Connecting Corpora and ISOcat

ISOcat workshop, LREC 2012 Istanbul, 21 May 2012

Franca Wesseling
(Meertens Institute, Utrecht University UiL OTS)
f.wesseling@uu.nl

Edisyn project

Edisyn (European Dialect Syntax) is a project on dialect syntax, its goals are twofold:

1. To establish a network of dialect syntacticians
2. To use this network to compile an extensive list of so-called doubling phenomena and to study them as a coherent object.

One of the deliverables of Edisyn is the ‘Edisyn search engine’.

Edisyn
European Dialect Syntax

Edisyn search engine

- The Edisyn search engine enables unified searches across corpora. Currently we have incorporated corpora on Scandinavian, English, Dutch, Estonian, Italian, Portuguese, Slovene dialects.
- It is possible to search for text, English glosses or PoS tags within these corpora and/or make a comparison hereof between the corpora.
- To make searches across corpora possible, these corpora need to be synchronized, i.e. correspond to the (structure of the) search engine.

Edisyn
European Dialect Syntax

Edisyn search engine

- A tag set has been developed which applies to all tagged items of each corpus.
- In constructing such a tag set it is important to keep the set up, content and theoretical views of the original corpus intact.
- Thus minimal deletion of tags. But: keep tag set transparent.

