Proper Nouns in Czech Corpora

Magda Ševčíková
Institute of Formal and Applied Linguistics
Charles University in Prague
Czech Republic
sevcikova@ufal.mff.cuni.cz

1 Introduction

Proper nouns are an inseparable part of natural language texts. They differ from common nouns especially by lacking a generic meaning. Their main function is to denote an individual, a thing, an institution *etc.* and to identify them as unique items. Besides semantics, also other characteristics distinguish them from common nouns. For example, a word can change its part-of-speech characteristics or some of its morphological categories when becoming a proper noun; e.g., the surname *Zelený* derived from the adjective *zelený* (green) is considered to be a noun, the Czech surname *Hlava* is animate although derived from the feminine *hlava* (head). As for syntactic characteristics, a complex proper noun such as the institution name *Vysoká škola ekonomická* (University of Economics) cannot be interrupted by other words. These and other aspects of proper nouns, e.g. their etymology, are traditionally studied by an autonomous discipline, onomastics. However, also other (not only linguistic) disciplines that deal with natural language in some way have to pay special attention to proper nouns.

Natural language processing (NLP) has been dealing with proper nouns for many years. To recognize that a word or word sequence functions as a proper noun in a natural language text is an important pre-condition of success of various NLP applications. For example, machine translation tools have to be able to discern between proper nouns and other words since the former are usually not translated into the target language, e.g., the sentence *Generálním ředitelem společnosti ABC je pan Zelený* should not be translated to *The CEO of ABC Company is Mr Green*. In other cases, translation of proper nouns follows special principles, *cf.* German town name *Frankfurt am Main* could be translated into Czech as *Frankfurt nad Mohanem* or occur in its original form in a Czech text, however, translation as neither *Frankfurt nad Main* nor *Frankfurt am Mohanem* is acceptable.

Text corpora contain a considerable amount of proper nouns. They could be therefore used not only for onomastic and other linguistic research but also for development of various NLP applications. In the following section, two large corpora of Czech are examined as sources of proper nouns. Whereas corpus SYN2000 does not include any explicit annotation of proper nouns and one has to combine several characteristics of proper nouns to find these words in this corpus (see Section 2.1), in Prague Dependency Treebank 2.0 (PDT 2.0) some basic annotation of proper nouns was performed (Section 2.2). However, the current annotation covers only some characteristics of proper nouns and does not allow for an assignment of other important phenomena of these words. In Section 3 we propose a complex annotation of proper nouns within the PDT 2.0. For this purpose, a proper noun classification for Czech and corresponding annotation requirements are introduced in Sections 3.1 and 3.2. In order to integrate them into the PDT 2.0 annotation scenario, the current annotation has to be enriched and modified (Section 3.3). Some concluding remarks are given in Section 4.

2 Proper nouns in corpora of Czech

2.1 Corpus SYN2000

Corpus SYN2000 is a representative corpus of contemporary Czech developed at the Institute of Czech National Corpus at the Faculty of Philosophy and Arts of Charles University in Prague.[1] Corpus SYN2000, which contains 100 million tokens, was enriched with morphological annotation: each token was assigned a lemma and a positional morphological tag, which is a string of 15 positions. Part-of-speech and other morphological categories of the token in question (such as number, gender or tense) are specified in corresponding tag positions.

Proper nouns were not explicitly marked in this corpus. Although proper nouns differ from common nouns in capitalization in Czech (what can be considered as an advantage in comparison, e.g., to German where all nouns are capitalized), this is not a sufficiently distinctive feature to find these words in a corpus since also each sentence begins with a capitalized word. In order to find proper nouns in this corpus, also other characteristics have to be taken into consideration. One of them is that proper nouns occur in certain contexts. Examples of queries based on such contexts and capitalization with corresponding results from SYN2000 are displayed in Table 1. In spite of the high precision of the results (see the fourth column of the table) it is obvious that many occurrences of proper nouns cannot be found in this way.[2]

