



Methodologische uitdagingen voor een longitudinaal corpus historisch Nederlands

Evie Coussé
Universiteit Gent



Plan voor vandaag

1. Achtergrond: Waarom ben ik geïnteresseerd in historische corpora? Wat heb ik erover te vertellen?
 2. Eigen ervaringen met corpussamenstelling
 3. Lessen en wensen voor de toekomst
-



Achtergrond





In het jaar 2004...

Start van doctoraatsonderzoek: de geschiedenis van de werkwoordsvolgorde in het Nederlands

= historische taalkunde

Onderzoeksmateriaal: historische teksten in het Nederlands vanaf de vroegste bronnen tot vandaag ⇒ historische trends ontdekken



Inventaris van digitale corpora

- ***Corpus Gysseling***
ambtelijke en literaire teksten uit de 13^{de} eeuw
- ***Corpus Van Reenen – Mulder***
ambtelijke teksten uit de 14^{de} eeuw
- ***Digitale Bibliotheek voor de Nederlandse Letteren***

Meer bronnen in Coussé (2007)



Problemen

- Bronnen op verschillende informatiedragers / internetlocaties ⇒ **centraliseren** nodig
- Bronnen in verschillende digitalen formaten (XML, HTML) en verschillend (of niet) geannoteerd
- Bronnen bevatten verschillende teksttypes / genres, regionale achtergrond, dialect vs. standaardtaal ⇒ representatieve **steekproef** nodig



Een eigen corpus samenstellen





Methodologische uitdagingen

Doel: samenstelling van een evenwichtig longitudinaal corpus historische teksten vanaf de vroegste bronnen tot vandaag

Uitdagingen: de externe geschiedenis van het Nederlands en literaire ontwikkelingen bemoeilijken de vergelijkbaarheid van teksten over de eeuwen heen



Externe geschiedenis Nederlands

- **Verschriftelijking (vanaf 13^{de} eeuw)**
in regionale schrijfcentra (bv. stedelijke kanselarijen) begint men in de volkstaal te schrijven met behulp van het Latijnse alfabet
 - Spellingsvariatie
 - Dialectaal gekleurd: klanken – woordenschat – grammatica
 - Aanzetten tot uniformering



- **Standaardisering (vanaf 16^{de} eeuw)**
in het noorden van het taalgebied probeert men de schrijftaal te uniformiseren, cf.
 - Boekdrukkunst creëert grotere afzetmarkt
 - Humanisme streeft naar een gecultiveerde taal naar het voorbeeld van de klassieken
 - Culturele bloei in Holland met invloedrijke schrijvers die navolging vinden (de zgn. ‘Gouden Eeuw’)
 - Vertaling van Statenbijbel: taalpolitiek



Literaire geschiedenis Nederlands

- **Middeleeuwse literatuur**
 - orale traditie: soms pas veel later opgeschreven
 - anonieme auteurs: dialectaal kleuring van teksten onzeker + latere aanpassingen door kopiisten
 - relatief beperkt register: epische gedichten, liederen, kluchten, receptenboeken, ambtelijke teksten



- **Ontwikkelingen sinds renaissance**

- geschreven traditie + snelle verspreiding door boekdrukkunst ⇒ nauwkeurige datering
- auteurs bekend ⇒ regionale herkomst beter te controleren + sociale achtergrond
- emancipatie van het Nederlands als taal voor politiek, hogere cultuur, wetenschap ⇒ nieuwe genres (wetenschappelijke, politieke en religieuze traktaten, tragedies, sonetten)
- verschuiving naar prozaliteratuur (roman)



Mijn antwoord op de geschetste uitdagingen (binnen de mogelijkheden van een éénmansproject met relatief weinig geld en weinig tijd)

Het **Compilatiecorpus Historisch Nederlands** (Coussé 2010)

- + 600 000 woorden
 - compilatie van bestaande bronnen
 - twee deelcorpora met ambtelijke / narratieve teksten
-



Deelcorpus ambtelijke teksten

- Beoogt het regionaal gekleurde taalgebruik te representeren dat typisch is voor de geschreven taal uit de Middeleeuwen
- Bevat lokale ambtelijke teksten uit Vlaanderen, Brabant en Holland van 1250-1800
- Bronnen: Selectie van het Corpus Gysseling (13^{de} eeuw), Corpus Van Reenen-Mulder (14^{de} eeuw) en scans van papieren oorkondenuitgaven (15-18^{de} eeuw)



