Systematic Parameterized Description of Pro-forms in the Prague Dependency Treebank 2.0*

Magda Ševčíková Razímová, Zdeněk Žabokrtský

Institute of Formal and Applied Linguistics
Charles University, Prague
E-mail: {razimova,zabokrtsky}@ufal.mff.cuni.cz

1 Introduction

A pro-form is a word that is used to replace or substitute other words, phrases, clauses, or sentences etc. Besides pronouns one can also distinguish pro-adjectives, pro-numerals, pro-adverbs, and pro-verbs.[1]

Pro-forms are related to a wide range of linguistic phenomena, from word-formative principles, through negation and quantification, to anaphoric and deictic functions. As it was recognized a long time ago (among others in Montague grammar), pro-forms are extremely important for studying natural language semantics, even if they constitute only a closed class.

Recently, a lot of work has been invested into developing large data resources for exploring natural language semantics, e.g. in the fields of predicate-argument structures or lexical databases. However, the treatment of pro-forms and related phenomena receives only a relatively minor attention in this data-dominated era (perhaps with the only exception of data for anaphora resolution, mostly limited to personal and demonstrative pronouns). Even if tag sets used in various corpora and treebanks clearly indicate some differentiation within the set of pro-forms (e.g.

*The research reported in this paper was supported by the projects 1ET101120503, GA-UK 352/2005 and GD201/05/H014.

[1]The well known difficulties with the heterogeneity of the criteria for delimitating the 'traditional' parts of speech lead to terminological confusion here: pronouns are often considered to span not only pro-nouns, but also some of the other pro-forms, especially pro-adjectives and also pro-adverbs, often denoted as pronominal adjectives and pronominal adverbs. The term pronominal nouns is used less frequently, perhaps because it sounds pleonastic. The term pronominal verb is mostly used to denote not a pro-verb, but a verb accompanied with a reflexive particle.

Hajič, J. and Nivre, J. (eds.): Proceedings of the TLT 2006, pp. 175–186.

wh-words in English), they do not allow us to directly observe and employ many semantically significant analogies present in the pro-form systems.[2] This concerns, e.g., the fact that *nobody*, *never*, and *nowhere* share certain semantic feature in their meanings, as well as *everybody*, *always*, and *everywhere* do, and that the two features are mutually exclusive.

Moreover, the present tag sets in some cases do not distinguish expressions which have the same surface shape but which significantly differ in semantic or pragmatic aspects. For instance, personal pronouns used in formal (esteemed) speaking may be homonymous with other pronouns (e.g. third person plural in German or second person plural in Czech). Similarly, interrogative and relative pronouns are known to be ambiguous in many Indo-European languages and they also usually obtain the same POS tag, although the difference between them would become crucial when constructing e.g. a dialog system.[3]

In this paper we present a formal linguistic system for the annotation of pro-forms which has been developed and implemented in the framework of the tectogrammatical layer of the Prague Dependency Treebank 2.0. The main motivation of our approach is the following: if there is a semantically relevant regularity within a certain subset of pro-forms, then it is more useful – at least from the viewpoint of treebank users interested in natural language semantics, in conversions into logical forms etc. – if such information is available in the treebank in an explicit, machine-tractable form. In this case, the semantic features originally present in the word form (given its context) are extracted and stored as values of inner parameters of tectogrammatical nodes corresponding to the given word form. Metaphorically, this can be seen as snatching pieces from the lexical space and reshaping them into multidimensional orthogonal blocks in which the semantics of each element can be derived from the semantics of its coordinates in an entirely compositional fashion.

Of course, the question of regularities in the pro-form systems is by far not new; various attempts at systematizing (at least certain subsets of) pro-forms can be found e.g. in [8], [7], [1], or in Wikipedia.[4] What we believe is new here is that the presented system is explicit and implementable, incorporated into the elaborated system of deep-syntactic analysis, and at the same time, applied (and verified) on large data.

[2] Besides the set of morphological tags used in the Prague Dependency Treebank, we have studied also rules for tagging pro-forms in Penn Treebank ([10]), Tiger Treebank ([6]), MULTEXT-East projects ([4]), and BulTreeBank ([13]) from the perspective of pro-forms. The last one seems to be the most developed in this aspect.