Searching SYN2000 for	**Query**	**Number of occurrences in SYN2000**	**Precision in 500 randomly selected occurrences (in %)**
names/surnames (or their parts)	lemmas *pan/paní/slečna* (Mr/Mrs/Miss) followed by a capitalized token	41,574	99.4
names/surnames (or their parts)	(un)capitalized short versions of Czech academic titles *doc./dr./ing./JUDr./MUDr./prof./RNDr.* followed by a capitalized token	26,394	96.0
names/surnames (or their parts)	lemmas of academic titles *doktor/profesor/docent/ inženýr* (doctor/ professor/docent/engineer) followed by a capitalized token	9,123	94.6
town names (or their parts)	digit combination corresponding to Czech zip code format followed by a capitalized token	7,954	92.6
street/square names (or their parts)	lemmas *ulice/náměstí* (street, square) followed by a capitalized token	6,233	87.6
street names (or their parts)	(un)capitalized abbreviation *ul.* (for street) followed by a capitalized token	696	87.4
company names (or their parts)	abbreviation *s.r.o.* (for Ltd.) preceded by a capitalized token (and optionally by a comma)	2,554	100.0
company names (or their parts)	abbreviation *a.s.* (for PLC) preceded by a capitalized token (and optionally by a comma)	4,274	99.6

Table 1: Examples of SYN2000 queries based on context patterns and capitalization

[1] http://ucnk.ff.cuni.cz

[2] Context patterns are used also for automatic extraction of proper nouns from corpora, for details see, e.g., Collins and Singer (1999) or Kang and Woo (2003).

2.2 Prague Dependency Treebank 2.0

2.2.1 Basic characteristics

Prague Dependency Treebank (PDT) is a richly annotated corpus of Czech newspaper texts. PDT, whose multilayered annotation scenario has been built on the theoretical basis of Functional Generative Description (*cf.* Sgall, 1967; Sgall, Hajičová and Panevová, 1986), was developed at the Institute of Formal and Applied Linguistics at the Faculty of Mathematics and Physics of Charles University in Prague. In the first version of this treebank (PDT 1.0), texts were annotated at two layers, at the morphological layer and at the so called analytical layer (Hajič *et al.*, 2001). At the morphological layer, each token is assigned a morphological lemma and a positional tag (similarly to SYN2000 corpus, see Section 2.1). At the analytical layer, a sentence is represented as a dependency tree the nodes of which are labelled with word forms and the edges with surface-syntactic functions (such as subject, object *etc.*). Each analytical node corresponds to exactly one token of the surface shape of the sentence (including punctuation marks).

In the second version of PDT (PDT 2.0), texts were additionally annotated at the deep-syntactic (so called tectogrammatical) layer (Hajič *et al.*, 2006). At this layer, a tree structure of a sentence is labelled with tectogrammatical lemmas (t-lemmas, in some cases different from the morphological ones) and deep-syntactic relations (so called functors, e.g., actor, patient) and enriched with valency annotation, annotation of co-reference, topic-focus annotation and annotation of semantically relevant grammatical meanings (so called grammatemes).[3]

2.2.2 Proper noun classification at the morphological layer

A very basic classification of proper nouns is already contained in the morphological annotation of PDT. Proper noun type is indicated by a value of a special flag which was attached to lemmas of proper nouns by a separator ;_ (e.g., Jan_;Y or Zelený_;S). Seven flag values corresponding to seven proper noun types were defined (Hana and Zeman, 2005), another 13 values of this flag correspond to scientific or professional terms rather than to proper nouns (*cf.* the flag j for legal terms such as *dissolution* or b for financial terms such as *dollar*). Values assigned to proper noun lemmas with lemma examples and numbers of occurrences in PDT 2.0 are displayed in Table 2. The numbers quoted there relate to morphologically annotated train data of PDT 2.0 of the following size: 1,171,191 tokens, 68,495 sentences in 4,264 documents.

Whereas one-word proper nouns were annotated satisfactorily by the lemma flags, application of this strategy to complex proper nouns led to false interpretations. For example, the town name *Frankfurt nad Mohanem* was analyzed as two separate geographical names with a preposition in between (assigned lemmas: Fankfurt_;G nad Mohan_;G), the institution name *Vysoká škola ekonomická* (University of Economics) as a sequence of one proper noun and two common nouns (lemmas: Vysoký_;K škola ekonomický). However, even if the flag was assigned to the lemmas of each part of these proper nouns, the annotation would not be sufficient since it is not indicated whether the flagged tokens form together a complex proper noun or, for example, whether they are a sequence of several one-word proper nouns.

[3] In this contribution we further deal only with the morphological and tectogrammatical layer of PDT 2.0. The analytical layer is omitted here.

