Experimental Linguistics

Session C: Corpus Tools

Lecturer: Roland Mühlenbernd

Corpus Analysis: overview

Sources

- Text corpora (e.g. newspaper)
- Corpora of spoken language

Analysis steps (3A)

- Annotation (pos, lemmata, morphology, etc.)
- Abstraction (mapping from scheme to model)
- Analysis (statistics, generalization, etc.)

Corpus Analysis: Annotation

Common to particular annotation concepts

- part of speech (e.g. noun, verb, adjective, ad-verb, pronoun, preposition, conjunction, article)
- lemma: canonical form of a (set of) words
- morphological features (number, tense, etc...)
- syntactic dependencies (e.g. tree structure according to dependency grammar theories)
- Metaphor types and metaphor signals

Annotated Corpora & Tools

Examples

- Universal Dependencies: 150 treebanks/90 lang.
- Sketch Engine: corpora of multiple languages
- VU Amsterdam Metaphor Corpus Online
- UAM corpus tool: annotation & analysis
- Corpus-analysis.com: overview of many tools

Universal Dependencies Project

- cross-linguistically consistent treebank annotation for many languages
- research from a language typology perspective
- annotation scheme is based on an evolution of (universal) Stanford dependencies (UD)
- uses Google universal part-of-speech tags
- philosophy is to provide a universal inventory of categories/guidelines for consistent annotation of similar constructions across languages
- allowing language-specific extensions

Sketch Engine

Features:

- Parallel corpora for translation studies
- Statistical analysis
- Morphological analysis and pos tagging
- Time annotation in historical corpora
- Corpora of more that 100 languages/dialects
  - e.g: pos, lemmatization, collocations for the languages English, Italian and Greek

VU Amsterdam Metaphor Corpus

- largest available corpus of hand-annotated metaphorical language use
- uses metaphor identification protocol MIPVU
- covers 190.000 lexical units from BNC-Baby (subcorpus of British National Corpus)
- annotated for different relations to metaphor, personification, and metaphor signals

UAM Corpus Tool

- annotation of texts using annotation scheme of your design
- search engine for instances across levels
- basic statistic tools
- annotations are saved in xml file format

UMA Corpus Tool: Exercise

Part 1: adding a text corpus

- Start a new project with name tutorial
- Add files to corpus, paste from clipboard
- Add the example text, save as Obama1.txt
- Make sure that language is set to English and encoding is set to utf-8

UMA Corpus Tool: Exercise

Part 2: adding an annotation layer

- The layer specifies your annotation/abstraction
- Go to layers → add a new layer
- Manual ann. → design your own → segments → special layer? No → automatic? no
- Layer name: Participant

UMA Corpus Tool: Exercise

Part 3: editing the layer scheme

- Layers → edit scheme
- participant-1 → rename feature → human
- participant-2 → rename feature → organization
- organization → add system with features company, government-body, media
- Entity → add system → rename to form
- rename/add features common, proper, pronoun

UMA Corpus Tool: Exercise

Part 4: manual and automatic annotation

- Files → Layers: participants
- Select segments and choose type
- Save file
- Layers overview → add a new layer → automatic annotation → pos → unextended → POS
- Files → POS

UMA Corpus Tool: Exercise

Part 5: Corpus statistics

- Statistics → Lexical Patterns → show results
- Statistics → Feature Patterns → show results (change between POS and Participant and some subcategories)
- Statistics → compare datasets (POS vs else)
- Add new file with spanish Obama text, save as Obama-sp.txt (language setting: Spanish)
- Statistics → describe each file (ENG vs ESP)