STSM: Semantic and Syntactic Patterns of Multiword Names

Participant: Svetla Koeva

Institute for Bulgarian Language, BulgarianAcademy of Sciences, Bulgaria

Host: Tita Kyriacopoulou, Université Paris-Est Marne-la-Vallée, France

Reference number: COST-STSM-ECOST-STSM-IC1207-260217-082014

1. Purpose of the STSM

This report describes the PARSEME IC1207 STSM, which took place from February 26 to March 5, 2015 (one week) at the Université Paris-Est Marne-la-Vallée: (to visit Tita Kyriacopoulou and the research group at Laboratoire d'Informatique Gaspard-Monge). The goal of the STSM was to work on two relevant for the PARSEME IC1207 Cost action topics: the elaboration of semantic and syntactic patterns of multiword names and the developing of reliable resources for multilingual recognition and classification of names (both single- and multiword).

The mission took place in the context of an ongoing and planned long-term collaboration between the visiting researcher and the host research group. Tita Kyriacopoulou, Claude Martineau and Svetla Koeva are co-authors (together with Cvetana Krstev, Dusko Vitas and Tsvetana Diitrova) of the book chapter *Semantic and Syntactic Patterns of Multiword Names: A Cross-Language Study*. The activities described below were carried out in collaboration with Tita Kyriacopoulou and Claude Martineau in a close interaction with Cvetana Krstev and Dusko Vitas.

The previous collaboration resulted in a general model for description and classification of proper names in different languages. In particular, semantic patterns for personal, location and organisation names are formulated and the corresponding syntactic patterns of single- and multiword names in Bulgarian, English, French, Greek, and Serbian are described.

Names (personal, location and organization) are grouped into different semantic classes and subclasses with respect to the properties of their referents (explicated by triggers – common nouns, such as *director, uncle, street*). A name from a given semantic class selects triggers from a particular set of semantic subclasses. For example, a first name (but not a family name) can be extended with a kinship term (i.e., *his beautiful step-daughter from France, **Anne Nicole***) and the kinship term can be specified in particular ways (different from the permissible specifications of other semantic subclasses), thus the respective semantic pattern is: (modifier: referent specification phrase) – trigger: kinship term – (modifier: possessor phrase) – (modifier: location phrase) – personal name. The permissible combinations between types of names (proper nouns, multiword expressions (MWEs)), and semantic subclasses of triggers determine the semantic patterns applicable to the personal, location and organization names.

The initial plan envisaged the evaluation of semantic patterns and further elaboration of syntactic patterns of **personal names for Bulgarian, English, French, and Greek**. During the STSM these tasks were extended to: evaluation of semantic patterns for **names (persons, locations and organizations)** and detailed elaboration of corresponding syntactic patterns valid for **Bulgarian,**

English, French, Greek and Serbian. The ultimate aim is also to provide reliable resources for multilingual recognition and classification of names (both single- and multiword). The known approaches so far are concentrated on the recognition of named entities and usually do not recognize the complex syntactic structures containing names and their triggers. Our approach will contribute to the more precise identification of co-reference chains including names and triggers: *his beautiful step-daughter from France, **Anne Nicole; Anne Nicole; his** step-daughter; etc.*