Flag value	Description	Examples from PDT 2.0 data	Number of occurrences in PDT 2.0 data
Y	first names	*Josef, Oskar, Marcus*	12,256
S	surnames	*Novotný, Jágr, Dalí*	21,029
E	inhabitant names	*Francouz, Středoevropan, Afričan*	1,893
G	geographical names	*Praha, Evropa, Mars*	21,775
K	company/organization/ institution names	*Microsoft, Aero*	10,868
R	product names	*Boeing, Fiat*	1,832
m	other proper nouns		658

Table 2: Lemma flags used for annotation of proper nouns at the morphological layer of PDT 2.0

2.2.3 Basic principles for tectogrammatical annotation of proper nouns

Although no complex annotation of proper nouns was performed either at the tectogrammatical layer of PDT 2.0, several selected phenomena concerning proper nouns were described there. For this purpose, three special annotation means were introduced. Firstly, the node attribute is_name_of_person was defined for the identification of person names. The value 1 indicates that the node in question represents a person name (i.e., this value occurs at a first name as well as at a surname without any difference; see Fig. 2a), otherwise the value is 0. Secondly, the functor ID was introduced for the annotation of street names, book titles *etc.* in cases when they are accompanied by a generic noun such as *ulice* (street) or *kniha* (book) and do not decline (i.e., are always, regardless of the form of the generic noun, in the nominative form, in Czech syntax handbooks called nominative of identity), see Fig. 1a. If the name declines together with the generic noun (Fig. 1b) or occurs without the generic noun (Fig. 1c), it appears in the dependency tree structure without marking its proper noun character. Functor ID is also applied when a generic noun is missing but the title has a character of a prepositional group or of a sentence fragment. In such cases, an 'artificial' node labelled with the t-lemma #Idph is added into the tree structure, which can be considered as a third annotation means. Nodes corresponding to individual parts of the title are then labelled with the ID functor and represented as depending on the #Idph node (Fig. 2b; see Mikulová *et al.*, 2006, esp. Chapter 8.8).

However, besides the individual cases which were captured by the above stated means, proper nouns were treated as common parts of a sentence at the tectogrammatical layer of PDT 2.0. Thus, for example, the relations between a first name and a surname (*cf.*, Fig. 2a) or between a street name and a town name were represented as dependency relations in spite of the fact that they are not relations of dependency nature in the linguistic sense of this term.

How to modify and extend the available tectogrammatical means so that all relevant characteristics of proper nouns can be captured in an adequate way is discussed in the following section (see esp. Section 3.3). Before starting this discussion a classification of proper nouns, requirements on their assignment as well as annotation of sample data are introduced (Sections 3.1 and 3.2).

3 Proper noun annotation proposed for Prague Dependency Treebank 2.0

As already mentioned in Section 1, an adequate treatment of proper nouns is an indispensable part of many NLP tasks. Since from the NLP point of view proper nouns share certain characteristics with some temporal and numerical expressions, proper nouns are treated together with these expressions under an umbrella term 'Named Entities'. This broad view was also applied when we started to deal with this domain within automatic processing of Czech. As this contribution focuses on proper nouns, we omit here the part of our work concerning temporal and numerical expressions and talk only about proper nouns. However, the term Named Entities (NEs) occurs when we refer to the NLP domain or when we suggest new annotation means which should serve for annotation of proper nouns as well as of temporal and numerical expressions.

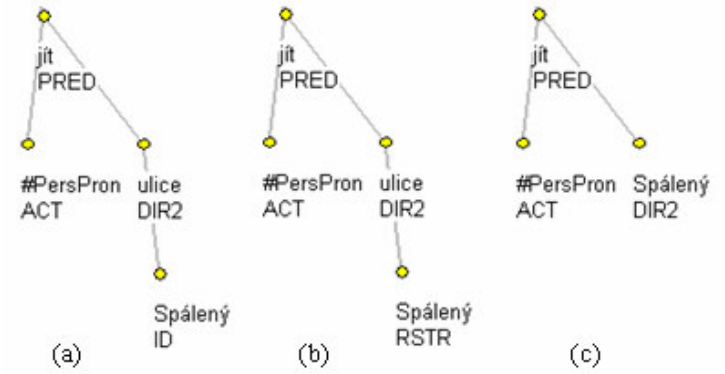

Figure 1: Representations of the sentences (a) *Šli jsme ulicí Spálená* (We walked through the street.*instr* Spálená.*nom*), (b) *Šli jsme ulicí Spálenou* (We walked through the street.*instr* Spálená.*instr*) and (c) *Šli jsme Spálenou* (We walked through Spálená.*instr*) at the tectogrammatical layer of PDT 2.0 (*instr* for instrumental case, *nom* for nominative case). The left-to-right order of the nodes follows conventions defined for PDT 2.0 (Mikulová *et al.*, 2006).

