

# DIXIT

TIJDSCHRIFT OVER TAAL- EN SPRAAKTECHNOLOGIE

## TAAL- EN SPRAAKTECHNOLOGIE voor de OVERHEID



**Een efficiëntere en meer klantvriendelijke  
overheid dankzij taal- en spraaktechnologie**

# INHOUD

## Algemeen

- TST voor de overheid 5
- Communiceren is een kerntaak van de overheid: TST kan daarbij helpen! 6

## Heldere communicatie

- Begrijpelijke brieven met Klinkende Taal 10

## Toegankelijke overheid

- Géén momentje geduld alstublieft! 13

## Wat wil de burger?

- Yournews: automatisch uw eigen nieuws 18
- DuOMAn: Dutch Language Online Media Analysis 20

## Instrumenten voor de organisatie

- TST voor de Rechtspraak, de politie en het NFI 23
- Automatisch vertalen en de overheid 25

## En verder

- Voorwoord 3
- Colofon 3

- Stijgend bezoek [www.notas.nl](http://www.notas.nl) 21

# COLOFON

**DIXIT:** Tijdschrift over toegepaste taal- en spraaktechnologie – 6e jaargang, editie TST voor de Overheid. DIXIT is een uitgave van Stichting NOTaS, Postbus 31070, 6503 CB Nijmegen, Tel. 024 - 352 88 88 Fax 024 -354 00 90 [www.notas.nl](http://www.notas.nl) **Redactieadres:** Stichting NOTaS, Postbus 31070, 6503 CB Nijmegen **Redactie:** Henk van den Heuvel: [h.vandenheuvel@let.ru.nl](mailto:h.vandenheuvel@let.ru.nl), Arjan van Hessen: [hessen@cs.utwente.nl](mailto:hessen@cs.utwente.nl), Remco van Veenendaal: [remco.vanveenendaal@inl.nl](mailto:remco.vanveenendaal@inl.nl), René Ouëndag: [rene.ouendag@q-go.com](mailto:rene.ouendag@q-go.com), Anne Wijnen: [notas@malta-online.nl](mailto:notas@malta-online.nl) **Gastredacteuren:** Folkert de Vriend: [fdevriend@taalunie.org](mailto:fdevriend@taalunie.org), Catia Cucchiarini: [ccucchiarini@taalunie.org](mailto:ccucchiarini@taalunie.org) **Advertenties:** Stichting NOTaS, Anne Wijnen: [notas@malta-online.nl](mailto:notas@malta-online.nl), 024 -352 88 88 **Abonnementen:** Voor een gratis abonnement kunt u zich wenden tot een van de NOTaS-deelnemers **Druk:** Leonard Nijmegen bv **Verantwoording:** DIXIT is een uitgave van Stichting NOTaS. Overname van de artikelen is alleen toegestaan met bronvermelding en na toestemming van Stichting NOTaS. Stichting NOTaS en de bij deze uitgave betrokken redactie en medewerkers aanvaarden geen aansprakelijkheid voor mogelijke gevolgen die zouden kunnen voortvloeien uit het gebruik van de in deze uitgave opgenomen informatie.

# Voorwoord

## De branche Taal- en Spraaktechnologie blijft innoveren

De economische uitdagingen van dit moment hebben nauwelijks effect op enkele sterke eigenschappen van de TST-branche: innovatiekracht en doorzettingsvermogen. Dit werd duidelijk zichtbaar tijdens de afgelopen STEVIN-dag op 4 september jl. op de campus van de Universiteit van Tilburg waar onderzoekers en ontwikkelaars uit kennisinstellingen en het bedrijfsleven bij elkaar kwamen voor de jaarlijkse kennisuitwisselingsdag.

## TST – Nice to have or need to have

Op dit moment kijken bedrijven vooral kritisch naar de efficiëntiewinst van mogelijke investeringen. Heldere communicatie, één van de hoekstenen van efficiënte bedrijfsvoering, speelt daarbij zowel in het bedrijfsleven als bij de overheid, een belangrijke rol. Ook de mogelijkheden van (gedeeltelijke) elektronische zelfbediening door burgers, worden gezien als een belangrijke stimulans voor een efficiëntere overheid. In deze DIXIT leest u meer over de wijze waarop TST op meerdere fronten een sleutelrol speelt binnen de overheid.

## Is investeren nog mogelijk?

Mede dankzij diverse subsidiemogelijkheden en het STEVIN-stimuleringsprogramma, blijven we als TST-branche investeren in “spraakmakende” applicaties. Zo kunnen wij ervoor zorgen dat er ook in de toekomst toonaangevende taal- en spraakoplossingen zullen zijn voor het Nederlands.

## NOTaS = Samen sterk

De bedrijven binnen de Nederlandse Organisatie voor Taal- en Spraaktechnologie (NOTaS) komen uit het MKB. Ze hebben in de regel beperkte middelen en menskracht beschikbaar voor in-house onderzoek- en ontwikkelingsactiviteiten en zijn dus sterk gebaat bij een nauwe samenwerking met (Nederlandse) kennisinstellingen. NOTaS faciliteert deze samenwerking en is alleen daarom al belangrijk. De samenwerking onderling, ook tussen bedrijven en kennisinstellingen die op andere fronten met elkaar concurreren, mag een voorbeeld zijn voor andere branches van hoe we samen veel sterker kunnen zijn, vooral in moeilijke tijden!

## Gastredactie

Bij deze DIXIT over TST voor de overheid zijn we bijzonder blij met de input van onze gastredacteuren van de Nederlandse Taalunie; Folkert de Vriend en Catia Cucchiarini. De artikelen die zij samen met de vaste redactie hebben gekozen, laten zien op hoeveel verschillende manieren TST de overheid kan ondersteunen bij het uitvoeren van haar taken.

Wij blijven innoveren zodat u in de toekomst dit ook kunt doen!

Namens de hele redactie en NOTaS wens ik u veel leesplezier!  
Debbie Kenyon-Jackson, Voorzitter Stichting NOTaS

# Taal- en spraaktechnologie voor de overheid

*De overheid zet momenteel volop nieuwe ICT in voor een betere dienstverlening voor burgers en bedrijven en een hogere efficiëntie van interne werkprocessen. Onder meer in het verlengde van de EU voeren Nederland, België en Vlaanderen gericht beleid voor deze 'e-overheid' (digitale overheid). Deze DIXIT laat zien hoe taal- en spraaktechnologie (TST) kan bijdragen aan de verwezenlijking van de e-overheid. In verschillende regio's komen burgers momenteel al met TST in aanraking wanneer ze het nieuwe netnummer 14+ bellen. Na het inspreken van de naam van een gemeente wordt de beller met spraakroutering automatisch doorgeschakeld naar het juiste loket. Er is met TST echter nog veel meer mogelijk. Het eerste artikel van dit nummer vertrekt vanuit een breder beleidsperspectief. Veel van wat Johan Van Hoorde in zijn tekst naar voren brengt, wordt in de rest van dit nummer geïllustreerd met concrete bedrijfs- en projectresultaten, o.a. afkomstig uit het STEVIN-programma. Deze DIXIT is verder opgedeeld in vier thema's die voor de overheid van belang zijn.*

## Heldere communicatie

Als de overheid goed en efficiënt met haar burgers wil communiceren dan moet ze heldere taal hanteren. Dirk Caluwé gaat daarom in op online beschikbare taalhulpmiddelen voor de schrijvende ambtenaar. Tigran Spaan en Oele Koornwinder bespreken vervolgens een product dat geïntegreerd kan worden in een tekstverwerker en aanwijzingen geeft om de leesbaarheid van de tekst te verbeteren.

## Toegankelijke overheid

De overheid moet verschillende communicatiekanalen faciliteren, ook voor mensen met een handicap. Its Kievits beschrijft hoe websites voor mensen met een visuele handicap automatisch kunnen worden voorgelezen met TST. Ook verwacht de burger steeds meer dat de overheid altijd bereikbaar is. Joop van Gent bespreekt daarom een prototype waarmee burgers 24/7, zonder wachttijden telefonisch te woord worden gestaan door een op TST gebaseerd systeem. De bijdragen van Henk van den Heuvel en René van Horik, ten slotte, gaan in op duurzame opslag van informatie waardoor toegang tot die informatie ook op de langere termijn gewaarborgd is.

