The Edisyn Search Engine

Jan Pieter Kunst and Franca Wesseling (Meertens Institute)

Structure of this talk

1. Linguistic content of the Edisyn Search Engine (Wesseling)
2. Technical aspects of the Edisyn Search Engine (Kunst)

Edisyn
European Dialect Syntax

Edisyn search engine

- Edisyn search engine enables unified searches across corpora.
- It is possible to search for text or PoS tags within these corpora and/or make a comparison hereof between the corpora.
- Searching on the basis of text is language specific.

Edisyn
European Dialect Syntax

Tag set Edisyn search engine

- To make searches across corpora possible, the tag sets of the different corpora must be made compatible with each other.
- A tag set has been developed which applies to all tagged items of each corpus. We are currently linking this tag set to the ISO standard ISOcat (ISO 12620).
- In constructing such a tag set it is important to keep the set-up, content and theoretical views of the original database intact.
- Thus minimal deletion of tags. But: keep tag set transparent.

Edisyn
European Dialect Syntax

Tag set Edisyn search engine

- Tag set of search engine consists of PoS tags and features.
- These may be combined (e.g. tag *V-fin-pres-1sg*) or searched for separately (e.g. tag *V* or feature *sg*).
- All PoS which are tagged in a corpus are included in tag set of search engine.
- But in some cases tags are subsumed under another tag, to maintain transparency. For instance, in SAND corpus (Dutch dialects) A(dim) corresponds to A in search engine.
- The tag set of the search engine is dynamic and can be expanded as needed (e.g. when a new corpus is added).

Edisyn
European Dialect Syntax

Important

- To enhance the comprehensibility of the search engine, ideally each corpus is also available in English.
- This is the case for the SAND corpus.
- A protocol, glossary of the tag set and an outline of the research design should be available (also in English).

Edisyn
European Dialect Syntax

To do

- Provide English glosses for each corpus.
- Make tags of search engine compatible with ISOcat.

Edisyn
European Dialect Syntax

Edisyn Search Engine

- Goal: concurrent search of various dialect-syntactical corpora with a single set of terms
- Will hopefully make it easier to find linguistic phenomena across dialects of different European languages
- A preliminary version is online at

  http://www.meertens.knaw.nl/edisyn/searchengine/

Edisyn
European Dialect Syntax

Edisyn Search Engine

The current version uses four corpora:

1. ASIS (Italian, University of Padua):

   http://asis-cnr.unipd.it
2. CORDIAL-SIN (Portuguese, University of Lissabon):

   http://www.clul.ul.pt/sectores/variacao/cordialsin/projecto_cordialsin.php
3. SAND (Dutch, Meertens Institute):

   http://www.meertens.knaw.nl/sand/
4. EMK (Estonian, University of Tartu) (partly):

   http://www.murre.ut.ee

Edisyn
European Dialect Syntax

Edisyn Search Engine

New corpora we are currently working on:

- Nordic Dialect Corpus (ScanDiaSyn)
- Afrikaans Variation Project (Rhodes University)
- Slovene Dialectical Syntax (M. Hladnik, University of Utrecht)

