Sentence Modality Assignment in the Prague Dependency Treebank

Magda Ševčíková and Jiří Mírovský

Charles University in Prague, Faculty of Mathematics and Physics,
Institute of Formal and Applied Linguistics

Abstract. The paper focuses on the annotation of sentence modality in the Prague Dependency Treebank (PDT). Sentence modality (as the contrast between declarative, imperative, interrogative etc. sentences) is expressed by a combination of several means in Czech, from which the category of verbal mood and the final punctuation of the sentence are the most important ones. In PDT 2.0, sentence modality was assigned semi-automatically to the root node of each sentence (tree) and further to the roots of parenthesis and direct speech subtrees. As this approach was too simple to adequately represent the linguistic phenomenon in question, the method for assigning the sentence modality has been revised and elaborated for the forthcoming version of the treebank (PDT 3.0).

Keywords: sentence modality, Prague Dependency Treebank, dependency tree, coordination root, coordinated clause

1 Introduction

Recognition of the contrast between declarative, imperative, interrogative, and possibly other types of sentences, which is referred to as sentence modality in the present paper, is a salient subtask needed for NLP applications in the domain of question answering, machine translation etc.: for instance, without distinguishing assertions vs. questions, it is not possible to choose the right verb form and place the words in the right order during the Czech-to-English translation. The present paper focuses on the assignment of sentence modality in the Prague Dependency Treebank (PDT), which is a richly annotated collection of Czech newspaper texts.

In Section 2, we explain how the term ‘sentence modality’ is used in the paper. Section 3 introduces the original annotation of sentence modality, released as a part of PDT 2.0 [3] in 2006 and yet not described in any published paper. As this annotation proved to be too simple to adequately represent the linguistic phenomenon in question, the sentence modality assignment has been revised and extended for the forthcoming version of PDT (PDT 3.0; Sect. 4).[1]

[1] PDT 3.0 is planned as a treebank consisting of the data of PDT 2.0 with corrections and revisions of several types, and of new data annotated in a comparable way. If we refer to the PDT 3.0 in the present paper, only the corrected and revised PDT 2.0 data are meant (i.e., PDT 2.0 and PDT 3.0 consist of the same texts and have the same size).

particle/wh-word	verbal mood	final punct. mark	sentence modality
Ø	indicative/conditional	. / Ø	declarative
Ø	indicative/conditional	?	interrogative (polar (yes/no) question)
wh-word	indicative/conditional	?	interrogative (non-polar (wh-)question)
Ø	imperative	! / .	imperative
Ø/ *at’, kéž, necht’*	indicative/conditional	! / .	desiderative
Ø	indicative	!	exclamative

Table 1. Means used for expressing sentence modality in Czech written texts

2 Sentence modality as a modal meaning of the sentence

In PDT as well as in the linguistic framework of Functional Generative Description (FGD; [12]), which the PDT annotation scenario is based on, sentence modality is understood as a modal meaning of the sentence; it is the function of the sentence to assert a content, ask a question, require that someone performs something etc.[2] In Czech written texts, these functions are conventionally expressed by combinations of formal means of different types, namely by the mood of the verb form, by the final punctuation mark, by the word order, and by modal particles *at’, kéž, necht’*.[3]

Five types of sentence modality are distinguished in PDT and FGD according to the Czech linguistic tradition (e.g. [13], [2]):
– declarative modality (e.g. *Ekonomika jde do vzestupu už letos.* ‘The economy rises already this year.’),
– interrogative modality (*Jaká je nezaměstnanost v této zemi?* ‘How big is the unemployment in this country?’),
– imperative modality (*Podívej se na mě!* ‘Look at me!’),
– desiderative modality (*At’ si provincie konečně oddychne.* ‘Let the province finally relax.’) or
– exclamative modality (*To nejsou špatně rozdané karty!* ‘The cards have been dealt not at all badly!’).

