

## Goesting in een puzzel?

*Localisatie van de Keesing-databanken voor de Vlaamse markt  
-Een onderzoekproject van Lessius Hogeschool en K.U.Leuven-*

*Persconferentie 16 juni 2011*

### **Een samenwerking tussen Keesing, Lessius-hogeschool en de K.U.Leuven**

Taalpuzzels zijn en blijven erg populair. Mensen vinden ze onder andere leuk omdat ze helpen om hun algemene kennis aan te scherpen: Het is gewoon erg bevredigend als je kan vaststellen dat je over een goede taal- en feitenkennis beschikt. Taalpuzzels zijn dan ook op een soort van impliciete kennis-canon gebaseerd: ze vragen naar dingen die iedereen wel kan of hoort te weten. Die canon is echter wel sterk cultuurgebonden en puzzelaars vinden het behoorlijk vervelend als er in een puzzel te veel vragen en woorden zitten die niet bij de canon van 'hun' gemeenschap aansluiten. Dan slaat de bevrediging al snel om in frustratie. In Vlaanderen stelt dit probleem zich met name bij puzzels die oorspronkelijk voor een Noord-Nederlands publiek ontwikkeld werden. Vlamingen hebben geen boodschap aan "een rivier in Gelderland" en vinden niet dat ze geacht worden woorden als chipknip of wethouder te kennen. De aantrekkelijkheid van taalpuzzels voor een Belgisch publiek zou beduidend verhogen als dit soort van 'hollandismen' vervangen worden door een Belgisch alternatief of expliciet als typisch Nederlands aangeduid worden (bv. in een vraag als "Benaming voor een schepen in Nederland": wethouder).

In een meesterproef aan Lessius Antwerpen (Liesbeth van Reeth, 2008) werd een deel van de puzzeldatabanken van Keesing onderzocht op hun vertrouwdheid voor een Belgisch publiek. De meesterproef identificeerde daarbij heel wat typisch Noord-Nederlandse woorden (bv. ansichtkaart of hartstikke) en culturele referenties (bv. namen van Nederlandse politici of plaatsnamen) die Belgische puzzelaars als storend ervaren. Voor een aantal daarvan werden ook Belgische alternatieven geformuleerd.

De meesterproef was een mooie eerste casestudy, maar kon natuurlijk slechts een klein stukje van Keesings omvangrijke puzzelbestand onder de loep nemen. Het project dat op deze persconferentie wordt voorgesteld (codenaam LoBKe: **Localisatie voor de Belgische markt van de Keesing-Databanken**) heeft de aanpassing ('localisatie') van de Keesing-databanken voor een Belgisch publiek op een veel grotere schaal uitgevoerd. Dat werd mogelijk gemaakt door het opsporen van typisch Noord-Nederlandse woorden en eigennamen deels te automatiseren en dus de taak van de databankreviewer te vergemakkelijken en te versnellen. Hiervoor heeft de afdeling Toegepaste Taalkunde van Lessius Antwerpen samengewerkt met haar vaste partner aan de Leuvense universiteit, de onderzoeksgroep Quantitative Lexicology and Variational Linguistics (QLVL). Terwijl Lessius een grote expertise heeft in het 'lokaliseren' van taalproducten in het algemeen, is QLVL dan weer gespecialiseerd in het onderzoek naar verschillen in woordgebruik tussen Nederland en België. De onderzoeksgroep heeft in de afgelopen jaren ook software ontwikkeld om die verschillen systematisch en op grote schaal te analyseren.

Hieronder beschrijven we kort hoe **QLVL en LESSIUS** de handen in elkaar geslagen hebben om de puzzeldatabank van Keesing door te lichten en uit te breiden. Daarna volgt nog een profiel van de twee onderzoekspartners.

