



SoNaR

Tender project written Dutch corpus

Consortium:

Radboud University Nijmegen, Tilburg University, University College
Ghent, KU Leuven, Twente University & Utrecht University

Outline

- Pilot project
- SoNaR project
 - 500 MW reference corpus: acquisition & IPR, composition & annotations
 - 1 MW subcorpus with semantic annotations
- Evaluation
- Dissemination
- Concluding remarks
- Recommendations

Dutch Language Corpus Initiative (D-Coi)

Pilot project 2005-2006

- aimed at the production of a blueprint for the construction of a large reference corpus of written Dutch
- entailed
 - the design of the corpus, incl. a user requirements study
 - the development/adaptation of protocols, procedures and tools needed for
 - sampling
 - cleaning up
 - converting file formats
 - marking up
 - annotating & post-editing
 - validating
- 50 MW pilot corpus to support developments and relieve strong need for data

Dutch Language Corpus Initiative (D-Coi)

Adhered and contributed to (inter)national standards and best practices

Results included

- Motivated design for large Dutch reference corpus
- D-Coi XML format
- Adapted version of CGN POS tagset & tagger/lemmatizer
- Adapted version of Alpino parser
- Annotation schemes for semantic role labeling and the annotation for spatio-temporal semantics
- 54 MW pilot corpus
- Adapted version of COREX corpus exploitation software

STEVIN Nederlandstalig Referentiecorpus (SoNaR) project

Directed at the actual construction of

a 500 MW reference corpus of contemporary standard written Dutch as encountered in texts originating from the Dutch speaking language area in Flanders and the Netherlands as well as translations published in and targeted at this area

Where possible, (re)use of formats, tools, protocols, and annotations/
annotation schemes previously developed (D-Coi but also COREA, DPC,
Lassy)

All data to be converted to a standard XML format and annotated
(automatically) for POS, lemmata, and named entities.

For 1 MW, semantic annotation – manually verified:

- named entities
- semantic roles
- co-reference
- spatio-temporal semantics

500 MW Reference corpus (SoNaR-500)

- includes native speaker language and the language of (professional) translators
- approx. two-thirds of the data originate from the Netherlands and one-third from Flanders
- only texts included from the year 1954 onwards
- comprises full texts rather than text samples
- includes texts from a wide range of text types incl. texts from the new media
- for (almost) all data IPR has been arranged
 - Tweets
 - chats

SoNaR-500: Composition

[table presenting overview of components/text types and amounts of text]

Acquisition and IPR

- Design served to guard the diversity of text types
- Acquisition process (incl. identifying data providers, negotiating IPR):
 - acquisition targeted at obtaining texts from many different sources
 - time-consuming
 - success difficult to predict
 - successful contacts from time to time yield much more data than envisaged in the design (e.g. Politics.be, subtitles) → surplus material

Note: on principle no payment for acquiring texts

- Expertise consolidated in an acquisition manual (for future reference)

IPR

Different types of arrangement

- Creative Commons, GPL, ...
- By arrangement of law (the public's right to information)
- (Model) Licence and standard agreement drawn up by HLT Agency legal advisor
- Implicit consent through donation

For certain data:

- use restricted to non-commercial use
- users to anonymize data in presentations

Corpus pre-processing

Directed at making the incoming data stream suitable for further upstream processing & involved

- text conversion:
of different file formats (pdf, html, MS Word, txt, ...) to standard XML format
- text tokenization and sentence splitting
- text normalization and correction
- language recognition

SoNaR-500: Linguistic annotations

- POS tagging and lemmatisation using FROG
 - for all data except texts from social media
 - tagset is essentially the set used in the CGN project (slightly extended)
 - accuracy lies around 96.5 % (correct tags; 98.5 on main tags)
- NE labeling by means of NE classifier developed on the basis of SoNaR-1

SoNaR-1: Corpus composition

Corpus includes wide range of text types:

	# words		# words
Administrative texts	28,951	Manuals	5,698
Autocues	184,880	Newsletters	5,808
Brochures	67,095	Newspapers	37,241
E-magazines & e-newsletters	12,769	Policy documents	30,021
External communication	56,287	Press releases	15,015
Instructive texts	28,871	Proceedings	6,982
Journalistic texts	81,682	Reports	20,662
Legal texts	6,468	Websites	32,222
Magazines	117,244	Wikipedia	260,533

