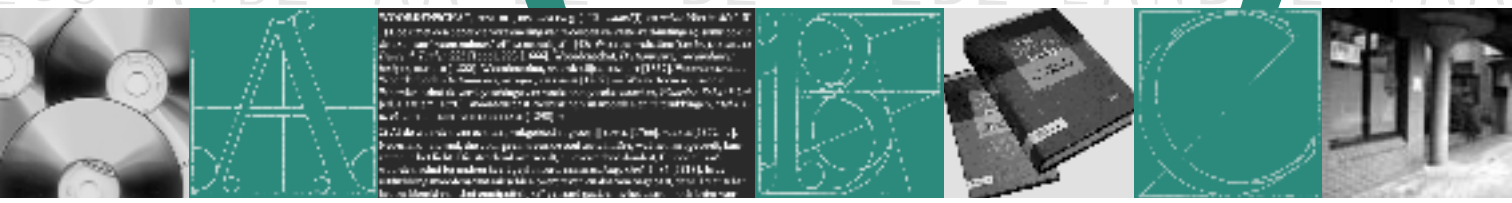
INL Nieuwsbrief

september 2007



Het ANW en de Sketch Engine

Eind april werd de Sketch Engine, een product van Lexical Computing Ltd., geïnitieerd door Adam Kilgariff, aangeschaft. De Sketch Engine is een webgebaseerd programma waarmee ten behoeve van de lexicografische bewerking diverse analyses op het ANW-corpus uitgevoerd kunnen worden.

De belangrijkste functies binnen de Sketch Engine zijn de 'Concordancer' en de 'Word Sketch function'.

Met de 'Concordancer' worden concordanties (regels tekst met links en rechts van het trefwoord de context) gegenereerd. Het programma bevat allerlei filter- en sorteermogelijkheden. Ook biedt het de mogelijkheid grote hoeveelheden concordanties tot een beperkte deelverzameling te reduceren. Daardoor kan de lexicograaf sneller greep krijgen op de betekenissen en gebruikskenmerken van zeer omvangrijke woorden.
De 'Word Sketch function' levert op één pagina een overzicht van de grammaticale relaties en collocaties van een woord. Bij zeer gebruikelijke woorden is het aantal concordanties doorgaans zo groot, dat men ze onmogelijk alle afzonderlijk kan bekijken. Met de word sketches hoeft dat dus niet meer, wat veel tijdsbesparing oplevert.

Om met de Sketch Engine te kunnen werken, is de getagde en gelemmatiseerde versie van het ANW-corpus in de software geladen en zijn er grammaticale patronen voor het Nederlands geschreven om de word sketches te genereren. Die patronen komen overeen met de patronen die in de artikelstructuur van het ANW zijn vastgelegd.

De Nederlandse Taalunie **[www.taalunieversum.org]** en het Instituut voor Nederlandse Lexicologie (INL) **[www.inl.nl]** hebben sinds 1986 nauwe banden. Beide organisaties hebben recent de intentie uitgesproken de samenwerking te intensiveren. Momenteel worden gesprekken gevoerd om vast te stellen of en hoe aan deze intentie uitvoering gegeven kan worden. De missie en doelstellingen van de Taalunie en het INL liggen voor een groot deel in elkaars verlengde. Beide organisaties verwachten dat nauwere samenwerking meerwaarde oplevert voor de taalgebruikers. Zo krijgt bijvoorbeeld de TST-centrale, het centrale loket voor digitale taalmaterialen voor het Nederlands, een solide basis bij het INL. In de komende periode informeren we u nader over de ontwikkelingen.

INL
INSTITUUT VOOR
NEDERLANDSE
LEXICOLOGIE

GiGaNT

Binnen het project Geïntegreerde Taalbank, werkt de Taalbank sinds kort aan de opzet van een computationeel lexicon, GiGaNT.

Het lexicon staat niet op zichzelf, maar wordt ontwikkeld in samenhang met een aantal hulpmiddelen voor de verwerking van natuurlijke taal (NLP-tools). Het gaat daarbij om programma's die (delen van) het lexicon inzetten bij verschillende vormen van tekstontsluiting, zoals het lemmatiseren, het taggen van woordsoorten (PoS-tagging), het herkennen van eigennamen (Named Entity Recognition) en het herkennen van schrifttekens (Optical Character Recognition). De ontwikkeling van deze tools en de data vindt plaats in het kader van de INL-deelname aan enkele nationale en internationale projecten. Op deze manier ontstaat een prachtige lexicale kruisbestuiving: de inzet van het lexicon bij het ontwikkelen van deze tools zorgt voor de opsporing van nieuwe woorden (types) en varianten, die weer tot een toename van de omvang en de diversiteit van het nagestreefde lexicon leiden. Daardoor kan het lexicon nog effectiever worden ingezet bij de toepassing van de tools, zodat weer nieuwe types en varianten opgedolven kunnen worden.

