# 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 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 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 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 tag set

#### Screenshot of Edisyn tag set:

			gs Manual Input						
	Fe	atures							
A	cas	se							
С		ab	abl	acc	ad	add	all	com	
Conj		dat	el	es	gen	ill	ines	inst	
Infmrk		loc	nom	partit	term	tr			
N		gender							
Р		f	m	neut					
Pron	nu	mber							
[]		1	2	3	pl	sg			
	pos	sition							
		encl	free	nomin	prep	post	postnomin	prenomin	
		procl							
	pro	nominal							
		art	def	dem	dim	exis	indef	num	
		pers	poss	quant	recipr	refl	rel	wh	
	ver	bal							
		act	asp	aux	caus	erg	fin		
		ger	imp	indic	infin	intrans	mod	participle	
		pass	past	perf	pres	subj	trans	unacc	
		unerg							
	oth	er							
		comp	coord	foc	neg	subord	sup		
							Move left	Move right	Delete
	C Conj Infmrk N P Pron	Conj Infmrk N P P Pron Infmrk Conj Infmrk P P P P P P P P P P P P P	C ab Conj dat Infmrk loc N gender P f Pron number [] 1 position encl procl pronominal art pers verbal act ger pass	C ab abl   Conj dat el   Infmrk loc nom   N gender nom   P f m   Pron 1 2   position encl free   procl pronominal art   art def pers poss   verbal act asp ger imp   pass past unerg other other other other	ab abl acc   Conj dat el es   lnfmrk loc nom partit   N gender     P f m neut   Pron 1 2 3   position encl free nomin   procl pronominal     art def dem    pers poss quant    verbal act asp aux   ger imp indic   pass past perf    unerg other	ab abl acc ad   Conj dat el es gen   Infmrk loc nom partit term   N gender  f m neut   P f m neut number number   Pron 1 2 3 pl   position encl free nomin prep   procl pronominal art def dem dim   pers poss quant recipr verbal   act asp aux caus ger infin   pass past perf pres unerg other	ab abl acc ad add   Conj dat el es gen ill   Infmrk loc nom partit term tr   N gender  f m neut   P f m neut  sg   Pron 1 2 3 pl sg   position  encl free nomin prep post   procl   att def dem dim exis   pers poss quant recipr refl verbal act asp aux caus erg   ger imp indic infin intrans pass past perf pres subj   unerg      act asp subj   inerg	ab abl acc ad add all   Conj dat el es gen ill ines   Infmrk loc nom partit term tr   N gender  m neut   P f m neut    Pron number    gender   m 1 2 3 pl sg   position     postnomin   procl      postnomin   procl          gender          f f m neut	ab abl acc ad add all com   Conj dat el es gen ill ines inst   Infmrk loc nom partit term tr inst inst   Infmrk gender r term tr term tr inst   P f m neut number r inst inst   Pron 1 2 3 pl sg sg r   position encl free nomin prep post postnomin prenomin   procl pronominal art def dem dim exis indef num   pers poss quant recipr ref1 rel wh   verbal act asp aux caus erg fin fut   ger imp indic infin intrans mod participle   pass past perf pres sub; trans



### 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 – ISOcat

# Screenshot of conversion table Edisyn – ISOcat – other corpora:

Category:	Edisyn search engine	ISOcat	SAND (Dutch dialects)	Cordial-Sin (Portuguese dialects)		i EMK (Estonian	(NDC (Nordic Dia
Infinitive marker	Infmrk	1896 infinitive particle	inf-mrk				inf-marker
Noun	Ν	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



# 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.



### **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.



# Thank you

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