Edisyn
European Dialect Syntax

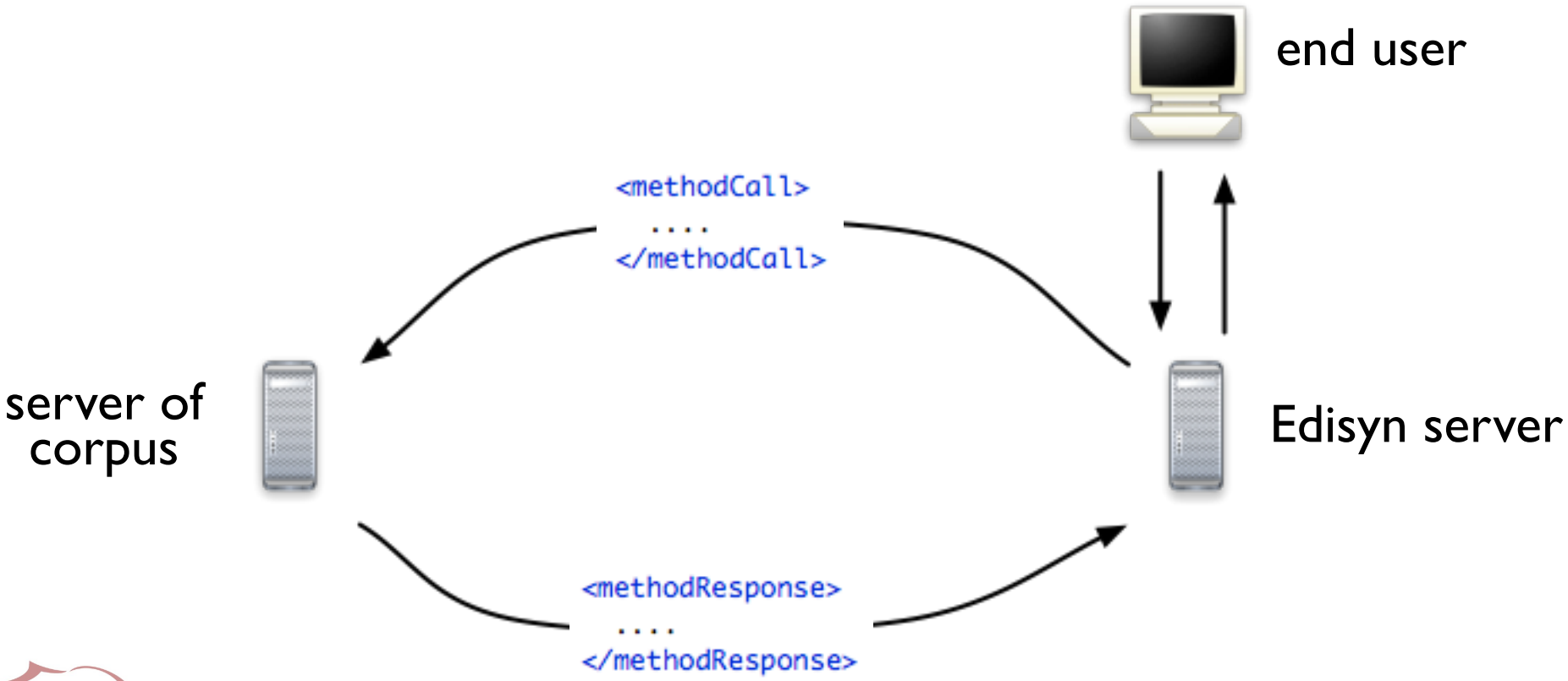
Ideal structure of Edisyn Search Engine

Each participating group hosts and maintains its own corpus, adding an extra web service interface to enable remote searching.

Edisyn
European Dialect Syntax

Ideal structure of Edisyn Search Engine

Each participating group hosts and maintains its own corpus, adding an extra web service interface to enable remote searching.



Edisyn
European Dialect Syntax

Ideal structure of Edisyn Search Engine

- The ideal structure with a distributed search engine is difficult to pull off in practice: the usual case is that organizations who have an interesting corpus don't have money or people available to set up a web service for their corpus
- For most corpora, the Meertens Institute uses local copies
- First corpus with a remote connection will be the Nordic Dialect Corpus
- But even with mostly local copies, we decided to stay with a web service architecture ...