[3] However, although the distinction between ambiguous relative and interrogative pro-forms is not explicitly marked e.g. in the Penn Treebank, it could be derived from the shape of the phrase-structure annotation with a high precision.

[4] http://en.wikipedia.org/wiki/Pro-forms

2 PDT 2.0 in a Nutshell

In the Prague Dependency Treebank annotation scenario, based on the theoretical framework of Praguian Functional Generative Description ([12]), three layers of annotation are added to Czech sentences (see Figure 1):[5]

Morphological layer (m-layer), on which each token in each sentence of the source texts is lemmatized and tagged with a positional POS-tag.[6]

Analytical layer (a-layer), on which a sentence is represented as a rooted ordered tree with labeled nodes and edges, corresponding to the surface-syntactic relations; each a-layer node corresponds to exactly one m-layer token.

Tectogrammatical layer (t-layer), on which a tree structure of a sentence is labeled with tectogrammatical lemmas (often different from the morphological ones) and dependency relations (semantic roles, functors) and enriched with valency annotation, annotation of coreference, topic-focus annotation and annotation of semantically relevant grammatical meanings (grammatemes) and related attributes for node classification such as sempos (semantic part of speech).

Annotations at all tree layers (m-layer, a-layer, and t-layer) are part of PDT 2.0. PDT 2.0 data consist of 7,110 manually annotated textual documents, containing altogether 115,844 sentences with 1,957,247 tokens (word forms and punctuation marks). All these documents are annotated at the m-layer, 75 % of them are annotated at the a-layer. 59 % of the a-layer data are annotated also at the t-layer (i.e. 45 % of the m-layer data; 3,165 documents, 49,431 sentences, 833,195 tokens). The CD-ROM including the final annotation of PDT 2.0 data, a detailed documentation as well as software tools has been publicly released by Linguistic Data Consortium in 2006 ([3]).[7]

3 Pro-forms in the PDT 2.0

At the m-layer, pronouns, pronominal adverbs, and pronominal numerals are treated, as usual, separately. The part-of-speech information is encoded in the first of the 15 tag positions: the upper case letter P stands for pronouns, D for adverbs and C for numerals. The part-of-speech information is further specified at the second tag position (see the left column in Figure 2). Neither pronominal adverbs nor

[5] See http://ufal.mff.cuni.cz/pdt2.0/ for a detailed documentation and sample data of PDT 2.0.

[6] Technically, there is also one more layer called w-layer (word layer) 'below' the m-layer; on this lowest layer the original raw text is only segmented into documents, paragraphs and tokens, and all these units are enriched with identifiers.

[7] The previous version of the treebank, PDT 1.0, was smaller and contained only m-layer and a-layer annotation (see [2]).

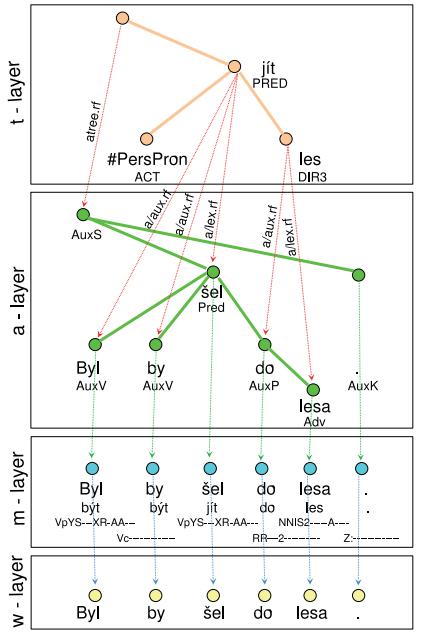

Figure 1: PDT 2.0 annotation layers and the layer interlinking illustrated (in a highly simplified fashion) on the sentence *Byl by šel do lesa* ([He] would have gone into [a] forest).

pronominal numerals are delimited as special subclasses of the respective parts of speech.

