Towards the Annotation of Named Entities in the National Corpus of Polish

Agata Savary[*†], **Jakub Waszczuk**[◇†], **Adam Przepiórkowski**[†◇]

[*]Université François Rabelais Tours, France
[†]Institute of Computer Science, Polish Academy of Sciences, Warsaw, Poland
[◇]University of Warsaw, Poland
agata.savary@univ-tours.fr, jw235843@students.mimuw.edu.pl, adamp@ipipan.waw.pl

Abstract

We present the named entity annotation task within the on-going project of the National Corpus of Polish. To the best of our knowledge, this is the first attempt at a large-scale corpus annotation of Polish named entities. We describe the scope and the TEI-inspired hierarchy of named entities admitted for this task, as well as the TEI-conformant multi-level stand-off annotation format. We also discuss some methodological strategies including the annotation of embedded, coordinated and discontinuous names. Our annotation platform consists of two main tools interconnected by converting facilities. A rule-based natural language processing platform *SProUT* is used for the automatic pre-annotation of named entities, due to the previously created Polish extraction grammars adapted to the annotation task. A customizable graphical tree editor *TrEd*, extended to our needs, provides an ergonomic environment for manual correction of annotations. Despite some difficult cases encountered in the early annotation phase, about 2,600 named entities in 1,800 corpus sentences have presently been annotated, which allowed to validate the project methodology and tools.

1. Introduction

The development of linguistic resources is one of the key aspects in natural language processing (NLP). Such resources include annotated corpora supporting both linguistic research and data-based applications. The on-going project of the National Corpus of Polish (NKJP for *Narodowy Korpus Języka Polskiego*, http://nkjp.pl/) is meant to create a large annotated versatile corpus of the Polish language. It is designed so as to be representative and balanced with respect to different genres (Przepiórkowski et al., 2009), and assumes several levels of annotation, one of which addresses named entities (NEs). To the best of our knowledge, this is the first attempt at a large-scale corpus annotation of Polish NEs. Its results are expected to boost the automatic Named Entity Recognition (NER) in Polish, similarly to other more resourced languages such as English or French, where NER has been a hot topic for over a decade.

We describe the scope and the taxonomy of NEs foreseen for this task, we recall the multi-level TEI P5-conformant (Burnard and Bauman, 2008) annotation standard (Przepiórkowski and Bański, 2009), and we present the annotation strategies assumed. Then we present the annotation platform constructed for the project. It consists of two main tools interconnected by converting facilities. A rule-based NLP platform *SProUT* is used for the automatic pre-annotation of named entities, due to the previously created Polish extraction grammars adapted to the annotation task. A customizable graphical tree editor, *TrEd*, extended to our needs, provides an ergonomic environment for manual correction of annotations. We show the first results and address difficult cases encountered in the early phase of the annotators' work.

2. Annotation Scope and Taxonomy

The National Corpus of Polish is meant as a general-use linguistic resource. Thus, the annotation scope and NE taxonomy are not oriented towards any particular NLP application. We are interested in the following main name categories:

- Personal names
- Geographical names
- Names of organizations and institutions
- Words related to (most often derived from) the above categories: relational adjectives, names of inhabitants and organization members
- Basic temporal expressions

Initially, we do not annotate other NEs taken into account in other projects: quantities and measures (*25,30 zł, 22,5 kg*), product and vessel names (*Danonki* 'Danone cheese', *Boeing 747*), titles of works (*Przeminęło z wiatrem* 'Gone with the wind', *Mona Lisa*), and events (*Targi Poznańskie* 'Poznań Fairs', *Święto Niepodległości* 'Independance Day', *Hugo* 'Hugo hurricane', *Czernobyl* 'Chernobyl disaster').

Similarly to most other existing approaches to modeling and identification of named entities, we have proposed a taxonomy of these semantically rich units, shown in Fig. 1. It is directly inspired by the TEI P5 guidelines, which is motivated by the fact that TEI is a de facto, constantly maintained XML standard for encoding and documenting textual data, with an active community, and supporting tools. It has particularly detailed and well presented guidelines[1]. Although its

[1] http://www.tei-c.org/release/doc/tei-p5-doc/en/html/index-toc.html

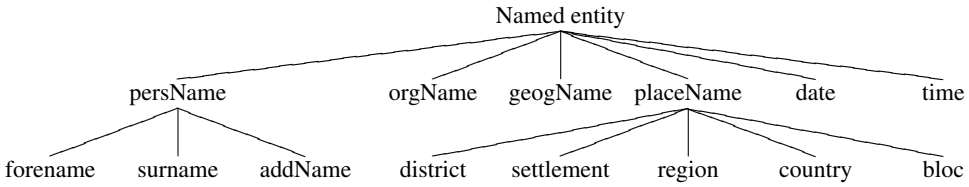

Figure 1: Hierarchy of Polish NEs

recommendations for the encoding of linguistic information are limited, they are sufficient for a corpus annotation project such as ours.

