From Sentence to Discourse: Building an Annotation Scheme for Discourse Based on Prague Dependency Treebank

Lucie Mladová, Šárka Zikánová and Eva Hajičová
Charles University in Prague
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
Malostranské nám. 25, 118 00 Prague 1, Czech Republic
{mladova, zikanova, hajicova}@ufal.mff.cuni.cz

Abstract

The present paper reports on a preparatory research for building a language corpus annotation scenario capturing the discourse relations in Czech. We primarily focus on the description of the syntactically motivated relations in discourse, basing our findings on the theoretical background of the Prague Dependency Treebank 2.0 and the Penn Discourse Treebank 2. Our aim is to revisit the present-day syntactico-semantic (tectogrammatical) annotation in the Prague Dependency Treebank, extend it for the purposes of a sentence-boundary-crossing representation and eventually to design a new, discourse level of annotation. In this paper, we propose a feasible process of such a transfer, comparing the possibilities the Praguian dependency-based approach offers with the Penn discourse annotation based primarily on the analysis and classification of discourse connectives.

1 Introduction

Annotation of discourse has become one of the burning issues of corpus annotation though there are only partial proposals available in linguistic literature. The most advanced and systematic work in the field of discourse corpus annotation has been carried out for English by the Penn University team of A. Joshi, B. Webber and others (see e.g. Miltsakaki et al., 2004; Joshi et al., 2006). The present contribution summarizes some preliminary results of an ongoing research in the area of discourse, based on the work of the Penn University team and on the Praguian dependency syntax. We propose to build a discourse annotation scheme for Czech and English on the basis of a consistent annotation scheme assigning sentences their underlying (tectogrammatical) structure in the form of dependency trees.

In the first part of the present paper, two linguistic resources of our research are introduced, i.e. the Penn Discourse Treebank (PDTB) and the Prague Dependency Treebank (PDT), with a special regard to how these annotation schemes can interact in finding the way from the sentence with its syntax and semantics to discourse annotation. In the second part, we discuss some specific linguistic issues we face when building the new discourse corpus, such as the set of discourse relations used for annotations or language-specific features showing up on discourse level of linguistic description.

2 Prague Dependency Treebank 2.0

The Prague Dependency Treebank (see Hajič et al., 2006) is conceived of as a multilayer annotation scheme of Czech journalistic texts (approx. 2 million word units) taken from the Czech National Corpus. The three annotation layers of the PDT 2.0 contain (i) full morphological annotation on the morphological layer (m-layer, the lowest level of description), (ii) superficial (surface) syntactic annotation on the analytical layer (a-layer, intermediate level of description), and (iii) deep or underlying syntactic annotation capturing the linguistic meaning on the tectogrammatical layer (t-layer, the highest level of description).

On the tectogrammatical layer, each sentence is represented by one dependency tree structure with the dependents concentrated around the predicate verb (see Fig. 1). A sentence, in the PDT view, may consist of one or more clauses. In principal, the annotation does not surpass the sentence boundaries, though sentences are annotated in their context rather than in isolation. This is reflected, first of all, in two respects:

(i) One of the attributes of the nodes in the tectogrammatical structures concerns the information structure of the sentence (Topic-Focus Articulation, TFA); each of the nodes of the dependency tree is assigned one of the TFA values 'non-contrastive contextually bound', 'contrastive contextually bound' and 'contextually non-bound'. On the basis of these values, the global bipartition of the sentence into its Topic (what the sentence is about) and Focus (what the sentence says about its Topic) is possible. A procedure has been proposed how to follow the 'activation' of the individual items throughout a text (document), which is assumed to help to resolve the assignment of pronominal reference. (Hajičová, 1993)

(ii) In addition to the tectogrammatical structure of the sentences, some basic coreference relations are being marked, especially those of grammatical coreference (in case of control, reflexive and relative pronouns) and some types of textual coreference; the latter annotation goes

already beyond the boundaries of the sentence (Nedoluzhko, 2007).

