Goed of Fout- Wat gebruikt men feitelijk?

Gertjan van Noord[*] G.J.M.VAN.NOORD@RUG.NL
Jan Odijk[*] J.ODIJK@UU.NL

[**] *RUG, Groningen, Nederland*

[*] *UiL-OTS, Utrecht, Nederland*

Outline

(1) Twee doelen met dit paper:
   a. CLARIN heeft het mogelijk gemaakt dat iedere taalkundige op eenvoudige wijze taalkundig verrijkte corpora kan gebruiken in zijn/haar onderzoek
   b. Illustratie hiervan: laten zien wat sprekers van het Nederlands feitelijk gebruiken in constructies waar een normatieve en een niet-normatieve variant van bestaan, op basis van gebruikersvriendelijke webapplicaties die het mogelijk maken te zoeken in geschreven en gesproken corpora

CLARIN

(2) CLARIN-NL heeft een grote reeks applicaties gemaakt om te zoeken in taalkundig verrijkte corpora, en om de zoekresultaten te analyseren

(3) Deze toepassingen hebben gebruikersvriendelijke interfaces, meestal meerdere, afgestemd op de expertise van de gebruiker en de complexiteit van de gewenste query

(4) Voorbeelden: OpenSONAR (5), PaQu (6), GrETEL (7)

(5) OpenSONAR `http://portal.clarin.nl/node/4195`
   a. geeft toegang tot het token-geannoteerde SONAR corpus, ca. 535 miljoen tokens (Oostdijk et al. 2013)
   b. 4 verschillende verbonden interfaces
   c. tokenannotaties waarop gezocht kan worden: woordvorm, woordsoort, uitgebreide woordsoort incl. inflectionele eigenschappen, lemma
   d. binnenkort ook toegang tot het hele CGN (OpenSONARCGN)

(6) PaQu `http://portal.clarin.nl/node/4182` (van Noord et al. 2013)
   a. geeft toegang tot LASSY-Klein, een deel van LASSY-Groot (van Noord et al. 2013), de Corpus Gesproken Nederlands (CGN) Treebank (Oostdijk et al. 2002), (en sinds kort, tot een treebank van de Nederlandse CHILDES corpora) (MacWhinney 2015)).
   b. LASSY-Klein en CGN zijn manueel gecontroleerd, de andere corpora automatisch ontleed
   c. 2 interfaces, een eenvoudige om te zoeken naar twee tokens op basis van hun lemma, woordsoort, of woordvorm, en hun onderlinge grammaticale relatie; en een volledige XPATH-interface
   d. PaQu biedt uitgebreide mogelijkheden voor analyse van de zoekresultaten:

   i. tellingen van voorkomens van eigenschappen van het hoofdwoord en het afhankelijke woord (lemma, woordvorm, woordsoort), en van voorkomens van de onderlinge grammatikale relatie.
   ii. tellingen van combinaties tussen de eigenschappen van hoofwoord en afhankelijk woord, en hun onderling grammatikale relatie. De resulterende tabellen bevatten zelf weer links naar queries voor speciale subgevallen.
   iii. analyse in termen van metadata (bijv. voor CGN, subcorpus, en sprekereigenschappen (leeftijd, geslacht, land, id)).
  e. PaQu stelt je in staat je eigen (Nederlandse) corpus op te laden, automatisch te laten ontleden, waarna je erin kan zoeken en de zoekresultaten analyseren.

(7) GreTEL `http://portal.clarin.nl/node/1967` (Augustinus et al. 2012)
  a. geeft toegang tot LASSY-Klein, CGN Treebank, en het automatisch ontlede SONAR-corpus (een deel van LASSY-Groot)
  b. twee interfaces: een voorbeeld-gebaseerde interface: een query wordt automatisch gemaakt op basis van een voorbeeldzin plus een indicatie wat in de voorbeeldzin essentieel is voor de te zoeken constructie; en een XPATH-interface

(8) Deze toepassingen kunnen direct gebruikt worden. Het is nuttig een cursus te volgen om deze toepassingen optimaal te benutten. Indien geïnteresseerd in een cursus, contacteer Jan Odijk (j.odijk@uu.nl)

(9) voor een volledig overzicht zie `http://portal.clarin.nl/clarin-resource-list-fs`