While m-layer annotation[8] was assigned only to pro-forms that are present in the surface shape of the sentence, at the t-layer we have to deal also with pro-forms that do not correspond to any word in the outer shape of the sentence – in the sequel, we call them restored nodes. If a restored node stands e.g. for a pro-dropped subject (as in Figure 1) that is not present in the surface sentence, it is considered to be a personal pronoun at the t-layer.[9]

At the t-layer, we have developed two different annotation schemes for pro-forms. The first scheme, which we present in Section 3.1, has been suggested for personal pronouns, taking into account the special character of these pronouns (e.g. they have a strictly deictic function, lacking a real lexical meaning). The second

[8]Note that the a-layer annotation does not add any information specifically related to pro-forms.

[9]Not only the new node is added to the structure, but also the values of its grammatical categories such as gender or number are reconstructed in the PDT 2.0 data (using e.g. subject-verb agreement or coreference relations).

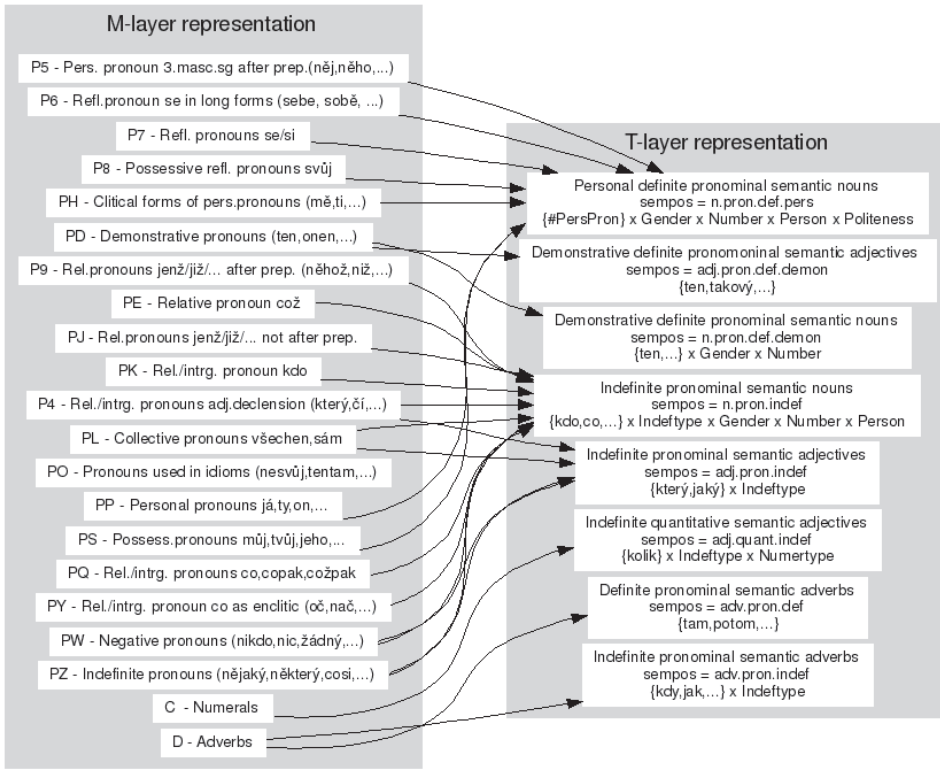

Figure 2: Rearrangement of pro-forms during the transition from m-layer to t-layer (note that arrows corresponding to other than pronominal numerals and adverbs are not displayed in the figure). At the m-layer, each pro-form is represented by its word form, morphological lemma and positional tag (values of the first one or two positions are specified in the entries in the left column). At the t-layer, pro-forms are represented as labels of t-nodes; for each of them, the attribute sempos (detailed semantic part of speech, see [11] for the explanation of the two-level t-node type hierarchy) specifies which other attributes are to be filled (besides tectogrammatical lemma, which is always obligatory). Thus the t-layer representation of a pro-form can be viewed as a vector from the space given by a Cartesian product corresponding to the given semantic part of speech.

scheme, which is introduced in Section 3.2, makes it possible to treat indefinite, interrogative, and other pronouns together with pronominal adverbs and numerals as there are many resemblances between them.

3.1 Personal Pronouns at the T-layer

All personal pronouns, no matter whether they are present in the outer shape of sentence or restored at the t-layer, are represented by nodes labeled with a single, 'artificial' lemma #PersPron. Information about person, number and gender that a personal pronoun expresses in a sentence is stored in node attributes called grammatemes (grammatemes person, number, gender).[10] Since there is a distinction between honorific and non-honorific forms in Czech, a special grammateme politeness was defined.

