented in Fig. 4. (Due to huge differences between maximal and minimal value it was necessary to use logarithmic scale for the y-axis.)

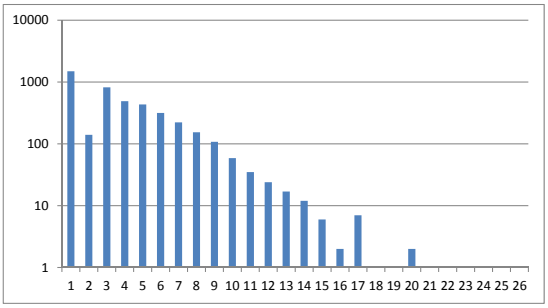

Figure 4: The table describing the numbers of tokens left after the reduction (logarithmic scale for the y-axis).

Fig. 5 contains the numbers of tokens reduced in a single step of AR. It is not surprising that the automatic procedure reduced only 1 token in a majority of steps (31,355 out of 40.521 cases). On the other hand, it seems that there is no natural upper boundary on the number of tokens which can theoretically be reduced in one step of AR because even in our limited sample we have identified sentences with more than 15 tokens reduced in one step (with the maximum of 19 tokens).

Sentences reduced to a single node

The 1,498 sentences which were completely reduced to only 1 node represent mostly sentences without any complicated phenomena and with a finite verb as a predicate. Other types of sentences include:

- non-verbal sentences (e.g., a heading or title) are reduced to a single noun (19 cases);

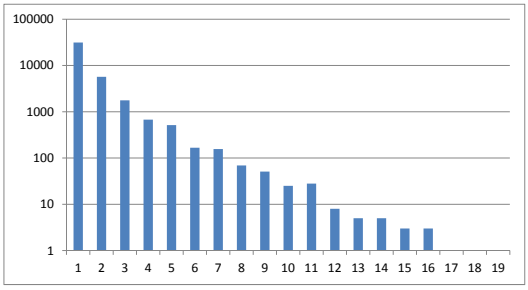

Figure 5: The table describing the number of tokens reduced in one step (logarithmic scale for the y-axis).

- colon substituting the verb (4 cases), as e.g. *Specializované výstavy: Cestovní ruch, sport, volný čas.* (Specialized exhibitions: tourism, sport, free time.);
- sentences containing the words *lze/nelze* (it is (not) possible) which have a peculiar status in Czech: standard linguistic classification marks them as adverbs but they behave syntactically more like verbs (although they have only 1 form and cannot be conjugated), therefore PDT annotates them as verbs.

Sentences reduced to two nodes

The table in Fig. 4 shows that the sentences reduced into two nodes constitute an exception – their low number (140) does not conform to the continually decreasing curve clearly visible in the table. The examples found in the corpus show that these sentences mostly contain complex verbal forms consisting of two tokens which cannot be reduced further:

- a finite verb and a clitic (49 cases including the reflexive particles *se/si)*;
- a false coordinating conjunction which functions more as a link to the previous text (59 cases); this category includes also sentences starting with a coordinating conjunction as, e.g., *A to představuje ...* (And this means ...); the same type of phenomenon also concerns subordinating conjunctions, but it is much less frequent (only 3 cases).
- a pair of a modal verb and lexical verb and an analytical form of a verb *mohli bychom/nebylo by* (we could/it would not) (11 cases) is not further reduced;
- the remaining cases to the total of 140 sentences are rare phenomena (or errors in annotation).

Sentences reduced to three nodes

This is the first group of sentences in this section which contain non-projective constructions. Due to the fact that our automatic procedure applies reductions in a projective way only, the non-projective constructions constitute an ultimate obstacle on the way towards a complete reduction of the sentence. Among the 851 sentences reduced to 3 nodes, there were 124 non-projective constructions. It is not surprising that none of them contained coordination because all three remaining nodes were involved in the non-projective construction. Let us now describe the issues encountered in the remaining sentences:

- By far the most numerous are constructions involving coordinations or appositions (494 cases). These represent the cases of coordinated clauses where an input sentence is reduced to a conjunction and two coordinated predicates (or heads of non-verbal clauses). These cases – as an analogy to a single-node reductions to a Predicate for sentences without coordination – must be viewed as successful reductions.
- The second largest group contains sentences with clitics (219 cases), sometimes combined with other phenomena, as, e.g., the false coordinations (17 cases).
- All remaining types of phenomena involved are rare (or – in several cases – incorrectly annotated).

Complicated sentences reduced to 9 and more nodes

These sentences cannot be neglected because they constitute as many as 111 cases (2.55% of the total). Several interesting cases have been identified:

- A sentence starting with a long coordination which is followed by a clitic – the coordination cannot be reduced because clitics cannot occupy the sentence first position in Czech, the reduction would result in an ungrammatical sentence and thus the correctness preserving principle of AR would be violated.
- A non-verbal sentence containing a long list (shopping list, football line-up etc.).
- A complex sentence with two clauses, each of which contains a clitic.

Reduction of a large number of nodes in one step

Apart from the total number of nodes remaining after AR it is also interesting to look at constructions which enable a drastic reduction of a sentence in one step. Let us present an example:

Example (4)

Stále je nejjistějším dokladem a zprávou o životě , o lásce , o smrti , o víře a touze , o metafyzickém strachu , o složitosti lidské existence .

still - is - the most certain - evidence - and - message - about - life - , - about - love - , - about - death - , - about - faith - and - desire - , - about - metaphysical - fear - , - about - complexity - (of) human - existence

(It is still the most certain evidence and message about life, about love, about death, about faith and desire, about metaphysical fear, about the complexity of human existence.)