(10) Voordelen van het gebruik van corpora en de bijbehorende toepassingen:
  a. ze bieden een extra onderzoeksinstrument voor taalkundigen
  b. het is *eenvoudig* relevante data te vinden (geen download van data, geen download en installatie van software, gebruikersvriendelijke interfaces)
  c. men kan relevante data *sneller* vinden dan voorheen
  d. men kan zijn onderzoek op *meer data* baseren dan met handmatige vergaring van data
  e. het betreft daadwerkelijk gebruikte taaldata

(11) Beperkingen / nadelen van het gebruik van corpora:
  a. Alle corpora (ook de manueel gecontroleerde) bevatten fouten,
  b. Alle corpora bevatten enkele inconsistente annotaties (maakt het moeilijker alle relevante gevallen te vinden)
  c. Corpora bevatten niet altijd de taalkundige informatie die noodzakelijk is (bijv. semantische informatie ontbreekt meestal)
  d. veel taalverschijnselen zijn zeer zeldzaam, dus corpora zijn heel vaak te klein...
  e. het betreft daadwerkelijk gebruikte taaldata

Constructies met een normatieve en een niet-normatieve variant

(12) normatieve en niet-normatieve varianten van constructies
  a. we zoeken naar constructies van het Nederlands met twee varianten, (meestal) een normatieve en een niet-normatieve
  b. in geschreven taal (door te zoeken in LASSY-Klein), ca 1 miljoen tokens
  c. en in gesproken taal (door te zoeken in de CGN Treebank), ca 1 miljoen tokens
  d. zowel in het Nederlands van Nederlanders als dat van Belgen (Vlamingen)

Constructie	correct?	eigen gebruik	gebruik in omgeving
hun hebben	8	8	8
jij kan	3	3	2
een aardige meisje	10	10	10
U hebt	1	2	4
hele mooie	2	1	1
hij heb	9	9	9
het boek wat	6	6	6
z'n eigen (refl)	7	7	7
de vrouw waarvan	4	4	5
een aantal mensen staan	5	5	3

Table 1: De rangorde van de centrale verschijnselen per schaal (1 = hoogst en 10 = laagst)

e. we gebruiken vooral PaQu, soms GrETEL voor hulp bij het creëren van de juiste query
f. de hoeveelheid data uit Nederland is groter dan de data uit Vlaanderen, hiervoor is beneden gecorrigeerd

(13) Dit onderzoek is complementair aan het onderzoek van (Bennis and Hinskens 2014)
a. (Bennis and Hinskens 2014) analyseerden een aantal van dergelijke constructies via een webenquête met proefpersonen o.a. aan de hand van vragen zoals
i. wat vindt u van het gebruik van de constructie? (correct)
ii. gebruikt u de constructie zelf? (eigen gebruik)
iii. hoort u het gebruik van de constructie in uw omgeving? (gebruik in omgeving)
b. resultaten, zie Tabel 1 (Bennis and Hinskens 2014, 150)

Belangrijkste bevindingen

(14) **NL**= het Nederland-deel van het corpus; **VL** = het België(Vlaanderen)-deel van het corpus

Bennis & Hinskens (B&H) voorbeelden

(15) *hun hebben*
a. B&H rangordes: **correct**=8; **eigen gebruik**=8; **gebruik in omgeving**=8
b. komt uitsluitend voor in het gesproken corpus
c. zeer weinig in verhouding tot de normatieve varianten *zij* en *ze* (factor 220)
d. komt uitsluitend voor in NL

(16) *jij kan*
a. B&H rangordes: **correct**=3; **eigen gebruik**=3; **gebruik in omgeving**=2
b. komt voor in het geschreven corpus en in het gesproken corpus
c. *jij kunt* komt vaker voor, zowel in het geschreven (factor 4,6) als in het gesproken corpus (factor 2,4)
d. komt vaker voor in NL dan VL (factor 2,1)

(17) *een aardige meisje*
a. B&H rangordes: **correct**=10; **eigen gebruik**=10; **gebruik in omgeving**=10
b. komt nauwelijks voor: 2 gevallen in het gesproken corpus

c. gevallen zonder determiner: 1 in het geschreven corpus, 2 in het gesproken corpus