This representation based on the combination of the (semantically empty) artificial lemma with grammatical and pragmatic (in case of honorification) features was complemented by the annotation of coreference, i.e. relations between nodes referring to the same entity (in our case, between the personal pronoun and the noun that is substituted by the pronoun).[11]

Possessive pronouns which correspond to personal pronouns (e.g. *jeho* (his), *náš* (our)) are treated in the same way as their personal counterparts at the t-layer. A t-tree representing a sentence which contains a (restored) personal pronoun as well as a possessive pronoun, both represented by #PersPron nodes, is shown in Figure 3. Also reflexive pronouns (including possessive reflexives) are treated similarly in specific cases (see [5]), however, this topic goes beyond the scope of this paper.

Also in case of other pronoun types we have aimed at finding a reduced (if not minimal) way of representation. However, we had to cope with features that are absent with personal pronouns.

3.2 Other Pronoun Types and Pro-forms at the T-layer

3.2.1 Indefinite, Interrogative, Negative, and Relative Pronouns

Neither indefinite pronouns nor other pronoun groups (i.e. interrogative, negative, and relative pronouns)[12] can be represented by means proposed for representing personal pronouns since each of the other pronoun groups has a special meaning (generally corresponding to a name of the respective group). However, in the Czech pronoun system single meanings are expressed regularly by means of a relatively small group of prefixes that combine with a small set of bases. Therefore, there is a

[10] The grammateme system that is used in PDT 2.0 framework has been described in detail in [11].

[11] Coreference relations are technically represented as pointers from the pronoun node to the coreferring antecedent node. In PDT 2.0, coreference relations are marked only with the 3rd person pronouns, not with the 1st and 2nd person pronouns (see [5]).

[12] The group of demonstrative pronouns is neglected in our contribution since a systematic semantic representation of this pronoun type has not been developed yet.

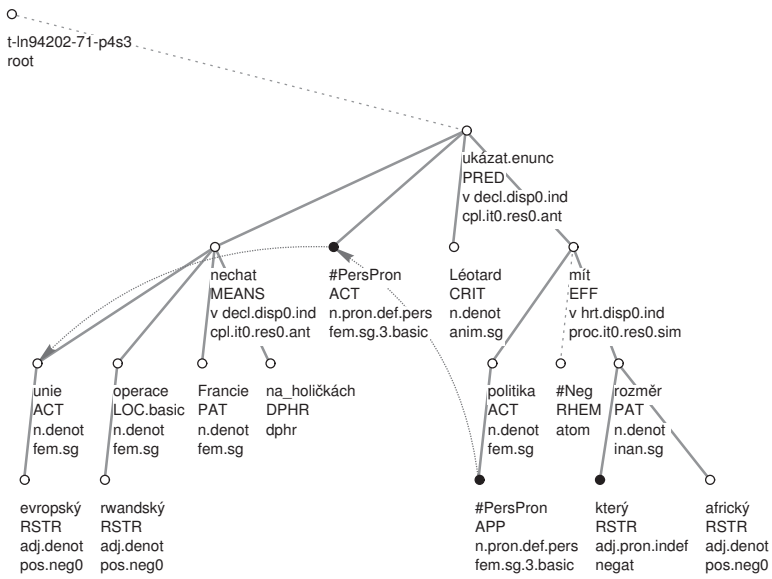

Figure 3: Tectogrammatical representation of the sentence *Tím, že Evropská unie nechala ve rwandské operaci Francii na holičkách, podle Léotarda ukázala, že její politika nemá žádný africký rozměr* (According to Léotard, by the fact that the European Union left France in the lurch concerning the Rwanda operation, [it] has shown that its politics has no African dimension). T-nodes corresponding to pro-forms are filled with black color. The #PersPron node which plays the semantic role of an actor (functor ACT) stands for the subject (*it*, i.e *European Union*) that is not present in the surface shape of the Czech sentence. Bellow the functor, the sempos value (here n.pron.def.pers, i.e. personal definite semantic noun) and the grammatical categories are specified. The second #PersPron node is labeled with functor APP (for appurtenance) and represents the possessive pronoun *její* (its). Each of these #PersPron nodes is linked by a coreference pointer with its antecedent node.

transparent correspondence between the meaning features and formal composition of pronouns, e.g. indefinite pronouns begin by prefixes *ně-* (e.g. *někdo* (someone), *něco* (something), *nějaký* (some)), negative pronouns begin by *ni-* (e.g. *nikdo* (nobody), *nic* (nothing)).

