THE EDISYN SEARCH ENGINE

JAN PIETER KUNST AND FRANCA WESSELING

Amsterdam

ABSTRACT

Edisyn (European Dialect Syntax) is a project on dialect syntax funded by the European Science Foundation. It runs at the Meertens Institute in Amsterdam from September 2005 until September 2010 (partially extended till March 2012). 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. One of the deliverables of the Edisyn project is a web-based search engine to search different linguistic corpora simultaneously and show the combined search results. This search engine is able to make differently structured databases comparable. Although the initial set up of the Edisyn project was to create similar standards for dialectal databases, in practice this has proven to be an unfeasable goal since most databases have a different structure and enrichment (we will come back to this below). Consequently, the Edisyn search engine has been created according to a more pragmatic philosophy and is able to handle databases of various structures.

[1] INTRODUCTION

The Edisyn project focuses on doubling phenomena in various languages. Since these phenomena primarily occur in non-standard varieties, their existence has gone largely unnoticed in the linguistic literature. Recent literature on Dutch dialects (SAND project) has revealed a wealth of doubling phenomena that do not appear in Standard Dutch. See for instance in the cases in (1)–(4) below¹.

- (1) Subject pronoun doubling and subject agreement doubling:
Ze peiz-n da-n ze ziender rijker zij-n
They think-3PL that-3PL they they richer are-3PL
'They think that they are richer.'

[1] PART = participle, PL = plural, 2 = second person, 3 = third person

- (2) WH-word doubling:
Wel denkst wel ik in de stad ontmoet heb
Who think-2PL who I in the city met have
'Who do you think I met in the city?'
- (3) Participial morphology doubling:
Zol hee dat edane hemmn ekund
Would he that done-PART have could-PART
'Could he have done that?'
- (4) Auxiliary doubling:
K-em da gezegd gehad
I-have that said-PART had-PART
'I have said that.'

Through the investigation of non-standard varieties, doubling phenomena can be adequately researched. The project therefore greatly enhances 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. By investigating doubling phenomena we are able to detect the pervasiveness and limitations hereof. The Edisyn project seeks to answer the question whether there are any limitations as to what kind of linguistic categories can be subjected to doubling. Furthermore an explanation is sought for any such restrictions. These answers will not only contribute to the characterization of micro-variation but will in turn have implications on how we look both at meso-variation (e.g. OV word order versus VO order) and macro-variation (e.g. polysynthetic versus non-polysynthetic).

To enhance cross-linguistic research on non-standard varieties a search engine - the so called Edisyn search engine- has been created enabling comparative research on dialect data of different languages. Until recently most dialectological work focused on variation within the non-standard varieties of one language, the availability of the Edisyn search engine, however, enables the investigation of dialects of various languages. At the moment of writing (March 2010) five databases containing data on non-standard varieties of a specific language have been combined within a single interface. The unified search interface allows the user to search different European linguistic corpora of dialect transcriptions simultaneously and shows the combined search results on a single results page. Searching for text strings and textual patterns should be possible, though this kind of search is of limited value when searching text across different languages. At the moment a basic search for strings is possible. The problems that arise when

attempting to connect linguistic corpora are outlined below.

[2] LINGUISTIC DESIGN OF THE EDISYN SEARCH ENGINE

[2.1] *Introduction*

Every database has its own, specific structure. This is due to various reasons. First of all, a dialect database differs according to the (type of a) language. The content of a database is dependent upon syntactic and morphological properties of a language. If a language has case marking, for example, the values hereof will be part of the tag set of a database. If a language does not assign case, these features will be absent in the database.

Second, the structure of a database depends on the kind of data that has been gathered. If the data consists of elicited speech different choices are made with respect to the structure of the database than if the data concerns, say, spontaneous speech. In a database containing elicited speech, the dialect data can be lined up with the question (or test) sentences, whereas this is not possible with spontaneous speech. In the latter case the data will be more difficult to parse and specific decisions need to be made concerning the desired way of presenting the data.

Furthermore, the theoretical views of the linguist(s) can alter the outlook of a database. If one is working within a generative framework the tags that are assigned may be theory-dependent. In the ASIt database (Italian dialects)² for example, the tag *raising* is assigned to certain parts of speech types. This tag is used to indicate verbs that do not assign an external theta role such as *appear* and *seem*, whereby the semantic subject of the lower clause verb is syntactically realised as a constituent of the higher clause. This term is highly theory-dependent for it is not used in non-generative frameworks.

