

1 Nederlandstalige corpora

1.1 Eindhoven Corpus (EHC) of Corpus Uit den Boogaart

Inhoud: de eerste verzameling van Nederlandstalige gesproken en geschreven teksten, 1960-1973

Omvang: geschreven taal: 600.000 woorden; gesproken taal: 120.000 woorden

Doelstelling: aanvankelijk om een frequentielijst van het Nederlands te kunnen samenstellen, later werd het corpus ingezet bij allerlei vormen van onderzoek binnen de taalkunde en de taaltechnologie. Anno 2006 is het corpus ook van belang voor diachroon taalkundig onderzoek.

Annotatie: het corpus is handmatig (vrijwel foutloos) geannoteerd, waardoor het corpus geschikt is als training- en testcorpus bij de ontwikkeling van part-of-speech-taggers

Beschikbaarheid: beschikbaar gratis via de TST-centrale:
<http://www.tst.inl.nl/producten/?MenuSelection=producten>

Info: <http://ww2.tst.inl.nl/index.php?option=content&task=view&id=370>
<http://www.ccl.kuleuven.be/about/ANNO/TEKST/spraakcorp.html>

1.2 38 miljoen woorden corpus (INL)

Inhoud: bestaat uit drie hoofdcomponenten: een gevarieerd samengestelde component, een component krantentekst (Meppeler Courant) en een juridische component, 1996

Omvang: 38 miljoen woorden

Doelstelling: voor allerlei types onderzoek (lexicaal, morfologisch, syntactisch, ...)

Annotatie: automatisch taalkundig verrijkt met een lemma (trefwoordvorm) en twee woordsoorttoekenningen: een globale (13 woordsoortcategorieën) en een verfijnde (met subcategorisatie) conform de MECOLB standaard

Beschikbaarheid: via het internet na de ondertekening van een Individuele gebruikersovereenkomst, te verkrijgen op de INL-site
http://www.inl.nl/index.php?option=com_content&task=view&id=57&Itemid=94

Info: http://www.inl.nl/index.php?option=com_content&task=view&id=57&Itemid=94
http://users.ugent.be/~tcollema/Website_heuristiek/Links_eletriek.html

1.3 PAROLE corpus

Inhoud: een verzameling modern-Nederlandse teksten, voornamelijk teksten uit kranten en tijdschriften, 2004

Omvang: ca. 20 miljoen woorden

Doelstelling: voor onderzoek van morfologische, lexicologische en - in beperkte mate - syntactische aspecten van het hedendaags Nederlandse taalgebruik en voor onderwijsgeevenden in de corpuslinguïstiek.

Annotatie: de teksten zijn verrijkt met een codering van typografie en tekststructuur. De woordvormen in de teksten zijn automatisch taalkundig verrijkt met een gedetailleerde woordsoortcode en een trefwoord (lemma)

Beschikbaarheid: kosteloos te raadplegen via het Internet met inachtneming van restricties http://parole.inl.nl/html/main_info_dutch.html

Info: http://parole.inl.nl/html/main_info_dutch.html
http://www.inl.nl/index.php?option=com_content&task=view&id=58&Itemid=95

1.4 CLEF datacollectie

Inhoud: bestaat uit volledige jaargangen 1994, 1995 van Algemeen Dagblad, NRC Handelsblad, omvat drie onderdelen: tekstuele documenten, zoekvragen ('queries'), en relevantiebepalingen (de 'goede antwoorden')

Omvang: 79 miljoen woorden

Doelstelling: als testcorpus voor de evaluatie van informatie-extractiesystemen, ontwikkeld aan de Universiteit Twente in samenwerking met NIST (Gaithersburg, VS), IZ Socialwetenschappen (Bonn), IEI-CNR (Pisa), UNED (Madrid) en Universiteit Hildesheim; het Nederlandse CLEF corpus is onderdeel van een groter meertalig testcorpus voor informatie-extractie, waarvan ook krantenmateriaal in het Duits, Engels, Frans, Italiaans en Spaans deel uitmaakt.