2. Description of the work carried out during the STSM

1. The semantic patterns of names were evaluated based on empirical evidences. In particular, specifications for domain and affiliation were included in the semantic pattern describing organization names with an internal trigger.
2. The syntactic patterns were elaborated to include the domain specification alternation, the location alternation and the noun modifier – prepositional phrase possessive alternation. For example, the English syntactic pattern for a personal name extended or substituted by a family name and combined with a trigger from one of the following semantic classes: legislative job title, executive job title, judicial position, academic position, academic title, military rank, and profession, is elaborated, as follows:
(((DefArt | GenDet | PossPron) Adj* Trigger (PP in|of = DomSpec)? ((PP at = Aff)* | (PP in = Loc) | (PP from = Org))?) | (DefArt Trigger (PP in|of = DomSpec)? ((PP at = Aff)* | (PP in = Loc) | (PP from = Org))?) | ((DefArt)? Trigger (PP in|of = DomSpec)? ((PP at = Aff)* | (PP in = Loc))?) | IndefTrigger) PerN
3. A parallel Bulgarian – English – French – Greek – Serbian corpus will be compiled and annotated with the semantic and syntactic patterns of proper names. The previous experience in the annotation of proper names was studied. The methods used for the development of a parallel corpus comprising the novel *Le tour du monde en quatre-vingts jours* (Jules Verne, 1872) and its translations in English, German and Serbian (Latin alphabet) (Lecuit et al. 2015) were discussed. In this work, the proper names (as well as the relational adjectives and nouns) have been annotated with previously developed transducers. The tags used to annotate the text are taken from the list of tags edited by the Text Encoding Initiative Consortium (TEI P5). For example, the following elements in the French text were annotated:
<name type="person">[human]</name> (1856 items)
<name type="org">[organization]</name> (115 items)
<name type="geographical">[natural geographical location]</name> (201 items)
<name type="building">[construction humaine]</name> (68 items)
<name type="place">[région administrative, ville]</name> (836 items)
<w type="relational noun">[relational noun]</w> (197 items)
<w type="relational adjective">[relational adjective]</w> (161 items)
Svetla Koeva, Cvetana Krstev, and Claude Martineau presented existing tools (regular expressions and transducers) for automatic annotation of proper names in Bulgarian, French, Greek and Serbian (Koeva and Genov, 2011; Krstev et al, 2013; Kyriacopoulou et al. 2013).
4. An annotation scheme based on the validated semantic patterns and elaborated syntactic patterns (for Bulgarian, English, Greek, Greek and Serbian) was discussed. The purpose of the annotation scheme is to develop an annotated multilingual resource that will serve for: validation of rule-based methods for multilingual named entity recognition and classification; training and testing corpus for machine learning methods for multilingual named entity recognition and classification (for example, using the MWEs toolkit – http://mwetoolkit.sf.net), including neural networks (for example, the open source platform TensorFlow – https://www.tensorflow.org/); training and testing

corpus for methods identifying co-reference chains. As the MWEs toolkit (and some other multiword expression parsers) allows adding external resources to guide tagging decisions – dictionaries of proper names and triggers developed for Bulgarian, English, French, Greek and Serbian can be used for this task.

3. Description of the main results obtained

Below the main outcomes of the STSM are described. We also list concrete outcomes that were not initially planned.

1. Eight semantic patterns of names (persons, locations and organizations) were evaluated.

2. Forty syntactic patterns were elaborated to include the domain specification alternation, the location alternation, the noun modifier – prepositional phrase possessive alternation, and the restrictive apposition of a proper noun whose omission changes the meaning of the sentence (i.e., *the new Professor of Law, **Chris Smith***).

3. The development of a large Bulgarian – English – French – Greek – Serbian corpus annotated for names (single-, multiword and phrases) is an ongoing work. The corpus will contain novels, i.e. *Le tour du monde en quatre-vingts jours* of Jules Verne, translated in five languages; wikipedia articles in five languages – https://dumps.wikimedia.org/, and news articles in five languages – http://www.monde-diplomatique.fr/, http://bg.mondediplo.com/, http://mondediplo.com/, www.monde-diplomatique.gr, mondo.rs/a989928/Info/Srbija/. Since it is difficult to collect enough parallel texts in five languages, some parts of the corpus will contain parallel texts in four languages (SETimes, the Parallel Corpus of English and South-East European Languages – http://nlp.ffzg.hr/resources/corpora/setimes/; EU legislative documents – http://eur-lex.europa.eu/collection/eu-law.html; etc.). The corpus will be semi-automatically annotated with the available tools for annotation, adjusted to mark the names and triggers with compliance to the annotation scheme.

4. The annotation scheme for semantic and syntactic annotation is elaborated. Different elements from the semantic and syntactic patterns are used as annotation attributes and values. The scheme consists of four levels.