## Wat wil de burger?

Dankzij eParticipatie kunnen burgers een adviserende rol in het beleidsproces spelen. Hun 'probleem- en contextkennis' kan gebruikt worden voor betere en creatievere oplossingen. Zo bespreekt Vincent de Klerk een project waarin van individuele burgers feedback over de dienstverlening werd verkregen met gebruik van TST. De ondersteuning van media-analisten bij het monitoren van de reputatie van hun organisatie wordt

door Valentin Jijkoun en Maarten de Rijke besproken. Joop van Gent beschrijft vervolgens een systeem waarmee het alsmaar groeiende nieuwsaanbod beheersbaar wordt gehouden, bijvoorbeeld voor een knipselkrant.

## Instrumenten voor de organisatie

Met het juiste instrument wordt ook de interne werking van de overheid efficiënter. Arjan van Hessen bespreekt het gebruik van TST bij Justitie en politie voor het doorzoeken van spraakopnames. Vincent Vandeghinste en Jaap van der Meer beschrijven de mogelijkheden van automatisch vertalen. Frieda Steurs gaat vervolgens in op betere ontsluiting van informatie door termextractie. Tot slot bespreken Christophe van Bael en Diana Binnenpoorte de inzet van spraakroutering.

Het is niet vanzelfsprekend dat TST ook met het Nederlands overweg kan. Daarom is enkele jaren geleden het STEVIN-programma opgezet, een meerjarig onderzoeks- en stimuleringsprogramma voor Nederlandstalige TST dat door de Vlaamse en Nederlandse overheid wordt gefinancierd ([www.stevintst.org](http://www.stevintst.org)). Verschillende van de in dit nummer besproken TST-oplossingen zijn binnen dit programma ontwikkeld. Als coördinator van STEVIN ziet de Nederlandse Taalunie het als haar taak om er voor te zorgen dat de resultaten van het programma ook zo veel mogelijk hun weg vinden naar concrete producten. Deze doelstelling sluit uitstekend aan bij de huidige overheidsbrede initiatieven in Nederland, België en Vlaanderen voor de realisering van de e-overheid.

**Folkert de Vriend  
en  
Catia Cucchiarini**  
Nederlandse  
Taalunie

# Communiceren is een kerntaak van de overheid: TST kan daarbij helpen!

*Taal- en spraaktechnologie (TST) is voor steeds meer mensen een alledaagse zaak. Zelfs dat die met Nederlands overweg kan. We vinden het normaal dat het navigatiesysteem in de auto Nederlands spreekt, of dat we onze mobiele telefoon kunnen vragen om naar kantoor te bellen. En wie staat er stil bij wat er allemaal komt kijken bij de kringeltjes onder slecht gebouwde zinnen in een tekst?*

**Johan Van Hoorde**  
Nederlandse  
Taalunie

Om dat soort dingen met een taal te kunnen doen, zijn basisvoorzieningen nodig, bijvoorbeeld uitspraakinformatie en programmatuur die zinnen in relevante delen kan opsplitsen. Die voorzieningen bleken voor het Nederlands vaak te ontbreken of van te bescheiden kwaliteit. Sinds meer dan tien jaar werkt de Nederlandse Taalunie hieraan, samen met vele partners in Nederland en Vlaanderen. Ontbrekende onderdelen worden ontwikkeld via het Vlaams-Nederlandse STEVIN-programma ([www.stevin-tst.org](http://www.stevin-tst.org)), dat door de Vlaamse (EWI, IWT en FWO) en Nederlandse (EZ, NWO en OCW) overheid wordt gefinancierd en door de Taalunie wordt gecoördineerd. De resultaten (corpora, lexica, taalmodules) worden beheerd en up-to-date gehouden door de TST-Centrale van de Taalunie ([www.tstcentrale.org](http://www.tstcentrale.org)). Via dat kanaal zijn ze ook beschikbaar voor gebruikers.

Veel ruwe grondstof als het ware, die zelf weinig in de kijker liep, maar er wel voor kon zorgen dat navigatiesystemen, telefoon-toestellen en tekstverwerkers natuurlijk ook met Nederlands overweg kunnen. Nu dat is bereikt kan directer worden ingespeeld op de behoeften van specifieke doelgroepen.

De Taalunie wil doelgroepen nu helpen om hun weg naar de voorzieningen te vinden. Centraal in die aanpak is de overheid zelf, niet alleen de centrale overheid maar ook provincies, gemeenten en instellingen binnen de publieke sector. Vrijwel alle overheden communiceren veel en vaak. In zekere zin kan men stellen dat communicatie een kernactiviteit van de overheid is. Ze is immers een organisator van maatschappelijke consensus en mediator tussen uiteenlopende inzichten en belangen in de samenleving. Hoe anders dan in communicatieve contacten met burgers en belangengroepen kan de overheid zorgen voor voldoende draagvlak voor haar maatregelen en besluiten?

Daarom wellicht dat de overheid zelf steeds meer belang is gaan hechten aan goede communicatieve vaardigheden en efficiënte communicatie-instrumenten.

Dat is ook het kader waarin Nederland, België en Vlaanderen e-governmentbeleid voeren, ook wel aangeduid als *eOverheid* of als *Digitale Overheid*. Door gebruik te maken van internet en ICT-toepassingen willen de overheden de kwaliteit en toegankelijkheid van hun informatie en dienstverlening verbeteren. Informatie komt sneller bij de burgers, wordt toegankelijk voor meer mensen en is minder afhankelijk van plaats en tijd. Bovendien kunnen burgers sneller en gemakkelijk hun zegje doen als het gaat om doelstellingen en inhoud van beleid. Anders gezegd: *eOverheid* zorgt voor meer interactiviteit. Er is niet alleen informatie van de overheid naar de burger, maar ook omgekeerd!

De Nederlandse Taalunie wil dat beleid ondersteunen met een actieplan Burger, Taal en Overheid. Het plan zal drie onderscheiden actielijnen omvatten:

- (a) het ondersteunen van e-government;
- (b) het ondersteunen van het bestaande overheidsbeleid voor *open standaarden* en *opensourcesoftware*;
- (c) het ondersteunen van overheidsinstanties bij het hanteren van correct, begrijpelijk en burgergericht Nederlands.

Eigenlijk is het niet meer dan vanzelfsprekend dat juist ook de overheid zelf mee kan profiteren van de resultaten van het TST-beleid. Toch vinden IT'ers uit de publieke sector niet vanzelfsprekend hun weg naar deze materialen. Men geeft zich zelfs onvoldoende rekenschap van de gebruiksmogelijkheden, terwijl de inzet van taal- en spraaktechnologie toch aanzienlijke kwaliteitsverbetering kan bieden, met name voor die doelgroepen die moeilijker bereikbaar

zijn, zoals voor mensen met communicatieve beperkingen. Denken we maar aan automatische voorleesmodules binnen overheidswebsites, die blinden en slechtzienden in staat moeten stellen schriftelijke – dus voor die groep moeilijk toegankelijke – informatie door de computer te laten voorlezen. Dit is slechts één voorbeeld maar het laat wel goed zien dat de overheid moet weten wat er op de TST-markt voorhanden is en hoe dit kan worden ingezet.

De TST-materialen kunnen ook het overheidsbeleid voor *open standaarden* en *opensourcesoftware* ondersteunen. Nederland, België en Vlaanderen willen het gebruik van open standaarden en open toepassingen bevorderen, omdat die verschillende systemen beter laten samenwerken en gebruikers minder product- en producentafhankelijk maken. Goede wil volstaat daartoe niet. Eén mogelijke hinderpaal is alvast de wijze waarop open toepassingen omgaan met het Nederlands, bijvoorbeeld of er spelling-, stijl- en grammaticacontrole is en hoe goed die werkt. Mensen die overstappen van gesloten naar open toepassingen zullen er zeker niet op achteruit willen gaan. Daarom wil de Taalunie de TST-materialen inzetten voor het ontwikkelen van taalmodules, zoals spelling- en stijlcontrole, synoniemen, maar ook tekst-naar-spraakmodules, terminologiebeheer en vertaalmodules. Onze basisvoorzieningen bieden voor Nederlandstaligen meer ondersteuningsmogelijkheden dan er nu zijn, terwijl bestaande functies aanzienlijk beter zouden kunnen!