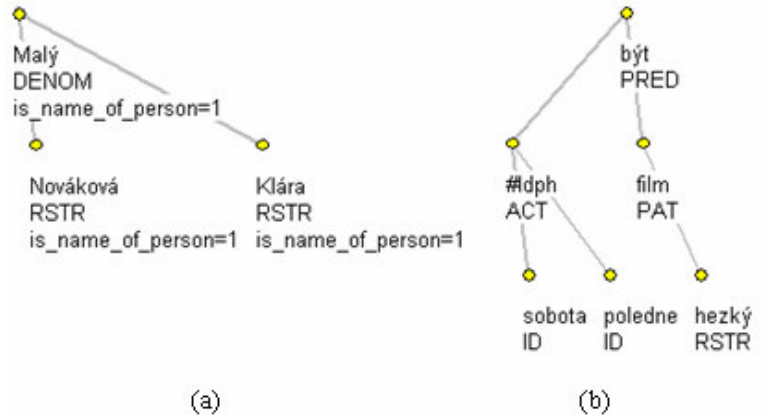

Figure 2: Representations (a) of the person name *Klára Nováková Malá* and (b) of the sentence *V sobotu v poledne je hezký film* (Lit.: 'On Saturday at Noon' is a nice film) at the tectogrammatical layer of PDT 2.0. For the left-to-right order of the nodes see caption of Fig. 1.

3.1 Proper noun classification for Czech

When proposing a classification of proper nouns for Czech we have drawn inspiration especially from English since processing of NEs has become a well established discipline in recent years. In spite of this fact there is no generally accepted typology of NEs. However, two main trends can be traced: there are typologies discerning just very few coarse-grained categories on the one hand; on the other hand very detailed classifications are proposed. The first type is represented for example by the typology developed for the 6th Message Understanding Conference (MUC-6; at this conference the NE detection and classification was defined as an autonomous task; Grishman and Sundheim, 1996). Within this typology, only three NE types, *entities*, *times* and *quantities*, are distinguished, they are further subdivided into seven subtypes (entities: *organization*, *person*, *location*; times: *date*, *time*; quantities: *money*, *percent*). As for the second typology type, Satoshi Sekine's Extended Named Entity Hierarchy originally contained about 150 NE types and has been further extended (Sekine, Sudo and Nobata, 2002). This hierarchy was published in 2002 with the aim of covering main NE types occurring in newspaper texts of all sorts.

For the annotation of proper nouns and other NEs in Czech texts, a two-level classification was proposed in order to meet the needs of both a robust and a more detailed processing. At the first level seven rough categories, so called supertypes, are distinguished: (1) personal names, (2) geographical names, (3) institution names, (4) artefact names, (5) media names and further (6) temporal expressions and (7) numerical expressions occurring in postal addresses. The second level corresponds to more detailed categories, called types. For example, street/square names, city/town names, state names etc. were distinguished within the supertype of geographical names. Each type is encoded by a unique two-character tag, e.g. `gs` for street/square names or `gu` for city/town names. A special tag, such as `g_`, makes it possible to leave the type underspecified.

With respect to the behaviour of proper nouns in Czech texts, the annotation scheme allows for two kinds of embedding. In the first case, a NE can be embedded in another NE (such as in *Frankfurt nad Mohanem* where the river name is part of the city name). In the second case, two or more NEs can form a complex structure, so called container. For example, a first name and a surname constitute a person name container such as in *Jan Zelený*. Containers are marked with a capital one-letter tag: `P` for person name containers, `T` for temporal expressions, `A` for addresses and `B` for bibliographic items. Recursive embedding is allowed here. A more detailed description of the proposed classification can be found in Ševčíková, Žabokrtský and Krůza (2007).

