Blois, May 3, 2007

Report on the work session of "Polonium" project

Date and place: IPI PAN, Warsaw, April 16-17, 2007

Were present: Béatrice Bouchou, Elżbieta Hajnicz, Małgorzata Marciniak, Denis Maurel, Agnieszka Mykowiecka, Adam Przepiórkowski, Agata Savary, Marcin Woliński, Joanna Rabiega-Wiśniewska, Łukasz Dębowski

This report is addressed to: the persons present, and Nathalie Friburger

This report has been edited by Agata Savary

1. Project workshop on Monday, April 16. List of presentations:
    a. Denis MAUREL - *Prolexbase: a Relational Multilingual Database of Proper Names*
    b. Béatrice BOUCHOU - *ISO Standard XML Models of Proper Names*
    c. Marcin Woliński - *The tools for morphological description of Polish developed and used at IPI PAN*
    d. Agnieszka Mykowiecka - *IE Grammars for Person and Institution Names in Polish*

2. Work session on Tuesday, April 17:
    a. Marcin Woliński – demonstration of the *Polish Grammatical Dictionary* (234 000 entries encoded up till now, a user friendly interface enables to check the encoded lemmas on the fly) and of the Prolog-driven syntactic analyzer of Polish.
    b. Marcin Woliński – demonstration of the existing resources for Polish proper names. About 5000 uni-word encoded entries have been extracted from the SQLite data base and sent to Agata. This resource needs to be further extended (by the Polish party). It will also serve as starting point for the encoding of multi-word proper names (via Multiflex on the French side).
    c. Denis Maurel – demonstration of the corpus processor Unitex : detecting the sentence boundaries, searching for concordances with regular expressions and graph patterns, creating and maintaining electronic dictionaries and lexicon grammars, morphological analysis and disambiguation, etc.

3. Perspectives :
    a. The next visit within the present project is to take place in **September/October 2007**. **Two persons** from the Polish party may come to **Blois** for 10 days at most. The precise dates and the list of persons should be

decided possibly early so that the necessary organizational job is done possibly before summer.

b. The Polish party is interested in a free large-coverage lexicon of named entities for Polish. It has already started encoding single-word named entities and will pursue in this direction. The encoding tool presented by Marcin is the main tool for this job.

c. The French party is interested in applying Multiflex formalism and tool to Polish compound named entities. The encoding of Polish multi-word proper names might start soon. Marcin, possibly in collaboration with Agnieszka, will prepare a corpus of proper name candidates extracted from the IPIPAN corpus by simple regular expressions via Poliquarp, or maybe by SPROUT rules.

   My own recent consideration is that an interesting issue would be to pipe the output of SPROUT's NE-rules into an interface module (to be created) which would allow a supervised encoding of Polish compound proper names within the Multiflex formalism. Some experience from the Web encoding interface (presented by Marcin) for Polish simple words might be used in the construction of the minimal questioning scenario for the encoding of compound NEs.

d. The French party is interested in a free access to the *Polish Grammatical Dictionary* for research projects. Denis Maurel proposes the consideration of the LGPL-LR licence (Lesser General Public Licence for Linguistic Resources) for this resource. One possible source of information on this type of a licence is http://www-igm.univ-mlv.fr/~unitex/lgpllr.html.

   The availability of the Polish lexicon would allow the French party to create a free Polish module within the multilingual Unitex platform.

e. The French party could be interested in using Unitex for Polish, provided that an English manual exists. I have checked that the English manual does exist and can be downloaded at http://www-igm.univ-mlv.fr/~unitex/manuel.html.

   Another useful feature would be the possibility of working on a corpus annotated beforehand.

f. The French party is clearly also interested in creating a Polish module within Prolexbase. A student project may be defined (and supervised by Adam ?) for the Fall session 2007, aiming at an automatic extraction of named entities from Wikipedia and their translations, in view of being inserted into the Polish module of Prolexbase. A possible starting point is the French contents of Prolexbase. If some of the French proper names can be automatically translated via Wikepedia into Polish, the resulting Polish module might automatically inherit most of the language-independent relations existing for the French entries.

g. The SPROUT formalism and its application to the named-entity extraction might be interesting for Nathalie Friburger who couldn't come to Warsaw but is a member of the project (her PhD concerned NE-extraction via rule-based finite-state tools within the Unitex framework).