Het lexicon wordt corpusgebaseerd. Dat wil zeggen dat er bijgehouden wordt uit welk tekstmateriaal de types komen en in welke context ze verschijnen. Hiermee levert GiGaNT een bijdrage aan de opsporing van lacunes in de beschrijving van de woordenschat van het Nederlands.
Het spreekt voor zich dat het GiGaNT-lexicon online voor het publiek beschikbaar komt als deel van de Geïntegreerde Taalbank.

De informatiecategorieën die ter verrijking aan de types worden toegevoegd zijn zorgvuldig geselecteerd. Daarbij hebben zowel de immense omvang als de diversiteit van het Nederlands door de tijd heen een grote rol gespeeld. De verrijking moet toepasbaar zijn op alle types en realiseerbaar zijn binnen een afzienbare tijd. De selectie bestaat vooralsnog uit de informatiecategorieën lemma, woordsoort, morfologische structuur, broninformatie, frequentie, tijd en lokalisering.
Op termijn zal in ieder geval semantische informatie worden toegevoegd, zodat de onderzoeker ook andere dwarsdoorsnedes van de taal te zien krijgt.

De eerste 'G' staat voor Giga en verwijst naar de geplande omvang: het lexicon zal de taal vanaf ca. 500 tot heden bestrijken. De tweede 'G' staat voor Geïntegreerd en verwijst naar de populatie: de woorden (types). Deze woorden zijn afkomstig uit bestaande en nog aan te boren lexicale bronnen en worden allemaal met de bijbehorende informatiecategorieën omgezet naar een eenduidig standaardformaat. De staartletters 'NT' staan voor de Nederlandse Taal; naast standaardtaal zullen ook streektaal en vaktaal hun plaats vinden binnen GiGaNT

ONW, EWN en iWNT

Op 23 november vindt in Amsterdam de jaarlijkse mediëvistendag plaats, met daarop deze keer ruim aandacht voor het Oudnederlands. Enkele leden van de redactie en de begeleidingscommissie van het ONW leveren een bijdrage aan deze dag. Zij zullen spreken over de oudgermanistiek in Nederland, de Wachtendonckse Psalmen en Oudnederlandse glossen.

ONW

Sinds het begin van dit kalenderjaar heeft de redactie meer dan duizend woordenboekartikelen geschreven. Hoewel de meeste van deze artikelen op een klein aantal vindplaatsen gebaseerd zijn, zitten er ook diverse grotere artikelen bij. Daartoe behoren onder andere persoonlijke voornaamwoorden als *ik* (onl. *ik*), *jij* (onl. *thu*) en *hij* (onl. *hi*) die samen meer dan 1500 keer voorkomen, en werkwoorden als *doen* (onl. *duon*), *hebben* (onl. *hebben*), *worden* (onl. *werthan*), *zijn* (onl. *sïn*) en *zullen* (onl. *sullon*), ook goed voor een totaal van meer dan duizend attestaties.

EWN

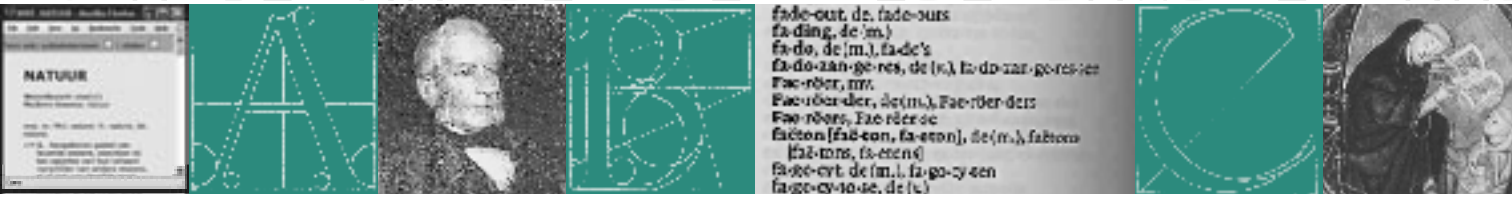
Het derde deel van het Etymologisch Woordenboek van het Nederlands zal vanaf half december in de boekwinkel te verkrijgen zijn. De lemmaschrijvers zijn inmiddels alweer bezig met het schrijven van lemma's voor het vierde en laatste deel van het woordenboek, waarin zo'n drieduizend woorden van de S tot en met de Z te vinden zijn.