	Vlaanderen					Brabant					Holland					Totaal
	Brug.	Ieper	Gent	Kotr.	Oud.	Antw.	Breda	Bruss.	Leuv.	Mech.	Amst.	Dordr.	Goud.	Haarl.	Leiden	
1250-74	22829	89	3770							2863						29551
1275-99	20255		16319		699			601		15242		11479			130	64725
1300-24	2153	234	531		1387	2428	650	1932	383			3894	239	425	686	14942
1325-49	1790	1734	1002		2575		1172	1948			360	1377	3230	662	2960	18810
1350-74	1910	1377	1717	286	3070	400	1769	1464			1879	1921	2858	1340	2384	22375
1375-99	1801	2442	1054	511	2677	1415	2507	1519			2786	2662	2909	2891	2653	27827
1400-24	1389	1528	2836	1760	233	82	1328	1420			1543	2735	2975	2797	628	21254
1425-49	2944	1008	1977		602		2860	1664			2767	2346	444	3312	2596	22520
1450-74	3159	309	2368	368			3083				2301	2730	2063	1991	2138	20510
1475-99	2740	423	3002	426	102		2104				2547	1558	1682	2304	2803	19691
1500-24	2045	3205	2864	517	1325	761					2845	2334	2268	2805	1786	22755
1525-49	1601	1618	1280	2311	434	997						2792	1919	2428	1624	17004
1550-74	1624	592	1385	2082	1241	1349		309				2885	1981	2704	2424	18576
1575-99	2422		2497	178	185	890						2197			2168	10537
1600-24	228	1270	1456		192						2890				1728	7764
1625-49	1233	122	2033		659						2886				2324	10255
1650-74		2763	2334					3607			2557				2254	13515
1675-99	790								1669						2228	4687
1700-24	825				442										2414	3681
1725-49	1265	2400	351						4812						1498	10326
1750-74	2877	1197	215												2880	7169
1775-99					380										2313	2693
Totaal	75880	22311	51156	8439	16203	8322	15473	15462	6864	18105	25361	40910	22568	23659	42619	393332



Deelcorpus narratieve teksten

- Beoogt de geschreven standaardtaal in Holland evenwichtig te respresenteren vanaf de zestiende eeuw tot vandaag (1575-2000)
- Bevat narratieve prozateksten (bv. romans, nouvelles, politieke traktaten)
- Bronnen: Selectie van het Digitale Bibliotheek voor de Nederlandse Letteren



Troeven van het corpus

- De zorgvuldige selectie van de teksten volgens een aantal welgedefinieerde criteria zoals de regionale herkomst en de datering van de teksten.
- Longitudinaal corpus maakt het mogelijk om taalkundig onderzoek te doen van 1250-2000.



Syntactisch onderzoek

- Woordvolgorde in tweeledige werkwoordclusters
- Extrapositie van subject en direct object
- Grammaticalisatie van de ambtelijke formule *zoals voorzien is*
- Grammaticalisatie van het *hebben*-perfectum

Morfologisch onderzoek

- Ontstaan van complexe persoonlijke meervoudspronomina (bv. *jullie*, *haarlieden*)
-



Beperkingen van het corpus

- Relatief klein: onderzoek naar zeldzamere constructies valt buiten mogelijkheden (bv. ditransitieve constructie)
 - Geen annotaties (annotaties van Corpus Gysseling en Van Reenen – Mulder gewist)
 - Afkortingen stilzwijgend opgelost in papieren oorkondenedities
-



Lessen en wensen voor de toekomst





Lessen voor de toekomst

- Meer zorg aan de annotatie van het corpus:
cf. eerste doel van het corpus was eigen
proefschrift, pas later groeide het besef dat het
corpus ook nuttig kon zijn voor een ruimer publiek

(binnenkort beschikbaar via TST-centrale)



Wensen voor de toekomst

- Vervolgcorpus op Corpus Gysseling en Corpus Van Reenen-Mulder maken: zelfde spreiding, nauwkeurigheid, annotatie (bv. mijn tekstscans waren noodoplossing)
 - Uniformeren van de corpora ambtelijke teksten
 - Interface bouwen dat doorzoeken vergemakkelijkt
 - Doorzoekbaarheid Digitale Bibliotheek voor de Nederlandse Letteren verhogen
-



Bedankt voor uw aandacht!

Voor meer informatie:

evie.cousse@ugent.be

<http://www.flw.ugent.be/eviecousse>
