Inleiding

Enkele vergelijkende bemerkingen:

- Afr.: AN samenstelling meer frequent en vaker geconcateneerd
  - Vb.: *witwyn* vs. *witte wijn*
    langverlof vs. *lang verlof*
    hardeskyf vs. *harde schijf*
- Ned.: meer links-hoofdige constructies
  - Vb.: *Zumaregering* vs. *regering-Zuma*
    Krügerkommissie vs. *commissie-Krüger*
- Afr.: veel onproductieve verbindingsklanken

Groter diachronisch verhaal

- Afr.: AN samenstelling meer frequent

  - Ned.: gelexicaliseerde frase
    1) puristische reactie tegen het Duits
    2) gevolg van minder inflectie (Hüning, 2010)

  - Afr.: puristische reactie tegen het Engels
    -> geconcateneerd (eigen hypothese)

Groter diachronisch verhaal

zwarte markt als frase in Ned. en Duits

- Duitse samenstelling is *Schwarzmarkt*
- Afr.: *swartmark*

schwarzer Markt	zwarte markt
Da ist der schwarze Markt	Daar is de zwarte markt
Der Wert des schwarzen Markts	De waarde van de zwarte markt
Ich verdanke das dem schwarzen Markt	Ik dank dat aan de zwarte markt
Ich suche den schwarzen Markt	Ik zoek de zwarte markt

Hüning, 2010

Groter diachronisch verhaal

- Ned.: meer links-hoofdige constructies

  Afr. vs. Ned.
  Vb.: *Zumaregering* vs. *regering-Zuma*
  Krügerkommissie vs. *commissie-Krüger*

  - Andere combinaties met *kabinet*, *commissie*, *regering*, *comité*, …

- Intuïtie: Franse invloed

  Vb. *cabinet Hollande, comité Martin*

  - Hypothese
    1) Overgenomen uit Frans als reactie tegen Duits
    2) Resultaat van Franse overheersing
    -> nieuwe onderzoeksvraag

Groter diachronisch verhaal

- Afr.: veel onproductieve verbindingsklanken
  - restanten van NL

-s-	fakulteit _ s + raad	faculteit _ s + raad
-e-	son(n) _ e + stelsel	zon(n) _ e + stelsel
-er-	kind _ er + moordenaar	kind _ er + moordenaar
-der-	been _ der + gestel	been _ der + gestel
-ns-	lewe _ ns + gebeure	leven _ s + gebeuren
-ens-	afkeur _ ens + waardig	afkeuren _ s + waardig
-n-	buite _ n + gewoon	buiten + gewoon
-e-	pan(n) _ e + koek	pan(n) _ en + koek
-ere	goed _ ere + trein	goederen + trein
-es-	gees _ tes + kind	geest _ es + kind

Definities

- Orthografie speelt kleine rol
  - Onderscheid functie en vorm (Booij, 2010)
  - Functie van samenstelling: naamgeving, iets als speciale categorie apart karakteriseren
  - Gelexicaliseerde frases dus ook functie van samenstelling
- Plausibel:
  - Hoe meer gelexicaliseerd, hoe vaker aaneengeschreven (heeft betekenisspecialisatie ondergaan)
    - Afrikaanse orthografie: Regel 14.31 (vb.: *geelwortel*)
  - Exocentrische samenstellingen vaker aaneengeschreven (vb.: *domkop*)

Linguïstische Analyses

Vier soorten:

1. Morpholexicologisch
2. Morphosyntactisch
3. Semantisch-syntactisch
4. Zuiver semantisch

Linguïstische Analyses

1. Morpholexicologisch (Productiviteit)

- Deictische samenstelling (deictic compound)
  - Voldoen aan een vluchtige behoefte in het discours
    - Vb. *stoelappel* (die appel die op de stoel ligt)
- Nieuwe samenstelling (novel compound)
  - Permanente naam voor de referent
    - Vb. *hoekstoel* (de stoel die gewoonlijk in de hoek staat)
- Gevestigde samenstelling (established compound)
  - Geaccepteerd door hele taalgemeenschap
  - Niet langer altijd letterlijk te begrijpen
    - Vb. *rooikop* (persoon met ros haar)
  - Meestal sprake van 'semantic drift'