Edisyn
European Dialect Syntax

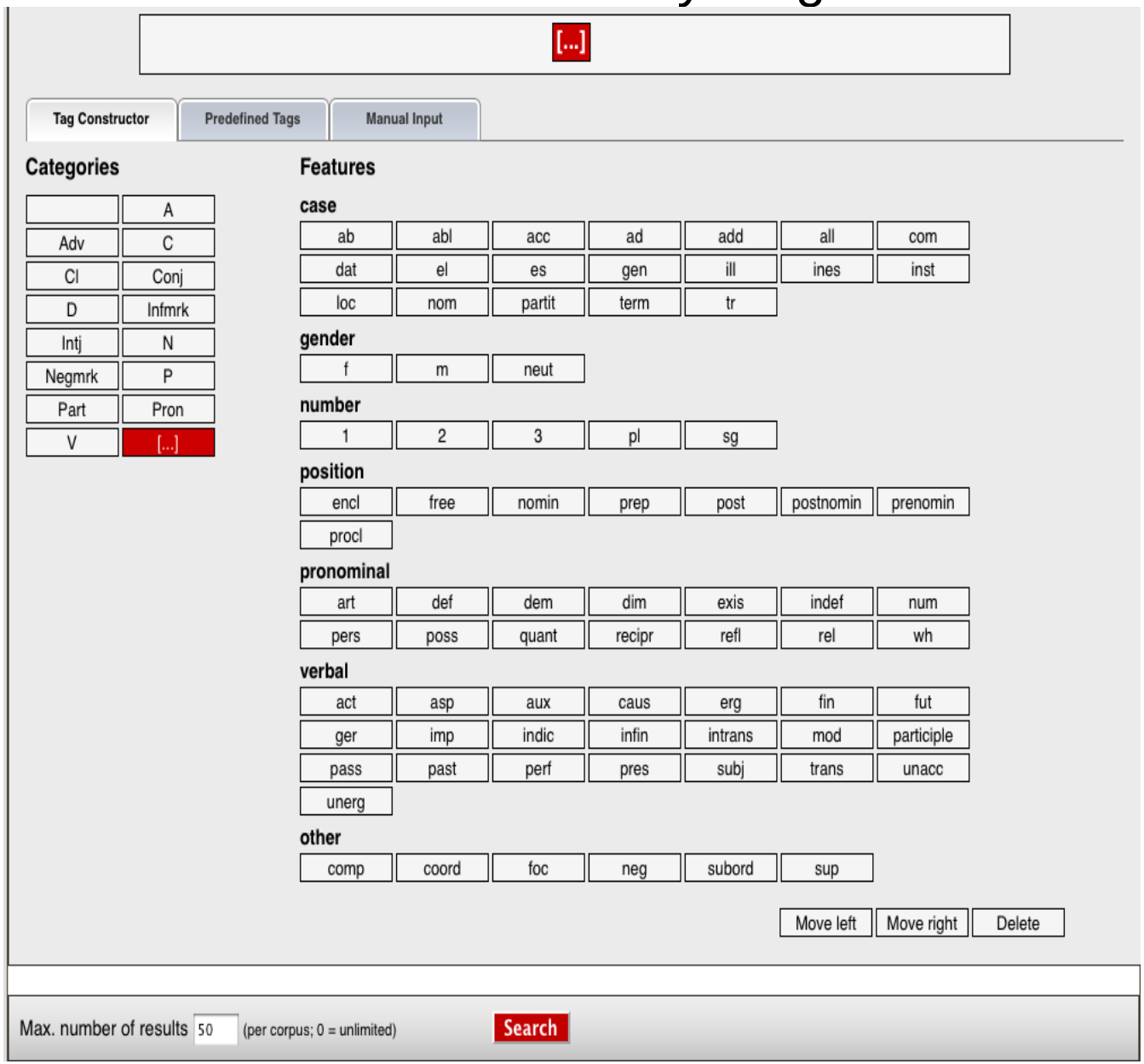
Edisyn tag set

- Edisyn tag set consists of Categories (word classes) and Features (specifications).
- These may be combined freely (e.g. Category ‘V’ with Features ‘fin’, ‘pres’, ‘1’, ‘sg’, resulting in ‘V(fin,pres,1,sg)’) or searched for separately (e.g. ‘V’ or ‘sg’).
- All tags that are used in a corpus can be mapped to Edisyn tags.

Edisyn
European Dialect Syntax

Edisyn tag set

Screenshot of Edisyn tag set:



Edisyn
European Dialect Syntax

Edisyn - ISOcat

Edisyn tags have been linked to ISOcat Data Categories (DC's) so that the Edisyn search engine is able to fit into the CLARIN research infrastructure.

Edisyn
European Dialect Syntax

Edisyn – ISOcat

Screenshot of conversion table Edisyn – ISOcat – other corpora:

Category:	Edisyn search engine	ISOcat	SAND (Dutch dialects)	Cordial-Sin (Portuguese dialects)	ASIt (Italian di	EMK (Estonian (	NDC (Nordic Dia
Infinitive marker	Infmrk	1896 infinitive particle	inf-mrk				inf-marker
Noun	N	1333 noun	N	N	N(proprio)	S / H	noun / prop
	N(sg)	3580 Noun(singular)	N(s)	N / NPR		S(sg) / H(sg)	noun(sg) / prop(s
	N(pl)	3581 Noun(plural)	N(pl)	N-P / NPR-P		S(pl) / H(pl)	noun(pl) / prop(p
	N(dim)	2225 diminutive noun	N(dim)				
Verb	V(infin)	1312 infinitive	V-infin	VB / SR / HV / ET / TR / VB-F / SR-F / HV-F / ET-F / TR-F	inf sempl/ 2 inf /	V(inf) / V(sup)	inf

Edisyn
European Dialect Syntax

Advantages of ISOcat

- Creates uniformity within proliferation of tag sets: diverse tag sets can be easily connected to ISOcat and thereby to each other.
- ISOcat DC's can easily be linked to existing tags / tag sets.
- We do not need everyone to use the same tag set (which is practically nearly impossible) but transparency is very welcome and needed.
- Also, in creating a new tag set for project X one can look at ISOcat DCS's and be inspired.

Edisyn
European Dialect Syntax

ISOcat - RELcat

- Disadvantage of ISOcat is impossibility to combine DC's. E.g. Edisyn tag 'V(fin,pres,1,sg)' has to be defined (created) in ISOcat.
- RELcat (in progress) does provide option to combine DC's, that is, DC's can be linked to each other.

Edisyn
European Dialect Syntax

Thank you

- Edisyn:
  www.dialectsyntax.org
- Edisyn search engine:
  www.meertens.nl/edisyn/searchengine

Edisyn
European Dialect Syntax