(18) *U hebt*

a. B&H rangordes: **correct**=1; **eigen gebruik**=2; **gebruik in omgeving**=4
b. *u heeft* komt in het geschreven corpus meer dan 2x zoveel voor als *U hebt*
c. *u heeft* komt in het gesproken corpus slechts 1,3x zoveel voor als *U hebt*
d. *u hebt* komt vaker voor in VL dan in NL (factor 3,9)
e. *u hebt* komt vaker voor bij vrouwen dan bij mannen (factor 6)
f. *u heeft* komt vaker voor in NL dan in VL (factor 3)
g. *u heeft* komt vaker voor bij mannen dan bij vrouwen (factor 4,9)

(19) *hele mooie*

a. B&H rangordes: **correct**=2; **eigen gebruik**=1; **gebruik in omgeving**=1
b. komt vooral voor in het gesproken corpus, maar ook in het geschreven corpus
c. komt vaker voor dan *heel mooie* in het gesproken corpus (factor 0,9)
d. komt heel weinig voor in het geschreven corpus (6x, factor 72,3)
e. komt meer voor in NL dan in VL (factor 8,4)

(20) *hij heb*

a. B&H rangordes: **correct**=9; **eigen gebruik**=9; **gebruik in omgeving**=9
b. 10 voorbeelden, waarvan er maar 8 correct zijn (de andere twee hebben foute analyses in de treebank)
c. uitsluitend in het gesproken corpus
d. uitsluitend in NL
e. uitsluitend in niet-voorbereide, spontane spraak
f. zeer weinig in verhouding tot *hij heeft* (factor 266,8)

(21) *het boek wat*

a. B&H rangordes: **correct**=6; **eigen gebruik**=6; **gebruik in omgeving**=6
b. komt voor zowel in het geschreven als in het gesproken corpus
c. *het boek dat* komt vaker voor dan *het boek wat* in het geschreven en het gesproken corpus
d. maar in het geschreven corpus betreft dat een factor 30,4 en in het gesproken corpus een factor 1,4

(22) *z'n eigen* (refl)

a. B&H rangordes: **correct**=7; **eigen gebruik**=7; **gebruik in omgeving**=7
b. komt niet voor in het geschreven corpus
c. *mezelf,..zichzelf* komt vaker voor in het gesproken corpus dan *m'n eigen...z'n eigen*: factor 10,9
d. *z'n eigen* komt vaker voor in VL dan in NL: factor 4

(23) *de vrouw waarvan*

a. B&H rangordes: **correct**=4; **eigen gebruik**=4; **gebruik in omgeving**=5
b. vereist informatie over de head-noun die niet in de corpora aanwezig is (het kenmerk *menselijk*)
c. vereist momenteel manuele filtering; extensies om externe bronnen (CELEX, Cornetto) te raadplegen staan gepland

d. Na manuele filtering:
   i. in het geschreven corpus komt *van wie* vaker voor dan *waarvan* (factor 2,4)
   ii. in het gesproken corpus komt *van wie* even vaak voor als *waarvan*

(24) *een aantal mensen staan*
   a. B&H rangordes: **correct**=5; **eigen gebruik**=5; **gebruik in omgeving**=3
   b. meervoud komt ongeveer even vaak voor als enkelvoud in het geschreven corpus
   c. meervoud komt (veel) vaker (factor 5.2) voor dan het enkelvoud in het gesproken corpus

Andere verschijnselen

(25) *hun / hen*
   a. *hen* komt vaker voor als meewerkend voorwerp dan *hun*, zowel in het gesproken corpus als in het geschreven corpus (geschreven: factor 5,3; gesproken: factor 2)
   b. *hen* komt vaker voor als lijdend voorwerp dan *hun* (geschreven: factor 83; gesproken: factor 2,6)
   c. *hen* komt vaker voor als object dan *hun*: (geschreven: factor 62,3; gesproken: factor 1,6)
   d. *hun* als object is zeer zeldzaam in het geschreven corpus

(26) *hem / 'm* als onderwerp
   a. komt uitsluitend voor in het gesproken corpus
   b. komt uitsluitend voor in VL
   c. niet in voorbereide spraak (radiocolumns/commentaar, toespraken, lezingen, voorgelezen spraak)

