Blois, December 20, 2007

Report on the work session of "Polonium" project

Date and place: Laboratoire d'Informatique, Blois, December 12-15, 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 Łukasz Dębowski, Nathalie Friburger, Elżbieta Hajnicz, Małgorzata Marciniak, Agnieszka Mykowiecka, Adam Przepiórkowski

This report has been edited by Agata Savary

1. Work session on Thursday, December 17 (Joanna, Marcin and Agata) – discussions on current projects in which the 3 persons are involved, connected to the subject of Polonium
    a. Joanna – LUNA European project
        i. Annotation of a dialogue corpus concerning bus, tram and metro communication in Warsaw
        ii. Different levels of annotation : morphology, phrase structures, semantic, dialogue acts, predicates, anaphoras
        iii. Aim : discovering the dialogue structure and creating a dialogue model
        iv. Available tools and resources :
            - Morphological dictionary of common words (cf. paper 2003, copyright questions to be discussed with the co-author)
            - AMOR morphological analyzer (cf. copyright questions to be discussed with the co-author)
            - List of 4400 canonical names of places in Warsaw
            - List of official names of bus, tram and metro stops in Warsaw
        v. Resources under construction :
            - Simple constituents of place-name list mentioned above, annotated morphologically.
            - List of simple and compound proper names, discovered in the dialogue corpus
    b. Marcin – motivations with respect to a large class of compounds, not only compound proper names
        i. Grammatical Dictionary of Polish (SGJP) – the second edition is in progress, words taken into account are limited to space-to-space sequences. Discussion on the availability of the SGJI for research purposes.

ii. ŚWIGRA – the syntactic analyzer used in particular to building the valence dictionary of Polish verbs, was able to analyze about 30% of corpus sentences. Most missed analyzes were due to compounds and proper names.

iii. Polish Treebank – a project for building such a corpus is under submission for a KBN (Polish Research Committee) grant. ŚWIGRA is to be used for this project. Same motivations with respect to compounds and proper names apply.

c. Agata -

i. Developing the interoperability of Multiflex – a formalism and a tool for the morphological description and treatment of compounds.

- The system is conceived so as to collaborate with different underlying morphological models and tools for simple words. Using the morphological tools of AMOR or SGJP would be an occasion to improve this interoperability.
- The lack of a large coverage morphological resource of simple words in Polish is the main problem hindering the development of a Polish module in Multiflex.

ii. Developing a Polish module in Prolex – the multilingual relational database of proper names

2. Tool demonstration – Friday December 17 a.m. :

a. Joanna – explaining the description model of the morphological lexicon of simple words used by AMOR

b. Marcin – demonstration of the SGJP interface

c. Agata – demonstration of Multiflex running under Unitex ; Multiflex-based description of interesting examples of compound proper names submitted by Joanna (*Białystok*, *Wielkanoc*, *Plac Bitwy Warszawskiej 1920 roku*) in which morphological issues are accompanied by variation (frequent in the corpus studied by Joanna) ; impact of the recognition of compounds on the sentence graph issued from the morphological analysis ; idea of some new operators in Multiflex (set inclusion, inclusion negation, nested description of compounds)

3. Project workshop - Friday December 17 p.m. :

a. Joanna - *Towards a description of Polish proper names*,

i. description of the LUNA project

ii. examples of proper names in Warsaw transportation system, and their behavior in spoken dialogues

b. Marcin – *Grammatical Dictionary of Polish*

i. Morphological model used in the dictionary, and its implementation in a relational database

ii. Discussion on extending the model to regular and/or irregular derivation

c. Both presentation slides will be attached to Polonium webpage (http://www.sir.blois.univ-tours.fr/~savary/Polonium/Polonium.html)

d. Denis' proposal of using Prolex terminology (aliases, instances, etc.) for the description of Polish proper names

4. Conclusions and perspectives:

   a. Extension of the present project for the second year is not conditioned by an intermediate report on the French side. The notification on the amount of funds allocated by EGIDE for 2008 will arrive per mail.

   b. Joanna

      i. continues the morphological description of simple components of proper names used in the LUNA project ; this resource may be freely available for the Polonium project

      ii. will also make clear the copyright constraints on the AMOR morphological analyzer and lexicon

      iii. may freely send to Agata the description of the tagset used in her morphological resources (its representation in Multiflex can thus be studied)

   c. Agata –

      i. will work on description of sample Polish compound proper names, until a morphological resource of simple components is available allowing a creation of a large coverage compound lexicon

   d. Marcin –

      i. will study the processing of compounds under ŚWIGRA

   e. The three persons agree on working on a common resource of Polish proper names of places and transportation terms in Warsaw.

   f. Next visit of Agata to Warsaw is planned around May 11, 2008.