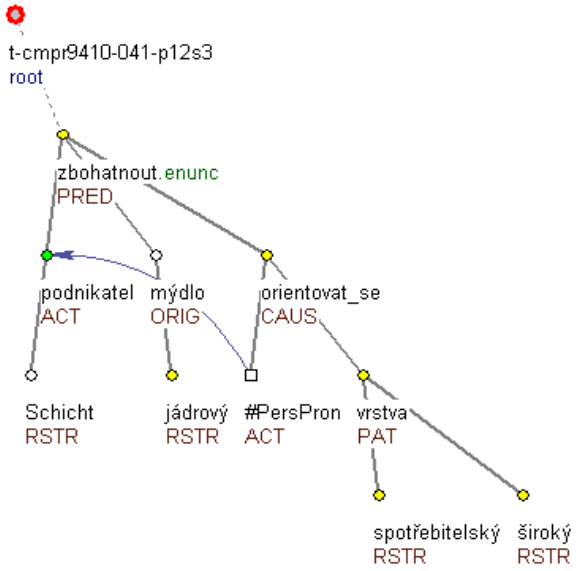

Figure 1: An example of a tectogrammatical tree (a single-sentence representation) with a coreference arrow, for the Czech sentence: *Podnikatel Schicht zbohatl na jádrovém mýdle, protože se orientoval na nejširší spotřebitelskou vrstvu. [The entrepreneur Schicht got rich on grain soap because he concentrated on the widest consumer rank.]*

3 The Idea of a Discourse Treebank

The annotation of discourse relations[1], as proposed in Section 6 of this paper, is meant to be an essential part of the future Prague Dependency Treebank 3.0 project. The PDT 3.0 will contain a new, fourth layer of annotation, which, unlike the PDT 2.0 annotation layers, will capture various types of relations going beyond the sentence boundary. Our research for PDT 3.0 concerns primarily those types of discourse relations which haven´t been marked yet explicitly as such: the syntactically motivated, i.e. 'connective' relations in discourse. These relations include coordinating relations and some of the subordinating relations within a sentence and, secondly, adjoining of discourse units across the sentence boundary.

4 The Penn Discourse Treebank as a Background for Praguian Discourse Annotation

Apart from the tectogrammatics in the Prague Dependency Treebank 2.0, the work on the future discourse layer of the PDT corpus is also widely inspired by the theoretical background of the Philadelphia Penn Discourse Treebank project (see footnote 1 in Lee et al., 2006) and formed on the basis of a critical comparison of the two linguistic approaches originated in Prague and Philadelphia. The Penn Discourse Treebank is a corpus of English texts from the Wall Street Journal (approx. 49 000 sentences) annotated for discourse relations. The annotation is based on the lexicalized grammar theory (Webber, 2004), its main point of interest being structuring of a text by lexical items – discourse connectives. In the annotation scheme, each connective is treated as a discourse-level predicate that takes two text spans (abstract objects) as its arguments. The discourse relations are annotated in the plain form of a text, which allows the annotation scheme to be independent from any syntactic theory and which is comfortable for annotators as well. The Penn Discourse Treebank is connected with the syntactic annotations of the same texts in the Penn Treebank. However, the Penn Treebank syntactic annotation does not surpass the sentence boundaries, therefore this connection often only refers to one of two discourse arguments of a discourse connective.

According to Asher (1993), the discourse arguments in the Penn Discourse Treebank are outlined as linguistic realizations of abstract objects, prototypically predications with finite verbs, but also gerunds and nominalizations. In relation to syntax, a discourse argument can be built by the whole sentence or by its part; the arguments of the connectives can be located at a distance from each other and they can be interrupted, too. The discourse relations have been classified into a detailed set of semantic labels ascribed to single discourse connectives in the context (The Penn Discourse Treebank 2.0 Annotation Manual, 2007).