With respect to other NE hierarchies, the subset of TEI chosen for NKJP has a medium degree of granularity appropriate for the scope of our pioneering (with respect to Polish NEs) project. It distinguishes more types and subtypes than MUC (Chinchor, 1997), but many less than the detailed multilevel taxonomies of dozens of categories and relations such as in ACE[2], in the GATE[3] default ontology, or in (Sekine et al., 2002) and (Maurel, 2008).

Differently form, e.g., (Chinchor, 1997), TEI P5 proposes to separate location names within two types called `placeName` and `geogName`. The former is meant for hierarchically-organized geo-political or administrative units (districts, regions, etc.), while the latter refers simply to objects having geographical features such as mountains or rivers. This distinction may be useful because names of administrative units frequently appear as metonyms (designating the inhabitants of the unit), in which case we regard them as organizations rather than locations (cf. section 4.).

In our hierarchy personal names ⟨`persName`⟩ refer to individual persons and families (as opposed to human collective names of organizations), both real and fictional or religious. While TEI defines 6 more detailed elements to distinguish various constituents of a personal name, we retain only those 3 of them which describe autonomous proper names:

- ⟨`forename`⟩ — possibly compound or diminutive: *Marcin*, *Krzysztofowie* 'Christophers'
- ⟨`surname`⟩: *Kaczmarek*, *Kowalscy* 'the Kowalski family'
- ⟨`addName`⟩ — pseudonym, nickname, dynasty, or additional epithet: *Grot* 'Spearhead', *Lwie Serce* 'The Lion Heart'

As mentioned above, names of geographical objects are divided into two TEI-conformant classes: geopolitical names ⟨`placeName`⟩, motivated by administrative territorial divisions, and other geographical names ⟨`geogName`⟩. The first class is further subdivided into 5 subtypes of increasingly general objects: ⟨`district`⟩ (*Żoliborz* 'a Warsaw district'), ⟨`settlement`⟩ (*Warszawa* 'Warsaw', *Nowa Słupia*), ⟨`region`⟩ (*gmina Pisz* 'Pisz commune'), ⟨`country`⟩ (*Republika Czeska* 'Czech Republic'), ⟨`bloc`⟩ (*Wspólnota Europejska* 'European Union'). Names of other geographical objects, such as historical regions, islands, rivers, city objects, forests, mountains, astronomical objects, etc., are not subdivided into fixed categories but can be occasionally specified via additional attributes provided, e.g., by specific lexical resources.

We use the term *NE-related words* with respect to adjectives, names of inhabitants and organization members. They are important for information extraction since they reflect the variability of NEs. For instance, the following expressions should ideally be recognized as equivalent:

(1) [Muzeum Narodowe
w [Warszawie]$_{\text{placeName:settlement}}$]$_{\text{geogName}}$
'National Museum in Warsaw'

(2) [warszawskie]$_{\text{adj}}$
[Muzeum Narodowe]$_{\text{geogName}}$
'Warsaw$_{\text{adj}}$ National Museum'

We annotate only the NE-related words referring to NEs belonging to any of the three previous classes: persons (*chopinowski* 'related to Chopin'), places (*skierniewicki* 'of Skierniewice city', *skierniewiczanin* 'inhabitant of Skierniewice'), and organizations (*microsoftowy* 'by Microsoft', *AK-owiec* 'member of the Interior Army'). Most often these words are morphological derivatives but this is not always the case, e.g. the relational adjective for the *USA* is *American*. Moreover we annotate only relational and not attributive adjectives (even if this distinction is sometimes unclear). For instance, no annotation will be introduced in *klucz francuski* 'wrench, literally: French key'. Note that automatic recognition of NE-related words is difficult due to the frequent lack of the initial capital letter.

As far as temporal expressions are concerned, the TEI guidelines suggest normalizing dates according to an ISO standard, and describing period intervals. A larger coverage of temporal phenomena is done by the TimeML standard (Saurí et al., 2006) addressing not only dates and time expressions, but also events, verb aspects, relative expressions, duration and temporal sets. In our project only the absolute time and date expressions and their normalization are taken into account (*25 marca 1969 r.* '25th March, 1969 ', *504 p.n.e.* '504 BC', *XXI wiek* '21st century', *latem 2000*

[2]http://projects.ldc.upenn.edu/ace/annotation/

[3]http://gate.ac.uk/

'in summer 2000', *12:05*, *pięć po dwunastej* 'five past twelve', etc.)
Note that the above hierarchy is non homogeneous. Personal subtypes correspond to parts of a personal name, while geographical subtypes refer to types of objects they name and their mutual relations. Such heterogeneity is common for many existing taxonomies.