Searching for a systematic representation, we have grouped together pronouns with the same base element (e.g. *někdo* (someone) together with *nikdo* (nobody) and *něco* (something) with *nic* (nothing)), thereby all indefinite, interrogative, neg-

Lemma:	*kdo*	*co*	*který*	*jaký*
indefype:				
relat	kdo	co	který,jenž	jaký
indef1	někdo	něco	některý	nějaký
indef2	kdosi,kdos	cosi,cos	kterýsi	jakýsi
indef3	kdokoli,kdokoliv	cokoli,cokoliv	kterýkoli,kterýkoliv	jakýkoli,jakýkoliv
indef4	ledakdo,leckdo	ledaco,lecco	leckterý,ledakterý	lecjaký,ledajaký
indef5	kdekdo	kdeco	kdekterý	kdejaký
indef6	kdovíkdo,čertvíkdo	kdovíco,...	kdovíkterý,...	kdovíjaký,...
inter	kdo,kdopak	co,copak	který,kterýpak	jaký,jakýpak
negat	nikdo	nic	žádný	nijaký
total1	všechen,vše	všechno	–	–
total2	–	–	každý	–

Table 1: The indeftype attribute has eleven values (1st column in the table). This makes it possible to represent all derivatives of the pronouns *kdo* (someone), *co* (something), *který* (that) and *jaký* (what) (in the 2nd, 3rd, 4th and 5th column) by only these four lemmas at the t-layer.

ative, and relative pronouns have fallen into four groups inside of each of which the same set of prefixes occur. Temporarily, each group of pronouns is represented by the lemma corresponding to the relative pronoun at the t-layer, i.e. for example, the indefinite pronoun *někdo* (someone) as well as the negative pronoun *nikdo* (nobody) are represented by the lemma *kdo* (who).[13]

The semantic feature (directly corresponding to the prefix) completing the reduced lemma is specified in the special attribute indeftype whose values indef1 to indef6 correspond to six types of indefinite pronouns, the value negat to negative pronouns etc. The pronouns with corresponding lemmas and values of the indeftype attribute are displayed in Table 1.

In Czech, possessive counterparts are available also for these pronoun groups. At the t-layer, they are again represented in an way analogous to how possessive pronouns corresponding to personal pronouns are treated. For example, *ničí* (nobody's) is a possessive counterpart of the negative pronoun *nikdo* (nobody) – the possessive form as well as the basic negative pronoun are represented by the lemma *kdo* (who) and the negat value of the attribute indeftype.[14]

[13] In some cases, such representation might look confusing because – in spite of the semantic symmetry – the selected representant is not morphologically related with all pronouns it represents (e.g. *žádný* represented by *který*). Due to this reason, it seems to be feasible to introduce an artificial lemma for each group in the future, similarly to the #PersPron lemma for personal pronouns (see Section 3.1).

[14] This annotation scheme, i.e. the treatment of possessive words as derived from other words, is

