Overview of Annotation Tools for Polish Dialogs

Joanna Rabiega-Wiśniewska, Małgorzata Marciniak, Agnieszka Mykowiecka
(Institute of Computer Sciences, PAS, Warsaw)

The aim of this presentation is to discuss available tools that we consider to use for a Polish dialogs corpus annotation within the LUNA project. We will briefly present goals of the project and collected data. Then, we concentrate on an overview of particular tools.

The LUNA – spoken Language UNderstanding in multilinguAl communication systems – is a STREP project which has started in September 2006. One of its task is to collect and annotate dialogs corpora in French, Italian and Polish. The Polish corpus will consist of 500 conversations between passengers and operators of Warsaw city transportation center. The dialogs refer to time schedules of public transport, finding a route between the given points in the city, trip duration etc. The corpus will contain representative samples of dialogs on the most popular topics.

The project has multilingual character so it is necessary to coordinate work on different languages. The same levels of annotation with the same description format have been accepted by all partners but everyone can select tools appropriate for its needs. We are looking for re-usable, free, open source tools that we could set together into one package. Therefore, we consider to use general purpose text processing system GATE or less-known platform SProUT. In this case we will create lacking resources for Polish within one of these systems, for example a chunker.

In our presentation we will discus advantages and disadvantages of following tools for subsequent levels of annotations:

Dialog transcription is done with Transcriber. It is comfortable tool for manual transcription of dialogs which produce an XML file as the output.

For the purpose of POS tagging we will adjust one of Polish morphological analyzers (AMOR or Morfeusz). We consider also to use one of Polish taggers (statistic tagger designed in IPI PAN or rule based Tagger KIPI). They have been designed for written texts so first we have to check their usability to dialogs.

The semantic levels of annotation are strongly bound to a domain ontology and all partners have to design the description of their domain. For LUNA it is agreed to use Protégé ontology editor for this purpose.

Attribute level of semantic annotation can be done with Semantizer.

Predicate structure annotation in LUNA is based on frames approach. Because of lack of Polish FrameNet, the domain verbs will be extracted from the corpus and frame-like patterns will be defined for them.

An annotation of anaphoric relations can be done with help of MMAX Annotation Tool. It is not a free application, however it is possible to work on an old version, stable but not developed any further.

Dialog acts description is based on DAMSL for which the DAT tool is available.

Although we have already tested the selected tools, the final decision are not yet made and if the programs do not fulfill our needs, we will consider creation of our own tools.