3. TEI-Conformant Annotation Format

As described in (Przepiórkowski and Bański, 2009), the annotation format defined for NKJP is a trade-off between the homogeneity of different annotation levels (morphosyntax, syntactic words, syntactic groups and named entities) as assumed by best practices and proposed standards such as Linguistic Annotation Framework (Ide and Romary, 2004) and PAULA (Dipper, 2005), and the multifarious possibilities given by the TEI guidelines (Burnard and Bauman, 2008). As seen in Fig. 2, entities at different annotation levels are marked as ⟨seg⟩ments, while their inclusions as ⟨ptr⟩ links. The NE annotation level ($\mathtt{L}_{named}$) is based on the level of syntactic words ($\mathtt{L}_{words}$), i.e., the scope of a given NE in text is defined by its ⟨ptr⟩ references to segments defined at the levels $\mathtt{L}_{words}$ and $\mathtt{L}_{named}$ (the latter, in case of embedded NEs). Here, the organization name *Radia Wolna Europa* 'Radio Free Europe', genitive, points to ⟨seg⟩ments `word31` and `word32` at the $\mathtt{L}_{words}$ level (in file `ann_words.xml`), and to ⟨seg⟩ment `ne_phr9` defined just above at the $\mathtt{L}_{named}$ level.
Various properties of annotated NEs are expressed by TEI- and ISO-conformant feature structures which may include:

- Basic type `@neType` and subtype `@neSubtype` (*personName*, *givenName*)
- Orthographic form `@orth` (*Stanów Zjednoczonych* 'United States$_{\text{genitive}}$')
- Base form (lemma) `@base` (*Stany Zjednoczone* 'United States$_{\text{nominative}}$')
- Normalized date or time `@when` (*2009-10-30*, *09:45*)
- Type of derivation `@derivType` (*relAdj* or *persDeriv*)
- Named entity a derivative stems from `@derivedFrom` (*Polska* 'Poland' for *polski* 'Polish')
- Annotation's degree of certainty `@cert` (*high*, *medium*, *low*)
- Comment to the degree of certainty `@certComment` (*Nie wiadomo gdzie jest lewa granica nazwy.* 'Not sure about the NE's left boundary')

Indicating the base form of a NE can be useful in different kinds of corpus studies. For instance, if the corpus is matched against external lexicographic or encyclopedic resources, base forms of NEs usually have the role of lookup entries. Moreover, as shown in (Piskorski et al., 2009), automatic lemmatization of multiword NEs is a particular challenge in highly inflected languages such as Polish. We believe that a rich resource such as the National Corpus of Polish can be a good training set for such methods.
The last two features listed above are introduced for methodological reasons. A 1-million word gold-standard subcorpus is being obtained by an automatic annotation and subsequent manual correction. Such a labor-intensive task needs validation in itself. The usual way to achieve good quality, adopted in the current project, is to cross the results produced by two different annotators. Additionally, an annotator may score her/his degree of confidence in her/his own judgment, and explain the source of doubts, in order to be able to discuss uncertain annotations within the project team.
Note that a NE often spans over the same fraction of text as some correct noun phrase. That's why some features important for NEs, notably the multi-word ones, will not be annotated at the $\mathtt{L}_{named}$ level but at the level of syntactic groups. Such features include the disambiguated morphosyntactic tag, and the pointer to the phrase's headword (if any). In case of no correspondence between a NE and any syntactic group, we suppose that the morphosyntax of the NE can be deduced later, largely automatically, from its lemma, and the underlying level of syntactic words.

4. Annotation Strategies

One of the interesting annotation strategies of our project is not to limit ourselves to the longest-match approach but to identify a NE together with all its embedded NEs. That makes the annotation more dense and informative with respect to reasoning about relations between different objects. Thus, further NLP tasks such as co-reference annotation and information extraction can be facilitated. In some cases, the embedding depth may be high, as in:

(3) [[Rozgłośnia [Polska]$_{\text{adj}}$]$_{\text{orgName}}$ [Radia Wolna [Europa]$_{\text{geogName}}$]$_{\text{orgName}}$]$_{\text{orgName}}$
'Polish Radio Station of Radio Free Europe'

We are also interested in a high-quality annotation of coordinated and discontinuous names. For instance, in