iWNT

De tweede update van het WNT online is gepland voor de laatste week van oktober. Ook deze keer zijn er weer vele citaten van een datering voorzien en is de bronnenlijst uitgebreid met een flink aantal nieuwe titels. Het gebruikersgemak is opnieuw toegenomen doordat bijna alle verwijzingen binnen het WNT zijn doorgelinkt naar het betreffende woordenboekartikel.
Het vrijwilligersteam heeft onder andere koppelingen aangebracht naar het Etymologisch Woordenboek van de Nederlandse Dialecten van prof. Weijnen en naar een gedeelte van het derde deel van het Etymologisch Woordenboek van het Nederlands. Verder hebben de vrijwilligers ervoor gezorgd dat een groot aantal religieuze termen in het WNT van een passende afbeelding is voorzien en dat de dierkundige illustraties uit de werken van Burgersdijk en Brehm-Huizinga zijn opgenomen. Bovendien wordt aan de informatie over het WNT een beknopte biografie van een aantal WNT-redacteuren toegevoegd.

400 jaar Cornelis Kiliaan

Om gepaste aandacht te geven aan het vierhonderdste sterfjaar van Cornelis Kiliaan organiseert de provincie Antwerpen van 9 november tot en met 6 januari een tentoonstelling in de Koningin Fabiolazaal van Antwerpen (**http://www.corneliskiliaan.be**). Hierbij wordt niet alleen de nodige aandacht besteed aan het leven en werk van Kiliaan zelf, maar is ook ruimte gemaakt voor een overzicht van de geschiedenis van het woordenboek in de Nederlanden. Op deze tentoonstelling is behalve de online versies van het EWN en het WNT ook een demonstratieversie van het ANW te aanschouwen.



Statenvertaling digitaal

In juni verscheen er in de media een oproep om met een aantal mensen de eerste druk (1637) van de Statenvertaling te gaan digitaliseren. Aan de hand van afbeeldingen van deze tekst, te vinden op http://www.bijbelgenootschap.nl/digibi/, zouden vrijwilligers elk een aantal pagina's overtikken, die vervolgens na correctie samengevoegd zouden worden tot een groot bestand. De respons was overweldigend, binnen de kortste keren waren er meer dan honderd mensen aan het werk en was er zelfs een wachtlijst!

Nicoline v.d. Sijs coördineert het project en Jaap Engelsman modereert de mailinglist, zodat de vrijwilligers gemakkelijk contact met elkaar kunnen hebben en vragen en problemen snel opgelost kunnen worden. Verdere ondersteuning wordt geboden door het Nederlands Bijbel Genootschap, het INL, de Digitale Bibliotheek voor de Nederlandse Letteren en de Nederlandse Taalunie. Op de INL-website (**www.inl.nl/onw/digistatenbijbel**) staat alle beschikbare informatie over dit project. Ook is daar te zien hoe het digitaliseringsproces precies verloopt en kan men een voorbeeld van een gedigitaliseerd bijbelboek (het boek Esther) bezichtigen.

Naar verwachting zal de tekst van de Statenvertaling 1637 nog dit jaar helemaal overgetikt en gecorrigeerd zijn. Inmiddels is een parallelgroep vrijwilligers gestart met het invoeren van de Delftse bijbel (1477) en zal een tweede groep binnenkort beginnen met de Leuvense bijbel (1548). Hierna staan de Lutherse bijbel (1648), de Deux-Aesbijbel (1562) en de Liesveltbijbel (1526) op het programma.

Als deze (en eventuele andere) bijbelteksten allemaal in digitale vorm beschikbaar zijn, kunnen we voor de Nederlandse taal beschikken over een immens bijbels tekstcorpus, waarmee niet alleen prachtig taalkundig onderzoek gedaan kan worden, maar wat ook op cultuur-historisch gebied een schat aan informatie zal bieden.

INL
INSTITUUT VOOR NEDERLANDSE LEXICOLOGIE

Colofon
Deze nieuwsbrief is een uitgave van de Stichting
Instituut voor Nederlandse Lexicologie en verschijnt tweemaal per jaar.
Postbus 9515, 2300 RA Leiden
t 071-5141648
www.inl.nl
Redactie: secretariaat@inl.nl
Ontwerp: Swantje Haage Ontwerp, Amsterdam