3.2 Sample annotation of proper nouns in Czech texts

In order to verify the applicability of the proposed classification, it was used for manual annotation of proper nouns and other NEs in Czech texts. For the sake of simplicity, texts were annotated in plain-text format editable in any text editor. The annotation was performed by two annotators in parallel. Their task had two parts: to delimit the one-word NE or the span of words belonging to a NE by symbols < and > and to define the type or the container by setting an appropriate tag immediately behind the < symbol; for example, <gu *Frankfurt nad* <gh *Mohanem*>>, <P<pf *Jan*> <ps *Zelený*>>. For this annotation, 2,000 sentences were randomly selected from the SYN2000 corpus from the result of the query ([word=".*[a-z0-9]"] [word="[A-Z].*"]).[4] NEs with differing annotations were checked and decided by a third person. In this annotation sample, 11,644 proper nouns and other NEs (types and containers altogether) were detected. For a sample of annotated text see Fig. 3.

[4] We searched the corpus for capitalized words preceded by another word or a figure (or a 'word' consisting of both figures and letters) in order to exclude capitalized words occurring at the sentence beginning.

In the next step, the data were enriched with morphological lemmas and tags used in PDT 2.0, were converted into XML format and divided into training, development test and evaluation test data (8:1:1). Such prepared data were used for training and evaluation of the automated annotation tool, so called NE tagger. The development of the NE tagger and its results are described in Ševčíková, Žabokrtský and Krůza (in press).

```
Britský multimediální umělec <p_ Sting> , vlastním jménem
<P<pf Gordon> <ps Sumner>> , který má vystoupit <T<td 14 .>
<tm června>> v pražské <ic Sportovní hale> , bude s největší
pravděpodobností bydlet se svým devětadvacetičlenným týmem
pod krycím jménem v některém z pražských hotelů .

V <ic Galerii <P<pf Václava> <ps Špály>>> bude dnes zahájena
výstava obrazů německého umělce <P<pf Herberta> <ps Achternbusche>> ,
připravená ve spolupráci s pražským <ic GoetheInstitutem> .
```

Figure 3: Sample of manually annotated text in which the span and the type of proper nouns and other NEs were assigned

3.3 Integration of proper noun annotation into the tectogrammatical layer

To sum up the previous section, the annotation proposed above makes it possible to annotate (i) one-word proper nouns, (ii) so called 'multi-word proper noun expressions', which are, in fact, continuous sequences of common nouns, adjectives *etc.* that function together as proper nouns (e.g., *Vysoká škola ekonomická*), (iii) 'complex proper noun expressions' in which other proper nouns are embedded (for example *Frankfurt nad Mohanem*), and (iv) containers consisting of two or more proper nouns (e.g., *Jan Zelený*). All these four annotation types are regarded as important and should be, in our opinion, captured in an annotated corpus.

However, if we reconsider the annotation layers of PDT 2.0 it is obvious that the morphological layer is considerably less convenient for the integration of proper noun annotation than the tectogrammatical one. Whereas the above mentioned problem concerning the annotation of the type *Vysoká škola ekonomická* (see 2.2.2; here marked as the type (ii)) could be solved for example by adapting the tokenization rules (the whole proper noun could be considered as being one token), it is not possible to capture either of the embedding types ((iii) and (iv)) without fundamental changes in the current annotation scheme of the morphological layer. On the other hand, the tectogrammatical layer already contains some special means for the annotation of proper nouns: node attribute is_name_of_person, functor ID and restored node labelled with the t-lemma #Idph. In order to integrate all the four types ((i) to (iv)) into the annotation we have to modify these means, to introduce some additional ones and to formulate corresponding annotation principles. Our suggestions are listed according to the four types stated above.

3.3.1 Annotation of one-word proper nouns

The present tectogrammatical annotation does not allow for a classification of proper nouns except for person names. We propose a new node attribute NE_roles whose value set corresponds to the above suggested classification. The proper noun represented by the node in question is both delimited and classified by filling in a value in this attribute. The is_name_of_person attribute becomes redundant then.