Also, the set up of a database is influenced by the subject matter of a research. If data is collected within a research project focused on the order of verbs in subordinate clauses, for instance, the ordering of verbs (and possibly other part of speech types) will be tagged. Other syntactic or morphological phenomena may then receive less attention/markings.

A crucial factor in the make up of a database is the kind of enrichment a database contains. A database may only have raw recordings, or these recordings may be lined up per sentence so that small parts of a conversation can be listened to. Furthermore, these recordings may be tagged with part of speech tags -per word- or keywords may be assigned to an entire phrase, which are in turn database specific. A database may also have both enrichment at the word level and contain syntactic parsing. In addition, the data may be translated into English, this can be done word by word, or apply to entire phrases.

[2] The ASIt database is available at <http://asis-cnr.unipd.it/>

The databases also differ with respect to the quality of the enrichment, that is, the assignment of tags can be very detailed or less thorough. In addition, the subject matter that is tagged can vary, for instance, question sentences can be enriched with linguistic tags (in the case of elicited data), or the answers can be tagged, or both question sentences and answers may be tagged. Also, databases will be dissimilar in the extensiveness of the English translations. Finally, the metadata is an important aspect that differs per database. This kind of information is often present to a limited extent, or not at all. Ideally, every database would provide information specifying the period in which the data has been gathered, the location(s) where the research has been undertaken, the kind of data that is presented in the database, the age of the informants, the people in charge of the research and database and their affiliation, et cetera. However, often none of these details are further specified, let alone in a similar fashion.

In summary, databases vary from one another in the following respects:

- (1) Type of data: type of language, elicited data versus spontaneous speech.
- (2) Enrichment of data: part of speech tags / syntactic labels / linguistic keywords / English glosses / a combination hereof.
- (3) Quality of the enrichment: the data is meticulously tagged / the data is tagged in a more general manner.
- (4) Quantity of the enrichment: only answers are tagged / only question sentences are tagged / both questions and answers are tagged / neither is tagged.
- (5) Metadata: information specifying the circumstances in which the data has been gathered is absent or the databases provide this information to a different degree.

In the attempt of making different databases interoperable via one search engine, these differences need to be considered. Ideally, each database would have similar standards with respect to the factors mentioned above, this is however never the case. Nevertheless, it is feasible to create a search engine that queries various databases which contain data that has been tagged and glossed, despite their external differences. This has been done in the Edisyn project, resulting in the Edisyn search engine. Via this search engine it is possible to search on the basis of part of speech tags and on the basis of strings of words, the latter search option being of course highly language specific.

Note that in the development of the search engine, the Edisyn team has no desire to change the configurations of any of the component databases. The aim of the search engine is simply to provide a tool via which it is possible to search dialect data of various languages through a single interface. Each database retains its own tag set and can be consulted individually at all times.

Abbreviation:	Category:
V	Verb
N	Noun
D	Determiner
Pron	Pronoun
A	Adjective
Adv	Adverb
Conj	Conjunction
Negmrk	Negation marker
P	Adposition
C	Complementizer
Part	Particle
Intj	Interjection

TABLE 1: Part of speech categories used in Edisyn search engine

[2.2] *The Edisyn Tag Set*

The first and perhaps most important step in connecting different databases is to equalize the different tag sets. Within the Edisyn project we have constructed a general tag set containing part of speech categories and linguistic features (this division will be elaborated upon below), as shown in Table 1 and 2. This tag set can be ‘translated’ to many different tag sets (note that the Edisyn tag set is dynamic and can be adjusted according to the needs of a database developer).

A *category* refers to commonly used parts of speech such as Verb, Noun, Adjective, etc. These can be combined with *features* such as ‘singular’ which results in a specific tag, such as a singular noun. A category can be combined with any and as much feature(s) as desired. Thus any tag can be created. However, not every query will generate a result because not every database has assigned the same tags to their data. This is clearly communicated to the user of the Edisyn search engine. That is, if a query has no result, the user is informed that the tag in question has not been assigned in the individual database.

Categories cannot be combined with other categories. Thus, the category Noun can be combined with the feature ‘nominative case’, but not with the category Adjective. It is possible to search for a sequence of categories, for instance Noun followed by Adjective. Tags can be either adjacent to each other -the default setting- or with an optional gap (zero or more words in between the tags).