Annotatie: door de 50 zoekvragen en bijbehorende relevantiebepalingen is het corpus geschikt voor het testen van zoeksystemen, filtersystemen, etc. die zich speciaal op de Nederlandse taal richten, daarnaast bevat het corpus handmatige annotaties van de uitgever van de documentcollectie (PCM Landelijke Dagbladen), zoals handmatig toegevoegde trefwoorden en classificatie naar onderwerp (bijv. sport, binnenland, financieel, etc.).

Beschikbaarheid: copyrighthouder van de documentcollectie is PCM Landelijke Dagbladen / het Parool. Het corpus wordt voor wetenschappelijke doeleinden beschikbaar gesteld enkel aan deelnemers van de officiële CLEF evaluatie. Meer informatie is te vinden op <http://parlevink.cs.utwente.nl/projects/clef.html> en <http://www.clef-campaign.org>

Info: <http://taalunieversum.org/taal/technologie/docs/daelemans-strik.pdf>
<http://parlevink.cs.utwente.nl/projects/clef.html>
<http://www.clef-campaign.org>
<http://www.let.rug.nl/~vannoord/alp/poster.pdf>

1.5 Corpus Gesproken Nederlands (CGN)

Inhoud: een databank van het hedendaags Nederlands zoals dat wordt gesproken door volwassenen in Nederland en Vlaanderen, 1998-2004

Omvang: 9 miljoen woorden

Doelstelling: voor ontwikkelingen in de taal- en spraaktechnologie en voor de taalkunde in brede zin. Tot nu toe waren alleen corpora van geschreven Nederlands beschikbaar. Dit heeft geleid tot een sterke focus op de beschrijving van aspecten van de geschreven taal, terwijl van het 'vluchtige' gesproken Nederlands vrijwel geen systematische kennis voorhanden is. Verder is een corpus gesproken Nederlands van belang voor het onderwijs.

Annotatie: Al het materiaal werd orthografisch getranscribeerd, terwijl er tevens een oplijning plaatsvond waarbij de orthografische transcriptie gekoppeld werd aan het spraaksignaal. De orthografische transcriptie vormde het uitgangspunt voor de lemmatisering en de verrijking van het materiaal met woordsoortinformatie. Verder werd er voor een selectie van één miljoen woorden een brede fonetische transcriptie vervaardigd, kwam er een geverifieerde oplijning op woordniveau beschikbaar en werd het materiaal door middel van een syntactische analyse verrijkt. Tenslotte werd een bescheiden deel van het corpus, circa 250.000 woorden, van een prosodische annotatie voorzien.

Beschikbaarheid: beschikbaar voor wetenschappelijk onderzoek en voor de ontwikkeling van niet-commerciële producten, beschikbaarheid al dan niet betaald voor verschillende doeleinden staat beschreven op http://ww2.tst.inl.nl/index.php?option=com_content&task=view&id=241&Itemid=380
<http://www.tst.inl.nl/producten/?MenuSelection=producten>

Info: <http://lands.let.kun.nl/cgn/>

1.6 Twente Nieuwscorpus

Inhoud: kranten, teletekst ondertiteling, autocues of broadcast nieuws shows en nieuws data van het WWW, 1999-2005

Omvang: 400 miljoen woorden

Doelstelling: ondersteuning van ontwikkelen en testen van statistische modellen

Annotatie: onbekend

Beschikbaarheid: beperkt, maar met de bedoeling om het corpus in de toekomst beschikbaar te stellen voor onderzoek, contact hltgroup@cs.utwente.nl ; prerelease versie beschikbaar onder voorbehoud op <http://www.vf.utwente.nl/~druid/TwNC/TwNC-main.html>

Info:

<http://www.vf.utwente.nl/~druid/TwNC/TwNC-main.html>
<http://www.vf.utwente.nl/~druid/TwNC/TwNC-main.html>

<http://parlevink.cs.utwente.nl/tkisis/onderwerp/Speech%20&%20Language%20Technology>
http://www.taskforce-archieven.nl/pdf/f_dejong_catch.pdf

2 Engelstalige corpora

2.1 BNC (British National Corpus)

Inhoud:

Geschreven gedeelte: een grote diversiteit aan teksttypes: fragmenten uit kranten (lokaal en nationaal), gespecialiseerde periodieke publicaties en tijdschriften, wetenschappelijke boeken en populaire fictie, (on)gepubliceerde brieven en memo's, school- en universiteitsessays
 Gesproken gedeelte: informele gesprekken, opgenomen in verschillende bevolkingsgroepen volgens leeftijd, regio en sociale klasse, en gesproken taal afkomstig van verschillende contexten variërend van zakelijk en overheid tot radio shows, 1991-1994

Omvang: 100 miljoen woorden: 90% geschreven, 10% getranscribeerd gesproken tekst

Doelstelling: te dienen als representatief sample van gesproken en geschreven Brits Engels vanaf de tweede helft van de 20ste eeuw tot heden

Annotatie: automatische part-of-speech tagger, een aantal structurele eigenschappen van de tekst (bv. titels, alinea, opsommingen etc.). Alle informatie over de classificatie, context en bibliografie is opgenomen bij elke tekst in een header.

Beschikbaarheid: beschikbaar op verschillende wijze: 'eenvoudig zoeken' voor gebruiksfrequenties is gratis op <http://www.natcorp.ox.ac.uk/using/index.xml.ID=simple>, daarnaast bestaat een abonnement op http://www.natcorp.ox.ac.uk/getting/index.xml.ID=order#order_BNCworld_sg beginnend met 50€ voor één gebruiker en afhankelijk van het aantal gebruikers/ computers.

Info: <http://www.natcorp.ox.ac.uk/corpus/index.xml.ID=intro>

2.2 ANC (American National Corpus)

Inhoud: gelijkaardig aan het BNC(zie 2.1) aangevuld met nieuwe genres van de laatste jaren zoals web blogs, webpagina's, chat, e-mail en rap lyrics, vanaf 1999 (2003 de eerste versie)

Omvang: er wordt nog aan gewerkt, het corpus zal ten minste 100 miljoen woorden bevatten

Doelstelling: taalkundig onderzoek, ontwikkeling van taaltechnologische toepassingen, lexicografie (woordenboeken en thesauri), een rijke bron voor gebruik in het onderwijs op alle niveaus

Annotatie: PoS tags, binnenkort ook aanvullende annotaties, inclusief syntactische annotaties, co-referentie en POS-annotaties met de 5 en 7 tags van CLAWS. Behalve annotaties zullen de documenten ook andere data bevatten zoals bigram and trigram lijsten om gratis te downloaden van de ANC-website. De nieuwe annotaties kunnen zowel samen worden gebruikt of geïntegreerd in de ANC-data verkregen via het LDC.

Beschikbaarheid: beschikbaar gesteld via het Linguistic Data Consortium (LDC) <http://www ldc upenn edu/> voor een symbolische prijs van \$75 voor niet-commerciële onderzoeksdoeleinden. Commercieel gebruik is beperkt tot de leden van het ANC Consortium (ANCC) tot de herfst 2008. Nieuwe commerciële leden kunnen zich op elk moment aanmelden bij het ANCC <http://www.americannationalcorpus.org/obtain.html>

Info: <http://www.americannationalcorpus.org/>
<http://www.americannationalcorpus.org/obtain.html>

2.3 Corpus of Professional, Spoken American English

Inhoud: een selectie van bestaande getranscribeerde gesprekken in professionele contexten die bestaat uit twee subcorpora van één miljoen woorden elk. Een subcorpus bevat hoofdzakelijk discussies uit de academische context zoals faculteitsraadvergaderingen en commissievergaderingen. Het tweede subcorpus bevat getranscribeerde persconferenties van het Witte Huis die bijna uitsluitend uit vraag-antwoord sessies bestaan, 1994-1998

Omvang: 2 miljoen woorden

Doelstelling: kan worden gebruikt voor lexicale en grammaticale analyse

Annotatie: de enige annotatie is de naam van de spreker voor de rest is er zoekmogelijkheid met MonoConc-concordancer

Beschikbaarheid: beschikbaar onder licentie \$49 (één gebruiker); \$179 (meerdere gebruikers), een gratis sample beschikbaar op de site <http://www.athel.com/cpsa.html>