Linguïstische Analyses

2. Morphosyntactisch

Indeling op basis van woordsoort

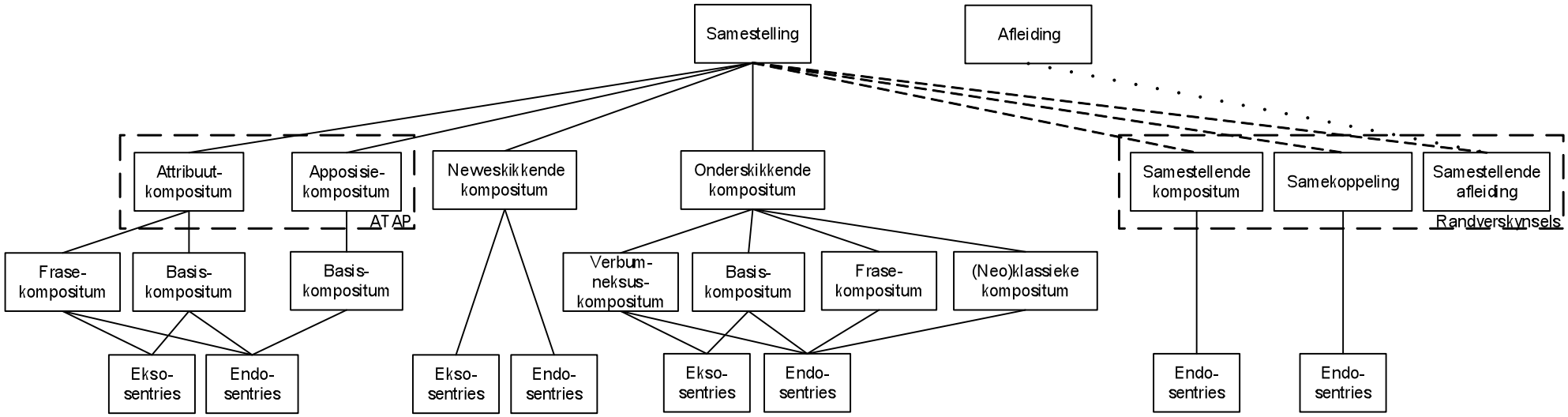
- N = nomen
- V = werkwoord
- A = adjectief
- P = voorzetsel
- Q = telwoord

NN-samenstellingen meest productief

Linguïstische Analyses

3. Semantisch-syntactisch

Nieuwe indeling uit Van Huyssteen (2014) op basis van Scalise & Bisetto (2009).



Linguïstische Analyses

3. Semantisch-syntactisch

- Nevenschikkend (CRD)
  - constituenten in 'EN'-relatie
    - Vb.: *skrywer-boer*
- Onderschikkend (SUB)
  - Complement-relatie tussen constituenten, C1 in semantische rol van C2
    - Vb.: *donderbui, slagtersmes*
- Attributief (ATT)
  - Modificeerder beschrijft eigenschap van het hoofd
    - Vb.: *fynkam*
- Appositief (APP)
  - Eigenschap van modificeerder beschrijft eigenschap van het hoofd
    - Vb.: *hoofbeginsel*

CompoNet

Beschrijvende database met voorbeelden voor alle mogelijke semantisch-syntactische categorieën.

Enkele voorbeelden:

	POS	Struct	Class	Sem Head	Synt Head	LE	Plural
blougroen	A	A+A	ATT	2	2		
blou-groen	A	A+A	CRD	12	2	-	
kwispelstert	N	V+N	SUB	2	2		2

Linguïstische Analyses

4. Zuiver semantisch

Beschrijvende benadering

Generatieve benadering

-> Falen beide door vele uitzonderingen

"Geordende chaos" (Ryder, 1994)

-> Problematisch om te schematiseren

-> Samenstellingen zijn niet semantisch karakteriseerbaar door enkele categorieën met enkele uitzonderingen te poneren. Het betreft een continuüm van minder tot meer productieve patronen.

