Introduction

- Topic: semantic annotation of Dutch and Afrikaans XN compound semantics (non-noun + noun)
- Broader context: AuCoPro project
  - University of Antwerp (Belgium)
  - Tilburg University (The Netherlands)
  - North-West University (South Africa)

  - Subproject 1: compound splitting
  - Subproject 2: compound semantics

- Purpose: both computational and descriptive linguistics

Previous Research

- Majority
  - English
  - Noun-Noun compounds
    - Most productive form

- Use of computational analysis
  - Machine translation
    - Location
      e.g. *Antwerp hostel* (Eng.) -> *auberge à Anvers* (Fr.)
    - Possession
      e.g. *room number* (Eng.) -> *numéro de chambre* (Fr.)

Previous Research

Computational approach

- Noun-noun compounds
- Supervised machine learning
  - Classification with six previously annotated classes, describing e.g. location or events
  - Input for learner is context of constituents or location in a semantic hierarchy

XN Compounds

This study:

- Afrikaans and Dutch
  - Other Germanic languages?
- Non-noun noun compounds
  - Second most productive types of compounds

- No research on semantic modelling so far
  - No annotation guidelines
  - No data sets

Some Examples

	Afrikaans (Afr.)	Dutch (Du.)	English (Eng.)
NN	*tafelblad* 'table top'	*pannenkoek* 'pancake'	*car key*
VN	*faksmasjien* 'fax machine'	*leesbril* 'reading glasses'	*skateboard*
AN	*geelwortel* 'carrot'	*geelzucht* 'yellow fever'	*lightweight*
PN	*onderrok* 'underskirt'	*achterlicht* 'back light'	*undertone*
QN	*agthoek* 'octagon'	*eenoog* 'cyclops'	*twoface*

NN = Noun-noun compound
VN = Verb-noun compound
AN = Adjective-noun compound
PN = Preposition-noun compound
QN = Quantifier-noun compound

General Principles for Annotation Protocol

- Follow the general principles and use the formalisms of Ó Séaghdha (2008)
  - Compatibility with work on NN compounds
  - Set standard for larger body of work
  - Describe relation between constituents

  - Differences
    - No directionality
    - More explicit paraphrasing
    - Treatment of lexicalised compounds

- So far: top-down

1. Verb-Noun Compounds

1. Event
2. Location
3. Composed of
4. Lexicalised

VN

Note: Linguistic Description

- Distinguish between V and N

English: *swimming pool*

ing-participle as nominalisation, referring to event in general

‘pool for the act of swimming’

Dutch: *zwembad*

verb interpretation, nominalisation would be infinitive ‘zwemmen’

‘bath where one swims’

Afrikaans: *swembad*

~ Dutch

1.1. Event

~ ACTOR & INST (Ó Séaghdha, 2008)

- **Subject**
  ('N that Vs; the goal of N is to V')
  Afr. *snydokter* **cut+doctor** 'doctor that cuts; surgeon'
  Du. *gloeilamp* **glow+lamp** 'lamp that glows; lightbulb'
- **Object**
  ('N that is (being) V-ed; VN is the result of V- INF; the goal of N is to be V-ed')
  Afr. *snyblomme* **cut+flowers** 'the goal of the flowers is to be cut'
  Du. *werpbal* **throw+ball** 'ball that is thrown'
- **Instrument**
  ('N is used to V-INF')
  Afr. *kapbyl* **chop+axe** 'axe used to chop down trees'
  Du. *leesbril* **read+glasses** 'glasses that are used to read; reading glasses'

1.2. Location

~ IN (Ó Séaghdha, 2008)

- **Space**
  ('V in (neighbourhood of) N; N where one Vs')
  Afr. h*erstelsentrum* ***r*ecover+centre** 'centre where people recover from injuries or operations'
  Du. s*laapkamer* ***s*leep+room** 'room where one sleeps; bed room'
- **Time**
  ('N during which one Vs')
  Afr. b*akleifase* **quarrel+phase** 'phase during which one quarrels'
  Du. r*egeerperiode* ***r*ule+period** 'period during which someone rules'

1.3. Composed of

~ HAVE: Part-whole & Group (Ó Séaghdha, 2008)

('N consists of V')

Afr. *skokterapie* **shock+therapy** 'therapy that consists of shocking the patient'

Du. *niesbui* **sneeze+shower** 'rapid succession of sneezes'

1.4. Lexicalised

- **Endocentric**
  Afr. *snyhou* **cut+stroke** ‘kind of tennis stroke’
  Du. *draaibal* **turn+ball** ‘ball that is kicked with a turning effect’
- **Exocentric**
  Afr. *speeltuin* **play+garden** ‘playground’
  Du. *verzamelwoede* **collect+anger** ‘urge or mania to collect things’