(4) *Ameryka Północna i Południowa* 'North and South America'

we wish to render two occurrences of geographical names, with one token being shared. Similar problem arises in duration expressions, which we presently do not annotate, but we still wish to detect temporal moments included in them, as in *od 23 do 25 marca* 'from the 25th to the 27th of March'. Other discontinuities occur when extra elements are inserted into a NE, as in:

```
<teiCorpus xmlns:xi="http://www.w3.org/2001/XInclude" xmlns="http://www.tei-c.org/ns/1.0">
 <xi:include href="NKJP_header-main.xml"/><TEI><xi:include href="header.xml"/><text><body>
      <!-- Previous paragraphs here -->
      <p xml:id="ne_p1" corresp="ann_words.xml#words_p1">
        <!-- Previous sentences here -->
        <s xml:id="ne_s3" corresp="ann_words#words_s3">
          <!-- Previous tokens here -->
          <seg xml:id="ne_phr9">
            <fs type="named">
               <f name="neType"><symbol value="geogName"/></f>
               <f name="orth"><string>Europa</string></f>
               <f name="base"><string>Europa</string></f>
               <f name="cert"><symbol value="high"/></f>
            </fs>
            <ptr target="ann_words.xml#word33"/>  <!-- Europa -->
          </seg>
          <seg xml:id="ne_phr10" corresp="ann_groups.xml#groups_g8">
             <fs type="named">
               <f name="neType"><symbol value="orgName"/></f>
               <f name="orth"><string>Radia Wolna Europa</string></f>
               <f name="base"><string>Radio Wolna Europa</string></f>
               <f name="cert"><symbol value="high"/></f>
             </fs>
             <ptr target="ann_words.xml#word31"/>  <!-- Radia -->
             <ptr target="ann_words.xml#word32"/>  <!-- Wolna -->
             <ptr target="#ne_phr9"/>              <!-- Europa -->
          </seg><!-- Next tokens here -->
        </s><!-- Next sentences here -->
      </p><!-- Next paragraphs here -->
 </body></text></TEI>
</teiCorpus>
```

Figure 2: Final Annotation Format for a sentence extract *Radia Wolna Europa* 'Radio Free Europe'

(5) *Wydział Matematyczny ówczesnej Akademii Krakowskiej* 'Department of Maths of the former Cracow Academy'

While the two embedded organization names can easily be bracketed pretty much as in example (3), the overall organization NE annotation should not contain the underlined segment.

As far as metonymy is concerned, we (unlike, e.g., the MUC standard) admit that the actual category to be annotated is the one that fits to the occurrence's context. For instance, in

(6) $[Niemcy]_{\text{orgName}}$ *zaatakowały* $[Polskę]_{\text{orgName}}$ 'Germany attacked Poland'

both country names represent human collectives (people living in both countries) rather than geopolitical territories. Thus, they are annotated as organizations, although their 'primary' category is a place name. Similarly, names of buildings, districts, cities, regions, blocs, etc., can be marked as geographical/geopolitical or as organization-bound, if the people living or working in them are concerned.

Note that personal names function in some contexts as names of historical periods and, as such, are not presently annotated:

(7) *kampingi w stylu późny Bierut, wczesny Gierek* 'campings in the late-Bierut-early-Gierek style' (where Bierut and Gierek are ex-leaders of the Polish leading communist party)

5. Difficult Cases

In the early phase of the annotation process we have encountered several linguistically difficult cases, which do not perfectly fit the admitted methodology. The type of a NE was sometimes hard to determine, despite the occurrence context. For instance, in example (8) *Princeton* could be annotated either as a settlement or as organization (elliptical variant of Princeton University). We chose the latter hypothesis as more informative.

Most other problems concern the annotation of relational adjectives and inhabitant members. The examples in (9) are clearly proper names but their attribution to a particular geopolitical unit seems problematic. We propose to mark them as personal derivations, without indicating the value of the `@derivedFrom` attribute. In (10) the morphological base is *America*, however semantically both adjectives can sometimes refer to the *United States*. The additional problem is to determine the canonical form of this country name (*United States of America, USA, US*, etc.). We propose to choose the

correct referent according to the context (if possible), and to indicate the non-acronym form judged as the most common (here: *United States*).

Example (11) refers to the area with its dominating point at the *Święty Krzyż* ('Saint Cross') summit. The names of both the mountain range (*Góry Świętokrzyskie*) and the surrounding province (*województwo świętokrzyskie*) stem from this base. When the adjective appears in other contexts, it hardly ever refers to the summit itself, but rather most frequently to the province. Thus, the value of the `@derivedFrom` attribute is a form that contains the very adjective that is described as 'stemming from' it. Similarly, example (12) can refer to the continent of Europe, but is more frequently used in relation with *Unia Europejska* ('European Union').

(8) *Pozostała mu jedynie praca naukowa w Princeton.* 'All that was left to him was the research activity at/in Princeton'

(9) *Żyd*, *Arab* 'Jew, Arab'

(10) *amerykański*, *antyametykański* 'American, anti-american'

(11) *świętokrzyski*

(12) *europejski* 'european'

6. Annotation Tools

The NKJP corpus is logically divided into two parts: a high-quality manually annotated 1-million-word subcorpus, and an automatically annotated 1-billion-word main corpus. Currently, most annotation efforts have concentrated on the 1-million word gold standard subcorpus. It is being automatically pre-annotated by knowledge-based methods, then the annotations are manually corrected and completed by linguistic experts.

