Morphologically related words and their syntagmatic properties: first ideas for phase II of SFB-732

Ulrich Heid, Kurt Eberle, Kerstin Eckart

Universität Stuttgart
Institut für maschinelle Sprachverarbeitung
– Computerlinguistik –
Azenbergstr. 12
D 70174 Stuttgart

SFB-Klausur, Blaubeuren, November 2008
Overview

- On corpus linguistics: two strands of research in B3
- Reminder: current state of B3 techniques
- Proposals for extending the techniques
- Proposals for extending the linguistic application domain
- Relationship with *Incremental Specification in Context*
Two strands of research in B3 and its successor

Basic research in corpus linguistics – (how) is it possible?

- Computational linguistic and corpus linguistic tools and methods
  - B3: ambiguity awareness in data extraction
    * representing ambiguities
    * dealing with ambiguities in extraction
- Linguistic issues to which corpus linguistic data extraction is applied
  - B3: context-dependency of the sortal interpretation of DE -ung-nominalizations
- Approach for follow-up:
  - keep the two strands
  - build on and enhance current methodology
  - address new, but related linguistic issues
Current B3: ambiguity-aware data extraction
Reminder: techniques

- Analysis of DE sentences
  - dependency grammar (Lingenio, Eberle 2002):
    broad coverage, large lexicon
  - underspecified representation of syntactic ambiguities
    based on FUDRT (Eberle 2004)

- Search functions, comparable with corpus search tools:
  - search for syntactic constellations possible,
    including search for a given type of underspecification situation
  - access to all features of all context partners of an item

- Storage in a database (under way)
Current B3: ambiguity-aware data extraction

Example of techniques

- **Assumption:**
  Local adjunct to *Messung*: e-indicator:
  *Messungen in Gomel sind nötig*

- **Ambiguous PP-attachement:**
  *Messungen des Instituts in Gomel*:
  
  (a) [Messungen [des Instituts]] [in Gomel] ⇒ e
  
  (b) [Messungen [[des Instituts] [in Gomel]]] ⇒ e/o

- **Sample tool outputs:**
Current B3: ambiguity-aware data extraction

Example of techniques
Extending B3’s corpus linguistic methods
Issues for the second phase

• To be kept, deepened, reused:
  – ambiguity-aware data extraction
  – representation of analysis results in a database
  – possibility to search analysis results flexibly

• New questions to be addressed:
  – instead of single sentence data, use data from multiple, systematically related contexts:
    (a) several analyses of the same sentence (>1 tool)
    (b) analyses of sentences for related words
  – how to improve reliability of data extraction through (a), i.e. by combining different analyses
  – how to relate and compare evidence from (b) and to generalize from observed context parameters
Extending B3's corpus linguistic methods
Issues for the second phase – optional additions

Representation

- So far: work on a database to store
texts – metadata – sentences – (competing) analyses
Plus: possibility to search any combination of these

- Possible extension of comparative approach:
  - Analysis of comparable monolingual corpora:
same event or fact $\rightarrow$ different news stories:
  (roughly) same semantics $\rightarrow$ different verbalizations
  - Analysis of multilingual comparable corpora

- In all cases:
  Emphasis on representation of (possibly underspecified) data:
  - Case-studies on a few phenomena (no broad coverage intended)
  - Work on the implications for the representation in a database of
    (comparable) texts – MD – (different) sentences – (related) analyses
Extending the linguistic application domain
Morphologically related words and their syntagmatic properties (1/3)

- Morphologically related words: e.g.

<table>
<thead>
<tr>
<th>adjective or verb</th>
<th>verb</th>
<th>nominalization</th>
<th>svc</th>
</tr>
</thead>
<tbody>
<tr>
<td>breit</td>
<td>verbreitern</td>
<td>Verbreiterung</td>
<td>V. erfahren</td>
</tr>
<tr>
<td>fallen</td>
<td>wegfallen</td>
<td>Wegfall</td>
<td>in W. geraten</td>
</tr>
</tbody>
</table>

- Syntagmatic properties
  - sucategorization: e.g. *an-singen, an-schreien,... + gegen*
  - selection properties: typical subjects, objects, etc.
  - distribution: e.g. with adverbials etc.
Extending the linguistic application domain
Morphologically related words: prefix/particle verbs (2/3)

- **Examples**
  - *reden/aufreden, drängen/aufdrängen, legen/zusammenlegen, schrumpfen/zusammenschrumpfen, ...* (particle)
  - *schlagen/zerschlagen, schreiben/(Tinte) verschreiben ...* (prefix)

- **Issues:**
  - syntactic (and semantic) regularity (*gegen* x *an-Ven*)
  - stability of syntactic, selectional, distributional properties across the derivational series
  - expected combinability of verb stems/roots with affixes, expected nominalizability:
    - verification in large amounts of corpus data
  - contribution of root vs. derivation to semantics and syntax: is particle/prefix verb formation compositional?
  - productivity of the patterns (cf. work by Baayen etc.)
Extending the linguistic application domain
Morphologically related words: data analysed (3/3)

• Comparing contexts of related words
  – breit + Straße, Weg, …
  – verbreitern + Straße, Weg, …, dito Verbreiterung

• Differences:
  – erfahren + daß/w-/ob
  – Erfahrung + daß
  – in Erfahrung bringen + daß/w-/ob (Lapshinova/Heid 2007)

• Regularities, depending on prefixes/particles
  – be–verbs with transitive readings (belegen/besprühen mit)
  – an–verbs with PP gegen (anschreien, ankämpfen, ancingen)
Extending the linguistic application domain
Morphologically related words: examples

- Parallel constructions: *verbreitern* – *Verbreiterung*

  ... hatte der Bahnkörper auf vier Gleise *verbreitert* werden müssen
  ... momentan provisorisch auf elf Meter *verbreitert* wird
  ... von bisher vier auf sechs Spuren *verbreitert* werde

  ... die Basis für die EG-Verträge [...] zu *verbreitern*
  ... die ihre Geschäftsgrundlage *verbreitern* wollten

  ... gegen die *Verbreiterung* der Leonberger Strasse auf vier Spuren
  ... die *Verbreiterung* auf drei Fahrstreifen je Fahrtrichtung

  ... die *Verbreiterung* des Produktprogramms
  ... Umschichtung und *Verbreiterung* der Publikumsbasis

- Parallel behaviour in subcategorization and selection

- But: quantitative relationship is different:
  verbs: roughly equal distribution – nouns: mainly Strasse etc.
Corpus linguistic techniques vs. linguistic issues
Analyses of contexts of morphologically related words

• Comparing syntagmatic properties
  – related complementation patterns:
    verb – nominalization – svc
  – heads of complements
  – distribution with adjuncts, etc.

• Guidance from:
  – expectations driven by theoretical considerations
    on semantics of the involved building blocks (work in B4)
  – hypothesis of “inheritance”
Relationship with *incremental specification in context*

Techniques

- Ambiguity awareness: data extraction as a specification process
- Comparing data from different sentences: to observe derivationally related items, we need to have means to compare context types; these act *together* as a context from where to extract interpretable data
- Comparing concurrent underspecified analyses: techniques for specification via comparison/voting: an issue of interoperability of representations
Relationship with *incremental specification in context*

Data under analysis

- Derivation processes: can they be seen as compositional?
- Which properties come from the root, which ones from the derivational processes: are the derivational processes specification processes – in which sense?
- Role of the context for the description of properties of words:
  - sentence context
  - comparable sentences for related words