The first level is morpho-syntactic and includes annotation of basic part of speech tags and grammatical characteristics of the following classes: proper nouns, nouns, adjectives, possessive pronouns, demonstrative pronouns, articles, prepositions, conjunctions, numerals and following categories where applicable: gender, number, definiteness, case. The annotation will be done automatically.

The second level is the syntactic annotation and it involves marking of noun phrases and prepositional phrases – parts of proper name phrase and trigger phrase. The shallow parsing will be done automatically with the existing tools accordingly adjusted if necessary. Manual post-editing will be performed.

The third level of annotation is semantic. The main outcome of the STSM is the definition of the 7 semantic tags for single personal names, 12 semantic tags for single personal name triggers (single words, multiword expressions or phrases), 6 semantic tags for personal name trigger specifiers (both single words or phrases), 5 semantic tags for multiword personal names; 2 semantic tags for single location names, 4 semantic tags for location name triggers (single words, multiword expressions or phrases), 4 semantic tags for location name trigger specifiers (both single words or phrases), 3 semantic tags for multiword location names; 2 semantic tags for single organization names, 5 semantic tags for organization name triggers (single words, multiword expressions or phrases), 5 semantic tags for organization name trigger specifiers (both single words or phrases), 3 semantic tags for multiword organization names. For example, the semantic tags for the organization name triggers are:

<specification type=“organization”>[referent]</specification>
<specification type=“organization”>[possessor]</specification>
<specification type=“organization”>[affiliation]</specification>
<specification type=“organization”>[domain]</specification>
<specification type=“organisation”>[location]</specification>
The shallow semantic parsing will be done automatically with the existing tools adjusted accordingly. Manual post-editing will be performed.
At the top level the whole phrase will be marked and the TEI compliant tags for the class of the proper name will be used.

4. Future collaboration with host institution

The STSM led to new directions of the joint research. They concern the semiautomatic development of multilingual annotated corpus, cross-lingual application of machine learning methods for detection of names (single- and MWEs), cross-lingual identification of co-reference chains. This determines a long-term collaboration between the visiting researcher and the host research group.
There will be at least one paper, which will be the direct outcome of the joint work. We aim at submitting it to LREC 2018. The paper will describe the development of the annotated corpus representing the semantic and syntactic patterns describing the structure of proper names.

5. Confirmation by the host institution of the successful execution of the STSM

Tita Kyriacopoulou: The present report was reviewed and accepted. Svetla Koeva spent one week at the Université Paris-Est Marne-la-Vallée. During the successfully executed visit, we have made progress on description of semantic and syntactic properties of multiword names. We began the development of a multilingual corpus annotated with single- and multiword names which will serve as training and testing corpus in named entity recognition and classification task.

References

Lecuit É., Maurel D., Vitas D. (2015). A Multilingual Corpus for the Study of Toponyms in Translation. In Schnabel-Le Corre B., Löfström J. *Challenges in Synchronic Toponymy: Structure, Context and Use*. Francke A. Verlag. 235-246.
Koeva., S. and Genov A. (2011) Bulgarian Language Processing Chain. *Integration of multilingual resources and tools in Web applications Workshop*. Hamburg, Germany.
Krstev, C., A. Zečević, D. Vitaš., T. Kyriacopoulou, (2013) *NERosetta* – an Insight into Named Entity Tagging, *Proceedings of 6th Language & Technology Conference*, December 7-9, 2013, Poznań, Poland, ed. Zygmunt Vetulani & Hans Uszkoreit, Fundacja Uniwersytetu im. A. Mickiewicza, Poznań. ISBN 978-83-932640-3-2, pp. 168-172.
Kyriacopoulou, T., C. Martineau, T. Mavropoulos. (2013) Les noms propres de personne en français et grec : reconnaissance, extraction et enrichissement de dictionnaires, *30e Colloque International sur le Lexique et la Grammaire*, Nicosie, Chypre, septembre 2011.