Ter voorbereiding van het plan heeft de Taalunie tal van contacten gelegd, vooral over ambtelijk taalgebruik. De ambtelijke taal was zelfs het centrale werkthema in 2008, onder het motto 'Burger, Taal & Overheid'. Onze contacten met overheidsinstanties wijzen steeds weer uit dat er grote behoefte bestaat aan een centrale, bemiddelende en faciliterende instantie. Vaak wordt erover geklaagd dat er teveel in gespreide slagorde wordt gewerkt, waardoor het resultaat van de inspanningen niet altijd optimaal is.

Voorlopig lokale overheden ontbreekt het aan middelen en knowhow. "We willen wel", is een veelgehoorde reactie, "maar we weten niet waar we moeten beginnen". Die reacties lijken overeen te komen met een belangrijke vaststelling in het rapport 'Burgers en eOverheid' dat Ernst & Young in het voorjaar presenteerde. Daarin stelt het onderzoeksbureau:

*Het verder centraliseren en inrichten van*

*uniforme ICT-voorzieningen voor overheidsinstanties kan de ontwikkeling van de digitale dienstverlening effectiever en efficiënter maken. Tot dusver worden ICT-projecten veelal decentraal uitgevoerd door afzonderlijke overheidsinstanties. Centrale, gezamenlijke, klantgerichte ICT-oplossingen voor meerdere instanties vormen nog de uitzondering. Bij verdere centralisatie zou de rijksoverheid meer de regie moeten nemen. Het denken vanuit de separate overheidsinstantie moet hierbij vervangen worden door het centraal stellen van de burger en de onderneming.*

De Taalunie is natuurlijk geen kandidaat om zelf de regie voor de *eOverheid* te voeren. Wel kan ze een belangrijke schakel vormen in een grensoverschrijdende samenwerking, omdat ze goed zicht heeft op de betrokken instanties in de beide landen en dicht bij de beide regeringen opereert. Veel pleit voor een grensoverschrijdende aanpak. De ervaring van de Taalunie heeft geleerd dat die tot betere resultaten én tot lagere kosten leidt. Zoals het geval was met het ambitieuze STEVIN-programma. Waarschijnlijk hadden de landen apart nooit dezelfde resultaten bereikt, noch naar omvang noch naar kwaliteit.

In een centrale regie ziet de Taalunie verder uitstekende kansen voor het maximaal inzetten van taal- en spraaktechnologie in innovatieve projecten. De overheid zal gemakkelijker zijn weg vinden naar TST als ze dat vanuit een centrale regiekamer kan doen.

De komende jaren zal de Taalunie taal- en spraaktechnologie nadrukkelijk op de agenda plaatsen van de overheid als gebruiker van informatiesystemen. Ze zal symposia opzetten, masterclasses organiseren, best practices in de schijnwerper plaatsen en pilotprojecten opzetten. Deze DIXIT laat zien hoe veel mogelijkheden daartoe bestaan!

Meer informatie over TST is te vinden op het Taalunieversum ([www.taalunieversum.org/taal/technologie](http://www.taalunieversum.org/taal/technologie))



e-OVERHEID



# Begrijpelijke brieven met Klinkende Taal

*Communiqueert uw organisatie klantgericht en open? Schrijven uw collega's mooie volzinnen... waar de burger niets van begrijpt? Volgt uw klant altijd wat u bedoelt? Gebruikt u heldere terminologie op uw website of vinden de lezers het onbegrijpelijk jargon?*

*Overheidsorganisaties en bedrijven willen helder communiceren. Een organisatie die helder communiceert laat zien dat zij open en klantgericht is. Heldere communicatie voorkomt problemen en bespaart kosten: minder klachten, minder telefoontjes en minder onduidelijkheid.*

**Tigran Spaan  
en  
Oele Koorwin-  
der  
GridLine B.V.**

Veel organisaties schakelen inhoudelijke experts in voor het schrijven van brieven, webpagina's en brochures. Dat zijn niet altijd mensen die gewend zijn te schrijven voor het algemeen publiek. Om deze experts bij het schrijven te ondersteunen heeft GridLine, mede dankzij financiering vanuit het STEVIN-programma, het product Klinkende Taal ontwikkeld ([www.klinkendetaal.nl](http://www.klinkendetaal.nl)).

Met Klinkende Taal kunnen professionals zich richten op de inhoud. Klinkende Taal is altijd aanwezig en met één druk op de knop aan te roepen tijdens het schrijven. Het programma geeft duidelijke aanwijzingen om de leesbaarheid van de tekst te verbeteren en af te stemmen op de doelgroep.

GridLine is een IT-bedrijf dat gespecialiseerd is in taaltechnologie voor het Nederlands. Doordat we gespecialiseerd zijn in het Nederlands, hebben wij alle mogelijke basis-modules en applicaties voor het Nederlands klaar liggen, zoals zoekmachine-optimalisatie, spelling, automatische classificatie, terminologiebeheer, opinion mining, spraakherkenning en ook leesbaarheids-beoordeling. Onze hardcore IT'ers integreren onze taalproducten in veelgebruikte software, zoals Microsoft Word, Sharepoint, FAST en Lucene.

## Overheidsbeleid

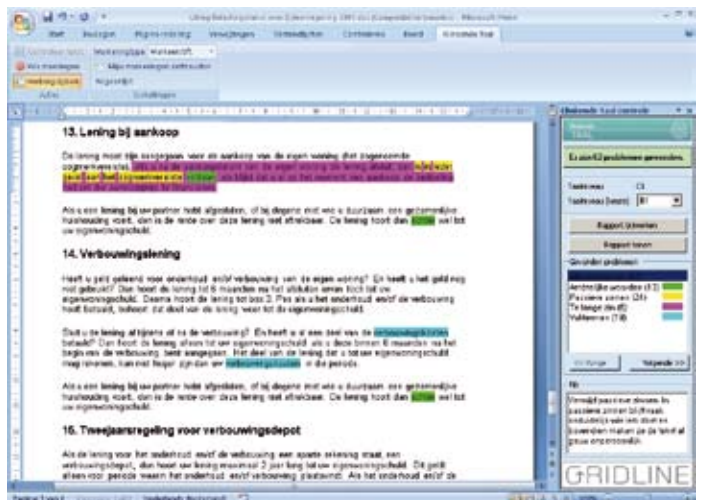
Klinkende Taal sluit aan bij een concrete overheidsbehoefte. De regering heeft namelijk bepaald dat de overheid zich moet inspannen om burgers begrijpelijk te informeren en binnen redelijke termijnen moet antwoorden op brieven, e-mails en bezwaarschriften. Deze doelstellingen zijn vastgelegd in verschillende documenten, waaronder de Webrichtlijnen-norm voor overheidswebsites, de richtlijn voor

Begrijpelijke Formulieren, de e-mailgedraglijn van burger@overheid en het manifest 'Overheid heeft Antwoord'. Het bevorderen van begrijpelijke taal staat bovendien in de Aanpak Top 10 van het Ministerie van Binnenlandse Zaken.

De Nationale ombudsman ziet scherp toe op de naleving van deze richtlijnen. In 2008 verscheen bijvoorbeeld een rapport met de titel 'Behandeling burgerbrieven kan behoorlijker'. Hierin concludeert de ombudsman dat de overheid in 2007 vooruitgang heeft geboekt met het beantwoorden van burgercorrespondentie (met name e-mails), maar dat de afhandeling van bezwaarschriften nog veel te wensen overlaat. De ombudsman heeft ook concrete brieven gecontroleerd op stijl en begrijpelijkheid. In het rapport valt te lezen hoe de onderzochte ministeries het resultaat kunnen verbeteren.

Klinkende Taal kan ambtenaren helpen bij het uitvoeren van deze richtlijnen en adviezen. Met Klinkende Taal kan een ambtenaar bijvoorbeeld snel nagaan of zijn brief het juiste taalniveau heeft. Bovendien vertelt de tool precies hoe hij de brief kan verbeteren. Dit is

Voorbeeld van een Klinkende Taalverbeteradvies bij een gecontroleerde tekst



niet alleen handig, maar werkt ook tijdbesparend. Niet onbelangrijk, gezien de eis dat de overheid binnen twee dagen antwoord moet geven op eenvoudige burgerbrieven en daarbij ook nog eens begrijpelijk moet zijn.