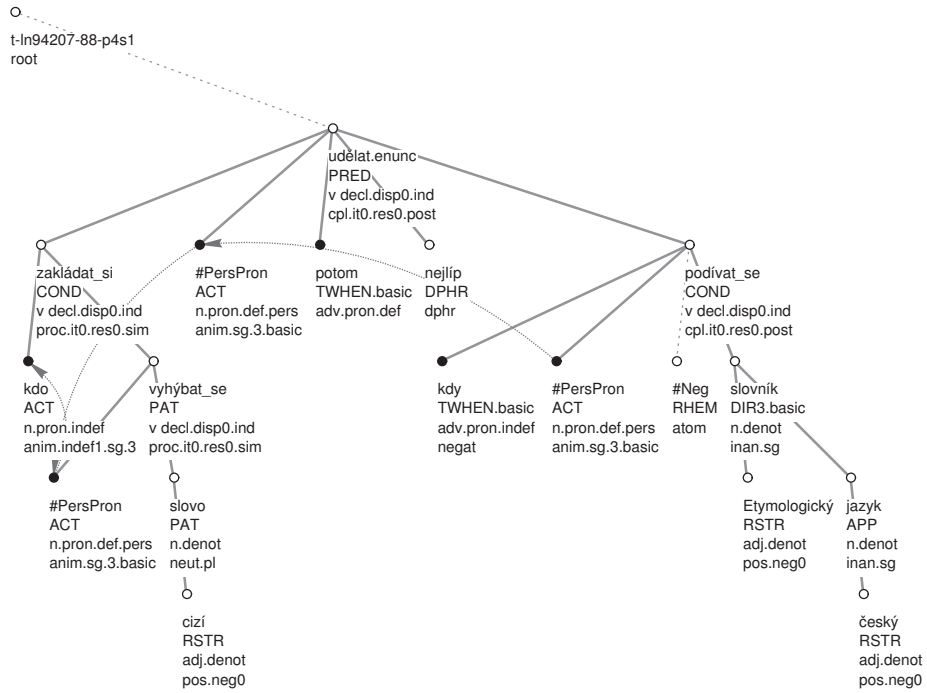

Figure 4: Tectogrammatical representation of the sentence *Zakládá-li si někdo na tom, že se vyhýbá cizím slovům, pak udělá nejlíp, když se nikdy nepodívá do Etymologického slovníku jazyka českého* (If someone finds it important that [he] avoids foreign words, then the best thing [he] can do is if [he] never looks in the Etymology Dictionary of Czech). The indefinite pronoun *někdo* (someone) and negative pronominal adverb *nikdy* (never) are represented by their relative counterparts (*kdo* (who) and *kdy* (when), respectively). Corresponding indeftype values (i.e. indef1 and negat) are displayed under the functor and sempos values (among grammateme values in the case of the pronoun). The three #PersPron nodes stand for subjects (*he*) that are not present in the surface shape of the Czech sentence.

3.2.2 Pronominal Adverbs and Numerals

Since other pro-forms (e.g. pronominal adverbs *nějak* (somehow) and *nikde* (nowhere) or the pronominal numeral *několik* (a few)) express the same semantic features (and

applied systematically also to possessive adjectives at the t-layer of PDT 2.0: e.g. the possessive adjective *matčin* (mother's) is represented by the tectogrammatical lemma *matka* (mother).

	English	English	German	German
Lemma	*who*	*what*	*wer*	*was*
indefype:				
relat	who	which	der	das
indef1	someone	something	jemand	etwas
indef2	–	–	irgendjemand	irgendetwas
indef3	whoever	whatever	–	–
inter	who	what	wer	was
negat	nobody	nothing	niemand	nichts
total1	all	everything	alle	alles
total2	each	each	jeder	jedes

Table 2: Selected English and German pronouns preliminarily classified according to the indeftype attribute.

show the same derivational relations) like certain types of pronouns in Czech, they are represented in the same way at the t-layer.

Another systematic relation can be seen between pronominal adverbs with directional meaning and those of location. E.g., the adverb *odněkud* (from somewhere) is represented by the lemma *kde* (where), the indef value of the indeftype attribute and the functor DIR1 capturing the directional meaning. A sample t-tree containing some pro-forms is displayed in Figure 4.

Indefinite, interrogative, and other pro-forms are unproductive classes with (at least to a certain extent) transparent derivational relations not only in Czech, but also in other languages. However, as it is obvious from the preliminary sketch of several English and German pronouns classified in Table 2,[15] the application of our scheme to other languages will not be straightforward and various subtle differences have to be taken into account. For instance, there is only one negative form *nikdo* corresponding to the lemma *kdo* in Czech, therefore the present system provides no means for distinguishing German negative pronouns *niemand* and *nirgendjemand*. A new question arises also in the case of English *anybody* when used in negative clauses, which has no direct counterpart in Czech or German.

