What good are ‘Nominalkomposita’ for ‘noun compounds’
Multilingual Extraction and Structure Analysis of Nominal Compositions using Linguistic Restrictors

Patrick Ziering  Lonneke van der Plas
Institute for NLP, University of Stuttgart, Germany
COLING 2014

Supported and funded by the Deutsche Forschungsgemeinschaft (DFG SFB 732-D11)

1. Controversy of Defining Compoundhood

General existence of compoundhood
There is virtually no reliable/universally accepted definition for compoundhood ([LS09])
- [Bau03]: formation of a new lexeme by adjoining two or more lexemes
- [Mar67]: No compounding word formation: EXPANSION

Distinction between compounds and phrases
Is tomato bowl a special kind of bowl (i.e., a lexeme)? (cf. deictic compounds ([Dow77]))

Solution: Linguistic tests for compoundhood
- Inseparability (black ugly bird)
- Inability to modify the modifier (very social policy)
- Spelling as one word (football, but waiting room)

2. First Extraction Iteration

Initial definition
An English word sequence is a compound, if it passes the following linguistic test:
- Spelling as one word - defined cross-lingually

Extraction process
1. Preselection of English compounds using PoS chunks (e.g., noun-prep-noun)
2. PoS error filter (e.g., stop words tagged as noun)
3. Word alignment filter (e.g., clipping determiners)
4. Closed compound restrictor for n languages CCR(n):
   The English word sequence has to be aligned to a closed compound in at least n languages
   (e.g., the German Wartezimmer for waiting room) → Optimal n for precision/recall trade-off

3. Experiment on First Iteration

Setup
- PoS tagging
- Sentence alignment (884K parallel sentences)
- Word alignment (GIZA++)
- Binary compound splitter (according to [SCA13])

Evaluation
g5 accepted/rejected samples for each database → Determination of true/false positives/negatives

<table>
<thead>
<tr>
<th>Database</th>
<th>Size</th>
<th>Precision</th>
<th>Recall</th>
<th>F-Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Basic database</td>
<td>3,178,661</td>
<td>38.0%</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>CCR(1)</td>
<td>795,518</td>
<td>84.0%</td>
<td>71.2%</td>
<td>77.1%</td>
</tr>
<tr>
<td>CCR(2)</td>
<td>495,837</td>
<td>92.0%</td>
<td>74.2%</td>
<td>82.1%</td>
</tr>
<tr>
<td>CCR(3)</td>
<td>316,330</td>
<td>98.0%</td>
<td>65.3%</td>
<td>78.4%</td>
</tr>
<tr>
<td>CCR(4)</td>
<td>143,121</td>
<td>98.0%</td>
<td>63.6%</td>
<td>77.2%</td>
</tr>
</tbody>
</table>

Controversial cases
→ further stimulate linguistic discussion:
For example: German A+N compounds (cf. [SH09]):
- strong wind ↔ de:Starkwind
- small car ↔ de:Kleinwagen
- used car ↔ de:Gebrauchtwagen

Collocation or compound?
⇒ Mostly: semantic specification (i.e., a lexeme)

4. Database case study: Bracketing compounds

The task
LEFT or RIGHT branching of tripartite noun compounds (e.g., [human rights] abuses or baby [bicycle seat])

Six aligned phrase patterns
For example:
- ADJ CNC
  geplante Bildungsreform ([education reform] plan)
- SN FC CNC
  verslagen over autoprijzen ([car prize] reports)
- SN ADJ FC SN
  consumo final de energia ([energy [end consumption]])

Method
LEFT baseline 71.1 %
Cross-lingual phrase patterns 91.6 %

References