5 From Tectogrammatics to Discourse

From the point of view of the Prague Dependency Treebank concept, the Penn approach to discourse, which is strongly oriented at syntax (having a linguistic realization of an abstract object as the core of the research), is very promising. It allows us to start from the present Praguian annotation of underlying (syntactico-semantic) relations on the tectogrammatical layer which, as a matter of fact, contains already some discourse relations, and to deepen and broaden it in a full and consequent annotation of the text relations. The original Penn set of semantic labels for discourse relations is being modified with respect to the present preliminary description of discourse relations on the tectogrammatical layer (see Section 6).

In contrast to the Penn Discourse Treebank, the Praguian discourse annotation is planned to be more complex. This annotation should not be separate from the annotations of other linguistic phenomena but it should be a part of a new layer catching also some other more or less textual features, such as e.g. coreference and TFA values mentioned above (these will be adopted and extended for the fourth layer). For the different types of the textual relations (connective discourse relations being one type of

[1] The term 'discourse' is used here to refer to either spoken (also in a dialog form) or written usage of the language as a system in the communication process. A text, or a discourse, is interlaced by a net of syntactic, semantic and pragmatic relations that contribute to its integrity and comprehensibility.

them), a new way of representation of a text will be introduced, connecting the representations of single sentences as they are on the tectogrammatical level into a large continuous representation of the whole document. Hence, in the PDT 3.0 annotation scenario, all sentences of one document will be interlinked by some type of "intersententional" relation. Technically, the tree structures representing separate sentences in PDT 2.0 scheme will be conjoined in a form of a megatree (see Fig. 3 at the bottom of this paper). As for annotators, they will have a possibility to work with a plain visual form of the document, yet the megatree-representation will be available for them, too. Like in the case of other layers of the Prague Dependency Treebank, the upper and lower layers will be interconnected – according to this principle, at the discourse layer no information contained in the lower layers should be lost, it will be accesible (though "hidden").

6 The Set of Discourse Relations in the Prague Dependency Treebank

In the Prague Dependency Treebank, some relations relevant for the discourse structure have been annotated already on the tectogrammatical layer within the syntactic relations of coordination, dependency and reference to the preceding context (see Fig. 2).

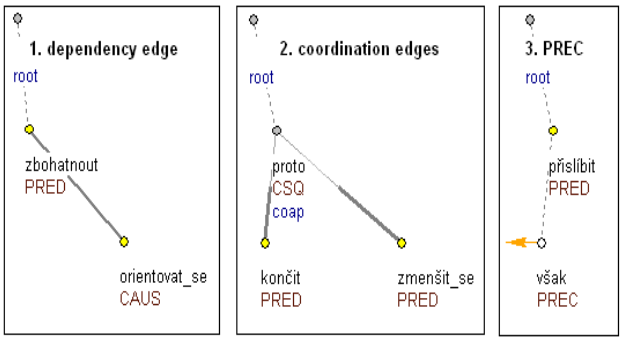

Figure 2: Three types of capturing a possible discourse relation on the tectogrammatical layer: 1. *get rich* (PRED) – *concentrate* (CAUS), 2. *end* (PRED) – *therefore* (CSQ) – *shorten* (PRED), 3. *however* (PREC) – *guarantee* (PRED)

In all of these cases, the discourse character of these relations is not marked explicitly. Thus, there is e.g. no difference in the tectogrammatical annotation between the dependency given by the valency frame of the verb (a non-discourse relation) and the dependency "outside" the valency frame (which prototypically indicates a discourse relation). With coordination, there is so far no special label assigned to abstract objects (a discourse relation) which would be different from the label for the coordination of minor units (like e.g. adjectives; a non-discourse relation). Therefore, just a subset of the annotated relations can be taken over to the discourse annotation.

The dependency and coordination edges of the tectogrammatical layer are classified according to their syntactico-semantic values. Having at our disposal the discourse subset mentioned above from the tectogrammatical layer, this could simplify the discerning of individual semantic labels in the discourse annotation.