6.1. Data Flow

The data flow in the 1-million-word subcorpus is shown in Fig. 3. The left-hand side presents different annotation levels in NKJP. Contrary to what was mentioned in section 3., the level of named entities is presently built upon the level of morphosyntax and not of syntactic words. This is because the annotation at the latter level (Głowińska and Przepiórkowski, 2010) is being performed in parallel to NE annotation. Note, however, that the annotation format includes links from words to tokens on one side, and from NEs to tokens on the other side (cf. Fig. 2). Thus, we plan to translate links from NEs to tokens into those from NEs to words as soon as both annotation tasks are complete. This process can be fully automatic unless some NEs cross the boundaries of syntactic words, which seems rather improbable.

A raw text taken from the corpus repository is processed by lexical resources and grammar rules within the *SProUT* platform, as discussed in section 6.2. The

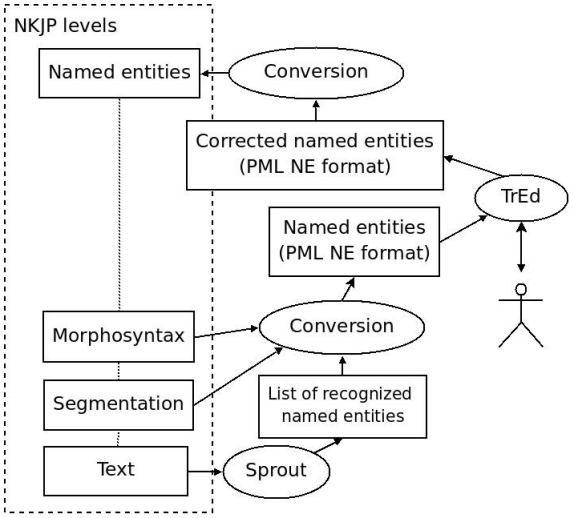

Figure 3: Data flow in the NE annotation task of the NKJP corpus

recognized NEs together with their embedded structures are included in an XML *SProUT*-proper output. This output is further converted into another XML format, called PML-NE, defined for the annotating platform *TrEd* (see section 6.3.). Since *SProUT* outputs the cardinal numbers of the beginning and ending characters of each recognized sequence, the converter consults the segmentation level of the text in order to translate text ranges into token identifiers. Moreover, for each token, its morphological tag and lemma are recopied from the morphosyntactic annotation of the text. As a result, the PML-NE input for *TrEd* contains perfectly correct morphological data (up to human annotators' mistakes at the morhosyntactic level) for tokens, and automatically created NE trees to be corrected.

Due to methodological constraints defined for the composition of the NKJP corpus, the files contained in the 1-million gold standard subcorpus are of a very variable length (from several to several thousand sentences). They do not correspond to complete texts taken from the 1-billion word corpus, but to randomly chosen paragraphs thereof. For the sake of ergonomy, it is important to present the human annotator with relatively small, thus easily manageable, portions of text. Therefore, the converter divides each text into files of a limited number of sentences corresponding to circa 1 hour of human annotation effort (presently about 100 sentences). Additionally, the division is designed so as to keep together all sentences appearing in one paragraph.

At the next stage two human annotators work on each corpus fragment. A super-annotator reviews the cases of disagreement and chooses the correct annotation. Each annotator and super-annotator works off-line with *TrEd* installed locally. She consults remote project repositories in order to get

new versions of NKJP extensions for *TrEd*. She also has an access to a remote private subversion repository, where files to be annotated are stored. Using *TortoiseSVN* (subversion client application, `http://tortoisesvn.tigris.org/`) she downloads new files, stores working versions and commits final results to dedicated directories. Despite no particular computing background of the annotators, they seem to install and operate the whole annotating platform rather easily. Except for the annotator's handbook, and the *TrEd* and *TortoiseSVN* user's guides, additional help to the annotators comes from regularly updated compilation of Frequently Asked Questions, as well as from a discussion list.

The last stage (presently under development) consists in converting the PML-NE format of the validated annotations into the final NKJP format described in section 3. Here, the subfiles have to be merged into files corresponding to the initial texts, embedded NEs get transformed into pointers, and links to morphological tokens are shifted to the level of syntactic words.

6.2. Automatic Pre-annotation with *SProUT*

The automatic pre-annotation is performed by a general-purpose multi-lingual NLP platform *SProUT* (Drożdżyński et al., 2004). It offers Unicode-capable processing components for basic linguistic operations (segmentation, morphological analysis, gazetteer lookup) and a cascaded unification-based finite-state grammar parser and interpreter. For each recognized sequence it outputs a feature structure whose attributes and types can be defined by the user. *SProUT* has been adapted to processing Polish texts by (Piskorski et al., 2004). The development of Polish lexical resources and grammars for extracting 'classical' named-entities (e.g., names of persons, organizations, locations, etc.) from Polish texts is addressed in (Piskorski, 2005).

