INL Nieuwsbrief
maart 2008

Samenwerkingsovereenkomst verzekert toekomst TST-centrale bij het INL

De Nederlandse Taalunie (NTU) en het Instituut voor Nederlandse Lexicologie (INL) hebben een samenwerking die al bestond per 1 januari 2008 geformaliseerd. Het INL blijft een zelfstandige stichting maar voert formeel het merendeel van haar activiteiten uit in opdracht en onder de koepel van de NTU. Ook de financiering van het INL loopt sinds 1 januari geheel via de NTU.

Onder impuls van de NTU en met subsidie van de Nederlandse en Vlaamse overheid wordt aan verschillende instellingen - waaronder het INL - gewerkt aan de ontwikkeling van digitale taalmaterialen. Om de resultaten daarvan te beheren en te distribueren, werd in 2004 een Centrale voor Taal- en Spraaktechnologie (TST-centrale) ondergebracht bij het INL. Het materiaal van de centrale is beschikbaar voor woordenboekmakers, wetenschappelijke instellingen en ontwikkelaars van software. Bedrijven kunnen het materiaal heel goed gebruiken wanneer zij producten willen ontwikkelen voor bijvoorbeeld het onderwijs of de zorg.

Met deze samenwerkingsovereenkomst is de toekomst van de TST-centrale verzekerd. Eén van de gevolgen is dat het INL en de TST-centrale samenwerken aan een geïntegreerde en verbeterde INL-website, die begin 2009 gelanceerd zal worden.

Goed nieuws is ook dat vanaf nu vrijwel alle taalmaterialen bij de TST-centrale gratis beschikbaar zijn voor gebruik zonder winstoogmerk. Dit is mogelijk omdat de NTU kortgeleden het nulprijsbeleid heeft ingevoerd.

Daarnaast worden in 2008 de eerste resultaten van het STEVIN-programma (Spraak- en Taaltechnologische Voorzieningen in het Nederlands) van de NTU beschikbaar gesteld. De projectresultaten zullen worden opgenomen in de productencatalogus, die op de website te vinden is onder **www.tst.inl.nl/productenlijst**. In de nieuwsbrief van de TST-centrale wordt hierop uitgebreider ingegaan.

Nederlandse Taalunie

De Nederlandse Taalunie (www.taalunieversum.org) is een beleidsorganisatie waarin Nederland, Vlaanderen en Suriname samenwerken op het gebied van de Nederlandse taal en letteren en het onderwijs in en van het Nederlands. De Taalunie ziet het als haar opdracht om ervoor te zorgen dat alle Nederlandssprekenden hun taal op een doeltreffende manier kunnen gebruiken.

INL
INSTITUUT VOOR NEDERLANDSE LEXICOLOGIE

INL werkt mee aan groot Europees project!

Hoe kan de toegang tot historische teksten zo verbeterd worden dat ze net zo toegankelijk worden als hun van oorsprong digitale tegenhangers?

Op 1 januari 2008 is het project IMProving ACcess to Text (IMPACT) gestart. Doel van het project is om massadigitalisering en toegankelijkheid van het Europese gedrukte cultureel erfgoed significant te verbeteren.

Massadigitalisering is de laatste jaren een belangrijk aandachtspunt voor bibliotheken in de hele wereld. Miljoenen pagina's worden gescand. Echter, als bibliotheken naast de afbeelding van een tekst ook de tekst zelf willen kunnen aanbieden, dan lukt dat voor historisch tekstmateriaal met de bestaande OCR-technieken (OCR: optische tekenherkenning) niet en overtikken van deze teksten is te tijdrovend en te kostbaar. Een consortium van vijftien instellingen uit Europa, Israël en Rusland (nationale en universiteits-bibliotheken, onderzoeksinstellingen en bedrijven) heeft zich verenigd om daar iets aan te doen.

Doel van het project is om de toegang tot historisch tekstmateriaal substantieel te verbeteren, niet alleen door de bestaande OCR-technologie te vernieuwen, maar ook door het ontwikkelen en inzetten van taal-technologieën om de historische taalbarrière te overbruggen. Er zal een Best Practiceleidraad gedefinieerd worden m.b.t. de operationele context voor digitalisering. De verschillende binnen IMPACT ontwikkelde technieken zullen 'interoperabel' zijn en er zal ook worden voorzien in een samenhangend programma van verspreiding, trainingen en demonstraties dat gericht is op capaciteitstoename zowel binnen als buiten de deelnemende instituten.