Edisyn
European Dialect Syntax

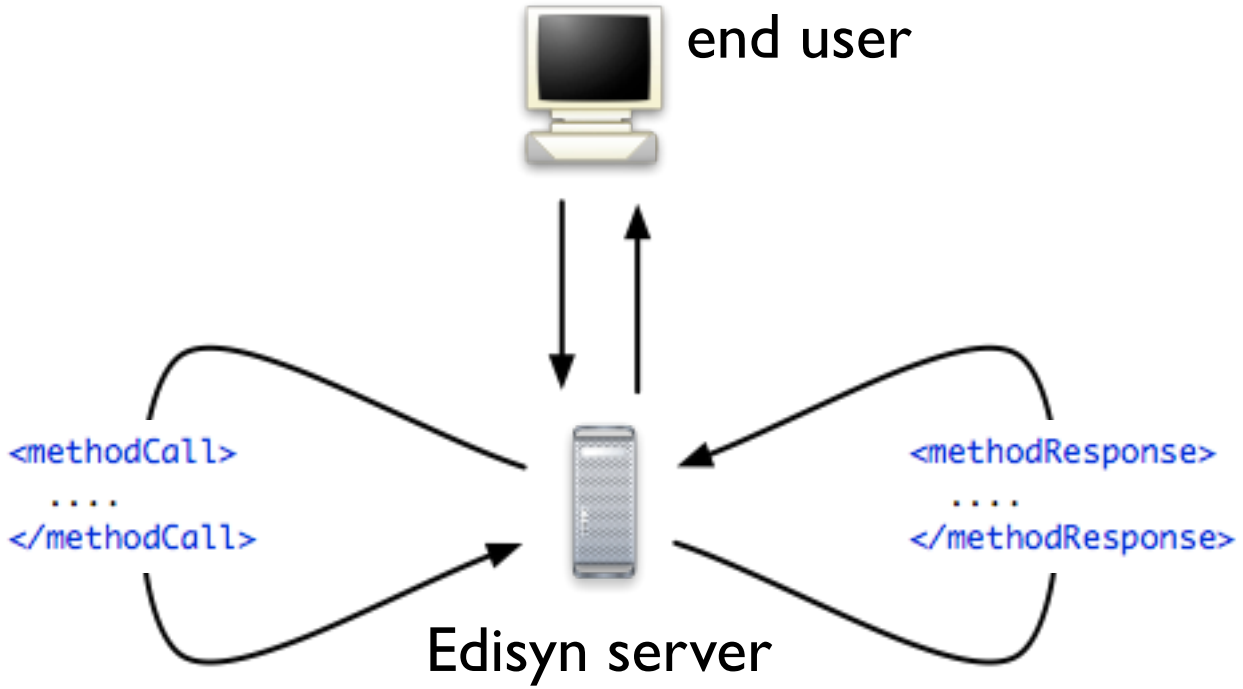
Current structure of Edisyn Search Engine

... the Edisyn server itself runs the web services for the corpora, so while the corpora are local, they are still very loosely coupled with the search engine. It would be trivial to switch this to using a remote web service — just change the URL for the service to point to a remote server instead of to the local host.

Edisyn
European Dialect Syntax

Current structure of Edisyn Search Engine

... the Edisyn server itself runs the web services for the corpora, so while the corpora are local, they are still very loosely coupled with the search engine. It would be trivial to switch this to using a remote web service — just change the URL for the service to point to a remote server instead of to the local host.



Edisyn
European Dialect Syntax

Technologies used

- Web service: XML-RPC
- programming language: PHP (PEAR XML-RPC library used)
- database on the backend: MySQL (storage mechanisms are abstracted away behind the web services, of course)
- JQuery Javascript library used for the user interface of the search engine

Edisyn
European Dialect Syntax

Screenshot of search page

Edisyn
European Dialect Syntax

Screenshot of search page

Organization of search terms

- Each corpus uses its own, native search terms
- On the Edisyn side, Edisyn-terms are converted to corpus-specific terms before the request is sent off to the corpus
- Mapping of Edisyn-terms to corpus-specific terms is done in XML-files. Example:

Edisyn
European Dialect Syntax

Organization of search terms

- Each corpus uses its own, native search terms
- On the Edisyn side, Edisyn-terms are converted to corpus-specific terms before the request is sent off to the corpus
- Mapping of Edisyn-terms to corpus-specific terms is done in XML-files. Example:

CORDIAL-SIN

```
<tag>
    <edisyn>N-sg</edisyn>
    <cordialsin type="equals">N</cordialsin>
    <cordialsin type="equals">NPR</cordialsin>
</tag>
```

SAND

```
<tag>
    <edisyn>N-sg</edisyn>
    <sand>
        <category>1</category><!-- Nomen -->
        <attribute>
            <name>5</name><!-- number -->
            <value>34</value><!-- sing -->
        </attribute>
    </sand>
</tag>
```

Edisyn
European Dialect Syntax

Search results

- Each corpus replies with a list of found sentences and with each sentence, various metadata
- E.g.: a geographical code, an English gloss (if available), a list of tags
- Our goal is to make the barrier of entry for corpora as low as possible: basically, expose your existing search facilities as a web service and let Edisyn handle the integration with the central search engine.

Edisyn
European Dialect Syntax

To do

- Add possibility to construct own tags (search terms) from categories and attributes
- Return geographical coordinates with search results (are already there in the databases, for the most part)
- Add map-making facilities (e.g. Google Maps)
- Add possibility to save and/or export search results
- Compliance with CLARIN infrastructure (ISOcat, metadata)
- ... probably lots more that I can't think of right now.

Edisyn
European Dialect Syntax

Thank you for your attention!

Edisyn
European Dialect Syntax