### Klinkende Taal

Klinkende Taal beoordeelt teksten op basis van het Common European Framework of Reference (CEFR), dat teksten indeelt in niveaus van A1 (heel makkelijk) tot C2 (heel moeilijk). Voor nadere informatie, zie de sectie over het CEFR op de Taaluniversity-website (taaluniversity.org). De praktische kennis van de afdeling Communicatie van de Gemeente Den Haag vormde een goede basis voor de eerste versie van de tool. Voor het tunen van Klinkende Taal hebben we onderzoek laten doen door Henk Pander Maat van de afdeling Taalbeheersing van de Universiteit Utrecht.

In het vervolgtraject hebben we Klinkende Taal in Microsoft Office Word geïntegreerd en met nieuwe functies uitgebreid. Hierbij hebben we goed naar onze gebruikers geluisterd. Dit heeft uiteindelijk een product opgeleverd dat niet alleen het niveau van een tekst kan beoordelen, maar ook praktische aanwijzingen geeft om de tekst zo te veranderen dat de doelgroep hem begrijpt.

### Taalcriteria

Klinkende Taal controleert zowel het woordgebruik als de zinsopbouw, de logische samenhang en de opmaak van een tekst. Bij de beoordeling van het woordgebruik markeert Klinkende Taal moeilijke woorden, specialistische taal, ambtelijk of ouderwets taalgebruik en abstract taalgebruik. Op zinsniveau kijkt Klinkende Taal onder meer naar zinslengte, bijzinnen, passiefconstructies, verbindingswoorden en tangconstructies. Op paragraafniveau beoordeelt Klinkende Taal de samenhang en de tekstopmaak.

Naast deze standaardcriteria kan Klinkende Taal ook aanvullende criteria meenemen, bijvoorbeeld uit de stijlwijzer van de klant. Een andere mogelijkheid is om Klinkende Taal te laten afstemmen op de eigen terminologie. Dit helpt om beter zicht te krijgen op het voorkomen van vaktermen en om de termkeuze te standaardiseren. Dit alles is mogelijk dankzij de Nederlandse taaltechnologie van GridLine.

### Productvormen

Klinkende Taal is verkrijgbaar als plugin in Microsoft Office Word en als laagdrempelige web-plugin. De schrijver van een brief

of webtekst roept Klinkende Taal met één druk op de knop aan. Er verschijnt dan een beoordeling van de moeilijkheid van de tekst (bijvoorbeeld B2) en duidelijke aanwijzingen om de leesbaarheid te verhogen (bijvoorbeeld naar B1). Klinkende Taal levert ook een Quick Scan. Deze tool kan in één klap een verzameling documenten of een complete website testen. Hierbij kijkt de tool onder meer naar de begrijpelijkheid voor de doelgroep en het jargongebruik, criteria die ook onderdeel zijn van de Webrichtlijnen. Met de Quick Scan kan een redacteur snel nagaan welke documenten en webpagina's extra aandacht nodig hebben.

Klinkende Taal is eenvoudig te installeren. Alle taaltechnologie staat op de server. Na het aanroepen van Klinkende Taal legt Microsoft Office Word via een webservice verbinding met de centrale taalserver. Deze aanpak heeft als voordeel dat het heel eenvoudig is om nieuwe versies door te voeren: hiervoor is een update op de server voldoende. De webversie roept dezelfde server aan. Het beheer van woordenlijsten kan de organisatie zelf doen via een eenvoudige webmodule.

### Gebruikers

Klinkende Taal is oorspronkelijk ontwikkeld voor gemeentes. De vier grootste Nederlandse Gemeentes hebben de tool als eerste in gebruik genomen. Andere overheidsdiensten zijn daarna gevolgd. Een nieuwe groep gebruikers zijn communicatie-bureaus. Zij vinden het handig om Klinkende Taal te gebruiken voor controledoeleinden, zodat de schrijvers zich kunnen richten op de inhoud en het creatieve proces. Bovendien biedt Klinkende Taal een meerwaarde voor hun klanten. Verzekerings-maatschappijen, woningcorporaties, banken en ziekenhuizen sluiten inmiddels aan bij de groeiende groep gebruikers van Klinkende Taal.

Het project Klinkende Taal werd mede gefinancierd door het STEVIN-programma. STEVIN is een meerjarig onderzoeks- en stimuleringsprogramma voor Nederlandstalige taal- en spraaktechnologie dat gezamenlijk door de Vlaamse en Nederlandse overheid wordt gefinancierd.

## **Instituut voor Nederlandse Lexicologie** **Schatkamer van de Nederlandse taal**



*Het Instituut voor Nederlandse Lexicologie (INL) is een Vlaams-Nederlands instituut dat Nederlandse woorden verzamelt, bestudeert, opslaat in databases, voorziet van allerlei (taalkundige) gegevens en daarmee wetenschappelijke woordenboeken maakt. Of u nu taalkundige bent of simpelweg geïnteresseerd in taal: met al uw vragen over woorden kunt u bij ons terecht.*

### **TST-Centrale**

**Vlaams-Nederlandse centrale voor beheer, onderhoud en distributie van digitale taalmaterialen**

**Bent u op zoek naar Nederlandstalige, digitale taalmaterialen?  
Bij ons bent u aan het juiste adres!**

De Centrale voor Taal- en Spraaktechnologie (TST-Centrale) stelt taalmaterialen beschikbaar die veelal met overheidsgeld zijn gefinancierd, waaronder materialen die op het INL ontwikkeld

worden en resultaten uit het STEVIN-programma (STEVIN: Spraak- en Taaltechnologische Essentiële voorzieningen voor het Nederlands). Daarnaast ondersteunt de TST-Centrale het gebruik van de materialen door uw vragen te beantwoorden en gastcolleges en workshops te organiseren.

*De TST-Centrale is een initiatief van en wordt gefinancierd door de Nederlandse Taalunie. De TST-Centrale is ondergebracht bij het Instituut voor Nederlandse Lexicologie met vestigingen in Leiden en Antwerpen.*

#### **Beschikbare materialen:**

- Historische en wetenschappelijke elektronische woordenboeken (o.a. Woordenboek der Nederlandsche Taal online)
- Gesproken, geschreven en multimediale corpora (o.a. Corpus Gesproken Nederlands)
- Mono- en bilinguale lexica (o.a. Referentiebestand (Belgisch-)Nederlands)
- Tools voor gesproken en geschreven teksten (o.a. AUTONOMATA-g2p-toolkit)

[www.inl.nl](http://www.inl.nl)  
[www.inl.nl/tst-centrale](http://www.inl.nl/tst-centrale)  
[www.inl.nl/producten](http://www.inl.nl/producten)

[servicedesk@inl.nl](mailto:servicedesk@inl.nl)

**nl | TST-Centrale**

# “Géén momentje geduld alstublieft!”

*“Er zijn nog 7 wachtenden voor u.”, “Heeft u een vraag over onze aanbiedingen, toets dan 7, heeft u een vraag over uw rekening, toets 8, voor overige vragen, toets een 9. Een momentje geduld alstublieft.” Wie kent ze niet, de ergerlijke telefoonbeantwoorders, die je simpelweg verwijzen naar een website of minutenlang – vaak tegen een stevig tarief – laten wachten met foute muziek, en je uiteindelijk doorverbinden met iemand die je niet verder kan helpen.*

Dit soort ‘kluitje-in-het-riet’-telefoontjes behoren binnenkort tot het verleden. De Delftse bedrijven Irion en Dutcheer hebben een prototype gebouwd, dat is gebaseerd op een spectaculaire en ingenieuze combinatie van taaltechnologie en spraakherkenning. Dit systeem – GemeenteConnect genaamd – levert de bellende burger meteen boter bij de vis: een bruikbaar antwoord op een concrete vraag, of automatisch doorverbinden met de juiste persoon bij de gemeente.

## Wat is de clou?

Het idee is eigenlijk simpel: mensen hebben geen volledige, correct geformuleerde zinnen nodig om een betekenis of bedoeling te begrijpen. Sterker nog, mondelinge communicatie verloopt zelden grammaticaal correct. Denk aan zinnetjes als “al in je nieuwe tent geslapen?” of “jij koffie?”. De zinnen zijn onvolledig, maar worden toch moeiteloos begrepen. Of denk eens aan een telefoongesprek via een belabberde, krakende verbinding (“sorry, ik zit in een tunneltje!”). Ook al versta je misschien maar een paar flarden, die flarden blijken vaak toch voldoende om de kern van de boodschap op te pikken. Dat komt doordat mensen in staat zijn deze flarden snel in een juiste context te plaatsen.