Nevertheless, not all of tectogrammatical relations can be transferred directly to the discourse annotation. First, the set of tectogrammatical relations does not correspond with the proposed set of discourse relations (see below). Furthermore, some of the present-day tectogrammatical labels relevant for the discourse relations should be re-classified again in a more detailed way. This applies for coordination (e.g. not every occurrence of the conjunction *but* has an adversative meaning) and especially for the functor PREC (reference to the preceding context). An expression marked with the PREC functor indicates a simple presence of a discourse relation but it doesn't mark the semantic type of the relation, so that the discourse annotation would be at this point underspecified.

Starting from the Penn Discourse Treebank hierarchy of senses of discourse relations (The Penn Discourse Treebank 2.0 Annotation Manual, 2007) and the former set of tectogrammatical functors of Prague Dependency Treebank (Mikulová et al., 2005), we try to set down a new hierarchy of discourse sense labels. From our point of view, the original Penn hierarchy could be improved in some details (e.g. by introducing new labels, such as 'purpose' or 'gradation', or by restructuring of the hierarchy – cf. the original position of 'concession' within the 'comparison' group, whereas it rather belongs to the same group as 'condition', i.e. to 'contingency'). On the other hand, the Penn hierarchy substantially enriches the spectrum of discourse relations discerned on the tectogrammatical layer in the Prague Dependency Treebank now, e.g. by meanings such as 'instantiation', 'restatement', 'list' etc. (see Zikánová, 2007).

Some of the Penn sense labels can be introduced directly to the Praguian discourse layer, being deduced from the corresponding tectogrammatical functors, cf. (1) and (2):

[Unit 1] ***discourse connective (DC)*** *[Unit 2]*

(1)
*[Jakou povahu jsi měl], **než** [jsi přišel o práci]?*
*[What had you been like] **before** [you lost your job]?*
DC = before
PDTB: temporal – asynchronous – precedence
PDT: functor TWHEN, subfunctor BEFORE

(2)
*[Buď půjdeme do kina], **nebo** [zůstaneme doma].*
*[Either we'll go to the cinema], **or** [we'll stay at home].*
DC = or (disjunctive meaning)
PDTB: expansion – alternative – disjunctive
PDT: functor DISJ

Other tectogrammatical functors need to be re-classified, cf. PREC (reference to the preceding context) mentioned above:

(3)
[…]. ***A*** *[potom odešel]*.
[…]. ***And*** *[then he left]*.
DC = and
PDTB: expansion – conjunction
PDT: functor PREC (no discourse semantics marked)

7 The Functor PREC and Discourse Connectives

Having proceeded from the surface shape of the sentence through the underlying structure to discourse relations, we analyzed also, like the Penn Discourse Treebank does, separate lexical items in Czech that can be of use for the description of discourse. The disadvantage of such a lexical approach is that the means of expressing connective relations, i.e. discourse connectives, are not obligatorily expressed in the sentence. On the other hand, the advantage is that once they appear explicitly in a sentence, they express the semantics of the connection between the conjoined units quite clearly, they are in fact the most significant indicators of discourse relations. This function is basically covered by conjunctions, some subjunctions, particles and adverbs, and marginally also by some other parts-of-speech. In the Prague Dependency Treebank, these lexical units are semantically subclassified in so far, as they connect or adjoin elements within one sentence (one tectogrammatical tree). Should they connect larger units (or, in other words, refer to a larger context), the underspecified functor PREC is assigned to them (ex.: Hence PREC, *I am happy.* However PREC, *isolated research cannot have good results*.). Still, regarding the size and extent of the discourse units, the functor PREC gives us one more piece of information: it applies primarily to units across the sentence boundary or even bigger text spans such as paragraphs (Mladová, 2008). Planning to annotate also such more complex discourse relations in the future, we have to accept the fact that the structure of the discourse, as shown on the example of the PREC functor, is hierarchical. In Figure 3, three levels of hierarchy in discourse are visible, indicated by the functors PREC, CSQ (consequence) and CAUS (cause).

