Universiteit Utrecht

The DUELME-LMF Project

Jan Odijk

CLARIN-NL Meeting 19 Feb 2010

Utrecht Institute of Linguistics OTS

Overview

- DUELME-LMF Project
- DUELME Database
- Curation of the Database
- References

DUELME-LMF Project

- 1 Mar 2010 – 31 Aug 2010
- Participants:
  - Uil-OTS, Utrecht (Nicole Grégoire, Jan Odijk)
  - INL, Leiden (Valentijn Geirnaert, Remco van Veenendaal)
- Data Curation Project
- Funded by CLARIN-NL (www.clarin.nl)

DUELME

- Lexical database of 5000 Dutch Multiword Expressions (MWEs) in accordance with parameterized Equivalence Class Method (ECM)
- MWEs (semi-automatically) selected from large text corpora:
  - TWNC02 (500M tokens)
  - CLEF corpus (80M tokens)
- MWE properties supported by (statistics of) occurrences in these corpora

DUELME

- ECM:
  - Standardized representation of MWEs
  - Method for incorporating MWEs into NLP-systems
- Additional:
  - Formalized syntactic tree model for each equivalence class
  - based on CGN syntactic structures (extended)
  - ➔ allows other methods than ECM for incorporation into NLP-systems
  - Incorporation experiments in Alpino

DUELME

- Data validated by CST, Copenhaguen
  - “a skillfully elaborated language resource of a high quality“
- Includes web-based GUI for searching, viewing and editing
- Data and Documentation available via TST-Centrale
  - Download
  - Online access (searching & viewing only)
  - http://www.inl.nl/index.php?option=com_content&task=view&id=611&Itemid=667

Research Questions

- Multi-word expressions
  - Disciplines
    - Lexicography
    - Theoretical linguistics
    - Psycholinguistics
    - Natural Language Processing
    - ….
  - Properties
    - Composition
    - Syntactic structure
    - Modifiability
    - Variation
    - …

DUELME Format

- DUELME format is idiosyncratic
  - Data (exchange) as a set of CSV files
  - Data (internally) as a Relational DB (mySQL)
  - LaTeX-based tree model notation
  - Attributes and values CGN-based but not identical and extended

DUELME Curation

- Converters
  - DUELME format ➔ LMF-compatible XML format
  - LMF-XML format ➔ DUELME format, or Relational DB format
    - ➔GUI remains useable
    - Investigate generic LMF editors (e.g. LEXUS)
- Find and apply XML-based standard tree (model) representation
- Apply the converter to create a curated resource

DUELME Curation

- Create CMDI Metadata for the curated resource
- Map data categories (attributes, values) to ISOCAT data categories or extend ISOCAT
- Report on any problems / desiderata / requirements for the CLARIN infrastructure or its components

Thanks for your attention

References

- CLEF Corpus:
  - http://clef-qa.itc.it/2004/resources.html
- ISOCAT: http://www.isocat.org
- DUELME http://duelme.inl.nl/
- ECM:
  - Odijk(2004) ‘Reusable Lexical Representations for Idioms’, LREC Proceedings pp. 903-906.
- Lexus: http://www.lat-mpi.eu/tools/lexus
- LMF: http://www.lexicalmarkupframework.org/
- TWNC02:
  - http://wwwhome.cs.utwente.nl/~druid/TwNC/TwNC-main.html

MWEs

BACK

- Multiword Expressions are combinations of words with linguistic properties that are not predictable from the individual components or their normal mode of combination