The home page of the search engine consists of an overview of the databases that can be consulted. By clicking on the box next to each database, the database selected will be included in the following query. It is also possible to search each database in its original layout, the link next to each database connects the user

Abbreviation:	Feature:	Abbreviation:	Feature:
ab	abessive case	m	masculine
abl	ablative case	mesacl	mesaclisis
acc	accusative case	mod	modality
act	active	neg	negative
ad	adessive case	neut	neuter
add	additive case	nom	nominative case
all	allative case	num	numeral
art	article	partit	partitive case
asp	aspect	pass	passive
aux	auxiliary	past	past tense
caus	causative	perf	perfective
cl	clitic	pers	personal
com	comitative case	pl	plural
comp	comparative	poss	possessive (case)
coord	coordinating	post	postposition
dat	dative case	pp	past articiple
def	definite	prep	preposition
dem	demonstrative	pres	present tense
dim	diminutive	presp	present participle
el	elative case	procl	proclisis
encl	enclisis	quant	quantitative
erg	ergative	recipr	reciprocal
es	essive case	refl	reflexive
f	feminine	rel	relative
fin	finite	sg	singular
foc	focus (marker)	subord	subordinating
fut	future tense	sup	superlative
gen	genitive case	term	terminative case
ger	gerund	tr	translative case
ill	illative case	trans	transitive
imp	imperative	unacc	unaccusative
in	inessive case	unerg	unergative
indef	indefinite	1	first person
infin	infinitive	2	second person
inst	instrumental case	3	third person
inter	interrogative (=wh)		
intrans	intransitive		

TABLE 2: Linguistic features used in Edisyn search engine

directly to that database. When using the Edisyn search engine the tag set described above and in Tables 1 and 2 is to be used, if one is querying an individual database the tag set of that specific database is of course employed.

After one or more database(s) has/have been selected one can start creating a tag, this is done by adding one or more features to a category, as described above. It is also possible to search for a category or feature by itself. When the appropriate tag has been selected, the search engine will present the results available for the selected database(s).

Note that when a query has been performed with the Edisyn search engine, the results contain the tags of the individual database. For example, if one wants to know if dialects of Portuguese and dialects of Dutch both have a way of marking a verb in the present tense for second person singular, one adds these databases to the search by selecting them. Then, one drags the category Verb to the search field, followed by the features 'pres', '2' and 'sg'. By clicking on *search* the query is started and the results will be shown. These results contain -in this example- the dialect sentences in Portuguese, with the tags provided by the Cordial-Sin database, and the data in Dutch with the tags used in the SAND database. These tags are easily interpreted by the user for all the tags used in the various databases are explained in a glossary.

The results are based on the conversion of the tag set of the Edisyn search engine to the tag set of each database. That is, at the backend, the tag used in the search engine is connected to the corresponding tag in each database. Every category and every feature has a corresponding tag in each of the databases, for instance, in the example above the Edisyn tag 'V(fin,pres,2,sg)' is linked to the Portuguese tag 'V-P-2S'.

With the Edisyn tag set available many databases can be interconnected via the search engine for each tag set can be translated into so called Edisyn tags. Again, we want to stress that we do not make any changes to the individual databases; we leave the structure and tag set of each database completely intact. Via the conversion of the Edisyn tag set to the tag sets of the databases it is possible to search various (dialect) databases at the same time, enabling a cross-linguistic comparison of dialect data.

[2.3] *Note on English Glosses*

It is of importance to add English glosses to a database, for this will enhance the accessibility of the search engine and it will allow more researchers to use the database. Most researchers will have (some) knowledge of the language and its dialects (s)he is working on, but this need not be the case for the other dialect databases which have been made interoperable in the Edisyn search engine. With the addition of English glosses however, all the dialect data is made comprehensible for every (English speaking) linguist, and may trigger their interest. By mak-

ing the content accessible to everyone in the field more research on dialects may even be stimulated.

Currently the database on Dutch dialects (SAND) contains English glosses, that is, there is a translation available for every word that is used in this database. The Cordial-Sin corpus (on Portuguese dialects) is working on the implementation of a word by word translation into English. Within the Nordic Dialect Corpus there is a possibility of translating every sentence by Google Translate. The other databases do not have an application to display the dialect data in English. This is work to be done in the future.