In Section 6, we have compared the set of Penn Discourse Treebank sense labels with the Prague Dependency Treebank set of functors. Furthermore, remaining on the level of lexical unit description, we can also compare the two sets of discourse connectives together with the sense labels assigned to them on each side. This is made possible by the existence of the Prague Czech-English Dependency Treebank (PCEDT), a collection of parallel texts in Czech and English annotated by the PDT tectogrammatical layer scenario (`http://ufal.mff.cuni.cz/pcedt/doc/PCEDT_main.html`). The texts in this corpus are the same Wall Street Journal texts as in the Penn (Discourse) Treebank, for the Czech subcorpus translated into Czech.

It has been observed that the delimitation of the group of discourse connectives is wider in the Penn Discourse Treebank. Some language expressions, in the first place some adverbs and particles like *in fact* or *indeed* are in PCEDT never marked with any of the possible discourse sense labels, they have rather a modal or a pragmatic characteristic.

Considering the fact that the Functional Generative Description focuses on the underlying syntactic structure rather than on surface shapes and always prefers transparent ways of capturing the linguistic information to the more complicated ones, for the graphical output we discuss the possibility to "hide" the connectives themselves while preserving their sense annotation on the discourse layer only. In spite of this, it will remain possible to view the discourse connectives and their attributes by the means of the links to the lower annotation layers.

8 Open Questions

By forming a new annotation level for discourse, many questions emerge that are not yet satisfactorily resolved in the Prague Dependency Treebank annotation guidelines or that are resolved in a way which seems inappropriate for discourse description. We mention some of these points in this section.

Parcelling. (*John lost his sock. The blue one.*) Dependency relations appear primarily within a single sentence, the governing node usually is not connected with its dependent node across the sentence boundary. If yet so, the dependent subtree is treated as parcelled: in most cases, it cannot stand independently from its governor. This phenomenon, nevertheless, occurs more than sporadically in dialogs and question answering. Whereas the present-day annotation guidelines instruct to mark an ellipsis of the governing predicate in such cases, and so almost every sentence of a dialog becomes elliptical, the future discourse annotation should be able to handle the spoken language with its particularities in a simple and unified way. Therefore, we focus our research besides the treebank texts also on a special set of recorded dialogs. Still, a lot of work is to be done in this respect.

Verbless clauses. According to the PDT annotation scheme, there are three types of clauses, that do not have a predicate verb as their governing node: subject-case clauses (*An important event.*; *You and your statistics!*), vocative clauses (*George!*) and interjectional clauses (*Oops!*). Treating the discourse units as based on the predicate verb or its modifications (see Section 4), it is still to be decided about the role of the verbless clauses in our notion of the discourse.

Parenthesis. (*The court, as it seems to me, has no opinion on the subject.*) Although there is a detailed study on types of parenthesis and their character in the tectogrammatical annotation manual (Mikulová et al., 2005), the problem of their partial syntactic independence is to be solved for the discourse annotation.

Nominalizations. The problem of nominalizations and other modifications of the verbal proposition has been already solved by the Penn team (Penn Discourse Treebank 1.0 Annotation Manual, 2006, p. 10–13). In Czech, however, there is a rich repertoire of deverbative affixes and other word-forming devices, so the question of the discourse unit delimitation is more complicated. As for the assignment of valency frames to deverbative nouns in tectogrammatics, we have worked with two Czech suffixes only: *-ní, -tí* which can (but do not have to) express the meaning of the verb almost without additional semantic features: *běžet – běhání (to run – the running)*. It is to be reconsidered, where to put a reasonable border to delimit a discourse unit for Czech in order not to overburden the discourse annotation.

In principle, it can be assumed that the connective discourse relations are language universal. Nonetheless, as signalled also in the previous paragraph, we expect the annotations to prove the existence of some language-specific phenomena on a very concrete level, e.g. in the repertoire and function of single discourse connectives in Czech and in English.

