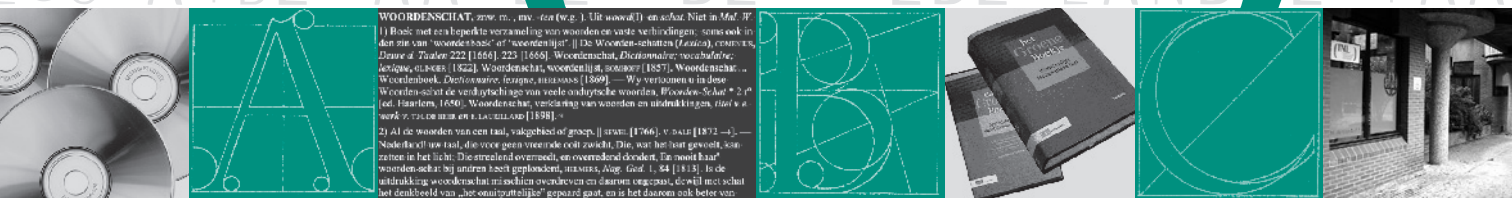
INL Nieuwsbrief

mei 2010



Zoeken in de schappen van de schatkist

Het INL als schatbewaarder van de Nederlandse taal past op een schatkist die voller en voller raakt. Taalmateriaal en speciaal daarvoor geschreven software blijven binnenstromen. Daarom wordt het steeds belangrijker de schatkist op orde te houden. De schatkist staat open voor alle geïnteresseerden in Nederland en Vlaanderen en natuurlijk moet iedereen snel kunnen vinden waarnaar hij of zij op zoek is.

De schatkist is toegankelijk via de website van het INL. Een website alleen is echter niet voldoende. Het INL heeft ook een servicedesk (**servicedesk@inl.nl**) in het leven geroepen. Via deze servicedesk kunnen deskundigen bereikt worden die technische en inhoudelijke vragen beantwoorden. Als iets niet duidelijk is, zal de servicedesk snel helpen.

Het materiaal in de schatkist is soms in huis ontwikkeld, maar veel materiaal is afkomstig van Nederlandse en Vlaamse universiteiten. Dat materiaal en de bijbehorende software is dikwijls verschillend van structuur en aanpak. Dit leidt ertoe, dat de schappen weliswaar netjes op orde zijn, maar dat er toch nog sprake kan zijn van verwarring en onduidelijkheid. Het is alsof je in een supermarkt zoekt naar “sinaasappelsap”, maar slechts het halve aanbod vindt, omdat de andere helft “jus d’orange” is genoemd. De pitloze druiven die gisteren op schap 12 stonden zijn opeens onvindbaar, omdat ze overeenkomstig nieuwe wetgeving verplaatst zijn naar schap 21. Aangekomen bij de kassa blijkt dat een deel van de boodschappen moet worden betaald met een pinpas, een deel met de chipknip en een deel contant.

Om aan dit soort problemen een einde te maken is het Europese project CLARIN gestart: Common Language Resources and Technology Infrastructure. Het INL draait hierin volop mee. Het project wil het voor onderzoekers in Europa mogelijk maken om na één enkele inlogactie en vanaf één scherm materiaal te raadplegen dat beschikbaar wordt gesteld in de deelnemende landen. Belangrijk is ook het structureren van de metadata (het oplossen van het “sinaasappelsap”-probleem) en het gebruik van zogenaamde Persistent Identifiers (waardoor het niet meer uitmaakt dat een product wordt verplaatst naar een andere schap). Het project zal nog tenminste drie jaar lopen.

INL
INSTITUUT VOOR NEDERLANDSE LEXICOLOGIE

Impact: 7 nieuwe talen

Even een opfrissertje: IMPACT gaat over OCR (tekenherkenning) en ontsluiting van historische teksten. Het INL maakt lexica die in een OCR-programma gebruikt worden en werkt mee aan het ontwikkelen van technieken om ervoor te zorgen dat het OCR-programma die lexica zo goed mogelijk kan gebruiken, en om die lexica zo efficiënt mogelijk te kunnen maken.
Verder proberen we ervoor te zorgen dat woorden in historische spelling *(waereld)* net zo makkelijk te vinden zijn voor gebruikers als moderne woorden *(wereld)*.