These resources and grammars, meant for an information retrieval (IR) task, had to be adapted to a corpus annotation task (Savary and Piskorski, 2010), in accordance with the needs of NKJP. In particular, we had to redesign the rules so that the output structures contain the features of all NEs embedded in the longest-match sequences. Conversely, the granularity of the output was reduced as far as fine-grained subtypes, and encyclopedic data (useful in IR but not in NKJP) are concerned. However, the particular impact on the correct lemmatization of compound names present in the previous grammars has been retained and reinforced in our grammars. We have also enlarged the Polish gazetteers for *SProUT* by adding circa 80,000 inflected forms of surnames, geographical and geopolitical names.

The first results show a precision of the adapted knowledge-based annotation ranging from 0.71 for organization names to 0.92 for personal names. The recall is much lower: from 0.14 for organizations to 0.71 for persons. When calculating these results an annotation was considered as correct only if all of its properties had been correctly determined: (i) its left and right boundaries, (ii) its type, (iii) its lemma. In terms of corpus pre-annotation prior to manual correction, this corresponds to a case when no action is required from the human annotator. Of course only partially correctly recognized NEs also help limit her intervention, as shown below.

While precision is essential for IR, a higher recall (even with a lower precision) is probably helpful in corpus annotation, since it comes easier to delete annotations done automatically than to add new annotations manually. Therefore, we wish to experiment with relaxing some constraints put on grammar rules. Together with further enrichment of gazetteers that should result in an even bigger reduction of human annotators' effort.

6.3. Manual Annotation with *TrEd*

The manual correction and completion of annotations is one of the most labor-intensive and expensive tasks in a corpus project, and needs optimal annotation tools. We have evaluated several annotation platforms before selecting the tree editor *TrEd* (Pajas and Štěpánek, 2008)[4] for the following facilities: (i) allowing for a pre-annotated input and maintaining a multi-level annotation (recall that in the NE annotation task we need to refer to the underlying annotation levels for segmentation and morphology), (ii) customizable open XML-based abstract data format, called PML (allows for free definition of feature structures), (iii) facility of building and modifying tree representations (needed for embedded structures), (iv) ability to represent coordinated and discontinuous structures such as (4) and (5), (v) ergonomic graphical user's interface customizable with macros and keyboard shortcuts, (vi) parallel edition of concurrent annotations (vii) rich documentation and good author's responsiveness, (viii) technically reliable installation, configuration, and operation.

Figure 4 shows the *TrEd* interface with a particularly difficult example sentence whose mostly erroneous annotation results from a conversion of the *SProUT* output to the PML-NE format. The lowest level of the main window contains sentence tokens (here: *Współpracował*, *z*, etc.). The highest level shows the longest-match NEs (here *France Nationale* of type *persName*, etc.). All other intermediate levels represent embedded NEs. Most essential node attributes — main type, subtype (if any), base form, and certainty level — are visible under the node. For instance, the father of node *Polską* has the main type *placeName*, the subtype *country*, the (incorrect) lemma *Polska*, and the certainty level *medium* (all annotations coming from *SProUT* initially have `@cert="medium"`).

Basic actions such as node selection, node or edge deletion, and going to the next or an arbitrary sentence, are available on *TrEd* installation. Moreover, our own extension macros allow to make the following actions