Wanneer een ontvanger al ongeveer weet binnen welke context hij iets moet begrijpen, is het interpreteren nóg gemakkelijker. Als u in een kamer zit, en iemand steekt zijn hoofd om de deur met de vraag “Koffie?”, is de bedoeling duidelijk. Was de vraag “Krokodil?”, dan zou u vermoedelijk niet begrijpend het hoofd schudden. In dit laatste voorbeeld zit de basisgedachte van het systeem opgesloten: als een beller maar een handjevol onderwerpen kán bedoelen, wordt het achterhalen van die bedoeling al snel een flink stuk makkelijker. Je belt een gemeente bijvoorbeeld niet voor een frietje satésaus of een huurauto. In een duidelijke context heeft een goed verstander maar een half woord nodig. En die goed verstander mag best een computer zijn.

## Proefkonijn

De resultaten: in de ‘proefkonijn’-gemeente Gilze-Rijen konden maar liefst 80% van de veel voorkomende vragen binnen 2 minuten door het systeem correct worden afgehandeld! Zelfs een test met Turkse inwoners leverde vergelijkbare resultaten op. De voordelen zijn duidelijk: bellers worden zonder wachttijden te woord gestaan, krijgen zonder op allerlei toetsen op hun toestel te hoeven drukken meteen een ‘digitale deskundige’ aan de lijn, en kunnen 24 uur per dag, 7 dagen in de week telefonisch met hun vragen bij de gemeente terecht! Ook voor de gemeente zelf biedt dit grote voordelen: er blijft meer tijd over om de complexere vragen af te handelen, de bereikbaarheid is drastisch verbeterd, en in veel gevallen zullen ook de kosten omlaag kunnen.

GemeenteConnect is een demonstratieproject uit de eerste ronde van het STEVIN-programma (zie [taaluniversity.org/taal/technologie/stevin/projecten/#gemeenteconnect](http://taaluniversity.org/taal/technologie/stevin/projecten/#gemeenteconnect))

**Joop van Gent**  
**Irion**



# Yournews: automatisch uw eigen nieuws

*Kort geleden zei iemand tegen me dat een mens tegenwoordig op één dag evenveel informatie moet verwerken als honderd jaar geleden in een heel jaar. Of dat waar is weet ik niet, maar zeker is dat de opkomst en kracht van de nieuwe media ervoor hebben gezorgd dat de tijd dat we voor het nieuws moesten wachten tot de krant op de mat viel voorgoed voorbij is.*

## Joop van Gent Irion



Met de nieuwe vormen van nieuwsvoorziening is ook het aanbod exponentieel gegroeid. Het nieuws wordt ons door alle poriën naar binnen geduwd. We kunnen moeiteloos en gratis de digitale versies van honderden kranten op onze PC krijgen, of zelfs op ons mobieltje. Maar naarmate de stroom van informatie aanzwelt, groeit net als bij de rivier ook de behoefte aan indamming.

We verliezen het overzicht en worden soms wanhopig van de overstelpende hoeveelheid informatie. Dat geldt zeker ook voor ambtenaren bij de overheid die nieuwsberichten verwerken voor bijvoorbeeld knipselkranten voor hun departementen. Een geslaagde poging van het op de juiste plaats en tijd in de juiste hoeveelheid aanbieden van informatie is Yournews.

### Het STEVIN-project Easyinfo

Yournews is een nieuwe nieuwssite, gebaseerd op de concepten uit het Stevin-project EasyInfo ([taalunieversum.org/taal/technologie/stevin/projecten/#easyinfo](http://taalunieversum.org/taal/technologie/stevin/projecten/#easyinfo)). Dit project, uitgevoerd door twee Nederlandse taaltechnologiebedrijven, Irion en Carp, in samenwerking met MD Info, een multimediale aanbieder van gepersonaliseerde zakelijke informatie, had tot doel het nut van taaltechnologie te bewijzen voor selectieve nieuwsvoorziening in een werkende applicatie.

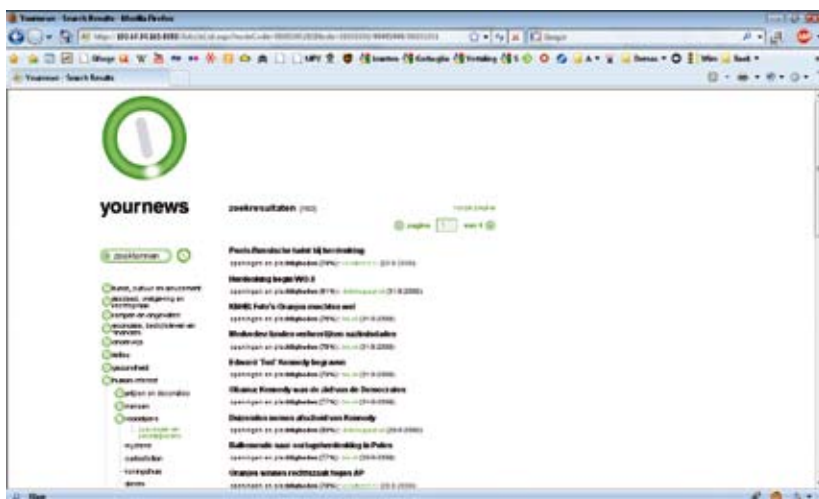
### Yournews

Yournews sorteert nieuws en stemt krantenberichten af op het behoefteprofiel van de gebruiker. Bovendien vat Yournews de nieuwsberichten ook nog eens samen. Het hele proces is automatisch, en abonnees krijgen precies het pakketje dat voor hen interessant is. Zo kan het dagelijks nieuws uit honderden, zo mogelijk duizenden bronnen op een veel efficiëntere manier bij de mens worden gebracht dan tot nog toe mogelijk was.

Bestond dit dan niet allang? Wel, ja en nee. Online nieuwssdiensten zijn inderdaad niets nieuws, er zijn er duizenden van. Maar bij deze diensten wordt het nieuws doorgaans met de hand uitgesorteerd, en worden de behoefteprofielen voor klanten ook met de hand gemaakt. En dat is meteen ook hun achilleshiel. Want dergelijk arbeidsintensief handwerk heeft alleen zin voor gebruikers die daarvoor goed willen betalen. Een alternatief model waarbij de inkomsten niet uit betalende klanten, maar uit advertentieinkomsten komen, levert onvoldoende op om het handwerk te bekostigen. Dat is dan ook de reden dat de meeste van deze diensten zich richten op de zakelijke markt. Voor veel overheidsinstanties is abonneren niet aantrekkelijk, temeer omdat ook de profielen meer zijn toegespitst op de zakelijke markt, en in veel mindere mate op de maatschappelijke thema's.

### Kostenbesparing

Doordat Yournews gebruik maakt van taaltechnologie, een technologie die het mogelijk maakt om tekstuele informatie automatisch door een computer te laten interpreteren, komt op maat gesneden nieuwsvoorziening beschikbaar voor een veel groter publiek. De technologie betekent een dermate grote kostenbesparing, dat Nederlandstalige nieuwsberichten niet alleen voor de zakelijke markt, maar ook voor burgers en overheid gemakkelijk binnen handbereik komen. Naar verwachting zal Yournews.nl aan het eind van dit jaar worden gelanceerd.



# DuOMAn: Dutch Language Online Media Analysis

Het STEVIN-project DuOMAn is gericht op het ontwikkelen van taaltechnologie ter ondersteuning van media-analisten, die zich richten op onderzoek naar de mediareputatie van organisaties. Een belangrijk doel daarbij is inzichtelijk maken hoe er door media, journalisten en in sociale media over hen gesproken wordt.

**Valentin Jijkoun**  
en  
**Maarten de Rijke**  
Universiteit van  
Amsterdam

De technologie die binnen DuOMAn ontwikkeld wordt kan ook ingezet worden bij informatievoorziening ten behoeve van de overheid enerzijds - bijvoorbeeld om antwoord te krijgen op de vraag welke meningen er over een gegeven onderwerp leven onder de bevolking - en ten behoeve van de burgers anderzijds - bijvoorbeeld om inzicht te krijgen in de belanghebbenden rondom een onderwerp.