Een voorbeeld uit het “Kort begrip der waereld-historie voor de jeugd” van J.F Martinet, Predikant te Zutphen, uit 1789.

A. De eerſte was de gevaarlykſte om de verlei-
ding aan 't Hof; de tweede de ſtilſte en veiligſte;
de derde de zwaarſte, daar hy byna drie millioenen
harde en onbeſchaafde Menſchen beſtieren moest.

Resultaat van de tekenherkenning aan het begin van het project:

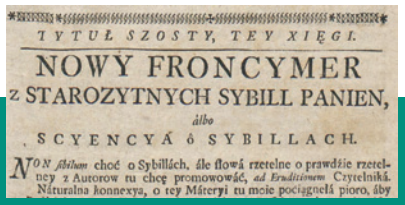
A. De eerde was de gevaarlykflti om de verlei¬
ding aan ‘t Hof; de tweede de ftillie en veiligde;
de derde de zwaarde, daar hy byna drie millioenen
harde en onbefchaafde Menfchen beftieren moest.

Resultaat begin 2010:

A. De eerste was de gevaarlykste om de verlei-
ding aan ‘t Hos; de tweede de stilste en veiligste;
de derde de zwaarste, daar hy byna drie millioenen
harde en onbeschaafde Menschen bestieren moest.

Voor het Nederlands zijn we dus al een heel eind opgeschoten. Ook voor bijvoorbeeld 16e-eeuws Duits zijn al aardige resultaten behaald. Een van de grootste uitdagingen voor de laatste twee jaren van het project is om dit voor niet minder dan 7 Europese talen (Engels, Frans, Spaans, Pools, Tsjechisch, Bulgaars, Sloveens) ook voor elkaar te krijgen.
De eerstgenoemde vier waren al voorzien in het werkplan, de laatste drie zijn het resultaat van een uitbreiding van het project die begin dit jaar door de Europese commissie is goedgekeurd (IMPACT-Enlarged European Union).
Het INL gaat bibliotheken en met name taalkundige partners uit de betreffende landen begeleiden bij het bouwen van historisch lexica voor hun talen.


ЦАРЕГРАДСКИ ВЪСТНИКЪ
1852.
№ 92


TYTUŁ SZOSTY, TEY XIĘGI.
NOWY FRONCYMER
z STAROZYTNYCH SYBILL PANIEN,
albo
SCYENCYÁ ó SYBILLACH.

Bulgaarse krant uit 1852

18de-eeuwse Poolse encyclopedie

FunQy: verbindende schakel

De computer houdt van duidelijkheid: hij wil pijnlijk gedetailleerde instructies voor zelfs de simpelste taken. De mens daarentegen wil dat de computer ophoudt met zeuren en gewoon doet wat er van 'm verwacht wordt.
Mens en computer zijn, kortom, niet bepaald voor elkaar geschapen. Het leek aanvankelijk een sprookjeshuwelijk, maar inmiddels vliegen de verwijten (en soms de borden) over de eettafel. Eigenlijk hadden ze al lang moeten scheiden, maar ja, de kinderen, hè...

Een van die kinderen is de zoekapplicatie van het *Algemeen Nederlands Woordenboek* (ANW). En zoals dat hoort bij een kind, houdt het evenveel van beide ouders: het probeert zowel te voldoen aan de hoge verwachtingen van de mens als begripvol te zijn over de moeizame communicatie met de computer.

Om die communicatie wat makkelijker te maken, is FunQy ontwikkeld. FunQy staat voor *functional query language* (functionele zoektaal). FunQy verzorgt binnen het ANW op internet de vertaalslag van invoer van de gebruiker naar zoekhandelingen voor de computer.
Het is een soort programmeertaal voor 'zoekintelligentie'.

Met behulp van FunQy kunnen de programmeurs en computerlinguïsten van het INL aangeven wat de computer naar aanleiding van de invoer van een gebruiker allemaal voor zoekhandelingen moet proberen: zoeken met EN, zoeken met OF, spelfouten corrigeren, zeer frequente woorden zoals lidwoorden en voorzetsels wegstrepen enzovoorts. Uiteindelijk worden de meest relevante resultaten aan de gebruiker getoond.