Although the sentence modality and thus the choice of formal means mirror the speaker’s intention to state something or to learn a piece of information etc. (cf. illocution in the Speech Act Theory by Austin [1] and Searle [11]), neither

[2] The terminology is far from uniform. Portner [8] speaks about sentential force, or simply about clause types or sentence types; all these terms are subsumed under discourse modality whereas the term sentential modality is used for isolated linguistic means operating “at the level of the whole sentence” Zaefferer [14] makes a terminological distinction between sentence mood (close to our usage of sentence modality) and sentential modality (underlying intention of the speaker).

[3] In spoken texts, prosodic features (esp. intonation) are reckoned for the most important means for conveying sentence modality. However, these features are not available in written texts, which we are concerned with.

the classification nor the annotation aim at capturing this intention since extra-linguistic factors (politeness conventions etc.) can play a crucial role in how the intention is expressed.[4] The theoretical delimitation of the five modality types as well as our annotation approach are based on linguistic means explicitly coded in the sentence; see Table 1 for the respective combinations.[5]

The relatively transparent relations between the sentence modalities and the formal markers seem to be a solid base for annotating the sentence modality in real language data. However, an essential question, namely which parts of the sentence are to be assigned the sentence modality, must be answered before any annotation starts. The sentence modality is often defined as a modal meaning of the whole sentence (cf. footnote 2), however, there are sentences with a more complicated structure, for which this definition is not satisfactory.

In FGD, the sentence modality is supposed to be a characteristic of the sentence as a whole if it involves just one main (syntactically independent) clause (see ex. (1)), but in a coordination structure, each of the syntactically independent clauses can have a different modality (ex. (2)). Similarly, an embedded but syntactically independent structure, such as direct speech, expresses its 'own' sentence modality, which may differ from the modality of the respective matrix clause (ex. (3)). The annotation of sentence modality in PDT 2.0 did not meet all these requirements (Sect. 3), they are reflected in the more advanced approach introduced in Section 4.

(1) *Neptejte*.imper *se mě, proč jsem přijel do Prahy.* 'Do not ask.imper me why I came to Prague.' (the modality is marked with the head of the respective structure, see Sect. 3.1 for the explanation of the values used)
(2) *Poprvé jste nastoupil*.enunc *v závěru zápasu v Benešově, jaké to bylo*.inter? 'For the first time you entered.enunc the game before the end of the match in Benešov, what was.inter it like?'
(3) *Kam se poděla*.inter *má bojovnost? ptala*.enunc *se sama sebe po utkání Martinezová.* 'Where did my fighting spirit disappear.inter? Martinezová asked.enunc herself after the match.'

3 Annotation of sentence modality in PDT 2.0

3.1 Sentence modality as a part of the deep-syntactic annotation

PDT 2.0 is a treebank of Czech written texts enriched with a complex annotation of three types (at three layers): the morphological layer (where each token was assigned a lemma and a POS tag), the so-called analytical layer, at which

[4] Cf. the examples of asking *Is there any salt?* or stating *It is cold here*, which both can be meant as a request (to pass over the salt and close the window, respectively).

[5] Some of the means for expressing sentence modality (esp. verbal mood, modal particles) are used, in combination with further ones, to recognize the so-called factuality of events in FactBank [10] or the factuality of conditions within the annotation of discourse relations in the Penn Discourse TreeBank 2.0 [9].

the surface-syntactic structure of the sentence (subject, object etc. relations) is represented as a dependency tree, and the tectogrammatical layer, at which the linguistic meaning of the sentence is represented. Nodes of the tectogrammatical tree represent auto-semantic words whereas functional words (such as prepositions, auxiliaries, subordinating conjunctions) and punctuation marks have no node of their own in the tree.[6] The nodes are labeled with a tectogrammatical lemma, with a functor (dependency relation; e.g. Actor ACT, Patient PAT, Location LOC) and other attributes; see [4]. One of the node attributes is the attribute sentmod, capturing the sentence modality of the respective syntactic structure. For this attribute, five values were defined: enunc for declarative modality (enunciation), inter for interrogative modality,[7] imper for imperative modality, excl for exclamative modality, and desid for desiderative modality.