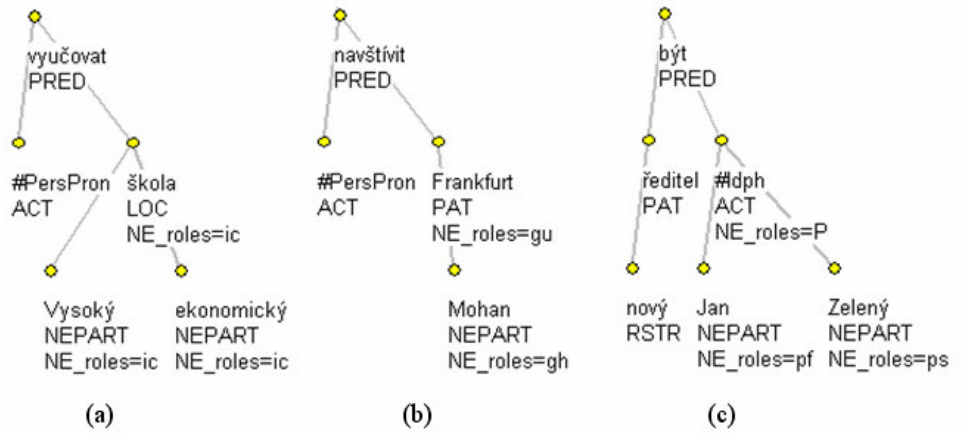

Figure 4: Proposed tectogrammatical representations of sentences (a) *Vyučuje na Vysoké škole ekonomické* (He teaches at University of Economics), (b) *Navštívil Frankfurt nad Mohanem* (He visited Frankfurt am Main) and (c) *Novým ředitelem je Jan Zelený* (Jan Zelený is the new director)

3.3.2 Annotation of multi-word proper noun expressions

At the tectogrammatical layer, we propose that every constituent of a multi-word proper noun expression (such as *Vysoká škola ekonomická*) has a node of its own. These nodes are formed into a sub-tree as follows: the governing node is chosen according to syntactic dependency principles (e.g., the governor of the multi-word proper noun expression *Vysoká škola ekonomická* is the noun *škola*), all the remaining nodes (here, *Vysoká* and *ekonomická*) depend on this node, their left-to-right order follows the surface word order. With all the nodes, the same value occurs in the NE_roles attribute. The fact that the nodes form together a single proper noun expression (and that there are not several independent proper nouns of the same type) is marked by a special functor NEPART which appears at the depending nodes. Functor of the governing node specifies the function of the whole proper noun expression in the sentence structure. E.g., the multi-word proper noun expression plays the role of a local complementation (functor LOC in PDT 2.0) in the sentence *Vyučuje na Vysoké škole ekonomické* (He teaches at University of Economics; tectogrammatical representation of this sentence using the proposed means is displayed in Fig. 4a). In order to mark which nodes belong to which multi-word proper noun expression when recursive embedding occurs, another node attribute NE_coindex is introduced.

Another solution which we consider as acceptable for annotation of these proper noun expressions is to represent the whole expression by only one tree node at the tectogrammatical layer. If we choose this solution, a multi-word proper noun expression is both detected and classified simply by setting a value in the node attribute NE_roles. Whereas in this case multi-word proper noun expressions would be treated in the same vain as one-word proper nouns, the first solution is consistent with the proposal of the next point.

3.3.3 Annotation of complex proper noun expressions

In a complex proper noun expression such as *Frankfurt nad Mohanem*, we distinguish the main part (here *Frankfurt*) and an embedded part (*Mohan*) or parts of such an expression. Each part should be represented by a separate node. The node corresponding to the main part is represented as governing the node corresponding to the embedded part. The type of the complex proper noun expression is indicated by the value of the NE_roles attribute at the governing node. The type of the embedded proper noun is marked by the value of this attribute at the dependent node. The function which the complex proper noun expression plays in the sentence structure is captured by the functor of the governing node. The new

functor NEPART which occurs at the dependent node is used for indicating that the node forms a single complex proper noun expression together with its governing node (to distinguish this case from two mutually independent proper nouns). Recursive embedding is indicated by the value of the NE_coindex attribute. For a sample tectogrammatical representation which follows the proposed annotation principles see Fig. 4b.

3.3.4 Annotation of containers

For the tectogrammatical annotation of a container which consists of two or more proper nouns (such as a person name container whose parts are the first name and the surname), the existing annotation means, the #Idph node, is used. This node becomes the governing node of the whole container. The container type is captured by the value of the NE_roles attribute at this node (i.e., in this attribute type tags as well as container tags can occur). The functor of this node indicates the role of the whole container in the sentence structure. All constituents of the container occur as nodes depending on the #Idph node. Proper noun types of the constituents are defined by the values of their belonging NE_roles attributes. The constituents obtain the new NEPART functor. Annotation of a container is displayed in Fig. 4c.