### Goesting in puzzels zonder gesappel?

Om gericht en efficiënt de Keesing-databank te kunnen aanpassen en uitbreiden voor een Belgisch publiek, moesten we eerst weten hoeveel woorden een te Noord-Nederlands karakter hebben en waar ze zich in de databank bevinden. Omdat de Keesing-databanken behoorlijk omvangrijk zijn (zo'n 400.000 unieke woorden) werd deze **analyse geautomatiseerd** aan de hand van statistische methodes die door QLVL ontwikkeld zijn. Kris Heylen en Dirk De Hertog van QLVL hebben alle puzzelwoorden uit de databank onderworpen aan een zogenaamde **Stable Lexical Marker Analyse** (SLM-analyse). In een heel grote verzameling van Belgische en Nederlandse teksten (meer dan 1,5 miljard woorden uit kranten tijdschriften) konden ze vaststellen hoe gebruikelijk elke woord is in respectievelijk België en Nederland, en of verschillen in het gebruik statistisch significant waren. Dat resulteerde dan in een toewijzing van een "regio-index" aan elk woord en aan de hand van deze index werd het woord één van de volgende labels toegewezen worden:

1. Gebruikelijk in Nederland, maar onbekend in België
2. Eerder Noord-Nederlands, maar gekend in België
3. Even gebruikelijk in Nederland en België
4. Eerder Belgisch, maar gekend in Nederland
5. Gebruikelijk in België, maar onbekend in Nederland

Geautomatiseerde analyses laten toe om grote hoeveelheden data in een keer te verwerken, maar computers blijven computers en die spreken natuurlijk geen Nederlands. Om zeker te zijn dat de resultaten betrouwbaar waren werd een steekproef van het verzamelde materiaal **manueel gevalideerd en geïnterpreteerd** door Ken De Wachter, projectmedewerker bij Lessius. Daarbij gebruikte hij niet alleen zijn eigen vakkennis over het gebruik van de woorden in België en Nederland, maar nam ook een enquête af bij Belgische studenten en medewerkers bij Keesing in Nederland om te peilen naar hun vertrouwdheid met de woorden.

Uit de automatische en manuele analyses samen konden we **de lokalisatiebehoefte** in de Keesing-databank inschatten, m.a.w. hoeveel woorden wel bruikbaar zijn in Belgische puzzels en hoeveel woorden echt wel te Hollands waren zoals sappelen (ploeteren, afjakkeren) of pinautomaat. Alles bij elkaar bleken dat er zo'n 6% te zijn. Die informatie werd in de puzzeldatabank geïntegreerd zodat te Hollandse woorden voortaan geweerd kunnen worden in puzzels die voor de Belgische markt bedoeld zijn.

Uit de analyses bleek echter ook dat heel wat typisch Belgisch-Nederlandse woorden niet in de databank aanwezig waren zoals jobstudent of onthaalmoeder. Zo'n **5000 woorden** die uit de statistische analyse als typisch Belgisch naar voren kwamen maar die ontbraken in de databank werden door de projectmedewerkers dan ook van een puzzelomschrijving voorzien en met het juiste label **aan de databank toegevoegd**. Voortaan kunnen deze in puzzels voor het Belgisch publiek opgenomen worden. Nog een paar voorbeelden:

fuijzaal	heemkring	containerpark	verloning
speelpleinwerking	handelszaak	speeltuig	klissen
wegenwerken	foorkramer	fietszoektocht	zoekertje
gemeenteraadszitting	kinesist	bankkaart	parochiecentrum
infoavond	schoolomgeving	waterkansje	bruggepensioneerde
intercommunale	sluikstorten	loonkost	sluikverkeer
cultuurschepen	maaltijdcheque	ploegkoers	mossselfestijn

**Profiel: LESSIUS Hogeschool, Antwerpen, Dep. Toegepaste Taalkunde**

**Lessius**

LESSIUS ANTWERPEN vzw - TOEGEPASTE TAALKUNDE  
Campus Sint-Andries - Sint-Andriesstraat 2 - BE-2000 Antwerpen  
tel. +32 (0)3 206 04 91 - fax +32 (0) 206 04 99  
sintandries@lessius.eu - www.lessius.eu



**PROFIEL**

Het departement Toegepaste Taalkunde, geassocieerd met de faculteit Letteren van de KULeuven, heeft binnen het onderzoeksdomein Translation Studies een grote expertise opgebouwd op het gebied van de lokalisatie. Het lokaliseren is één van de belangrijkste onderdelen van de huidige vertaalwereld: teksten, producten, brochures, websites, etc. moeten worden aangepast aan lokale markten, rekening houdend met terminologie, culturele verschillen, wetgeving, benadering van een doelpubliek, marktgevoeligheden etc. Zo werken we aan onderzoeksprojecten voor de automobieliindustrie (Europees project voor een Japans automerk, lokaliseren van handleidingen in 14 talen en een tweede project voor het lokaliseren voor een Duits automerk naar Nederlands en Frans voor de Belgische markt), onderzoek voor de industrie van computerspellen etc.

