SEMANTIC AND SYNTACTIC PATTERNS OF BULGARIAN MULTIWORD NAMES

Svetla Koeva, Tsvetana Dimitrova

Department of Computational Linguistics, Institute for Bulgarian Language

The Task

We aim at extending the resources for named entity recognition and classification for Bulgarian by developing a set of rules, expanding available gazetteers and building a corpus with manually annotated NEs.

Bulgarian Multiword Names

Proper nouns are logically connected to the object to which they refer and inherit its characteristics which may be morphologically expressed if the language structure allows it.

The triggers semantically depend on the referent of the proper noun.

Semantic classes

Persons, locations, organisations and their triggers are grouped into semantic classes sharing common semantic patterns (language independent, to a great extent).

Names (example)

Personal names that select triggers – academic titles, can be only (first) names modified with or substituted by family names, while the personal names that select aristocratic titles, can be only first names.

Triggers (example)

The triggers denoting academic positions are related to the semantic pattern: in-something; at-somewhere (profesor po himiya v Sofiyskiya universitet 'professor in chemistry at the Sofia University').

The triggers of person NEs

(1) Legislative job positions: ministăr-predsedatel 'prime minister';

(2) Executive job positions: izpălnitelen direktor 'executive officer';

(3) Aristocratic titles: princ 'Prince';

(4) Judicial positions: sădiya 'judge';

(5) Academic positions: glaven asistent 'senior assistant';

(6) Academic titles: doktor na filolologičeskite nauki 'Doctor of Philology';

(7) Military ranks: general mayor 'Major General';

(8) Religious titles: mitropolit 'Metropolitan'; etc.

The Rules

The rules define NE class, PoS and grammatical subclass of the head noun, PoS of NE constituents, dependencies between constituents, cliticization (possessive and interrogative clitics), contiguity, word order permutations.

The rules account for: single or compound proper nouns, proper noun phrases without triggers, and proper noun phrases extended with triggers.

The Gazetteers

Different lexicons are constructed, each containing a set of words with common semantic, morphological and syntactic features, i.e., nobility particles, feminine family names, military titles etc.

Example

```
<group><or><star><group>
<star><e pr="A.*"/></star><e d="TITLES_DEF"/><star><group>
<or><e w="по"/><e w="в"/><e w="при"/><e w="от"/><e w="на"/><e
w="към"/></or>
<star><e pr="A.*"/></star><e pr="N.*"/></group></star>
<star><group><star><e pr="A.*"/></star><e
d="TITLES_INDEF"/></group></star>
</group></star></or>
<star><e d="PERSONAL_MASC_NAME"/></star>
<e d="FAMILY_MASC_NAME"/>
<star><e d="FAMILY_MASC_NAME"/></star>
</group>
```

The Corpus

The annotated corpus is excerpted from the Bulgarian National Corpus (texts of genres such as news, fiction, popular science, subtitles have been selected), and, at the moment, it consists of 101,507 tokens. Sentence boundaries, PoS tags and lemmas are automatically annotated. The MWNs (721) are manually annotated with their major constituents and features – proper names, triggers, proper name heads, triggers heads, and semantic classes of proper names and triggers.