A tagset for the morphosyntactic tagging of Arabic

Shereen Khoja
Computing Department,
Lancaster University,
Lancaster, LA1 4YR.
s.khoja@lancaster.ac.uk
Tel: 01524 592329
Fax: 01524 593608

Roger Garside
Computing Department,
Lancaster University,
Lancaster, LA1 4YR.
rgg@comp.lancs.ac.uk
Tel: 01524 593803
Fax: 01524 593608

Gerry Knowles
Department of Linguistics and
Modern English Language,
Lancaster University,
Lancaster, LA1 4YT.
g.knowles@lancaster.ac.uk
Fax: 01524 843085

A morphsyntactic tagging system for Arabic has been around for centuries. When analysing sentences, Arabic grammarians give to each word a detailed morphosyntactic tag or part-of-speech. These tags contain a large amount of information, perhaps more than what would be found in Indo-European tags. This is because Arabic words are formed by following fixed patterns, and by knowing the tag of the word, we know its pattern, and can thus predict various properties and meanings of the word. It is from this tagging system that we derive our Arabic tagset.

The Arabic tagset we describe does not follow the EAGLES recommendations for the morphosyntactic annotation of corpora, but this is to be expected since Arabic is very different from the languages for which EAGLES was designed, and belongs to the Semitic family rather than the Indo-European one. Following a normalised tagset and the EAGLES recommendations would not capture some of Arabic's relevant information. However, the Arabic tagset can be mapped onto the EAGLES recommendations.

In Arabic we have three major categories, these are Noun, Verb and Particle. To handle other common features of Arabic we have included another three categories, Residual (foreign words, abbreviations, etc), Numerals, and Punctuation (comma, period, etc).

The EAGLES recommendations include the major categories described above (except the Particle) plus eight other categories. These are Adjective, Pronoun/Determiner, Article, Adverb (which in Arabic are all considered to be subcategories of the Noun), and Adpositions, Conjunctions, and Interjections (which are subcategories of the Particle in Arabic). The Unique category in the EAGLES recommendations is applied to categories with a unique membership, which do not follow any of the standard part-of-speech tags.

In this paper we will show that although the Arabic tagset we have derived is very different to any Indo-European tagset, it is appropriate for Arabic. A detailed description of the tagset, along with examples of how the tags map onto real words will be included in this paper. We will also describe the results of using this tagset with an Arabic part-of-speech tagger to tag an Arabic corpus.