The Edisyn Search Engine

Jan Pieter Kunst and Franca Wesseling (Meertens Institute, Amsterdam)

Edisyn (European Dialect Syntax) is an ESF-funded project on dialect syntax. It runs at the Meertens Institute in Amsterdam from September 2005 until September 2010. It aims at achieving two goals.
One is to establish a European network of (dialect)syntacticians that use similar standards with respect to methodology of data collection, data storage and annotation, data retrieval and cartography. The second goal is to use this network to compile an extensive list of so-called doubling phenomena from European languages/dialects and to study them as a coherent object. Since these phenomena primarily occur in non-standard varieties, their existence has gone largely unnoticed in the linguistic literature. The project will therefore greatly enhance the empirical basis of syntactic research. Cross-linguistic comparison of doubling phenomena will enable us to test or formulate new hypotheses about natural language and language variation.

One of the deliverables of the Edisyn project is a centralized search engine which can search different corpora simultaneously, and show the combined search results. The basic requirements for this search engine can be summarized as follows:

- A single unified search interface for different European corpora of dialect transcriptions.
- A single mapping solution (map of Europe showing data from different corpora).
- Searching for text strings and patterns across different corpora.
- Searching for PoS tags across different corpora. The necessary assumption here is that the grammatical tagging of different corpora will be similar enough for a unified search to be possible and useful.

The ideal architecture of such a search engine would, in our view, be a distributed one: each research group hosting, maintaining, and being responsible for its own corpus, and exposing its search interface via a web service. The central search engine then calls the different corpora via these web service interfaces, and shows the aggregated results on its own results page.

In practice, research groups often don't have the resources to add and maintain a web service interface to their existing corpora. In those cases we host local copies of the corpora on our own server. Of course, this makes things like handling updated versions of corpora more complicated.

An experimental version of the Edisyn Search Engine is online at http://www.meertens.knaw.nl/edisyn/searchengine/. Currently four corpora are part of the Edisyn Search Engine: SAND (Syntactic Atlas of the Dutch Dialects), CORDIAL-SIN (Corpus Dialectal para o Estudo da Sintaxe; Portuguese), ASIS (Syntactic Atlas of Northern Italy), EMK (Tartu University’s Estonian Dialect Corpus) (partly).

Searching for PoS tags is enabled via a central tagset (visible in the 'tags' menu on the search page). The user can either search for complete tags or for features. For each corpus, there is an XML file which translates the tags from the central tagset into the native tag set of the corpus.

In the current tagset the major categories are represented, adding to such a PoS tag a specific feature may yield another PoS tag (linguistic category). For example, the tag 'D' *determiner/pronoun* may be combined with the feature (dem) or (rel), in Dutch, thereby referring to a demonstrative or a pronoun, respectively. The twofold way of tagging provides a solution for Dutch (and other languages) where the same form, in this case *die* 'that', can represent a determiner, as in *Ik zag die fiets* 'I saw that bike', or a pronoun, e.g. *Ik heb gister een film gezien. Die was erg goed.* 'I saw a movie yesterday. It was very good.'

Another possibility could be to insert two seperate PoS tags, namely that of determiners and that of pronouns, however not all databases make the same distinction between these linguistic categories, if they make the distinction at all. By using a tagsystem which combines categories and features it is most clear to the user what the tagged item represents.

When combining linguistic databases it is important to keep the set up of each database in its original state as much as possible. In order to achieve this, the list of features which is used in the Edisyn search engine contains all frequently occurring features. Features such as masculine, negative or transitive are included, but not all possible values of case systems since these are often not tagged in the individual databases.

In order to make the databses as useful as possible for all linguists, a tranlsation of the dialect corpus is desired. One of the ways to achieve this is to list all occurring words according to frequency. In this manner the lexical items which are used most often can be translated into English.

One of the future aims of the Edisyn search engine is to make its tags compatible to the Category types which are applied in ISOcat (www.isocat.org). ISO 12620 provides a framework for defining data which is compliant with the ISO/IEC 11179 family of standards.