Info: <http://www.athel.com/cpsa.html>
http://www.athel.com/English_corpora.html

2.4 Wall Street Journal (WSJ) - corpus

Inhoud: bestaat uit voorgelezen zinnen uit Wall Street Journal, afkomstig uit het testset van de WSJCAM0 database, opgenomen in de special ingerichte vergaderingsruimtes. De zinnen worden voorgelezen door veel verschillende sprekers (45) met verschillende accenten, waaronder een aantal niet-Engelstalige sprekers, 1989

Omvang: 78 000 zinnen (1 miljoen woorden in de Penn Treebank)

Doelstelling: ontwikkeling van microfoon array ASR (Array Switch Regulator) processing, audiovisuele ASR, audiovisuele persoonherkenning, integratie van audiovisuele persoonherkenning in microfoon array ASR processing, herkenning van Engelstalige en niet-Engelstalige sprekers, herkenning van overlappingsen in spraak

Annotatie: binnen Penn Treebank geannoteerd in de Penn Treebank stijl, syntactische annotaties

Beschikbaarheid: beschikbaar tegen betaling op de LDC-site <http://www ldc upenn edu/Catalog/CatalogEntry.jsp?catalogId=LDC94S13A>, ook als

onderdeel van Penn Treebank

<http://www ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC99T42>

Info: http://www.idiap.ch/mmm/corpora/ami_wsj

<http://www ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC94S13A>

<http://www ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC99T42>

2.5 TEC

Inhoud: bevat moderne geschreven teksten vertaald naar het Engels vanuit verschillende Europese en niet-Europese talen, bestaat uit vier subcorpora: fictie, biografie, nieuws en magazines van vliegmaatschappijen, vanaf 1996-1997

Omvang: 10 miljoen woorden

Doelstelling: om te gebruiken voor onderzoek op twee gebieden: opsporen van verschillen in taalgebruik in vertaalde en niet-vertaalde teksten binnen één taal en stilistische variaties bij verschillende vertalers

Annotatie: onbekend, wel metadata: geslacht, nationaliteit, beroep van de vertaler, vertaalrichting, bron/ doeltaal, uitgever van de vertaalde tekst enz

Beschikbaarheid: gratis beschikbaar voor onderzoeksdoeleinden op

<http://www.llc.manchester.ac.uk/Research/Centres/CentreforTranslationandInterculturalStudies/ResearchProgrammesPhDMPHil/TranslationEnglishCorpus/>

Info:

<http://www.llc.manchester.ac.uk/Research/Centres/CentreforTranslationandInterculturalStudies/ResearchProgrammesPhDMPHil/TranslationEnglishCorpus/>

http://www2.umist.ac.uk/ctis/research/TEC/tec_home_page.htm

2.6 LOB (The Lancaster-Oslo/Bergen Corpus)

Inhoud: een tegenhanger van het Amerikaanse BROWN-corpus, bestaat uit fragmenten van 15 dezelfde types tekst als BROWN, namelijk: persmateriaal, teksten over religie, teksten over vaardigheden, ambachten, hobby's, populaire kennis, biografieën, essays, fictie, wetenschappelijke teksten, andere (overheidsdocumenten, bedrijfs- en andere zakelijke rapporten, universiteitscatalogi), 1961

Omvang: 1 miljoen woorden

Doelstelling: referentiecorpus en ook voor onderzoek op het gebied van taalkunde

Annotatie: PoS tags

Beschikbaarheid: beschikbaar voor onderzoek tegen betaling via ICAME (International Computer Archive of Modern English) <http://nora.hd.uib.no/whatis.html>

Info: http://clwww.essex.ac.uk/w3c/corpus_ling/content/corpora/list/private/LOB/lob.html

2.7 BROWN

Inhoud: geschreven teksten Amerikaans Engels, bestaat uit fragmenten van 15 teksttypes, namelijk: persmateriaal, teksten over religie, teksten over vaardigheden, ambachten, hobby's, populaire kennis, biografieën, essays, fictie, wetenschappelijke teksten, andere (overheidsdocumenten, bedrijfs- en andere zakelijke rapporten, universiteitscatalogi), 1961

Omvang: 1 miljoen woorden

Doelstelling: referentiecorpus en ook voor onderzoek op het gebied van taalkunde

Annotatie: PoS

Beschikbaarheid: beschikbaar voor onderzoek tegen betaling via ICAME (International Computer Archive of Modern English) <http://nora.hd.uib.no/whatis.html>