IMPACT streeft ernaar om 1. OCR-software en -technologieën te ontwikkelen die de nauwkeurigheid van de huidige state-of-the-art software substantieel verbeteren, en die het voor het eerst mogelijk zullen maken grote hoeveelheden gedigitaliseerde historische teksten in elektronische tekst om te zetten. 2. Een software-systeem te leveren dat de implementatie mogelijk maakt van nieuwe ideeën op het gebied van webgeba-seerde collaboratieve correctie. 3. Taaltools en lexica te ontwikkelen om onafhankelijk van de historische varianten van een taal toegang te bieden tot historische teksten. 4. Toepassers van deze tools te steunen zodat meer Europese historische lexica gebouwd kunnen worden. 5. Een aantal kleinere modules te ontwik-kelen, zoals toolkits voor beeldverbetering en beeldsegmentatie, functionele parsers etc., met als doel de automatische tekstherkennning en/of toegang tot historisch tekstmateriaal te ondersteunen.

De Taalbank leidt de werkpakketten rondom lexiconbouw en –toepassing ten behoeve van tekstont-sluiting, zal een lexicon bouwen voor het Nederlands en zal ook verantwoordelijk zijn voor de training dienaangaande. Daarnaast wordt ook meegewerkt aan het ontwikkelen van technieken om de OCR te verbeteren met behulp van taalmodellen en historische lexica.

IMPACT wordt gecoördineerd door de Koninklijke Bibliotheek en gefinancierd binnen het Zevende Kaderprogramma van de Europese Commissie (FP7). De doorlooptijd is vier jaar. Meer informatie op de Impact project website: **http://www.impact-project.eu**

Gouverneur van Antwerpen ontvangt Etymologisch Woordenboek van het Nederlands

Op 14 december jl. heeft de heer Camille Paulus, gouverneur van de provincie Antwerpen en voorzitter van het INL-bestuur het eerste exemplaar van het derde deel van het *Etymologisch Woordenboek van het Nederlands* (Ke-R) in ontvangst genomen.

Bij deze overhandiging waren dr. Willy Pijnenburg, voorzitter van de begeleidingscommissie van het EWN, dr. Frans Debrabandere en dr. Tanneke Schoonheim, leden van de hoofdredactie van het EWN aanwezig.

Het *Etymologisch Woordenboek van het Nederlands* wordt uitgegeven door de Amsterdam University Press. Het bestaat inmiddels uit drie delen, het laatste en vierde deel (S-Z) zal eind 2009 verschijnen. Het is verkrijgbaar bij de boekhandel, maar kan ook rechtstreeks bij de uitgever besteld worden: **www.aup.nl**. Het is ook mogelijk om een licentie te verkrijgen voor de digitale versie van het EWN, die te vinden is op **www.etymologie.nl**.

Lezingen op het INL

In het kader van de lezingencyclus Capita Selecta Lexicologie, met als thema 'Onomasiologie en/in lexicologie/lexicografie' zal prof. Henk Verkuyl van de Universiteit Utrecht op 11 maart een lezing verzorgen over ''de' en 'het': telbaarheid in een woordenboek'. Op 17 april houdt mw. dr. Annette Klosa van het Institüt für Deutsche Sprache in Mannheim een lezing over 'Das elexiko-Worterbuch'. Op 15 mei spreekt prof. Fons Moerdijk, hoofdredacteur van het Algemeen Nederlands Woordenboek over 'Onomasiologie in het elektronische woordenboek'. De lezingen vinden plaats op het INL, in het Verbarium, Matthias de Vrieshof 3, Leiden.
Alle belangstellenden zijn welkom!

SCHATBEWAARDER DER NEDERLANDSE TAAL

Het ANW-corpus op internet

Vanaf eind maart zal het ANW-corpus, voorzien van een uitgebreide gebruikershandleiding, op internet (http://anw-corpus.inl.nl) beschikbaar zijn. Dit corpus is de verzameling elektronische teksten die het fundament vormt voor een nieuw online woordenboek van het eigentijdse Nederlands: het Algemeen Nederlands Woordenboek (ANW).

Het corpus is 'taalkundig verrijkt': de teksten zijn van allerlei coderingen voorzien, die het mogelijk maken om op velerlei taalvragen antwoord te krijgen. Met zijn ruim 100 miljoen woorden is het ANW-corpus het grootste taalkundig verrijkte corpus van het geschreven Nederlands in Nederland en in Vlaanderen.

Het ANW-corpus werd aangelegd in de periode 2000-2005 en bevat taalmateriaal vanaf 1970. Het is zeer divers van samenstelling en inhoud. Het bevat materiaal uit kranten, literaire teksten, essays, internetteksten uit vele samenlevingsgebieden (van 'aangeboren afwijkingen' tot en met 'zwemsport') en tal van neologismen.

De eerste resultaten van de bewerking van het woordenboek zelf zijn op internet te verwachten in het voorjaar van 2009.

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