Overleeft het Nederlands het digitale tijdperk?

Jan Odijk
Universiteit Utrecht

Minstens 21 Europese talen lopen grote risico's om het digitale tijdperk niet te overleven. Daarvoor waarschuwen vooraanstaande taaltechnologische experts uit heel Europa in een nieuwe studie. Op 26 september, de Europese dag van de talen, is een reeks van 30 witboeken oftewel taalrapporten gepresenteerd, die per taal de risico's inzichtelijk maken. De studie is uitgevoerd door META-NET, een Europees excellentienetwerk met 60 onderzoeksinstellingen in 34 landen. Voor het Nederlands is de studie uitgevoerd door de Universiteit Utrecht (Odijk, 2012).

Ruim 200 experts hebben voor 30 van de ongeveer 80 Europese talen vastgesteld in hoeverre zij digitaal worden ondersteund met taaltechnologie. De conclusie luidt dat de digitale ondersteuning voor 21 van de 30 talen *niet-bestaand* is of op zijn best *zwak*. Bekende voorbeelden van taaltechnologische toepassingen zijn programma's voor spelling- en grammaticacontrole, interactieve persoonlijke assistenten op smartphones (zoals Siri op de iPhone), gesproken telefoonmenu's, automatische vertaalsystemen, zoekmachines op het web en de stemmen in autonavigatiesystemen.

Slechte taaltechnologische voorzieningen

Voor iedere taal is de taaltechnologische ondersteuning op vier verschillende gebieden vastgesteld: automatisch vertalen, spraakinteractie, tekstanalyse en de beschikbaarheid van taalbronnen. Verschillende talen, bijvoorbeeld IJslands, Lets, Litouws en Maltees krijgen de laagste score op alle gebieden. In totaal scoren 21 talen slecht op minimaal één gebied. Opmerkelijk is dat geen enkele taal de categorie *excellente ondersteuning* krijgt. Alleen het Engels wordt beschouwd als een taal met *goede ondersteuning*, gevolgd door talen als het Nederlands, Frans, Duits, Italiaans en Spaans met *beperkte ondersteuning*.

De situatie van het Nederlands

De situatie van het Nederlands is helemaal niet zo slecht. Het Nederlands doet mee in de categorie van talen als het Frans en het Spaans. Dat lijkt op het eerste gezicht verrassend, maar er is toch een eenvoudige verklaring voor. De *beperkte ondersteuning* voor het Nederlands is vooral te danken aan de bewuste politiek van de afgelopen tien jaar om het Nederlands te versterken in de digitale informatiemaatschappij. Die politiek heeft geresulteerd in een aantal programma's die Nederland en Vlaanderen gezamenlijk uitgevoerd hebben.

In de eerste plaats moeten we hierbij het project noemen dat het *Corpus Gesproken Nederlands* (CGN) gecreëerd heeft. Het CGN is een groot corpus van gesproken Nederlands met zeer rijke annotaties op orthografisch, fonetisch, morfologisch, lexicaal en syntactisch vlak. Het corpus was opgezet om bruikbaar te zijn voor zowel taalkundig onderzoek als taal- en spraaktechnologische toepassingen. Over het *Corpus gesproken Nederlands* is recent nog bericht in Neerlandia/Nederlands van Nu (Van Eerten & Depoorter, 2011).

Nog belangrijker dan het CGN was het STEVIN-programma. Het STEVIN-programma was een samenwerkingsproject van Nederland en Vlaanderen met als belangrijkste doelstelling ervoor te zorgen dat het Nederlands behouden kan blijven en versterkt kan worden in de digitale informatiemaatschappij. In het programma, waarin kennisinstellingen en het bedrijfsleven nauw samenwerkten, is een grote reeks van taalbronnen (data en basisprogrammatuur) voor het Nederlands gemaakt. Dergelijke taalbronnen zijn essentiële 'grondstoffen' voor het ontwikkelen van taal- en spraaktechnologie. Die taalbronnen zijn niet alleen *aangemaakt*, maar er is ook voor gezorgd dat ze *beschikbaar* zijn en *gebruikt* kunnen worden door onderzoekers en in het bedrijfsleven. Hiervoor is de zogenaamde TST-Centrale in het leven geroepen, die allerlei taalbronnen voor het Nederlands beheert en distribueert. Verder zijn ook kwesties rond intellectueel eigendomsrecht (auteursrecht, kopierecht etc.) vanaf het begin goed geregeld, zodat die geen obstakel vormen voor de beschikbaarstelling.

Het STEVIN-programma heeft drie belangrijke taalbronnen opgeleverd. In de eerste plaats is er het SoNaR-corpus geschreven Nederlands van meer dan 500 miljoen woorden, dat verrijkt is met allerlei annotaties. Het wordt nog aangevuld door het Lassy-corpus, dat bij iedere zin een volledige taalkundige ontleding levert. Vervolgens is er de lexicaal-semantische databank CORNETTO, die woorden bevat met niet alleen de bijbehorende lexicale en grammaticale eigenschappen, maar ook voor iedere betekenis van een woord de relaties met andere betekenissen, bijvoorbeeld synoniemen (dezelfde betekenis), hyponiemen (een specifieker betekenis), hyperoniemen (een minder specifieke betekenis) en veel ander betekenisrelaties. Het SPRAAK-systeem ten slotte is software waarmee onderzoek gedaan kan worden naar spraakherkenning en waarmee onderzoekers Nederlandse spraakherkenning kunnen inbouwen in toepassingen.