4 Final remarks

In order to make the text corpora more useful for onomastic research as well as within the NLP domain, an explicit annotation of proper nouns is needed. While several corpora such as BulTreeBank[5] (Osenova and Simov, 2004) or MultiNet semantic networks (Helbig, 2006) allow for proper noun annotation, Czech corpus linguistics has been paying scant attention to this area.

In the present paper, we proposed a complex annotation of proper nouns within the tectogrammatical layer of Prague Dependency Treebank 2.0. The suggested annotation scheme makes it possible to assign one-word proper nouns as well as multi-word proper noun expressions (e.g., *Vysoká škola ekonomická*, University of Economics) but also relatively complicated groups of proper nouns, so called complex proper noun expressions (such as *Frankfurt nad Mohanem*) and containers (for example, *Petr Novák*). In this scheme, the type of a proper noun is captured by a value of the node attribute NE_roles. The inner structure of proper noun expressions or containers is described by the new functor NEPART. Also some other means were introduced. However, some questions concerning the annotation proposal remain still open, e.g., concerning t-lemmas of annotated proper nouns. Besides these topics, also treatment of proper nouns at the morphological layer of PDT 2.0 should be addressed in the near future.

Acknowledgements

I would like to thank Professor Eva Hajičová, Professor Jarmila Panevová and Jarka Hlaváčová for valuable comments on the draft of this paper. Special thanks go to my colleague Zdeněk Žabokrtský for technical assistance and extensive discussion on the presented topic. Work reported on in the paper has been supported by the grants 1ET101120503, GD201/05/H014 and GA UK 7643/2007.

[5] http://www.bultreebank.org

References

Collins, M. and Y. Singer (1999) Unsupervised Models for Named Entity Classification. *Proceedings of the Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC 1999)*, 189-196.

Grishman, R. and B. Sundheim (1996) Message Understanding Conference - 6: A Brief History. *Proceedings of the 16th International Conference on Computational Linguistics (COLING 1996)*, Volume I, 466-471.

Hajič, J. *et al.* (2001) Prague Dependency Treebank 1.0. Philadelphia: Linguistic Data Consortium.

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

Hana, J. and D. Zeman (2005) Manual for Morphological Annotation. ÚFAL Technical Report 2005-27. Prague: MFF UK.

Helbig, H. (2006) Knowledge Representation and the Semantics of Natural Language. Berlin: Springer.

Kang, S.-S. and C.-W. Woo (2003) Unsupervised Learning of Pattern Templates from Unannotated Corpora for Proper Noun Extraction. *Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing, Proceedings of the 9th International Conference (RSFDGrC 2003).* Lecture Notes in Computer Science. Berlin – Heidelberg: Springer, 623-626.

Mikulová, M. *et al.* (2006). Annotation on the tectogrammatical level in Prague Dependency Treebank. Reference book. ÚFAL Technical Report 2006-32. Prague: MFF UK.

Osenova, P. and K. Simov (2004) BTB-TR05: BulTreeBank Stylebook. BulTreeBank Project Technical Report Nr. 05. Available on-line from http://www.bultreebank.org/TechRep/BTB-TR05.pdf (accessed 16 May 2007)

Sekine, S., K. Sudo and C. Nobata (2002) Extended Named Entity Hierarchy. *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC 2002)*, 1818-1824.

Sgall, P. (1967) Generativní popis jazyka a česká deklinace. Prague: Academia.

Sgall, P., E. Hajičová and J. Panevová (1986) Meaning of the Sentence in Its Semantic and Pragmatic Aspects. Dordrecht: D. Reidel Publishing Company.

Ševčíková, M., Z. Žabokrtský and O. Krůza (2007) Zpracování pojmenovaných entit v českých textech. ÚFAL Technical Report 2007-36. Prague: MFF UK.

Ševčíková, M., Z. Žabokrtský and O. Krůza (in press) Named Entities in Czech: Annotating data and Developing NE Tagger. *Proceedings of the 10th International Conference Text, Speech and Dialogue (TSD 2007).*