Info:

http://clwww.essex.ac.uk/w3c/corpus_ling/content/corpora/list/private/brown/brown.html

<http://nora.hd.uib.no/whatis.html>

<http://khnt.hit.uib.no/icame/manuals/brown/INDEX.HTM>

3 Franstalige corpora

3.1 Corpus de la littérature narrative

Inhoud: het Corpus de la littérature narrative is het eerste deel van B.A.S.I.L.E. (Base internationale de Littérature Électronique). Het is een elektronische bibliotheek van literaire proza

Omvang: 30 miljoen woorden

Doelstelling: een elektronische bibliotheek voor onderwijsinstellingen die alle werken bevat die in Franse onderwijscurricula staan van middelbaar tot hoger onderwijs

Annotatie: onbekend

Beschikbaarheid: beschikbaar onder licentie via het ARTFL –project

<http://www.lib.uchicago.edu/efts/ARTFL/databases/champion/basile/access.html>

Info: <http://www.lib.uchicago.edu/efts/ARTFL/databases/champion/basile/>

3.2 ARTFL(The Project for American and French Research on the Treasury of the French Language)/ FRANTEXT (vroeger le Trésor de la Langue Française)

Inhoud: een grote variëteit aan teksten: klassieke Franse literatuur van 16 t/m 20 eeuwen, journalistieke, essayistische, technische, en wetenschappelijke teksten, alsook toneel, brieven, literaire recensies (1957-1981)

Omvang: 114 miljoen woorden

Doelstelling: oorspronkelijk doel: samenstelling van een databank van tekstvoorbeelden voor de redacteurs van TLF (lexicografie), later binnen ARTFL: samenstelling van een representatief corpus van Franse teksten voor onderzoek op allerlei domeinen

Annotatie: PoS

Beschikbaarheid: beschikbaar onder licentie voor onderwijs- en onderzoekinstellingen in Europa via ATILF (Analyse et Traitement Informatique de la Langue Française), contactadres voor individueel en groepsgebruik <http://www.atilf.fr/frantext.htm>

Info: <http://atilf.atilf.fr/frantext.htm>
<http://www.lib.uchicago.edu/efts/ARTFL/databases/TLF/index.html>
<http://humanities.uchicago.edu/orgs/ARTFL/artfl.flyer.html>

4 Parallele corpora

4.1 EUROPARL Parallel Corpus

Inhoud: de verslagen van het Europees Parlement met versies in 11 Europese talen: Romaanse (Frans, Italiaans, Spaans, Portugees), Germaanse (Engels, Nederlands, Duits, Deens, Zweeds), Grieks and Fins, 1996-2003

Omvang: 20 miljoen woorden

Doelstelling: voor automatische vertaling

Annotatie en alignering: geen annotatie, automatische alignering van documenten en op zinsniveau zonder handmatige verifiëring

Beschikbaarheid: gratis beschikbaar op <http://people.csail.mit.edu/koehn/publications/euoparl/>

Info: <http://people.csail.mit.edu/koehn/publications/euoparl/>
<http://webpace.isi.edu/mt-archive/MTS-2005-Koehn.pdf>

4.2 JRC-Acquis Multilingual Parallel Corpus

Inhoud: de huidige EU-wetgeving, bestaand uit geselecteerde teksten van de periode tussen de jaren 50 en 2005, namelijk tekstenverzameling die bekend is onder de naam Acquis Communautaire (AC) en bestaat uit ongeveer acht duizend wettelijke teksten op een aantal domeinen. Bestaat als een parallelle tekst in 20 officiële EU-talen sinds 2005

Omvang: 8,825,544 miljoen woorden

Doelstelling: oorspronkelijk doel: als voorwaarde voor de toetreding tot de EU moesten de kandidaat-staten dit corpus vertalen naar hun eigen taal en accepteren, doelstelling nu: voor (computationele) linguïstiek en voor andere doeleinden en toepassingen

Annotatie en alignering: annotatie onbekend, automatische alignering op zinsniveau zonder handmatige verifiëring

Beschikbaarheid: gratis beschikbaar op <http://langtech.jrc.it/JRC-Acquis.html>

Info: <http://langtech.jrc.it/JRC-Acquis.html>

4.3 UN Parallel Text

Inhoud: archieven in Engels/ Frans/ Spaans van the Office of Conference Services van de Verenigde Naties in New York voor de periode van 1988 tot 1993

Omvang: 165 miljoen woorden, verdeling per taal: Engels 59 mln., Frans 58 mln., Spaans 48 mln.