[4] http://ufal.mff.cuni.cz/~pajas/tred/

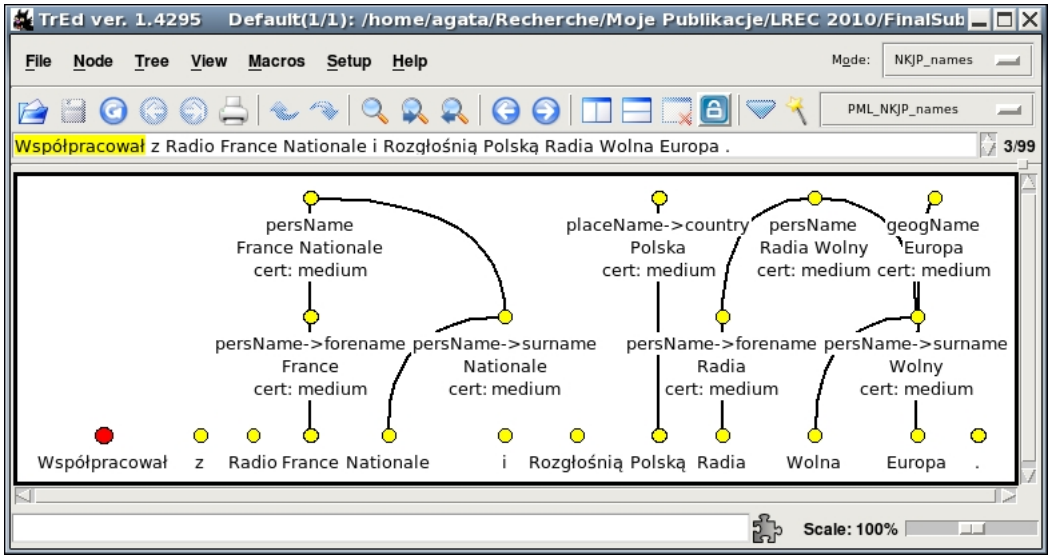

Figure 4: *TrEd* interface showing a *Sprout*-pre-annotated sentence 'He collaborated with Radio France Nationale and the Polish Broadcasting Station of the Free Europe Radio'

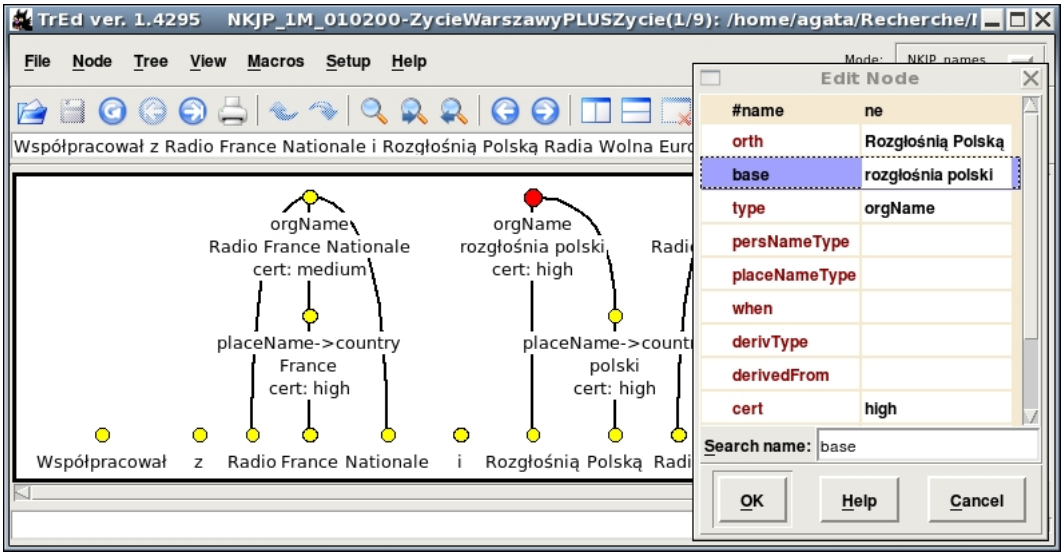

Figure 5: Editing node attributes in *TrEd*. Base form components are recopied from the morphosyntactic level

with simple combinations of a mouse click and/or a keyboard shortcut: inserting a node with a predefined type, adding or deleting a subtype, adding a new father to several nodes at a time, and adding or deleting a secondary edge (e.g., in a coordinated structure). Attribute values belonging to a fixed list can be changed rapidly by a cyclic mouse-clicking. Otherwise attributes can by edited in the attribute window (see Fig. 5). The correction of text attributes, such as the base form, are facilitated due to the fact that lemmas of components are automatically recopied from the underlying level of the morphosyntactic annotation (here, *rozgłośnia polski* is the concatenation of lemmas of *Rozgłośnią* and *Polską*, and should be corrected to *Rozgłośnia Polska*). Figure 6 shows the complete correction of the initial annotation with multiply embedded NEs.

7. Perspectives

The annotation of NEs in the National Corpus of Polish is in progress. Until mid-March 2010 the first corpus fraction processed by two annotators (not yet validated by the super-annotator) contains about 27,500 tokens, 1,800 sentences and 2,600 named entities. Since the difficult linguistic and technical cases arising in this early phase could be successfully solved with no change to the admitted format and strategies, we consider the initial design of the annotation task as validated. However, in case some minor evolution of the admitted methodology is necessary, we must allow backtracking to previously annotated fragments. This aspect, together with the management of the annotation throughput and quality is one of our main future challenges.