9 Conclusion

In the previous sections we have discussed the possibilities of building an annotation scenario for discourse on the basis of the Praguian underlying (tectogrammatical) syntax formalism. We have argued that the current version of the Prague Dependency Treebank already captures some types of discourse relations within the syntactic relations of dependency, coordination and reference to preceding context (PREC). These relations and their sense labels (functors) can be transferred to the discourse level of annotation. However, for the time being, the labels are assigned to relations within separate sentences (tectogrammatical tree structures) only. The functor PREC, which indicates a discourse relation going over the sentence boundary, needs to be subclassified. Further, the comparison with the scenario of the Penn Discourse Treebank has shown the pros and cons of the dependency-based approach. We hope we will be able to take the advantage of the Prague Dependency Treebank tree structure properties, such as projecting linguistic information across the annotation layers, in the future megatree-representation of discourse. Such a representation, for Czech as well as for English, will allow the researchers to carry out experiments going across linguistic domains from morphology to discourse and across languages.

10 Acknowledgments

The research reported in this contribution has been carried out under the grant projects of the Center for Computational Linguistics – CKL (LC536), EU Companions (207-55/6694), GA 405/06/0589 of the Grant Agency of the Czech Republic, and the Ministry of Education, Youth and Sports Czech Republic (MSM-0021620838) "Modern methods, systems and structures of informatics".

11 References

Asher, N. (1993). Reference to Abstract Objects in Discourse. Kluwer Academic Publishers, Dordrecht.

Hajič, J. et al. (2006). Prague Dependency Treebank 2.0. Linguistic Data Consortium, Philadelphia.

Hajičová, E. (1993). Issues of Sentence Structure and Discourse Patterns. Charles University, Prague.

Hajičová, E. et al. (2006). An Annotated Corpus as a Test Bed for Discourse Structure Analysis. In Proceedings of the Workshop on Constraints in Discourse, National University of Ireland, Maynooth, Ireland, pp. 82–89.

Joshi, A. et al. (2006). Discourse Annotation: Discourse Connectives and Discourse Relations. Tutorial at the Association for Computational Linguistics, Sydney.

Lee, A. et al. (2006). Complexity of Dependencies in Discourse: Are Dependencies in Discourse More Complex Than in Syntax? Proceedings of the 5th International Workshop on Treebanks and Linguistic Theories. Prague, Czech Republic

Mikulová, M. et al. (2005). Annotation on the Tectogrammatical Layer in the Prague Dependency Treebank: Annotation Manual. Universitas Carolina Pragensis, Prague.

Miltsakaki, E. et al. (2004). The Penn Discourse Treebank. In Proceedings of the Fourth International Conference on Language Resources and Evaluation. LREC 2004, Lisbon, Portugal.

Mladová, L. (2008). Diskurzní vztahy v češtině a jejich zachycení v anotovaném korpusu. [Discourse Relations in Czech and their Representation in an Annotated Corpus of Texts.] Diploma thesis. Charles University, Prague.

Nedoluzhko, A. (2007). Zpráva k anotování rozšířené textové koreference a bridging vztahů v Pražském závislostním korpusu. [Report about the Annotation of the Extended Text-Coreference and Bridging Relations in the Prague Dependency Treebank.] Technical report. Institute of Formal and Applied Linguistics, Charles University, Prague.

Sgall, P. et al. (1969). A Functional Approach to Syntax in Generative Description of Language. American Elsevier, New York.

The Penn Discourse Treebank 1.0 Annotation Manual. (2006). `http://www.seas.upenn.edu/~pdtb/papers/pdtb-1.0-annotation-manual.pdf`

The Penn Discourse Treebank 2.0 Annotation Manual. (2007). `http://www.seas.upenn.edu/~pdtb/PDTBAPI/pdtb-annotation-manual.pdf`

The Prague Czech-English Dependency Treebank 1.0 http://ufal.mff.cuni.cz/pcedt/doc/PCEDT_main.html

Webber, B. (2004). D-LTAG: Extending Lexicalized TAG to Discourse. Cognitive Science, 28 (5), pp. 751-779.

Zikánová, Š. (2007). Possibilities of Discourse Annotation in Prague Dependency Treebank (Based on the Penn Discourse Treebank Annotation). Technical report. Institute of Formal and Applied Linguistics, Charles University, Prague.