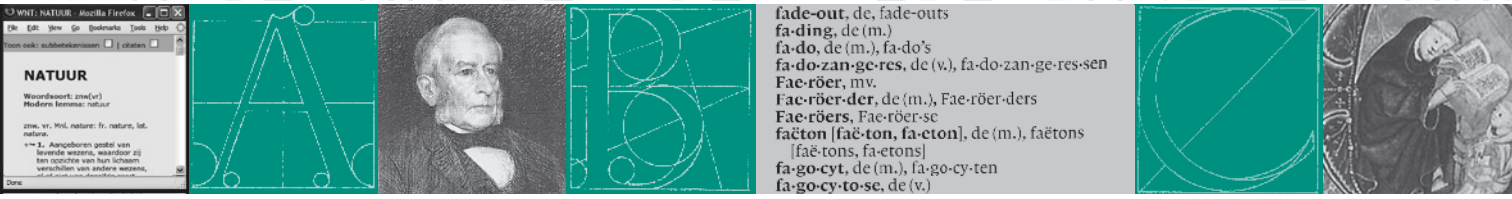
FunQy maakt het mogelijk om het zoeken in het ANW beter af te stemmen op de verwachtingen van gebruikers. Zo vormt het een verbindende schakel in de getroebleerde relatie tussen mens en computer. Het bewijs dat er ook iets goeds voort kan komen uit een slecht huwelijk!

In de ANW-applicatie is een artikel met technische informatie over FunQy opgenomen. Klik op de 'help'-tab en scrol helemaal naar onderen.

TST-Centrale bij 'Overheid en ICT'

Tijdens het evenement 'Overheid en ICT' heeft de TST-Centrale zich, samen met vier partners, in een zgn. Taal- en Spraakpaviljoen gepresenteerd en bezoekers geïnformeerd over de mogelijkheden van de hedendaagse taal- en spraaktechnologie.

In juni verschijnt de nieuwsbrief van de TST-Centrale met daarin uitgebreid verslag van dit evenement. De TSTC-nieuwsbrief is dan ook op onze website te vinden onder 'over het INL'.



Demo ANW in een ander jasje

Sinds 7 december 2009 staat een demoversie van het *Algemeen Nederlands Woordenboek* (ANW) online. Daarin 914 artikelen die de gebruikers een indruk moeten geven hoe het ANW eruit gaat zien en op welke wijze daarin gezocht kan worden. Na eenmalige registratie via anw.inl.nl kan iedereen gratis gebruikmaken van het gloednieuwe woordenboek.

Terwijl er hard wordt gewerkt aan een eerste release van het ANW, met daarin duizenden woorden, wordt er ook voortdurend gesleuteld aan de demoversie. Volgens een vast schema wordt de demo eens in de twee maanden geüpdatet: zo worden opmerkingen van recensenten en gebruikers verwerkt, gesignaleerde foutjes worden hersteld, de vormgeving wordt aangepast aan de wensen van gebruikers en de literatuurlijst wordt voortdurend geactualiseerd.

In de update van 6 april 2010 zijn veranderingen aangebracht die meer in het oog springen: voortaan worden standaard drie voorbeelden getoond, de rubrieken 'Combinaties' en 'Verbindingen' zijn omgedoopt in 'Combinatiemogelijkheden' en 'Vaste verbindingen' en naast de al aanwezige link naar Wikipedia is er nu ook een link naar Google. Als een ANW-artikel eenmaal geopend is, volstaat één klik om direct het materiaal van Wikipedia of Google bij het betreffende woord te bekijken.

Er zijn nog meer ingrijpende veranderingen op komst. Er is besloten om in toekomstige updates ook telkens een nieuwe reeks woorden toe te voegen. Dat is goed nieuws voor de ANW-gebruikers, want zo valt er voor hen telkens iets nieuws te beleven!

INL
INSTITUUT VOOR NEDERLANDSE LEXICOLOGIE

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