A Taxonomy for Afrikaans and Dutch Compounds

Gerhard B van Huyssteen
Centre for Text Technology
North-West University
Potchefstroom, South Africa
gerhard.vanhuyssteen@nwu.ac.za

Ben Verhoeven
CLiPS - Computational Linguistics
University of Antwerp
Antwerp, Belgium
ben.verhoeven@uantwerpen.be
1. CompoNet
2. Existing taxonomies
3. Our taxonomy for Afrikaans and Dutch
4. Future work
CompoNet

<table>
<thead>
<tr>
<th>Compound</th>
<th>Script</th>
<th>Cat</th>
<th>Struct</th>
<th>Class</th>
<th>Endo</th>
<th>Head C</th>
<th>Head S</th>
<th>Const 1</th>
<th>Cat 1</th>
<th>LE 1</th>
<th>Const 2</th>
<th>Cat 2</th>
<th>LE 2</th>
<th>Const 3</th>
<th>Cat 3</th>
<th>Plural</th>
</tr>
</thead>
<tbody>
<tr>
<td>able bodied</td>
<td>A</td>
<td>[A+N+extA]</td>
<td>ATT</td>
<td>3</td>
<td>0</td>
<td>able</td>
<td>A</td>
<td>body</td>
<td>N</td>
<td>ed</td>
<td>extA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>acrylic fiber</td>
<td>N</td>
<td>[A+N]</td>
<td>ATT</td>
<td>2</td>
<td>2</td>
<td>acrylic</td>
<td>A</td>
<td>fiber</td>
<td>N</td>
<td></td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>aero drome</td>
<td>N</td>
<td>[N+sN]</td>
<td>SUB</td>
<td>0</td>
<td>2</td>
<td>aero</td>
<td>N</td>
<td>drome</td>
<td>sN</td>
<td></td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>aero naut</td>
<td>N</td>
<td>[N+sN]</td>
<td>SUB</td>
<td>0</td>
<td>2</td>
<td>aero</td>
<td>N</td>
<td>naut</td>
<td>sN</td>
<td></td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>aircraft carrier</td>
<td>N</td>
<td>[N+N]</td>
<td>SUB</td>
<td>2</td>
<td>2</td>
<td>aircraft</td>
<td>N</td>
<td>carrier</td>
<td>N-Vder</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>all-goes-well atmosphere</td>
<td>N</td>
<td>[CP+N]</td>
<td>ATT</td>
<td>2</td>
<td>2</td>
<td>all-goes-well</td>
<td>CP</td>
<td>atmosphere</td>
<td>N</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>apron string</td>
<td>N</td>
<td>[N+N]</td>
<td>SUB</td>
<td>2</td>
<td>2</td>
<td>apron</td>
<td>N</td>
<td>string</td>
<td>N</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>around-the-world flight</td>
<td>N</td>
<td>[PP+N]</td>
<td>ATT</td>
<td>2</td>
<td>2</td>
<td>around-the-world</td>
<td>PP</td>
<td>flight</td>
<td>N</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>astro physics</td>
<td>N</td>
<td>[sN+N]</td>
<td>SUB</td>
<td>2</td>
<td>2</td>
<td>astro</td>
<td>sN</td>
<td>physics</td>
<td>N</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>baby care</td>
<td>N</td>
<td>[N+N]</td>
<td>SUB</td>
<td>2</td>
<td>2</td>
<td>baby</td>
<td>N</td>
<td>care</td>
<td>N-Vder</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>back swimmer</td>
<td>N</td>
<td>[N+N]</td>
<td>SUB</td>
<td>2</td>
<td>2</td>
<td>back</td>
<td>N</td>
<td>swimmer</td>
<td>N-Vder</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

http://componet.sslmit.unibo.it/
Existing (amongst others)
   English = 163
   Dutch = 188

56 revisions

New

Afrikaans = 144
Problems with 2005 annotation guidelines

Often vague
  • Only one article: Bisetto & Scalise (2005)
  • Some language-specific guidelines for some languages on website

Outdated
  • Revised taxonomy: Scalise & Bisetto (2009)
  • Applications and critiques:
    • Lieber (2009a, 2009b)
    • Arcodia et al. (2009)
    • Vercellotti and Mortensen (2012)
    • Arnaud and Renner (2014)

Insufficient for language specific phenomena
  • Problematic for annotators
Bisetto & Scalise (2005)

- Head-complement (argumental) relation
- NN compounds: often of-relation
- Also synthetic and (neo)classical compounds

- Head-modifier (non-argumental) relation
- AN compounds

- Two semantic heads, one categorial head
- and-relation

Morphosemantic level
- Mostly semantic head

Semosyntactic level

Subordinate
- endo
- exo

Attributive
- endo
- exo

Coordinate
- endo
- exo

Endo
- endo
- exo

Exo
- endo
- exo

Mushroom soup
- mushroom

Mushroom cloud
- mushroom

Poet-painter
- poet
- painter

Doormat
- doormat

Other
Scalise & Bisetto (2009)