Annotation at all three layers is available for 3,168 documents, containing altogether 49,442 sentences with 833,357 tokens (word forms and punctuation marks). The statistics reported in this paper have been measured on the training set of these data (2,533 documents, 38,727 sent., 652,544 tokens).[8]

3.2 Semi-automatic assignment of sentence modality

Due to the large amount of data and a limited amount of time, a simplified approach to the sentence modality was carried out in PDT 2.0. The simplification consisted in that only one sentence modality value was determined for the whole syntactic structure; the fact that coordinated clauses can have different sentence modalities was intentionally omitted. Two types of embedded syntactic structures (direct speech and parenthesis) were assigned a separate sentmod value.[9]

As the first step of the sentence modality assignment in the PDT 2.0 data, the set of candidate nodes to be assigned a sentmod value was delimited as follows:
(a) child nodes of the technical root node, i.e. nodes representing the main verb or noun and the root nodes of coordination structures (corresponding to a conjunction or punctuation; 'coordination roots' in the sequel);
(b) root nodes of subtrees representing a direct speech; these nodes were identified on the basis of the node attribute is_dsp_root, which had been assigned before the sentmod annotation was carried out;
(c) root nodes of parenthesis subtrees (labeled with the functor PAR).

[6] There are certain, rather technical exceptions, e.g. coordinating conjunctions used for representation of coordination constructions are present in the tree structure.

[7] The difference between polar (yes/no) questions and non-polar (wh-)questions is not captured by the sentmod value but by the non/presence of the wh-word in the tree.

[8] For searching in the PDT 2.0 data and for data manipulation, we used the Netgraph query language [5] and the PMLTQ extension [7] to the Tree Editor TrEd [6].

[9] It means, a sentence with an embedded direct speech or parenthesis was assigned two sentmod values: one value was specified for the sentence as a whole (and assigned to the child node of the technical root node; see under (a) bellow), one value for the direct speech or parenthesis as a whole (see under (b) and (c)). The inner structure of the direct speech subtrees and parenthesis subtrees was not analyzed here.

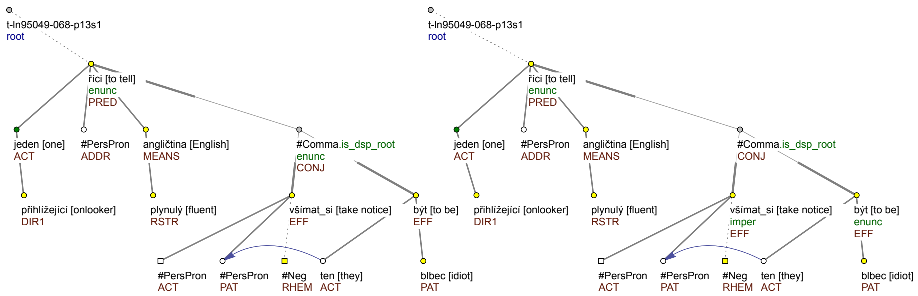

Fig. 1. The tectogrammatical tree for the sentence *"Nevšímejte si jich, jsou to blbci," řekl mi plynulou angličtinou jeden z přihlížejících.* ("Do not take notice of them, they are idiots," told me one of the onlookers in fluent English.), in which two clauses with different sentence modalities are coordinated within a direct speech (the coordination root is assigned the functor CONJ and the attribute is_dsp_root). Within the original PDT 2.0 annotation (on the left side), the CONJ node was assigned the enunc value, the imperative modality of the first clause of the direct speech was omitted. The new annotation specifying the modality for each clause of the direct speech (imper & enunc) as well as for the matrix clause (enunc) is on the right side.

With the nodes identified as (a), (b) or (c), the value of the sentmod attribute was filled in according to the following 'algorithm', taking advantage of the links between the tectogrammatical, analytical and morphological annotation:
1. if the node represented an imperative verb form (i.e., technically, if one of the morphological tokens which the node was interlinked with was assigned the tag Vi.* (imperative verb form)), the node was assigned the sentmod value imper;
2. if the syntactic structure to which the node belonged ended with a question mark (technically, if the node corresponded to an analytical node that had a question mark among its child nodes), the sentmod value inter was filled in;
3. from the rest of the nodes, nodes that were a part of a sentence introduced by the particles *at', kéž, necht'* and/or ended with an exclamation mark were identified (92 occurrences in the training data of PDT 2.0) and assigned manually one of the sentmod values desid, excl or imper;
4. the remaining nodes were assigned the sentmod value enunc.