Hier volgt een beknopte selectie uit onze speerpunten in onderzoek rond vertalen, tolken en vertaaltechnologie

- Terminologieleer en technologische ontwikkelingen voor efficiënt meertalig terminologiebeheer, onder andere in samenwerking met TermNet ([www.Termnet.org](http://www.Termnet.org))
- Terminologiebeheer en corpusonderzoek van juridische teksten voor een efficiënt beheer van juridische vertalingen in de federale overheidsdiensten; (IOF kennisplatform "TermWise")
- Termextractie in samenwerking met *Tèmis* en de *Taalunie*; corpuslinguïstiek voor vertaaltechnologische toepassingen [alignering, vertaalgeheugens];
- Corpuslinguïstisch en cognitief taalkundig onderzoek;
- Onderzoek in vertaalevaluatie; testen van Translation Quality Assurance [TQA] software en Ontwikkeling van TQA tool
- Ontwikkeling van CALL tools voor taalverwerving, vertalen, tolken en meertalig communiceren; in samenwerking met ISO/TC 37 en NEN-ISO TC 37 opstellen en updaten van internationale normen inzake terminologieleer, concept modelling, socioterminologie, vertaaldiensten
- TSB : Translation Studies Bibliography <http://www.benjamins.com/online/tsb/>
- HoTS : Handbook of Translation Studies (<http://www.benjamins.com/online/hts/>)
- Lessius is stichtend lid van EULITA (European Legal Interpreters and Translation Association)
- Coördinator en deelnemer aan een aantal Europese projecten rond gerechtsvertalen en –tolken (o.a.) Avidicus (Assessment of Videoconference Interpreting in the Criminal Justice Services) (E. Hertog, K. Balogh, Y. Vanden Bosch)
- Tolkonderzoek rond juridisch tolken, tolken voor de gezondheidszorg en conferentietolken

Lessius is lid van het prestigieuze EMT-netwerk voor de top-vertaalopleidingen (European Master in Translation), en lid van CIUTI (internationale conferentie van universitaire vertaal- en tolkopleidingen).

We behoren tevens tot het Optimale Erasmus Academic netwerk (Optimising Professional Translator Training in a Multilingual Europe).

Ons departement is actief lid van TermNet, het internationale netwerk voor terminologie ([www.termnet.org](http://www.termnet.org)) en biedt een groot aantal gecertificeerde opleidingen en bijscholingen aan voor professionele vertalers (SDL Universities, TILP : the institute for localisation

### **MEDEWERKERS BETROKKEN BIJ HET PROJECT**

Supervisie: Prof. Dr. Frieda Steurs

Uitvoering: Ken De Wachter

### **CONTACT**

Prof.dr. Frieda Steurs  
Departementshoofd Toegepaste Taalkunde  
Lessius/KULeuven  
Sint Andriesstraat 2  
B-2000 Antwerpen  
tel +32(0)3 2060488  
fax +32(0)3 2060499  
mobiel +32 478 28 33 70  
[Frieda.Steurs@lessius.eu](mailto:Frieda.Steurs@lessius.eu)  
[www.lessius.eu](http://www.lessius.eu)



**Profiel QLVL, K.U.Leuven**

Quantitative Lexicology and Variational Linguistics  
K.U.Leuven

**PROFIEL**

De onderzoeksgroep *Quantitative Lexicology and Variational Linguistics (QLVL)* onder leiding van Prof. dr. Dirk Geeraerts en Prof. dr. Dirk Speelman maakt deel uit van de Faculteit Letteren van de Katholieke Universiteit Leuven en heeft een lange traditie op het gebied van onderzoek naar variatie in woordbetekenis en woordgebruik, met bijzondere aandacht voor de verschillen in woordenschat tussen Nederland en België. Daarbij werd taalkundige theorievorming over lexicale variatie (Geeraerts et al. 1994<sup>1</sup> is ondertussen een standaardwerk) steeds gekoppeld aan gedegen empirisch onderzoek op basis van grote tekstcorpora (specifiek Geeraerts et al. 1999 over variatie in het Nederlands en de veranderende verhouding tussen het Nederlands in Vlaanderen en dat in Nederland).