[3] IMPLEMENTATION OF THE SEARCH ENGINE

[3.1] *Ideal Architecture for a Search Engine*

The ideal architecture of 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, i.e. an interface for computer programs, as opposed to human users, to access the corpus. The central search engine then calls the different corpora via these web service interfaces, and shows the combined results on its own results page. In practice, such an ideal architecture is difficult to realize. Some linguistic corpora do not have a search interface as such, but are simply made available as downloadable text files. Other corpora do have a web-based search interface, but strictly one for human users. In those cases the research groups responsible for the corpora usually do not have the resources to add the needed features to their existing corpora.

In those cases we opted for the pragmatic solution of hosting copies of the corpora locally on our own server. Of course, this makes problems like handling updated versions of corpora more complicated than in a web service-based solution, but that is a necessary trade-off in this situation, because otherwise there would not be a search engine at all. In the case of the Nordic Dialect Corpus we access the corpus remotely, at the moment of writing not yet via a true web service but by doing normal http requests with a *curl* library and 'screen scraping' the returned pages with results. We hope to convert this system to a real web service connection in the future.

But, even if, in many instances, we have to work with locally hosted corpora out of necessity, we still built the search engine using a web service architecture, with *localhost* URLs for the corpora. This makes it relatively easy to switch to a remote web service for a corpus if the opportunity arises: change the URL to point to the remote host instead of localhost. It is unlikely that the interface will be exactly the same as the one we created ourselves for our localhost web services, so probably some additional fine-tuning will be needed, but that will certainly be less work than converting a platform-specific local connection for a corpus to a web service connection.

[3.2] *Current State of the Search Engine*

At the moment an experimental version of the Edisyn Search Engine is online at <http://www.meertens.knaw.nl/edisyn/searchengine/>, with five corpora included: SAND (Syntactic Atlas of the Dutch Dialects, Dutch, Meertens Institute), CORDIAL-SIN (Syntax-oriented Corpus of Portuguese Dialects, Portuguese, University of Lissabon), ASIt (Syntactic Atlas of Northern Italy, Italian, University of Padua), EMK (Estonian Dialect Corpus, Estonian, University of Tartu) and NDC (the Nordic Dialect Corpus, Scandinavian Languages, ScanDiaSyn). With the exception of the Nordic Dialect Corpus, all corpora are hosted locally at the Meertens Institute.

Searching for POS tags is enabled via a central Edisyn tag set (visible in the 'tags' menu on the search page; see fig. 1 for a screenshot). The user can search for complete tags, partial tags, or features. For each corpus, there is an XML file which translates the tags from the central tag set into the native tag set of the corpus. So the central search engine is quite 'shallow' and does not know anything about the tag sets of the corpora it uses, in turn, the participating corpora only see search requests with their native tag sets and do not know anything about the Edisyn tag set. This set up makes it possible to add new corpora to the search engine without affecting the existing system.

[3.3] *Technical Details of the Search Engine*

User Interface

As mentioned before, the Edisyn search engine is web based and should work in any reasonably modern browser. The user interface consists of standard XHTML pages enriched with JavaScript via the JQuery library. We use JQuery to create a drag-and-drop interface for constructing search queries, in order to make the potentially tedious process of entering POS tags in queries as streamlined as possible; and we use an AJAX interface (also provided by JQuery) to the server to avoid unnecessary page reloads.

Server-side Technologies

The Edisyn Search Engine is written in object-oriented PHP. The web page containing the search form is created by a class `EdisynPage`. This class creates the search form and checks if it has been submitted; if it is, it fetches the search results and adds them to the page; if not, it just shows the form.

Fetching the results is done by instantiating search classes for each checked corpus, called `Edisyn_Search_<corpusname>`. As their name implies, these classes are corpus-specific; they are child classes of an abstract class `Edisyn_Search` which contains general, non-corpus-specific methods and properties. The knowledge about how the searches are performed is encapsulated in the search classes; the `EdisynPage` class just feeds the form data to the search classes and calls a

FIGURE 1: Screen shot of Edisyn search engine

getResults() method on them.

[4] FUTURE PROSPECTS FOR THE SEARCH ENGINE

The Edisyn Search Engine in its current state is not finished. We list some features and enhancements which will be added in the future in this section.