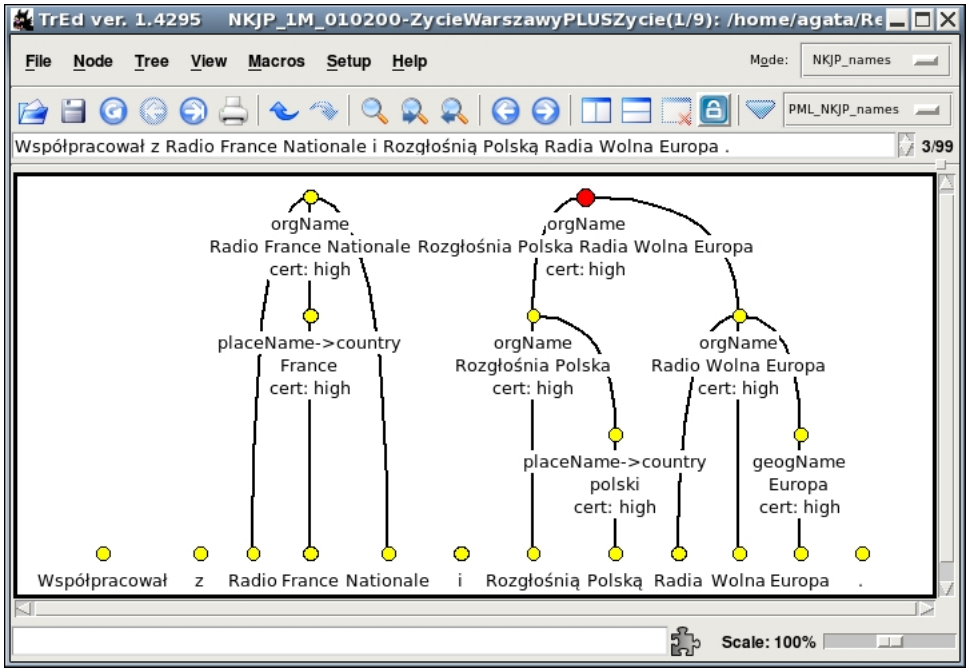

Figure 6: Complete sentence annotation with multiply embedded NEs

8. Acknowledgement

The research was funded in 2007-2010 by a research and development grant R17-003-03 from the Polish Ministry of Science and Higher Education.

9. References

Lou Burnard and Syd Bauman, editors. 2008. *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. Oxford.

Nancy Chinchor. 1997. MUC-7 Named Entity Task Definition. In *Proc. of MUC-7*.

Stefanie Dipper. 2005. Stand-off Representation and Exploitation of Multi-Level Linguistic Annotation. In *Proceedings of Berliner XML Tage 2005 (BXML 2005)*, pages 39–50, Berlin.

Witold Drożdżyński, Hans-Ulrich Krieger, Jakub Piskorski, Ulrich Schäfer, and Feiyu Xu. 2004. Shallow Processing with Unification and Typed Feature Structures – Foundations and Applications. *Künstliche Intelligenz*, 1/04.

Katarzyna Głowińska and Adam Przepiórkowski. 2010. The Design of Syntactic Annotation Levels in the National Corpus of Polish. In *Proceedings of LREC 2010, Malta, to appear*.

Nancy Ide and Laurent Romary. 2004. International standard for a linguistic annotation framework. *Natural Language Engineering*, 10:211–225.

Denis Maurel. 2008. Prolexbase. A multilungual relational lexical database of proper names. In *Proceedings of LREC'08, Marrakech, Marocco*.

Petr Pajas and Jan Štěpánek. 2008. Recent Advances in a Feature-Rich Framework for Treebank Annotation. In *Proceedings of COLING'08, Manchester*.

Jakub Piskorski, Petr Homola, Małgorzata Marciniak, Agnieszka Mykowiecka, Adam Przepiórkowski, and Marcin Woliński. 2004. Information Extraction for Polish Using the SProUT Platform. In *Proceedings of IIS'04, Zakopane, Poland*.

Jakub Piskorski, Karol Wieloch, and Marcin Sydow. 2009. On knowledge-poor methods for person name matching and lemmatization for highly inflectional languages. *Information Retrieval*, 12(3):275–299.

Jakub Piskorski. 2005. Named-Entity Recognition for Polish with SProUT. In *LNCS Vol 3490: Proceedings of IMTCI 2004, Warsaw, Poland.*

Adam Przepiórkowski and Piotr Bański. 2009. Which XML standards for multilevel corpus annotation? In *Proceedings of LTC'09*, Poznań, Poland.

Adam Przepiórkowski, Rafał L. Górski, Marek Łaziński, and Piotr Pęzik. 2009. Recent developments in the National Corpus of Polish. In *Proceedings of Slovko'09, Smolenice/Bratislava, Slovakia*. Tribun.

Roser Saurí, Jessica Littmana, Bob Knippen, Robert Gaizauskas, Andrea Setzer, and James Pustejovsky. 2006. TimeML Annotation Guidelines Version 1.2.1.

Agata Savary and Jakub Piskorski. 2010. Lexicons and Grammars for Named Entity Annotation in the National Corpus of Polish. In *IIS'10, Siedlce, Poland (submitted)*.

Satoshi Sekine, Kiyoshi Sudo, and Chikashi Nobata. 2002. Extended named entity hierarchy. In *Proceedings of LREC'02*, Canary Island, Spain.