QLVL beschikt over een databank met teksten uit kranten, tijdschriften en internetfora (en zelfs met gesproken materiaal), die meer dan 1,5 miljard woorden beslaat. De laatste 10 jaar werden geavanceerde statistische technieken ontwikkeld om systematisch en op grote schaal woordgebruik en woordbetekenis te bestuderen in deze tekstcorpora. De onderzoeksgroep werd zo trouwens een van de grootste gebruikers van de supercomputerinfrastructuur van de K.U.Leuven<sup>2</sup>. Kortom, QLVL staat in de internationale academische wereld aan de top op het terrein van het corpuslinguïstische betekenis- en woordenschatonderzoek.

Met projecten zoals de hier voorgestelde samenwerking met Keesing en Lessius willen we die expertise ook maatschappelijk valoriseren. Zo is de techniek van *Stable Lexical Marker Analysis* (Speelman et al. 2006), die de typische woorden voor een bepaalde taalvariëteit identificeert, oorspronkelijk ontwikkeld om taalverschillen tussen België en Nederland aan sich te bestuderen. Dit project toont echter dat de techniek ook heel praktisch toegepast kan worden om kruiswoordpuzzels aan te passen aan een specifieke markt. Andere statistische modellen die de onderzoeksgroep ontwikkeld heeft, kunnen dan weer verschillen in de betekenis van woorden opsporen (bv. verschil in betekenis van *patat* of *zetel* tussen België en Nederland). Ze laten ook toe om automatisch equivalente uitdrukkingen in de twee variëteiten terugvinden (bv. *wethouder* en *schepen* of *proton* en *chipknip*). Een greep uit het andere onderzoek van QLVL:

- *Attitudes tegenover taalvariëteiten: welke accenten van het Nederlands vinden sprekers uit andere regio's makkelijk verstaanbaar en welke vinden ze sympathiek klinken? (In samenwerking met Universiteiten van Nijmegen en Groningen)*
- *Anglicismen in het Nederlands: wat bepaalt of een Engelse term ingeburgerd raakt in het Nederlands?*
- *Labeling van Nederlands- Nederlandse woorden in een woordenboek (in samenwerking met Prisma)*

<sup>1</sup> Zie de website voor meer informatie over de publicaties: <http://www.ling.arts.kuleuven.be/qlvl/>

<sup>2</sup> zie *De Standaard*, 26 maart 2009, <http://www.standaard.be/artikel/detail.aspx?artikelid=K0285LEV>

- *Juridische terminologie: welke juridische termen worden inconsistent gebruikt in de Belgische wetgeving en welke juridische concepten worden anders benoemd op Vlaams, Belgisch en Europees niveau? Worden ze altijd correct uit het Frans vertaald? (In samenwerking met Lessius Antwerpen, Lessius Campus De Nayer en KULEuven dep. Computerwetenschappen)*
- *Categorisatieonderzoek: hoe benoemen sprekers concepten als ze de keuze hebben uit meerdere woorden? Zeggen ze zomereik, eik of gewoon boom? (Samenwerking met het KULEuven Dep. Psychologie)*
- *Kleurennamen: waar halen bedrijven de mosterd vandaan als ze nieuwe kleuren moeten benoemen van bv. lippenstift of auto's?*
- *Sociolinguïstisch onderzoek: welke groepen in de maatschappij gebruiken welke woorden en in welke situatie? Spreken vrouwen anders dan mannen? Hoogopgeleiden anders dan laagopgeleiden? Ouderen anders dan jongeren? En spreken die dan hetzelfde onder alle omstandigheden?*
- *Causatieve werkwoorden: wanneer **laat** je iemand lachen en wanneer **doe** je iemand lachen? Ofte, kan je met een statistisch model voorspellen wanneer sprekers doen en laten gebruiken?*

#### **MEDEWERKERS BETROKKEN BIJ HET PROJECT**

Supervisie: Prof. dr. Dirk Geeraerts, Prof. dr. Dirk Speelman

Uitvoering: Dr. Kris Heylen, Dirk De Hertog

#### **CONTACT**

Dr. Kris Heylen

Quantitative Lexicology and Variational Linguistics (QLVL)

Subfaculteit Taalkunde, K.U.Leuven

Blijde-Inkomststraat 21/3308

3000 Leuven (Belgium)

tel: +32 16 324819

fax: +32 16 324767

mobiel: +32 475844166

kris.heylen@arts.kuleuven.be

<http://www.ling.arts.kuleuven.be/qlvl/>