- **Primary compounds**
- **Secondary compounds**
- **Deverbal components**

- **Uninflected components**

- **Morphosyntactic level**
  - **SUB**
    - **ground** (endo: endo, exo: exo)
    - **verbal-nexus** (endo: endo, exo: exo)
  - **ATAP**
    - **attributive** (endo: endo, exo: exo)
    - **appositive** (endo: endo, exo: exo)
  - **COORD**

- **Non-head expresses property of head by means of noun acting as attribute**
- **Often metaphorical**

- **Examples**
  - *mushroom soup*
  - *taxi driver*
  - *atomic bomb (ATT) vs mushroom cloud*
Scalise & Bissett (2009)

- Secondary compounds
- Deverbal components

- Primary compounds
- Uninflected components

Morphosyntactic level

- Endo
- Exo

• Non-head expresses property of head by means of noun acting as attribute
• Often metaphorical

atomic bomb (ATT) vs mushroom cloud

Fábregas and Scalise (2012)
1. “is unclear how many languages would need this category, given the difficulty distinguishing the category”
2. “‘appositive’ is already in the literature as a type of coordinate compound”
Arnaud & Renner (2014)

NN composites

- subordinative
  - relational ←→ attributive
    - generic-specific
    - analogical ←→ equative

- coordinative
  - multifunctional
  - hybrid
  - additional
Van Huyssteene & Verhoeven (2014)
- Appositives differ sufficiently from coordinates and attributives
Compounding

Subordinate compound

Verbal-nexus compound
- Exo-centric
- Endo-centric

Ground compound
- Exo-centric
- Endo-centric

Phrasal compound
- Exo-centric
- Endo-centric

(Neo-)classical compound

Attributive compound

Phrasal compound
- Exo-centric
- Endo-centric

Ground compound
- Exo-centric
- Endo-centric

Appositive compound

Phrasal compound
- Exo-centric
- Endo-centric

Ground compound
- Exo-centric
- Endo-centric

Coordinate compound

Phrasal compound
- Exo-centric
- Endo-centric

Ground compound
- Exo-centric
- Endo-centric

Parasynthetic compound

Reduplication

Word-formation

Derivation

- Kind of component
- Category of component
Compounding

Subordinate compound

Coordinate compound

Classical compound

Attributive compound

Appositive compound

Endocentric compound

Verbal-nexus compound

Ground compound

Phrasal compound

(Neo-)classical compound

Phrasal compound

Ground compound

Phrasal compound

Ground compound

Ground compound

Phrasal compound

Ground compound

Phrasal compound

Ground compound

Phrasal compound

Ground compound

Phrasal compound

Ground compound

(hierarchical)

Word-formation

Derivation

Language specific/marginal

Semosyntactic level

Morphosyntactic level

Phrasal compound

Ground compound

Word-formation process

Endocentric compound

Exocentric compound

Morphosemantic level

Exocentric compound

Endocentric compound

(Morpho)semantic head

Knip+oog snip+eye ‘to wink’

.rooi+kop red+head ‘ginger (derogatory)

Only two kinds of exocentric compounds
• $[[[a]_{\text{Adj}} [b]_{\text{N}}]_{\text{NP}} [c]_{\text{N}}]_{\text{N}}$ volcanic-mountain climber
• $[[[a]_{\text{Num}} [b]_{\text{N}}]_{\text{NP}} [c]_{\text{N}}]_{\text{N}}$ three-tooth giant
• $[[[a]_{\text{P}} [b]_{\text{N}}]_{\text{PP}} [c]_{\text{N}}]_{\text{N}}$ in-house expertise
- Compounding through derivation
  *groot+skaal-s* ‘large-scale’
Well-known phenomenon in Germanic languages

\textit{op}+\textit{zoeken} \textit{up}+\textit{look} ‘look up/search for’
Compounding

Subordinate compound

Attributive compound

Appositive compound

Coordinate compound

Endo-centric compound

Phrasal compound

Ground compound

Word-formation

Derivation

Reduplication

Verbal-nexus compound

(Neo-)classical compound

Exo-centric

Endo-centric

(S)j of SEM

[[x],[y]]_k
Hence...

• rigorous/explicit annotation guidelines
  • Corrected anomalies and discrepancies
  • Formalised (Construction Morphology)
  • Terminology disambiguated and translated

• comprehensive for all compounds in Dutch and Afrikaans
  • All patterns in De Haas & Trommelen (1993) and Afrikaans literature
  • Illustrated what/how to adapt for specific languages

• Next:
  • Re-annotate data in CompoNet
  • Future research
A Taxonomy for Afrikaans and Dutch Compounds

Gerhard B van Huyssteen
Centre for Text Technology
North-West University
Potchefstroom, South Africa
gerhard.vanhuyssteen@nwu.ac.za

Ben Verhoeven
CLiPS - Computational Linguistics
University of Antwerp
Antwerp, Belgium
ben.verhoeven@uantwerpen.be

Acknowledgements

- **Funding:** Nederlandse Taalunie (Dutch Language Union), South African Department of Arts and Culture (DAC), South African National Research Foundation (NRF), European Network on Word Structure (NetWordS) (European Science Foundation)
- **Benito Trollip,** who populated the first version of the Afrikaans section in the CompoNet database
- **Anonymous reviewers** for comments and suggestions