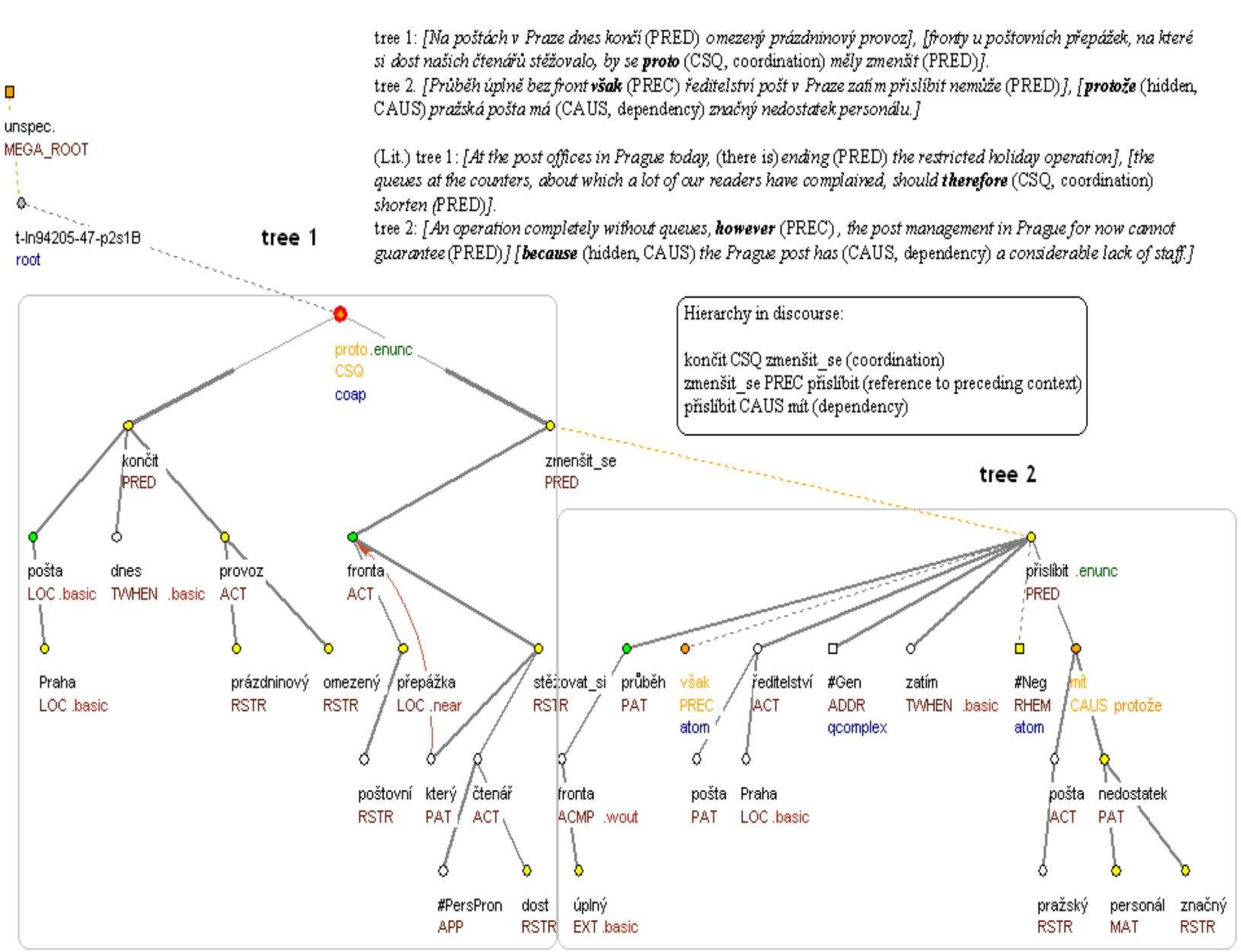

Figure 3: A megatree demonstrating the hierarchy in discourse