Waar het gaat om overheidsinformatie komt er steeds meer online materiaal, zowel geredigeerd, in de vorm van bijvoorbeeld persberichten, officiële websites, interviews of krantenberichten, als user generated in de vorm van blogs, tweets of bijdragen aan discussiefora van politici, vertegenwoordigers van de overheid en burgers. Analyses van dergelijk materiaal zijn nuttig in vele scenario's waarin overheid en burger elkaar treffen.

Hieronder noemen we er enkele.

- Een overheidsorganisatie wil graag een overzicht van de belangrijkste standpunten en argumenten rondom een gegeven onderwerp, bijvoorbeeld 'fietsverlichting' of 'OV-chipkaart'.
- Een burger wil achterhalen wat de recente discussies in het parlement waren rondom bijvoorbeeld de kosten van immigratie, en wil daarbij inzicht krijgen in de belangrijkste experts en belanghebbenden (plus hun standpunten).
- Het voorspellen van de mogelijke reacties en impact van beslissingen rondom gevoelige onderwerpen ('het verhogen van de pensioengerechtigde leeftijd').
- Het meten van publieke opinies in sociale media is duidelijk relevant voor politieke partijen, de overheid en de burger.

Binnen het DuOMAn-project wordt voornamelijk gewerkt aan technologie waarmee de

*Peilend.nl: de publieke demonstrator van het DuOMAn-project*

The screenshot shows the Peilend.nl website interface. At the top, there's a search bar with the query 'amsterdam chipkaart' and a date range from 08/15/2009 to 09/15/2009. Below the search bar is a table of search results with columns for 'Datum', 'Bronnen', 'Artikel', and 'Reacties'. The table lists several news items related to the introduction of the OV-chipkaart in Amsterdam. To the right of the table is a 'Woordenwolk' (word cloud) containing terms like 'Amsterdam', 'GVB', 'chipkaart', 'metro', 'OV', 'problemen', 'stations', and 'strippenkaart'.

Datum	Bronnen	Artikel	Reacties
27 aug 18:5	V	[20:58] D-day voor ov-chip in Amsterdamse metro	0
27 aug 16:4	7	Invoeren chipkaart Amsterdam verloopt soepel AMSTERDAM (ANP) - Het afschaffen van de strippenkaart en de invoering van de ov-chip	0
27 aug 16:0	7	Alleen nog OV-chip in metro Amsterdam Vanaf donderdag is de strippenkaart passé in de Amsterdamse metro. Reizigers kunnen er	0
27 aug 11:2	7	Rover tevreden over einde strippenkaart Amsterdam ... tevreden over de overgang van de strippenkaart op de OV-chipkaart in de Amsterdamse	0
27 aug 11:1	7	Invoering ov-chipkaart in Amsterdam verloopt soepel Invoering ov-chipkaart in Amsterdam verloopt soepel 27 augustus 2009, 13:11 uur   FD.nl	0
27 aug 10:5	7	GVB: invoering chipkaart gaat soepel ... GVB is vooralsnog tevreden over de invoering van de ov-chipkaart in de Amsterdamse	0
27 aug 10:5	7	Rover positief over invoering ov-chipkaart AMSTERDAM (ANP) - Reizigersvereniging Rover vindt dat de invoering van de ov-chipkaart	0
27 aug 9:48	7	Amsterdamse metro stapt over op ov-chipkaart AMSTERDAM (ANP) - Rijen voor verkoopautomaten en toegangspoortjes, maar geen extre	0
27 aug 8:44	7	GVB: invoering chipkaart gaat soepel AMSTERDAM (ANP) - Het openbaarvervoerbedrijf GVB is vooralsnog tevreden over de inv	12
27 aug 8:16	7	Einde aan strippenkaart Amsterdamse metro	0

Woordenwolk: Amsterdam, CBS, GVB, Het, GVB, OV-Chipkaart, Rb, Rotterdam, VS, Volgens, aantal, afschaffen, alle, alleen, apparaat, automaten, basistarief, bijna, bus, chipkaart, dag, druk, eerste, enkele, extra, gaat, goed, goedkoper, grote, honderd, inchecken, ingezet, in, jaar, januari, juli, kaart, keer, komen, kopen, kost, kwartaal, laat, lange, maanden, maken, medewerkers, mensen, metro, mobiele, mogelijk, ondergrondse, openbaar, openbaarvervoerbedrijf, OV, passagiers, poortje, poortjes, problemen, procent, reis, reizen, reizigers, retourtje, retourtjes, rijen, soepel, stations, steden, stonden, strippenkaart, stromen, Organisaties, Locaties, Andere, instellingen

eerste twee scenario's ondersteund kunnen worden. (In het door NWO gefinancierde project Tracking News Events and their Impact (TNT), dat binnen dezelfde onderzoeksgroep wordt uitgevoerd als DuOMAN, ligt de nadruk juist op het derde scenario.)

Binnen DuOMAN kunnen we verschillende ontwikkelingslijnen onderscheiden die van belang zijn waar het gaat om analysemiddelen voor zowel overheid als burgers. Om te beginnen werkt DuOMAN aan het verzamelen en linken van geredigeerd en user generated materiaal. Voor deze links wordt gebruik gemaakt van named-entity-herkenning en named-entity-normalisatie, waarmee belangrijke belanghebbenden kunnen worden geïdentificeerd. Daarnaast kunnen sentimenten en opinies geanalyseerd en geaggregeerd worden. Op de linker pagina is de publieke demonstrator van het DuOMAN-project te zien, met daarin het resultaat scherm voor de zoekvraag 'Amsterdam chipkaart', werkend op artikelen vergaard van nieuwssites. In dit scherm zien we links een overzicht van de resultaten, met de duplicaten gegroepeerd, met vermelding van het aantal reacties waartoe een bericht aanleiding heeft gegeven. Rechts is een kleur-gecodeerde woordenwolk waarin de kleuren overeenkomen met verschillende soorten van named entities (personen, organisaties, etc.).

Na doorklikken op een zoekresultaat heeft men toegang tot (een samenvatting van) het desbetreffende nieuwsartikel, met een woordenwolk waarin de voornaamste entiteiten in het artikel worden getoond. In de nabije toekomst zullen hier entiteiten en opinies te zien

zijn welke in de tekst gemarkeerd worden. Naast peilend.nl omvatten de te verwachten resultaten van het DuOMAN-project verschillende taalkundige bronnen (lexicons, geannoteerde corpora), software en taaltechnologische webservices die beschikbaar zullen zijn voor zowel commercieel als niet-commercieel gebruik. In peilend.nl maakt DuOMAN gebruik van online beschikbare nieuwsbronnen, discussiefora en binnenkort ook blogs. Hiermee kunnen aan zowel media-analisten als het brede publiek zeer gerichte publieke opinies geleverd worden over mensen, producten en onderwerpen.

DuOMAN is een samenwerkingsverband tussen de Universiteit van Amsterdam, Trend-Light Netherlands B.V., GridLine B.V, de Rijksuniversiteit Groningen en de Hogeschool Gent. Het project wordt gefinancierd door het STEVIN-programma. STEVIN is een meerjarig onderzoeks- en stimuleringsprogramma voor Nederlandstalige taal- en spraaktechnologie dat gezamenlijk door de Vlaamse en Nederlandse overheid wordt gefinancierd.



Gedetailleerde blik op een zoekresultaat

## Stijgend bezoek [www.notas.nl](http://www.notas.nl)

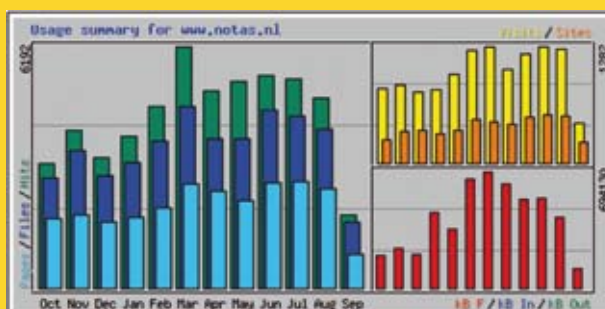
Zowel mens als machine hebben hun weg naar de NOTaS-website weten te vinden, gezien het stijgend aantal bezoekers. Via de vernieuwde website kunnen deelnemers van NOTaS zelf nieuws posten, deelnemerspagina's bijhouden, bestanden uploaden en vacatures plaatsen.