SoNaR-1: Annotation of named entities

- Goal was to create a balanced data set labeled with NE information, which would allow for the creation and evaluation of supervised machine learning NE recognizers
- Applied to wide variety of text types and genres in order to allow for a more robust classifier and better cross-corpus performance
- Guidelines developed on the basis of annotation schemes developed in ACE and MUC, and the work on metonymy from Markert & Nissim

SoNaR annotation scheme characterized by

- finer granularity than other schemes (incl. NE main type, subtype, usage and – where applicable – the metonymic role)
- differentiation between the literal and metonymic use of entities

SoNaR-1: Annotation of co-reference

- Based on guidelines developed in COREA project
- MMAX2 used as annotation environment
- Annotation of co-reference links between nominal constituents using four relations:
 - Identity (NPs referring to the same discourse entity)
 - Bound
 - Bridge (as in part-whole, superset-subset relations)
 - Predicative
- Flagging special cases, viz. negations and expressions of modality, time-dependency and identity of sense

Example:

Het is een eer om hier te zijn op *MGIMO* [id="1"]. *Deze prachtige universiteit* [id="2" ref="1" type="ident"] is *een kweekvijver voor diplomatiek talent* [id="3" ref="1" type="pred"]. *Deze instelling* [id="4" ref="1" type="ident"] heeft hechte contacten met Nederland.

SoNaR-1: Annotation of semantic roles

- Adaptation of the PropBank annotation scheme
- Annotation builds upon manually corrected (Alpino) dependency trees
- TrEd used as annotation environment
- Bootstrapping approach: annotation task is in fact verification task

Examples:

Nederland(Arg0) | gaat | de bestrijding van het terrorisme (Arg1) | anders en krachtiger (ArgM-MNR) | aanpakken (PRED).

Minister Donner van justitie (Arg0) | krijgt (PRED) | verregaande bevoegdheden in die strijd (Arg1).

Binnen in de gymzaal (ArgM-LOC) | plakken (PRED) | gijzelaars (Arg0) | de ramen (Arg1) | af en |plaatsen (PRED)| ze (Arg0) | explosieven(Arg1)| aan de muur (Arg2).

SoNaR-1: Ann. of temporal and spatial entities

STEx (Spatio Temporal Expressions) annotation scheme

- Combines rules with large spatio-temporal knowledge base, the Varro toolkit and TiMBL
- Builds upon information available through previous syntactic and semantic layers
- Automatic annotation + manual verification

Example:

Zij hebben hun zoon *gisteren* [temp type="cal" ti="tp-1" unit="day" val="2008-05-22"] in Amsterdam [geo type="place" val="EU::NL::-::NH::Amsterdam::Amsterdam" coord="52.37,4.9"] *gezien* [temp type="event" value="vtt" rel="before(ti,tp)"]

Quality Control

- User questionnaire; results presented 4 Oct. 2010
- Internal validation of manual annotations
- External validation by CST
 - primarily directed at:
 - does the corpus contain what is documented
 - does the documentation properly describe the corpus contents

Dissemination

All results will be available through the Dutch HLT Agency:

- SoNaR-500 — multi-purpose corpus in XML format
POS tagging & lemmatization & NE*
- SoNaR-1 — subset of SoNaR-500
[syntactic annotation (D-Coi, Lassy)]
semantic annotations (SoNaR)
 - NE labeling
 - co-reference
 - semantic roles
 - spatio-temporal semantics

* No linguistic annotations for data from social media

Concluding remarks

SoNaR project has definitely yielded value for money:

- It has filled major gaps in Dutch language resources infrastructure
- It has contributed to the development and consolidation of (de facto) standards
- It has yielded tools and procedures that were/are being used in various other projects

Points for criticism:

- Sparseness of metadata
- Lack or shortage of certain types of text: e.g. Email, sms, chats, private letters, ...

Recommendations

Future investment in

- the development/adaptation of corpus exploitation software
- development/adaptation of tools for the linguistic annotation of texts from the social media
- the processing of all surplus material (substantial amounts already processed) and making these data available
- making the corpus fully CLARIN compatible