Doelstelling: voor onderzoek op het gebied van automatische vertaling

Annotatie en alignering: geen annotatie, wel metadata, automatische alignering van documenten en gedeeltelijk op zinsniveau zonder handmatige verificering

Beschikbaarheid: beschikbaar tegen betaling via LDC

<http://www ldc upenn edu/Catalog/CatalogEntry.jsp?catalogId=LDC94T4A>

Info: <http://www ldc upenn edu/Catalog/CatalogEntry.jsp?catalogId=LDC94T4A>

http://www ldc upenn edu/Catalog/readme_files/un.readme.html

4.4 OPUS

Inhoud: vrij beschikbare en vrij van copyright teksten van het Internet in 60 talen die uit afkomstig zijn van drie belangrijkste bronnen: OpenOffice.org documentatie (<http://www.openoffice.org>), KDE handleidingen (K Desktop Environment is een gratis grafische bureaubladomgeving voor UNIX systemen) inclusief KDE systeem berichten (<http://i18n.kde.org>), PHP handleidingen (Hypertext Preprocessor is gratis beschikbare computertaal) (<http://www.php.net/download-docs.php>), parallele teksten niet beschikbaar voor alle talen

Omvang: 30 miljoen woorden en wordt alsnog constant aangevuld

Doelstelling: aanmaken van een linguïstische databank voor allerlei doeleinden

Annotatie en alignering: gedeeltelijk PoS annotatie voor een deel in Engels, Frans, Duits, Italiaans en Zweeds, gedeeltelijk syntactische annotaties en lemmatisering voor het Engels, automatische alignering op zinsniveau zonder handmatige verificering

Beschikbaarheid: gratis beschikbaar op <http://logos.uio.no/opus/>

Info: http://stp.ling.uu.se/~joerg/paper/opus_lrec04.pdf

<http://logos.uio.no/opus/>

4.5 The Hansard Corpus

Inhoud: verslagen van het Canadese Parlement in Engels en Frans, van het midden jaren 70 tot 1988

Omvang: 26 miljoen woorden

Doelstelling: voor automatische vertaling

Annotatie en alignering: annotatie onbekend, automatische alignering op zinsniveau zonder handmatige verifiëring

Beschikbaarheid: training en testing sets beschikbaar ook voor vermenigvuldiging mits men de data nauwkeurig weergeeft en het Canadese Parlement niet in zijn eer aantast, zie <http://www.isi.edu/natural-language/download/hansard/>, het hele corpus is beschikbaar tegen betaling op <http://www ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC95T20>

Info: <http://www.isi.edu/natural-language/download/hansard/>
<http://www ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC95T20>
<http://spraakbanken.gu.se/pedant/parabank/node6.html>

4.6 BAF corpus

Inhoud: talen: Engels-Frans, hoofdzakelijk ambtelijke teksten: verslagen van het Canadese Parlement (zie Hansard), VN-documenten, documenten van het Hoger gerechtshof maar ook wetenschappelijke en literaire teksten, 1997

Omvang: 800 000 miljoen woorden (400 000 miljoen per taal)

Doelstelling: voor onderzoek en praktische toepassingen

Annotatie en alignering: annotatie onbekend, automatische alignering op zinsniveau zonder handmatige verifiëring

Beschikbaarheid: beschikbaar, versie 1.1 te downloaden op <http://rali.iro.umontreal.ca/Ressources/BAF/>

Info: <http://rali.iro.umontreal.ca/Ressources/BAF/>
<http://rali.iro.umontreal.ca/Ressources/BAF/Description.html>
<http://209.85.135.104/search?q=cache:gjCs0NiWF0gJ:rali.iro.umontreal.ca/Publications/urls/rec1998-fe.ps+%22BAF+corpus%22&hl=nl&gl=be&ct=clnk&cd=3>