Op de korte termijn vindt klein onderhoud plaats aan de website van NOTaS. Zo zullen bijvoorbeeld de layout en inhoud worden opgefrist. Op de iets langere termijn wordt onderzocht of de deelnemers additionele wensen hebben voor interactieve componenten op de website.

Naast de website van NOTaS is sinds augustus 2009 ook een NOTaS-groep op de sociale netwerksite LinkedIn ([www.linkedin.com](http://www.linkedin.com)) beschikbaar. Via deze

groep kunnen NOTaS-deelnemers, belanghebbenden en andere geïnteresseerden contact houden en ideeën uitwisselen.

U wordt van harte uitgenodigd feedback te geven op de NOTaS-website en deel te nemen aan de NOTaS-groep op LinkedIn.



# TST voor de Rechtspraak, de politie en het NFI

*Net als de rest van de maatschappij, ontkomen ook Justitie en politiediensten niet aan de algemene tendens om sneller en efficiënter met de beschikbare geluidsinformatie om te gaan. Om iets met de geluidsopnamen te kunnen doen, is het noodzakelijk dat de opnamen voorzien worden van beschrijvende, tijdsynchrone data (metadata) die iets vertellen over de opnamen op een bepaald moment. Nog mooier zou het zijn wanneer de spraak in zo'n opname automatisch omgezet kan worden in tekst zodat je een vorm van ondertiteling krijgt.*

Naarmate de TST beter gaat werken, komt dit ideale einddoel beter in zicht. Bovendien is het zo, dat de mogelijkheid om spraak automatisch om te zetten in tekst, nieuwe toepassingen mogelijk maakt om data met spraak in te voeren.

## Politie en Justitie

Het gesproken woord vormt nog steeds een van de belangrijkste bronnen van informatie voor Justitie. Vooral voor de politie en rechtbanken vormt het mondelinge verhoor de kern van het proces van waarheidsvinding. Maar ook op andere niveaus binnen deze organisaties is het gesproken woord een reële bron van informatie en dus is het zinvol te inventariseren waar TST, en dan vooral spraakherkenning, een zinvolle bijdrage kan leveren aan het ontsluiten van deze informatie.

## Rechtbanken

In het STEVIN-demonstratieproject RechtSpraakHerkenning (2008-2009) is onderzocht in hoeverre het herkennen van alle spraak tijdens de rechtszitting een zinvolle bijdrage kan leveren in het sneller en beter verwerken van de rechtszitting. Hoewel de evaluatie van het project op dit moment nog loopt, is duidelijk dat er een aantal significante voordelen zitten aan het gebruik van TST. Tijdens de rechtszitting wordt de spraak van iedere spreker op een apart kanaal opgenomen waardoor precies duidelijk is wie wanneer aan het woord is. Vervolgens wordt alle spraak herkend en omgezet in een tijdsynchrone tekst zodat ook duidelijk is wanneer wat gezegd werd. Omdat per kanaal bekend is wie er sprak, kan voor de bekende sprekers een eigen akoestisch profiel gebruikt worden zodat de herkenning optimaal is. Uiteraard wordt de spraak niet foutloos herkend en is het zeker niet zo dat er geheel automatisch een ondertiteling van de rechtszitting gemaakt kan worden.

Het grote voordeel van het systeem zit in het kunnen zoeken in de opname op spreker,

gesproken woord en deel van de rechtszitting. Gevonden passages kunnen dan met één druk op de knop worden beluisterd. Vooralsnog is het systeem bedoeld om het werk van de griffier te ontlasten.

**Arjan van Hessen**  
Universiteit Twente  
en Telecats



## Politie

Net zoals bij de rechtbank, worden ook bij de politie veel verhoren afgenomen. Anders dan bij de rechtbank, is het verhoor bij de politie meer één-op-één (verdachte en ondervrager) en is het taalgebruik minder formeel. Wanneer het om een incidentele ondervraging gaat waarbij een verdachte, getuige of expert maar één of tweemaal wordt verhoord, is het weinig zinvol om veel tijd te spenderen aan het creëren van een eigen akoestisch model. Anders ligt dat bij mensen die gedurende lange tijd vaak worden ondervraagd. In die gevallen levert de initiële inspanning van het maken van een eigen akoestisch profiel voldoende voordeel op (betere herkenning) om die inspanning ook te rechtvaardigen.

## NFI

De meest minimale vorm van verhoren is die waarbij men zichzelf bevraagt middels een beschrijvend verslag. Dit is wat de experts van het NFI doen wanneer ze sporenonderzoek doen op een plaats delict. Om vervuiling van de sporen te voorkomen, mogen alleen bepaalde vooraf gescreende mensen

naar binnen. Zij doen het sporenonderzoek en nemen met een camera op het hoofd alles op wat ze doen. Soms maken ze schriftelijke aantekeningen en na afloop wordt er een verslag van het geheel gemaakt. Het zou zoveel makkelijker zijn wanneer ze commentaar konden geven tijdens het sporenonderzoek en wanneer deze spraak dan realtime, of na afloop, door de herkenner zou worden gehaald. Ook hier is het resultaat een synchrone tekst en spraak die het mogelijk maakt om in de opnamen te gaan zoeken. Op dit moment zijn er gesprekken met het NFI gaande om te zien of een proof of concept mogelijk is.

### Tapgesprekken

Met enige fantasie zou men het afluisteren van een telefoongesprek óók als een vorm van verhoor kunnen definiëren. Hoewel er door de politie geen vragen gesteld worden, wordt de opgenomen spraak wel gebruikt bij de waarheidsbevinding en uiteindelijk ook gebruikt in het proces. Echter, de spraak van de meeste opgenomen telefoongesprekken is meestal zo slecht, dat automatische herkenning niet mogelijk is. Toch kan spraakherkenning hier zinvol gebruikt worden! Niet de telefoonspraak wordt door de herkenner gehaald, maar de door de politieman of -vrouw nagesproken conversatie. Dit herkennen van nagesproken spraak heet respeaking en wordt al met veel succes door de omroepen gebruikt bij het live ondertitelen van tv-programma's.

Op dit moment wordt een tapgesprek binnen 14 dagen nadat het is opgenomen door de politiebeambte zo letterlijk mogelijk uitgeschreven, iets dat de meeste beambten niet heel erg plezierig vinden. Dit uitschrijven kost ongeveer 6 tot 8 keer meer tijd dan het gesprek zelf. Door het gesprek gewoon na te spreken en deze nagesproken tekst te herkennen, wordt de benodigde tijd voor het uitschrijven sterk bekort. Een bijkomend voordeel is dat er vervolgens ook in de tekst gezocht kan worden omdat de nagesproken spraak en de herkende tekst op millisecondenniveau aan elkaar gekoppeld zijn. Omdat bij het naspreken ook de originele spraak beschikbaar is, krijgen we uiteindelijk drie aan elkaar gekoppelde 'informatielagen': de originele opgenomen spraak, de nagesproken spraak en de herkende/uitgeschreven tekst. De drie lagen zijn op tijdsniveau aan elkaar gekoppeld. Hierdoor kan er gezocht worden in de tekst en kan naar believen een van de twee audiolagen worden beluisterd.

Maar als dit mogelijk is, dan moet het ook mogelijk zijn om dit te doen bij tapgesprekken waarbij de originele taal geen Nederlands

is. Op dit moment wordt er voor niet-Nederlandse tapgesprekken een beëdigde tolk-vertaler gebruikt die de gesprekken moet uitschrijven. Door deze tolk-vertaler de opgenomen spraak direct in het Nederlands te laten naspreken en dan de Nederlandse spraak door de herkenner te halen, krijgen we net als hiervoor een gekoppeld drielaagsmodel. Het enige verschil is dat de brontaal in dit geval geen Nederlands is.

### Kentekenlijn

Omdat spraakherkenning, zeker bij beperkt woordgebruik, ook goed mogelijk is met mobiele toepassingen, werd in 2007 in het kader van het STEVIN-demonstratieproject Kentekenlijn door de politie in Utrecht een experiment uitgevoerd met het herkennen van nummerborden van geparkeerde auto's ([taaluniversum.org/taal/technologie/stevin/projecten/#SNRT](http://taaluniversum.org/taal/technologie/stevin/projecten/#SNRT))

In plaats van het intoetsen of scannen van de nummerborden (waarvoor altijd een relatief zwaar apparaat nodig is) werd onderzocht of het mogelijk is om de nummerborden gewoon via de mobiele telefoon in te spreken en te herkennen. De proef slaagde geheel en momenteel wordt onderzocht of de pilot verder wordt uitgerold.