4 Conclusion and Future Work

In the paper we have presented a formal system used for tectogrammatical representation of pro-forms in the Prague Dependency Treebank 2.0. The main features of the system are the following:

[15] We chose English and German, because, first, the two languages are the most familiar to us, and second, certain experiments concerning their t-layer have already been performed, see e.g. [9].

1. All pro-forms in Czech are divided into two groups: (a) personal and possessive pronouns (be they present in the surface shape of the sentence or pro-dropped), all represented by tectogrammatical lemma #PersPron and a vector of values for person, gender, number, and politeness, and (b) other pro-forms, clustered into blocks parameterized by their semantic part of speech, type of indefiniteness etc.

2. Our system covers not only the 'traditional' pronouns, but also other pro-form expressions, originally tagged as adverbs or numerals on the morphological layer.

3. Unlike in traditional part-of-speech tagging approaches, several pro-form analogies (e.g. those concerning the type of indefiniteness) crossing the boundaries between semantic nouns, semantic adjectives, or semantic adverbs, are explicitly marked in the annotation.

4. Expressions which differ only due to grammatical reasons (e.g. personal and corresponding possessive pronouns, or indefinite pronouns and their possessive counterparts) are represented identically. On the other hand, homonymous expressions bearing a semantically or pragmatically relevant difference are distinguished (e.g. interrogative and relative pro-forms).

In the future, we plan to study relationships between pro-forms and several semantic phenomena (quantifier scope, multiple negation, etc.). We will also consider further compactification of the lexical space on the tectogrammatical layer, using the same idea of multidimensional parametrization, probably in combination with the framework of lexical functions ([14]).

Acknowledgements

We would like to thank professors Eva Hajičová, Jarmila Panevová, and Petr Sgall for numerous comments on the draft of the paper.

References

[1] Igor A. Bolshakov. Treatment of personal pronouns based on their parameterization. In *Proceedings of CICLing 2001*, pages 80–92, 2001.

[2] Jan Hajič et al. Prague Dependency Treebank 1.0. Linguistic Data Consortium, CAT LDC2001T10, ISBN 1-58563-212-0, 2001.

[3] Jan Hajič et al. Prague Dependency Treebank 2.0. Linguistic Data Consortium, CAT LDC2006T01, ISBN 1-58563-370-4, 2006.

[4] Ludmila Dimitrova et al. Multext-East: Parallel and Comparable Corpora and Lexicons for Six Central and Eastern European Languages. In *COLING-ACL '98*, pages 315–319, Montréal, Québec, Canada, 1998.

[5] Marie Mikulová et al. Anotace Pražského závislostního korpusu na tektogramatické rovině: pokyny pro anotátory. Technical report, ÚFAL MFF UK, Prague, Czech Republic, 2005.

[6] Sabine Brants et al. The TIGER treebank. In *Proceedings of the Workshop on Treebanks and Linguistic Theories*, Sozopol, 2002.

[7] Hermann Helbig. *Die semantische Struktur natürlicher Sprache*. Springer-Verlag, Berlin, Heidelberg, New York, 2001.

[8] Miroslav Komárek. *Příspěvky k české morfologii*. SPN, Prague, 1978.

[9] Ivona Kučerová and Zdeněk Žabokrtský. Transforming Penn Treebank Phrase Trees into (Praguian) Tectogrammatical Dependency Trees. *Prague Bulletin of Mathematical Linguistics*, (78):77–94, 2002.

[10] Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. Building a Large Annotated Corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330, 1994.

[11] Magda Razímová and Zdeněk Žabokrtský. Annotation of Grammatemes in the Prague Dependency Treebank 2.0. In *Proceedings of the LREC 2006 Workshop on Annotation Science*, pages 12–19, 2006.

[12] Petr Sgall, Eva Hajičová, and Jarmila Panevová. *The Meaning of the Sentence in Its Semantic and Pragmatic Aspects*. D. Reidel Publishing Company, Dordrecht, 1986.

[13] Kiril Simov, Petya Osenova, and Milena Slavcheva. BTB-TR03: BulTreeBank Morphosyntactic Tagset. BTB-TS version 2.0. Technical report, Bulgarian Academy of Sciences, Sofia, Bulgaria, 2004.

[14] Leo Wanner, editor. *Lexical Functions in Lexicography and Natural Language Processing*. John Benjamins Publishing Company, 1996.