After reducing the words *Stále* (still), *nejjistějším* (the most certain), *metafyzickém* (metaphysical), *lidské* (human) and *existence* (existence) in five reduction steps, AR continues in a single reduction step during which 19 nodes are deleted:

– deleting coordinating conjunctions ‘,’ and *a* (and);
– deleting prepositions *o* (about), 6 times;
– deleting coordinated members *životě* (life), *lásce* (love), *smrti* (death), *víře* (faith), *touze* (desire), *strachu* (fear), *složitosti* (complexity);

– deleting punctuation marks ‘,’ (6 times).

After the deletions, only the reduced sentence *je dokladem a zprávou.* ([It] is an evidence and a message.) remains. Now, the coordination *dokladem a zprávou* (an evidence and a message) and the final punctuation are reduced – as a result, only the verb *je* ([It] is) remains.

Conclusions

In our previous papers we have used the method of AR to capture relationships between dependency and word order. Here we have naturally extended this method in order to capture also the third basic syntactic concept, coordination.

The main goal of our experiments with the automatic procedure of analysis by reduction consisted in testing the applicability of this method to all three basic syntactic phenomena – dependency, word order, and coordination. We have also identified certain problematic constructions which may serve as a material both for further linguistic research as well as for further modification or refinement of the method of AR in the future. The observations and findings we have made during our experiments may also help to improve the formal description of natural languages.

Our experiments have demonstrated that even complicated linguistic constructions of coordination and apposition may be to a large extent reduced automatically. We have also identified several problematic constructions, out of which especially the connection between coordination and the location of clitics constitute issues which have not been sufficiently linguistically described.

Acknowledgments

The research reported in this paper has been supported by tbe GAČR grant No. GA P202/10/1333. This work has been using language resources stored and/or distributed by the LINDAT-Clarin project of MŠMT (project LM2010013).

References

Bejček, E.; Hajičová, E.; Hajič, J.; Jínová, P.; Kettnerová, V.; Kolářová, V.; Mikulová, M.; Mírovský, J.; Nedoluzhko, A.; Panevová, J.; Poláková, L.; Ševčíková, M.; Štěpánek, J.; and Zikánová, Š. 2013. *Prague Dependency Treebank 3.0*. Prague: Charles University in Prague, MFF, ÚFAL.

Kuboň, V.; Lopatková, M.; and Mírovský, J. 2013. Automatic Processing of Linguistic Data as a Feedback for Linguistic Theory. In Castro, F.; Gelbukh, A.; and González, M., eds., *Proceedings of the 12th Mexican International Conference on Artificial Intelligence (MICAI 2013)*, volume 8265 of *LNCS*, 252–264. Berlin Heidelberg: Springer-Verlag. volume 1.

Lopatková, M.; Plátek, M.; and Kuboň, V. 2005. Modeling Syntax of Free Word-Order Languages: Dependency Analysis by Reduction. In *Proceedings of TSD 2005*, volume 3658 of *LNAI*, 140–147. Berlin Heidelberg: Springer-Verlag.

Lopatková, M.; Plátek, M.; and Sgall, P. 2007. Towards a Formal Model for Functional Generative Description: Analysis by Reduction and Restarting Automata. *The Prague Bulletin of Mathematical Linguistics* 87:7–26.

Mareček, D., and Žabokrtský, Z. 2012. Exploiting reducibility in unsupervised dependency parsing. In *Proceedings of EMNL-CoNLL 2012*, 297307. ACL.

Mel’čuk, I. A. 1988. *Dependency Syntax: Theory and Practice*. Albany: State University of New York Press.

Oliva, K. 2011. Linguistics behind the Mirror. In Lopatková, M., ed., *Information Technologies – Applications and Theory*, 1–6. Košice, Slovakia: Univerzita Pavla Jozefa Šafárika v Košiciach.

Pajas, P., and Štěpánek, J. 2008. Recent Advances in a Feature-Rich Framework for Treebank Annotation. In *Proceedings of CoLING 2008*, volume 2, 673–680. Manchester, UK: The Coling 2008 Organizing Committee.

Pajas, P., and Štěpánek, J. 2009. System for Querying Syntactically Annotated Corpora. In *Proceedings of the ACL-IJCNLP 2009 Software Demonstrations*, 33–36. Singapore: ACL.

Plátek, M.; Mráz, F.; and Lopatková, M. 2010. (In)Dependencies in Functional Generative Description by Restarting Automata. In *Proceedings of NCMA 2010*, volume 263 of *books@ocg.at*, 155–170. Wien, Austria: Österreichische Computer Gesellschaft.

Sgall, P.; Hajičová, E.; and Panevová, J. 1986. *The Meaning of the Sentence in Its Semantic and Pragmatic Aspects*. Dordrecht: Reidel.

Štěpánek, J. 2006. *Závislostní zachycení větné struktury v anotovaném syntaktickém korpusu (nástroje pro zajištění konzistence dat)*. Ph.D. Dissertation, MFF UK, Prague.

Tesnière, L. 1959. *Eléments de syntaxe structurale*. Paris: Librairie C. Klincksieck.