### Conclusie

De afgelopen jaren werd er op verschillende niveaus door Justitie en politie onderzoek gedaan naar de mogelijkheden om taal- en spraaktechnologie in te zetten in het dagelijks werkproces. Behalve bij de Kentekenlijn waar het eigenlijk ging om het mondeling bevragen van een database, zijn alle applicaties gericht op het doorzoeken en vervolgens afluisteren van opgenomen spraak. We begonnen deze bijdrage bij de ietwat formele spraak in de rechtbank. Vervolgens werden de mogelijkheden onderzocht om dezelfde technologie in te zetten op verschillende niveaus van het politiewerk. De toepassing voor het NFI is hier weer direct aan gekoppeld.

# Automatisch vertalen en de overheid

*Elke overheid heeft nu en dan nood aan vertalingen van documenten. Het kan hierbij gaan om documenten van die overheid die ter beschikking moeten komen van anderstaligen of het kan gaan om documenten van anderstaligen (eventueel andere overheidsinstanties) die vertaald moeten worden naar de taal van de overheid zelf. De vertalers van de Belgische federale overheid maken geen of nauwelijks gebruik van automatische vertaling om eigen documenten in andere talen beschikbaar te maken. Wel wordt er gebruik gemaakt van vertaalgeheugens en tools zoals Trados of EURAMIS.*

## Vertaalgeheugens

Een vertaalgeheugen is een collectie reeds manueel vertaalde fragmenten. Via softwaretools wordt het vertaalgeheugen geïntegreerd in tekstverwerkers en in de werkomgeving van de vertaler, zodat deze snel en gemakkelijk al vertaalde fragmenten kan opsporen en zelf kan beslissen welke hij gebruikt en of er aanpassingen moeten gebeuren. Aangezien in veel documenten bepaalde zinnen of fragmenten vaker terugkomen, is het goed dat deze niet telkens opnieuw manueel vertaald worden, maar door middel van een vertaalgeheugen een consistente vertaling krijgen.

## Vertaalbureaus

Volgens een recente studie van Forbes maken vertaalbureaus wel gebruik van automatische vertaling, maar hier wordt niet bij vermeld hoe. Wordt de automatische vertaling systematisch gebruikt in het vertaalproces in combinatie met post-editing? Wordt automatische vertaling slechts gebruikt om af en toe dingen op te zoeken? Of is de workflow geautomatiseerd zonder dat de feitelijke vertaling geautomatiseerd is?

## NL-Translex

In 1999 zette de Nederlandse Taalunie een project op om het gebruik van automatische vertaling door de overheid van en naar het Nederlands mogelijk te maken, en zocht ze daarvoor een private partner. NL-Translex werd ontwikkeld door Systran, maar lijkt achteraf bekeken niet vaak gebruikt te worden door overheden, onder andere door de hoge kosten voor aanpassing van de woordenboeken en integratie in de workflow. Er zijn wel een aantal grote bedrijven, zoals Fortis AG, die claimen tot 18% productiviteitswinst gehaald te hebben door het gebruik ervan.

## Europa

De Europese overheid is bij ons de enige overheid die echt zelf automatische vertaling gebruikt. De Europese Commissie heeft

jarenlang Systran gebruikt, van midden jaren 70 tot nu, en ontwikkelde zelf haar eigen variant van het systeem: ECMT (European Commission Machine Translation). NL-Translex werd nooit aangepast voor of door de Commissie. In 2006 werd besloten de verdere ontwikkeling en verbetering van ECMT stop te zetten, omdat het niet vaak genoeg gebruikt werd door de vertalers van de Commissie. Woordenboeken kunnen wel nog steeds door de vertalers zelf worden aangevuld.

Ook al gebruiken de vertalers het systeem niet of niet vaak, het blijkt wel regelmatig gebruikt te worden door niet-vertalers om een idee te krijgen waarover een bepaalde anderstalige tekst gaat ("gisting" genoemd). De Commissie heeft het ECMT via een webservice opengesteld voor de overheidsinstanties van de lidstaten en de andere gebruikers van hun websites. Op die manier wordt het systeem dagelijks gebruikt.

Ondertussen voert het Euromatrixplus-project in opdracht van de Commissie experimenten uit met statistische vertaalsystemen, die getraind worden op de gigantische vertaalgeheugens van de Commissie. Aangezien meertaligheid het officieel beleid van de Commissie is, zal automatische vertaling alsmat belangrijker worden, misschien meer nog bij andere diensten dan bij de vertaalafdeling zelf.

## PaCo-MT

Binnen het STEVIN-project PaCo-MT (Parse and Corpus-based Machine Translation), bouwen we verder op het succes van vertaalgeheugens binnen de vertaalsector. We bouwen een machinevertaalsysteem dat gebruik maakt van dezelfde principes als een vertaalgeheugen, maar dat zelf beslist welke vertalingen er gekozen worden en welke aanpassingen er nog moeten gebeuren.

Hierbij gaan we als volgt te werk: we nemen de verslagen van het Europees parlement

**Vincent Vandeghinste**  
**Centrum voor Computerlinguïstiek,**  
**K.U. Leuven**

van 1996 tot 2006 in de verschillende talen waarmee we willen werken, in ons geval Nederlands, Frans en Engels. Deze verslagen worden heel vaak gebruikt in datagedreven automatische vertaalsystemen (zoals statistische vertaalsystemen), omdat ze gratis beschikbaar zijn in alle officiële talen van de lidstaten die op dat moment lid waren van de EU. In deze verslagen worden de zinnen gealigneerd: er wordt aangeduid welke zinnen vertalingen van elkaar zijn. Zinsalignering is eveneens een belangrijke component in een vertaalgeheugen. Daarnaast wordt ook woordalignering uitgevoerd: welke woorden zijn vertalingen van elkaar? Uit deze gegevens kunnen woordenboeken afgeleid worden als we weten hoe vaak welk woord als wat vertaald wordt. Let wel, al deze aligneringen gebeuren machinaal, zonder menselijke interventie, en zijn dus verre van perfect. We hopen deze imperfectie op z'n minst gedeeltelijk te compenseren door de grote hoeveelheden data die door het systeem in rekening gebracht worden.

Op basis van al de aligneringen wordt een eerste vertaalmodel opgesteld, dat het vertaalsysteem toestaat beslissingen te nemen omtrent welke vertaalde fragmenten er gekozen worden en hoe deze deelvertalingen gecombineerd worden tot een nieuwe zin. Als basissysteem gebruiken we de teksten van het Europees Parlement, omdat het hierin gehanteerde taalgebruik breed en gevarieerd is. Een mogelijk probleem hierbij is dat het niet per se om rechtstreekse vertalingen van elkaar gaat. Zowel bron- als doeltaaldocument kunnen vertaald zijn uit een derde taal.

Omdat niet elke overheid hetzelfde taalgebruik hanteert als het Europees Parlement is het belangrijk PaCo-MT aan te passen aan het domein waarvoor het gebruikt zal worden. Dit kan gebeuren door een reeds bestaand vertaalgeheugen in te laden en te bewerken zoals hierboven beschreven. Zo past het systeem zijn taalgebruik automatisch aan het type vertalingen aan dat ingeladen wordt.

Omdat we er niet onmiddellijk van uit gaan dat PaCo-MT altijd de gewenste vertaling oplevert, wordt er toch nog een fase van nabewerking voorzien, waarbinnen de vertaler de door de computer genomen beslissingen kan corrigeren. Deze correcties worden teruggevoerd naar het systeem zodat, na verloop van tijd, PaCo-MT zichzelf aanpast aan de vertaalgewoontes van de vertaler in kwestie.

De laatste paar jaar is er een zeer grote toename van het wetenschappelijke onderzoek naar automatisch vertalen, en er wordt dan ook een aanzienlijke vooruitgang verwacht op kwalitatief gebied. PaCo-MT hoopt daaraan bij te dragen door taalparen met het Nederlands te ontwikkelen. Eenmaal die taalparen kwalitatief goed genoeg zijn zal dat hopelijk leiden tot meer gebruik van automatische vertaling bij de overheid, en dus ook tot een efficiëntere overheid.

