

Dutch Language Corpus Initiative *or* D-Coi project

Consortium:

RU, UTwente, UvT, KU Leuven, RUG, RUU, Polderland

Overview

- Aims
- Results
- Evaluation
- Dissemination
- Summing up
- Follow-up

D-Coi project: aims

- to produce a blueprint for the construction of a large corpus of written Dutch
 - the design of a large (500 MW) corpus of written Dutch
 - the development of all protocols, procedures and tools needed for the compilation and annotation of the data
- to demonstrate the feasibility of the approach
 - the compilation of a 50 MW pilot corpus, partially enriched with linguistic annotations
- to adapt the COREX software for use with written data

Results: corpus design

A motivated design was made for a 500 MW reference corpus for use in different types of linguistic (incl. lexicographic) and HLT research and the development of applications; corpus is to include

- contemporary (post-1954), standard written Dutch texts
- texts originating from the Dutch speaking language area in Flanders and the Netherlands as well as Dutch translations published in and targeted at this area
- native speaker language and the language of (professional) translators
- conventional genres and texts from new media

Results: protocols, procedures and tools

Development/adaptation of protocols, procedures and tools for

- the conversion of different formats into a common XML format
- lexical normalization (e.g. run-on words, split words)
- tokenization
- POS tagging and lemmatization: adaptation of CGN manual; adaptation of tagger/lemmatizer
- syntactic annotation: adaptation of Alpino; documentation
- COREX (corpus exploitation software)

Results: D-Coi pilot corpus

- 54 MW corpus, partly enriched with linguistic annotations
 - XML format
 - normalized (lexically)
 - tagged for part-of-speech and lemmatized;
500,000 words manually verified
 - parsed (syntactically) with Alpino;
200,000 words manually verified

Results: semantic annotation

Two pilot studies:

1. Annotation of semantic roles
development of annotation scheme (Propbank-like);
annotation of 2,000 sentences from D-Coi corpus
2. Annotation of temporal and spatial semantics
development of annotation scheme;
annotation of 3,000 sentences with MiniSTex

External evaluation

Carried out by CST, Copenhagen
Final report expected shortly

Dissemination

All results - various protocols, tools* and the pilot corpus are/will be available through the Dutch HLT Agency

* in so far as these are not already available under GPL or similar license

Summing up

- D-Coi project has been very successful at what it set out to do, viz. to develop a blue-print for a 500 MW reference corpus and put in place all the necessary protocols, procedures and tools
- two tasks, however, were severely underestimated:
 - data acquisition, esp. arranging IPR
 - data conversion, esp. pdf and html

Summing up: IPR

IPR is sought

- to the fullest extent for every text in the corpus
- beyond “for research purposes only”
- also for commercial applications

Strategies:

- Mode 1: through dedicated staff
- Mode 2: drawing on personal networks

Summing up: IPR

Mode 1: through dedicated (hired) staff whose task it is

- to identify likely text providers (content owners)
- to contact them
- to get them to see the importance of the undertaking
- to get them to sign a (simple) contract
- to get them to deliver their texts – preferably in a ‘nice’ format

Mode 2: drawing on personal networks; task is the same as under mode 1

Summing up: Data conversion

There is no such thing as 'nice' text formats!

- all collections of texts, whatever format they are in, seem to have quirks
- conversion into the common XML format requires ad hoc solutions
- occasionally manual intervention on the text level cannot be avoided

Summing up: Data conversion

There is no such thing as 'nice' text formats!

- all collections of texts, whatever format they are in, seem to have quirks
- conversion into the common XML format requires ad hoc solutions
- occasionally manual intervention on the text level cannot be avoided

Follow-up

D-Coi results are (also) being used in

- in other STEVIN projects; e.g. Corea, DPC and Lassy
- in various other projects, incl. CONLL shared task, METIS II, NWO Vici project 'Implicit Linguistics'

The construction of the full (500 MW) reference corpus is undertaken in the SoNaR project