Een andere taalbron opgeleverd door het STEVIN-programma en speciaal van belang voor de meertaligheid is het *Dutch Parallel Corpus* (Vandeweghe, 2012).

De META-NET-experts bijeen in Berlijn op 21 november 2011

Het bestaan van deze (en veel andere) taalbronnen en hun goede beschikbaarheid verklaren in belangrijke mate waarom het Nederlands niet onaardig scoort in de META-NET-studie op het punt van de digitale ondersteuning. Een overzicht van de STEVIN-resultaten vindt u op http://taalunieversum.org/taal/technologie/stevin/etalage.

Blijvende inspanningen nodig voor ondersteuning van het Nederlands

Het STEVIN-programma liep van 2004 tot 2011 en is nu afgelopen. Veel van de taalbronnen zijn pas tegen het eind van het programma opgeleverd. Hoewel eerdere versies van deze taalbronnen al gebruikt zijn binnen het programma, is het potentieel dat de taalbronnen leveren voor onderzoek en ontwikkeling, nog nauwelijks benut. Daarom moeten de Lage Landen na deze succesvolle programma's doorpakken en hun inspanningen voor de ontwikkeling van taaltechnologische bronnen voortzetten en ze gebruiken om onderzoek, innovatie en ontwikkeling voort te stuwen. Anders is niet uit te sluiten dat ook het Nederlands in de gevarenzone komt.

Taaltechnologisch onderzoek moet versterkt worden

Als we de meertaligheid van Europa, een essentieel onderdeel van ons culturele erfgoed, in het digitale tijdperk willen behouden en zelfs versterken, dan is het van groot belang nieuwe mogelijkheden te creëren voor onderzoek naar en ontwikkeling van taaltechnologie. Het META-NET-excellentienetwerk heeft hiervoor een belangrijke aanzet gegeven op Europees niveau: het roept op tot een gezamenlijke inspanning van nationale overheden en de Europese overheid, van kennisinstellingen en bedrijfsleven. Als dat goed aangepakt wordt, dan kan Europa, met een relatief kleine investering, de meertaligheid zelfs tot een economische sterkte maken. META-NET heeft daartoe een strategische onderzoeksagenda opgesteld, die in november 2012 gepubliceerd is en die aangeeft wat de belangrijke thema's voor onderzoek en ontwikkeling zouden moeten zijn voor de periode tot 2020.

Ook voor het Nederlands is er in opdracht van de Nederlandse Taalunie een plan opgesteld om een passend vervolg te definiëren voor het STEVIN-programma. Hierin trekken de kennisinstellingen en het bedrijfsleven in Nederland en Vlaanderen gezamenlijk op, en onderzoek naar taal- en spraaktechnologie wordt ingezet op een gebied met een enorm economisch potentieel: de hoeveelheid digitale data neemt ieder jaar enorm toe. Dagelijks stromen er grote hoeveelheden data binnen. *Big Data* is de trendy term die hiervoor gebruikt wordt. Meer en meer bedrijven moeten die data analyseren om er belangrijke gegevens of trends uit te destilleren. Een zeer groot deel van die data is in de vorm van natuurlijke taal (geschreven of gesproken). Daarom biedt taal- en spraaktechnologie de mogelijkheid enorme kostenbesparingen te realiseren bij de analyse van dergelijke data en zijn bedrijven bereid te investeren in taal- en spraaktechnologie. Maar we kunnen daarbij alleen succesvol zijn als er voldoende onderzoeksmogelijkheden zijn om een kans te hebben de belofte die taal- en spraaktechnologie hier inhoudt, ook daadwerkelijk in te kunnen lossen.

Het is nu aan de politiek, zowel op nationaal als Europees niveau, om beslissingen te nemen over de aanbevelingen die gedaan worden door META-NET op Europees niveau en door de Nederlandse Taalunie op nationaal niveau, en ervoor te zorgen dat de gevaren die de Europese meertaligheid bedreigen, afgewend worden en omgebogen worden tot een sterkte.

Literatuur

- Odijk, J.E.J.M. (2012). *Het Nederlands in het Digitale Tijdperk: The Dutch Language in the Digital Age* (META-NET White Paper Series. Georg Rehm and Hans Uszkoreit (Series Editors)). Berlin: Springer. URL: http://www.meta-net.eu/whitepapers.
- Van Eerten, L. & Depoorter, G. (2011). Het Corpus Gesproken Nederlands: Een waardevolle bron van hedendaagse spraak uit Nederland en Vlaanderen. *Neerlandia/Nederlands van Nu*, 115 (3), pp. 31-33.
- Vandeweghe, W. (2012). Dutch Parallel Corpus. *Neerlandia/Nederlands van Nu*, 116 (2), pp. 38-40.