A tectogrammatical tree with sentmod values assigned according to this algorithm is displayed in Fig. 1 (on the left side). The distribution of the sentmod values in PDT 2.0 is given in Table 2.

4 An extended approach to sentence modality in PDT 3.0

4.1 Identification of weak points of the annotation

The main motivation for revision of the annotation of sentence modality was the insufficient treatment of coordination structures. However, at the very beginning of the revision, we wanted to find out whether there are, in addition to direct speech and parenthesis, further types of embedded structures which express a sentence modality on their own and thus require a separate sentmod value. A simple test, based on the direct interconnection between imperative mood and imperative sentence modality,[10] has pointed out to one more type of such structures, namely to sentence-like titles assigned the functor ID in the annotation (see ex. (4): the title *Pohlad'te si králíčka* 'Stroke a bunny rabbit' expressing the imperative modality is embedded in a matrix clause with declarative modality).

(4) *Zítra bude u příležitosti III. výročí české a slovenské edice Playboy otevřena.*enunc *výstava Pohlad'te.*imper *si králíčka sestavená z ilustrací pro časopis Playboy.* 'An exhibition Stroke.imper a bunny rabbit consisting of illustrations for the magazine Playboy will be opened.enunc tomorrow on the occasion of the 3rd anniversary of the Czech and Slovak editions of Playboy.'

4.2 Redesigning the assignment process

When considering the relation between the sentence modality annotation in PDT 2.0 (which concerned the nodes listed under (a) to (c) in Sect 3.2) and the new aim to specify a sentmod value for each clause in coordinations as well as for the title structures, it was not possible to preserve the current annotation and just to add sentmod values to the new candidates, since the decision to deal with coordinations affects all current subgroups (a) to (c). Another reason in favor of repeating the annotation was the fact that errors of several types were corrected during a systematic revision of the PDT 2.0 annotation carried out in the recent two years. Therefore, the sentmod values available in the PDT 2.0 data were canceled and the assignment process has been redesigned for PDT 3.0 and applied to the data from the scratch.

First of all, the set containing the candidate nodes (a) to (c) was extended by (d) the root nodes of title subtrees (functor ID). Secondly, from all these candidates, coordination roots were extracted and handled separately (see Sect. 4.3). Thirdly, for the remaining (non-coordination) nodes the steps described under 1 to 4 in Sect. 3.2 were applied; manual annotation (step 3) was needed for 82 nodes (in the training data).

4.3 Assigning coordinated clauses with sentence modality

Coordinations were handled as a homogeneous group, regardless which of the subgroups (a) to (d) they belonged to. On the basis of the extracted list of coordination roots, the set of root nodes of coordinated clauses which were to be assigned a sentmod value was delimited: 17,320 roots of coordinated clauses (governed by 7,598 coordination roots) were identified in the training data.

[10] In Czech, the imperative mood occurs exclusively in sentences with the imperative sentence modality and, the other way round, the imperative modality is mostly expressed by sentences with an imperative verb form.

sentmod value	frequency in PDT 2.0	frequency in PDT 3.0	with coordinated clauses
enunc	41,949	57,608	17,106
inter	777	828	130
imper	175	271	64
desid	17	13	2
excl	84	62	18
total	43,002	58,782	17,320

Table 2. The sentmod values in the PDT 2.0 and PDT 3.0 training data. The values of the last column are involved in the values of the third column.

For the sake of specification of the sentmod value for the root of each coordinated clause, the step 1 of the algorithm could be applied "locally", i.e. just for the particular clause of the coordination structure, not for all the clauses in a coordination: 64 root nodes of the individual coordinated clauses (in the training data) were assigned the value imper since they represented an imperative form.