(27) *ie* als onderwerp
   a. komt voor in het geschreven en het gesproken corpus
   b. komt meer voor in NL dan in VL (factor 3.4)
   c. in vrijwel alle subcorpora

(28) groter dan / groter als
   a. *groter als* komt nauwelijks voor in het geschreven corpus
   b. *groter als* komt meer voor in het gesproken corpus, maar veel minder (factor 8) dan *groter dan*
   c. niet in voorbereide spraak (radiocolumns/commentaar, toespraken, lezingen, voorgelezen spraak)
   d. meer in NL dan in VL (factor 2)
   e. meer door mannen dan door vrouwen (factor 2)

(29) *U is / U bent*
   a. *U is* komt alleen voor in VL
   b. maar is afkomstig van slechts 1 spreker

Conclusies

Normatieve en niet-normatieve constructies

(30) Niet-normatieve vaker/ even vaak als de normatieve variant:
   a. *een aantal mensen staan* vaker dan *een aantal mensen staat* in het gesproken corpus

b. *hele mooie* vaker dan *heel mooie* in het gesproken corpus

c. *de vrouw waarvan* even vaak als *de vrouw van wie* in het gesproken corpus

(31) Verschillen met (Bennis and Hinskens 2014)

a. *een aantal mensen staan*: vaker dan verwacht(B-H rangordes 5,5,3)

b. *het boek wat*: vaker dan verwacht in het gesproken corpus (B-H rangordes: 6,6,6)

c. *de vrouw waarvan*: vaker dan verwacht in het gesproken corpus (B-H rangordes: 4,4,5)

CLARIN

(32) CLARIN

a. biedt nieuwe onderzoeksinstrumenten voor taalkundigen

b. die het mogelijk maken *snel*, *eenvoudig* en *op gebruikersvriendelijke wijze* de empirische basis voor taalkundig onderzoek te verbreden

c. en daarmee beter en efficiënter taalkundig onderzoek mogelijk maken

References

Augustinus, Liesbeth, Vincent Vandeghinste, and Frank Van Eynde (2012), Example-based treebank querying, *in* Calzolari, Nicoletta, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Asunción Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC 2012)*, European Language Resources Association (ELRA), Istanbul, Turkey.

Bennis, Hans and Frans Hinskens (2014), Goed of fout: Niet-standaard inflectie in het hedendaags Standaardnederlands, *Nederlandse Taalkunde* **19** (2), pp. 131–184. DOI: 10.5117/NEDTAA2014.2.BENN.

MacWhinney, Brian (2015), Tools for analyzing talk, electronic edition, part 1: The CHAT transcription format, *Technical report*, Carnegie Mellon University, Pittsburg, PA. `http://childes.psy.cmu.edu/manuals/CHAT.pdf`.

Oostdijk, N., M. Reynaert, V. Hoste, and I. Schuurman (2013), The construction of a 500 million word reference corpus of contemporary written Dutch, *in* Spyns, Peter and Jan Odijk, editors, *Essential Speech and Language Technology for Dutch: Results by the STEVIN-programme*, Springer, Berlin, pp. 219–247. `http://link.springer.com/book/10.1007/978-3-642-30910-6/page/1`.

Oostdijk, N., W. Goedertier, F. Van Eynde, L. Boves, J.P. Martens, M. Moortgat, and H. Baayen (2002), Experiences from the Spoken Dutch Corpus project, *in* González Rodriguez, M. and C. Paz Suárez Araujo, editors, *Proceedings of the third International Conference on Language Resources and Evaluation (LREC-2002)*, ELRA, Las Palmas, pp. 340–347.

van Noord, Gertjan, Gosse Bouma, Frank Van Eynde, Daniël de Kok, Jelmer van der Linde, Ineke Schuurman, Erik Tjong Kim Sang, and Vincent Vandeghinste (2013), Large scale syntactic annotation of written Dutch: Lassy, *in* Spyns, Peter and Jan Odijk, editors, *Essential Speech and Language Technology for Dutch*, Theory and Applications of Natural Language Processing, Springer Berlin Heidelberg, pp. 147–164. `http://dx.doi.org/10.1007/978-3-642-30910-6_9`. http://dx.doi.org/10.1007/978-3-642-30910-6$_9$.