Beschrijvende benadering toch gebruikt in computationele taalkunde, weinig alternatieven

Computationeel Onderzoek

Soorten van semantische categorieën

- Parafrases (Underlying Sentences – Lees, 1963)
  - met preposisies
    Vb.: *blomsteel* = steel VAN blom
  - vrije parafrase
    Vb.: blomsteel = die steel wat deel is van ‘n blom

- Schema met klasses (Recoverably Deletable Predicates – Levi, 1978)

  klasse AGENT

  ‘X is performed by Y’

  Vb.: *werkerstaking* = staking word gedoen deur werkers

Computationeel Onderzoek

Nut en toepassingen

- Kennis over semantiek verbetert:
  - automatische vertaling
    - Vb.: *madeirakoek* (Afr.) vs. *gâteau de Madeira* (Fr.)
    - Vb.: *sjokoladekoek* (Afr.) vs. *gâteau au chocolat* (Fr.)
  - informatie-extractie
  - vraag-antwoordsystemen

Computationeel Onderzoek

Protocol (Ó Séaghdha, 2008)

	Afrikaans	**Nederlands**
BE	*skrywer-boer*	*dichter-muzikant*
HAVE	*blomsteel*	*autodeur*
IN	*nagaktiwiteit*	*tuinfeest*
ACTOR	*beerjagter*	*studentenprotest*
INST	*tapytborsel*	*hamerslag*
ABOUT	*kategismusboek*	*postzegelverzameling*

Andere categorieën: REL, LEX, MISTAG, UNKNOWN, NONCOMPOUND

Computationeel Onderzoek

Annotatie van NN-samenstellingen

Volgens voorgaand protocol

	Ned.	**Afr.**
Aantal	1802	1500
Bron	e-Lex corpus	CKarma corpus
Subset IAA	500	1500
IAA	60.2 %	53.4 %

Computationeel Onderzoek

Lexicale similariteit

Samenstellingen zijn semantisch gelijkaardig als hun respectievelijke constituenten semantisch gelijkaardig zijn.

(Ó Séaghdha, 2008)

Vb.: *hawersak* en *sorghumblik*

soort graan + container

Computationeel Onderzoek

Distributionele semantiek

"Set van contexten waarin een woord voorkomt is een impliciete representatie van de betekenis van dit woord."

(Harris, 1968)

	leiband	loop	eienaar	huisdier	blaf
hond	3	5	5	3	2
kat	0	3	2	3	0
leeu	0	3	0	1	0
lig	0	0	0	0	0
blaf	1	0	2	1	0
kar	0	0	3	0	0

Vertaald van Baroni (2008)

Computationeel Onderzoek

Machine learning

	Constituent 1					**Constituent 2**					**Klasse**
	die	op	het	jy	kort	die	op	het	jy	kort	
oondrak	1	0	1	0	0	1	1	1	1	0	HAVE
pannekoek	1	0	0	1	0	1	1	1	0	0	INST
soengroet	1	1	0	1	1	1	0	0	1	0	BE
kardeur	1	0	1	0	0	1	1	1	1	1	HAVE

Computationeel Onderzoek

Resultaten

	Afrikaans	**Nederlands**
Baseline (meest frequente klasse)	28.2 % (407/1439) ABOUT	29.5 % (428/1447) IN
Beste resultaat	51.1 %	49.0 %

Eindexperimentje

Doorloop de vier linguïstische analyses met volgende (fictieve) samenstelling:

papierplein

Erkenning

Onderzoek befondst door:

- Nederlandse Taalunie
- 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

Bedankt.

Voor suggesties en/of vragen, contacteer:

Ben Verhoeven, Gerhard van Huyssteen & Walter Daelemans

CLiPS, Universiteit Antwerpen, Belgium
ben.verhoeven@uantwerpen.be
walter.daelemans@uantwerpen.be

CTexT, Noordwes-Universiteit, Zuid-Afrika
gerhard.vanhuyssteen@nwu.ac.za

CLiPS
Computational Linguistics & Psycholinguistics
University of Antwerp

ctext™

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