Those non-imperative clauses which were coordinated with the imperative ones were extracted to be assigned a sentmod value manually. The second portion for manual annotation were roots of coordinated clauses that were part of a coordination structure ending with a question mark. Our assumption that the question mark occurring as the final punctuation mark of the whole coordination structure is to be interpreted as a signal of the sentence modality just for the final clause of the coordination structure (i.e. it does not mirror the sentence modality of the non-final clauses) proved to be true during the annotation. Roots of coordinated clauses which were part of a coordination structure ending with an exclamation mark or involving the particles *at'*, *kéž* and *necht'* were the third portion for manual annotation. The manual annotation thus concerned 268 roots of coordinated clauses in total. It was carried out by two annotators in parallel, with the inter-annotator agreement of 93.7% (Cohen's Kappa 0.89).

All the remaining coordination structures ended with a period (or without punctuation etc.) and involved only clauses with an indicative or conditional verb form. As in 100 coordination structures randomly selected from this group, only coordinated clauses with declarative modality were found, clauses in these coordination structures were automatically assigned the sentmod value enunc.

The distribution of the sentmod values in the training data of PDT 3.0 is listed in Table 2, besides the overall statistics (3rd column of the Table), the frequency of the values with the coordinated clauses is given as well (4th column). Lower frequency of the values desid and excl in PDT 3.0 in contrast to PDT 2.0 is due to some recent theoretical clarifications and corrections which were reflected in the manual annotation to be included in the PDT 3.0. The substantial increase with the other values is connected with the assignment of coordinated clauses. The differences between the sentence modality assignment in PDT 2.0 vs. 3.0 are illustrated in Fig. 1.

5 Conclusions

In the Prague Dependency Treebank, sentence modality is understood as a modal meaning of the sentence, or of each of its syntactically independent parts, and represented by a special node attribute sentmod in the tectogrammatical annotation. Within the original, simplified annotation of sentence modality, which was implemented in the PDT 2.0 data, sentence modality was assigned to the root node of each sentence and to the roots of parentheses and direct speech. Within the recent months, the sentence modality assignment has been elaborated with coordinated clauses and extended to embedded titles. The resulting annotation is thus more consistent and theoretically adequate, it will be released as a part of the PDT 3.0.

Acknowledgments

The research reported on in the present paper was supported by the grants GA ČR P406/2010/0875 and GA ČR P406/12/P175. This work has been using language resources developed and/or stored and/or distributed by the LINDAT-Clarin project of the Ministry of Education of the Czech Republic (project LM2010013).

References

1. Austin, J. L.: *How to Do Things with Words.* Harvard (2005)
2. Daneš, F., Hlavsa, Z., Grepl, M. et al.: *Mluvnice češtiny 3.* Praha (1987)
3. Hajič, J. et al.: *Prague Dependency Treebank 2.0.* CD-ROM. Philadelphia (2006)
4. Mikulová, M. et al.: *Annotation on the tectogrammatical level in the Prague Dependency Treebank. Annotation manual.* Tech. Rep. Nr. 2006/30. Prague (2006)
5. Mírovský, J.: *Searching in the Prague Dependency Treebank.* Prague (2009)
6. Pajas, P., Štěpánek, J.: Recent Advances in a Feature-Rich Framework for Treebank Annotation. In *Proceedings of Coling 2008* Manchester, pp. 673–680 (2008)
7. Pajas, P., Štěpánek, J.: System for Querying Syntactically Annotated Corpora. In *Proceedings of the ACL-IJCNLP 2009 Software Demonstrations.* Singapore, pp. 33–36 (2009)
8. Portner, P.: *Modality.* Oxford (2009)
9. Prasad, R. et al.: The Penn Discourse Treebank 2.0. In *Proceedings of LREC 2008.* Marrakech, pp. 2961–2968 (2008)
10. Saurí, R., Pustejovsky, J.: FactBank: a corpus annotated with event factuality. *Language Resources and Evaluation* 43, pp. 227–268 (2009)
11. Searle, J.: *Speech Acts.* Cambridge (1969)
12. Sgall, P., Hajičová, E., Panevová, J.: *The Meaning of the Sentence in Its Semantic and Pragmatic Aspects.* Dordrecht – Praha (1986)
13. Šmilauer, V.: *Novočeská skladba.* Third Edition. Praha (1969)
14. Zaefferer, D.: On the coding of sentential modality. In J. Bechert et al. (eds.): *Towards a Typology of European Languages.* Berlin – New York, pp. 215–237 (1990)