[4.1] Mapping

An option to show search results on a map will be added in the future. The ground-work is already there: almost all of the data which is hosted locally at the Meertens Institute is enriched with geographical coordinates, as is the Nordic Dialect Corpus, so enhancing the search results to include geographic locations is not a difficult problem. This will provide the user with the possibility to show the data from different corpora combined on a single map of Europe. We plan to use Google Maps as the web mapping solution to display these data.

[4.2] *Additional Corpora*

Some corpora which we plan on adding in the near future are: the Afrikaans Variation Project (Mark de Vos, Rhodes University), Slovene Dialectal Syntax (Marko Hladnik, University of Utrecht), Diversion in Dutch DP Design (DiDDD, University of Utrecht) and Freiburg English Dialect Corpus (FRED, University of Freiburg). In the distant future we also hope to give access to Lauseopin Arkisto on Finnish dialects (Kotus (Research Institute for the Languages of Finland)), COSER on Spanish dialects (Corpus Oral y Sonoro del Español Rural, Autonomous University of Madrid) and a database of Breton dialects (ARBRES, Melanie Jouitteau). At the moment data on Basque dialects are being gathered at the University of Bayonne (IKER), which will also be made interoperable by the Edisyn search engine.

It is our aim to add as many databases as possible, the requirements for a suitable database being rather limited, namely having reliable and useful data on any (European) dialect, which has been tagged and preferably contains English glosses.

[4.3] *Clarín*

The CLARIN project is a large-scale pan-European collaborative effort to create, coordinate and make language resources and technology available and readily usable. Standards for data and metadata for language resources are being developed in the CLARIN project. We plan to adhere to these CLARIN standards to preclude the Edisyn project being an isolated effort. For further information about CLARIN, see <http://www.clarin.eu>.

One of the standards which are being developed within CLARIN is the so called ISOcat category set. This allows linguists to tag their data with a dataset which has been approved by the ISO standard (ISO 12620 provides a framework for defining data categories according to the ISO/IEC 11179 family of standards). At this moment we are modifying the Edisyn tag set according to the standard of the ISOcat categories. This will lead to a more unified way of tagging which will make dialect databases more comparable.

Finally, we will develop and implement more user-friendly applications along the way. That is, more differentiated search options will be added and other enhancements which prove to be useful, will be put into effect.

REFERENCES

- Barbiers, S. & H. Bennis. 2007. The syntactic atlas of the dutch dialects. a discussion of the choices in the SAND-project. In K. Bentzen & Ø. Vangsnes (eds.), *Nordlyd*, vol. 34, 53–72.
- Barbiers, S., L. Cornips & J.P. Kunst. 2007. The syntactic atlas of the dutch dialects (SAND): A corpus of elicited speech and text as an on-line dynamic atlas. In

- J. Beal, K. Corrigan & H. Moisl (eds.), *Creating and digitizing language corpora: Vol. 1, synchronic database*, 54–90. Hampshire: Palgrave-Macmillian.
- Barbiers, S. et al. 2006. Dynamic syntactic atlas of the dutch dialects (Dynasand). URL <http://www.meertens.knaw.nl/sand/>. Amsterdam, Meertens Institute.
- Benincà, P. & C. Poletto. 2007. The asis enterprise: a view on the construction of a syntactic atlas for the northern italian dialects. In K. Bentzen & Ø. A. Vangsnes (eds.), *Nordlyd*, 34, 35–52.
- Johannessen, J. B., J. Priestley, K. Hagen, T.A. Åfarli & Ø. A. Vangsnes. 2009. The nordic dialect corpus - an advanced research tool. In K. Jokinen & E. Bick (eds.), *NEALT proceedings series*, vol. 4. Proceedings of the 17th Nordic Conference of Computational Linguistics NODALIDA.
- St.Laurent, S., J. Johnston & E. Dumbill. 2001. *Programming web services with xml-rpc*. Sebastopol: O'Reilly.

AUTHOR CONTACT INFORMATION

Jan Pieter Kunst
Meertens Institute
P.O. Box 94264
NL-1090 GG Amsterdam
The Netherlands
janpieter.kunst@meertens.knaw.nl

Franca Wesseling
Meertens Institute
P.O. Box 94264
NL-1090 GG Amsterdam
The Netherlands
franca.wesseling@meertens.knaw.nl