EAGLES compliant tagset for the morphosyntactic tagging of Esperanto

Antonio Toral, Sergio Ferrández and Andrés Montoyo
Natural Language Processing and Information Systems Group
Department of Software and Computing Systems
University of Alicante, Spain
(atoral, sferandez, montoyo)@disi.ua.es

Abstract
This paper presents the first stage of a research related to automatic morphosyntactic annotation in Esperanto. We present and justify a tagset which fulfills the EAGLES standard. This standard allows us to map our tagset with the tagsets developed for other languages. In future studies, an automatic tagger and a corpus will be developed using the proposed tagset.

Background
Esperanto

- Human-created International Language
- Neutral tool for global communication
- Supported by UNESCO resolutions (1954 and 1985)
- Its features regarding word formation facilitate its morphosyntactic annotation

PoS Tagging

- Important step for Natural Language Processing tasks (Question Answering, Information Extraction, etc)
- Classifies words in a natural language according to their Parts-of-Speech

EAGLES

- European Standard for PoS annotation
- Provides an Intermediate tagset with a set of features:
  - mandatory, PoS of words
  - recommended
  - optional
- Tagsets defined according to the standard must match the intermediate tagset
- Linkings can be made between tagsets defined with this standard by using the intermediate tagset

Tagset

- Follows the EAGLES standard
- The 13 mandatory attributes of EAGLES suit the Parts of Speech defined in Esperanto
- Just a subset of the recommended attributes that EAGLES considers has been necessary:
  - There are different values for the attribute case, but we only need one (accusative)
  - Grade for adjectives
  - Only two types of pronouns are needed
  - The attributes for the PoS article are not used as Esperanto has only one article (la)
- No optional attributes have been needed
- EAGLES does not provide any attribute for a feature of Esperanto called direction adverbs
- We have added tags to take into account this fact but we do not consider them for the intermediate tagset in order to maintain the compatibility with the standard

Conclusions and further work
A standard compliant tagset for the morphosyntactic tagging of Esperanto has been defined

Useful as a starting point to build NLP systems for this language

EAGLES provides a set of attributes that are able to represent the morphosyntactic features of Esperanto

One aspect of Esperanto (direction adverbs) is not considered by EAGLES
Due to morphosyntactic features of Esperanto, it has been possible to design a small tagset (86 tags). EAGLES compliant tagsets for other languages are bigger (i.e. 114 for English, 274 for Italian or 280 for Urdu)

Next steps are to develop a PoS tagger for Esperanto using this tagset and to build an annotated corpus