2. Adjective-Noun Compounds

1. Lexicalised
   - Endocentric
   - Exocentric

Note: Linguistic Description

- No productive AN compounding
- A + N = NP

- But compounding when there's extension of meaning
  - E.g. *blackboard*

- When compounded: concatenated form
  - Most frequent in Afrikaans
  - E.g. Afr. *witwyn*, Du. *witte wijn*, Eng. *white wine*

- All per definition lexicalised

2.1. Lexicalised Endocentric

- **Duration**
  (‘kind of N that is A’)

Afr. *langverlof* **long+leave** ‘kind of leave that is longer than what is normally taken’

Du. no examples found

- **Colour**
  (‘kind of N that is A’)

Afr. *geelrys* **yellow+rice** ‘kind of rice that is yellow’

Du. *rodekool* **red+cabbage** ‘kind of cabbage that is red’

- **Other qualities**
  (‘kind of N that has the quality expressed by A’)

Afr. *sterkstroom* **strong+current** ‘high voltage; the power current is strong’

Du. *hogeschool* **high+school** ‘school for higher education’

2.1. Lexicalised Exocentric

- **Attributive**

Afr. *luigat* **lazy+bottom** ‘person that is lazy’

Du. *kaalkop* **bald+head** ‘person that has a bald head’

- **Other**

Afr. *groenskrif* **green+script** ‘first draft of legislation; green paper’

Du. *blijspel* **happy+game** ‘theatre play that is supposed to amuse people’

3. Quantifier-Noun Compounds

1. Quantity-Object
2. Lexicalised

3.1. Quantity-Object

Quantifier specifies the quantity of N within a larger phrasal compound

$[\ [Q+N]_{NP}\ N\ ]_{N}$

Afr. *sewejaardroogte* **seven+year+drought** 'seven-year drought'
Du. *achtmansploeg* **eight+men+team** 'team with eight members'

3.2. Lexicalised

- **Endocentric**

No examples in Afrikaans or Dutch have been found yet

- **Exocentric – Attributive**
  (compound is 'entity that has Q number of N')

Afr. *vierkleur* **four+colour** 'flag of the old Transvaal Republic'

Du. *duizendpoot* **thousand+leg** 'centipede'

4. Preposition-Noun Compounds

1. Location
2. Process-based
3. Lexicalised

4.1. Location

- **Space**
  (‘N is spatially at position P relative to G’)

Afr. *onderrok* **under+skirt** ‘skirt worn under other skirt’

Du. *achterlicht* **behind+light** ‘light at behind of car or bike; rear light’

- **Time**
  (‘N is temporally at position P relative to G’)

Afr. *voormiddag* **before+noon** ‘forenoon’

Du. *nagesprek* **after+talk** ‘conversation after previous event’

- **Abstract/Metaphorical**
  (‘N is at abstract position P relative to G’)

Afr. *byverdienste* **by+income** ‘additional income to normal income’

Du. *overgewicht* **over+weight** ‘the weight that is over the normal’

4.2. Process-based

(‘N goes in direction P’)

Afr. *opmars* **up+march** ‘march’

Du. *overstap* **over+step** ‘transfer on public transport’

4.3. Lexicalised

- Endocentric

Afr. *optog* **up+trip** ‘procession’

Du. *uitgroeisel* **out+growth** ‘excrescence’

- Exocentric

Afr. *insig* **in+sight** ‘insight’

Du. *nageboorte* **after+birth** ‘afterbirth’

Discussion

- Protocol purposes
  1. Computational linguistics
     - Classification of compound semantics
  2. Comparative descriptive linguistics

- All AN and QN compounds seem to be lexicalised
  - To be confirmed in further, corpus-driven research
  - Computational experiments only on VN and PN

- Event-based VN categories
  - More general and informative approach

Future Work

- Continuous development of current protocol
- Annotation of Dutch and Afrikaans XN compounds
    - Afrikaans: CKarma data
    - Dutch: compound list from e-Lex corpus
  - Followed by protocol adjustments (bottom-up verification)
- Start collaborations for protocol expansion to other languages

Acknowledgement

Research funded by:

- Nederlandse Taalunie (Dutch Language Union)
- Departement of Arts and Culture (DAC) of South Africa
- National Research Foundation (NRF) of South Africa

@clipsua

AuCoPro
Automatic Compound Processing

http://www.tinyurl.com/aucopro

Thank you.

For suggestions and/or questions:

Ben Verhoeven & Gerhard van Huyssteen

CLiPS, University of Antwerp, Belgium
Ben.Verhoeven@ua.ac.be

CTexT, NWU, Potchefstroom, South Africa
Gerhard.VanHuyssteen@nwu.ac.za

COMPUTATIONAL LINGUISTICS & PSYCHOLINGUISTICS RESEARCH CENTER CLiPS

ctext™

NORTH-WEST UNIVERSITY
YUNIBESITI YA BOKONE-BOPHIRIMA
NOORDWES-UNIVERSITEIT
POTCHEFSTROOM CAMPUS