

# GEVAAR VOOR ONZE TALEN EN EEN UITDAGING VOOR TAALTECHNOLOGIE

We zijn getuige van een digitale revolutie die een dramatisch effect heeft op de communicatie- en informatiemaatschappij. Recente ontwikkelingen in de digitale informatie- en communicatietechnologie worden soms vergeleken met de uitvinding van de boekdrukkunst. Wat kan deze analogie ons vertellen over de toekomst van de Europese informatiemaatschappij en onze talen in het bijzonder?

---

We zijn getuige van een digitale revolutie  
vergelijkbaar met de uitvinding van de  
boekdrukkunst.

---

Na de uitvinding van de boekdrukkunst werden ware doorbraken in communicatie- en kennisuitwisseling verwezenlijkt door bijv. de vertaling van de Bijbel in de lokale taal. In de daarop volgende eeuwen werden culturele technieken ontwikkeld om beter om te gaan met taalverwerking en kennisuitwisseling:

- de orthografische en grammaticale standaardisatie van belangrijke talen maakte de snelle verspreiding van nieuwe wetenschappelijke en intellectuele ideeën mogelijk;
- de ontwikkeling van officiële talen stelde burgers in staat om te communiceren binnen bepaalde (vaak politieke) grenzen;
- het onderwijs en de vertaling van talen maakte uitwisseling over talen heen mogelijk;

- de creatie van uitgevers- en bibliografische richtlijnen verzekerde de kwaliteit en beschikbaarheid van gedrukt materiaal;
- de creatie van verschillende media zoals kranten, radio, televisie, boeken, en andere formaten bedienden verschillende communicatienoden.

In de laatste twintig jaar heeft de informatietechnologie eraan bijgedragen veel processen te automatiseren en makkelijker te maken:

- desktop publishing software heeft typen en zetten vervangen;
- Microsoft PowerPoint heeft transparanten voor overheadprojectors vervangen;
- e-mail verstuurt en ontvangt documenten sneller dan een fax-machine;
- Skype biedt goedkope Internet telefoonoproepen aan en verzorgt virtuele ontmoetingen;
- Audio- and videocoderingsformaten maken het makkelijk om multimedia-inhoud uit te wisselen;
- zoekmachines leveren trefwoordgebaseerde toegang tot webpagina's;
- online diensten zoals Google Translate produceren snelle, ruwe vertalingen;
- platforms voor sociale media zoals Facebook, Twitter, and Google+ maken communicatie, samenwerking, en het delen van informatie makkelijker.

Hoewel zulke hulpmiddelen en applicaties nuttig zijn, zijn ze nog niet in staat om een duurzame meertalige Europese maatschappij voor iedereen te ondersteunen met vrij verkeer van informatie en goederen.

## 2.1 TAALGRENZEN STAAN DE EUROPESE INFORMATIEMAATSCHAPPIJ IN DE WEG

We kunnen niet precies voorspellen hoe de toekomstige informatiemaatschappij eruit gaat zien. Maar het is zeer waarschijnlijk dat de revolutie in de communicatietechnologie mensen die verschillende talen spreken op nieuwe manieren bij elkaar zal brengen. Dat legt druk op individuen om nieuwe talen te leren en vooral op ontwikkelaars om nieuwe technologische toepassingen te maken om wederzijds begrip en toegang tot deelbare kennis te verzekeren.

---

Een globale economische en informatieruimte confronteert ons met verschillende talen, sprekers en inhoud.

---

In een globale economische en informatieruimte is er toenemende interactie tussen verschillende talen, sprekers en inhoud dankzij nieuwe mediatypes. De huidige populariteit van sociale media (Wikipedia, Facebook, Twitter, YouTube, and, recentelijk, Google+) is maar het topje van de ijsberg.

We kunnen vandaag de dag in een paar seconden gigabytes tekst rond de wereld sturen voordat we ons realiseren dat de tekst in een taal is die we niet begrijpen. Volgens een recent rapport van de Europese commissie 57% van de Internetgebruikers in Europa goedere en diensten aan in andere talen dan hun moedertaal (Engels is de meest gebruikte vreemde taal, gevolgd door

Frans, Duits en Spaans). 55% van de gebruikers lezen inhoud in een vreemde taal terwijl slechts 35% een andere taal gebruikt om e-mails te schrijven of om commentaren te plaatsen op het Web [2]. Een paar jaar geleden was het Engels waarschijnlijk de lingua franca van het Web – de overgrote meerderheid van inhoud op het Web was in het Engels – maar de situatie is nu drastisch veranderd. De hoeveelheid online inhoud in andere Europese talen (en talen uit Azië en het Midden Oosten) is explosief toegenomen.

Het is verrassend dat deze overal aanwezige digitale tweedeling niet veel publieke aandacht gekregen heeft; maar het doet toch een prangende vraag rijzen: Welke Europese talen zullen gedijen in de genetwerkte informatie- en kennismaatschappij, en welke zijn gedoemd te verdwijnen?

## 2.2 ONZE TALEN IN GEVAAR

Hoewel de drukpers ertoe bijdroeg de uitwisseling van informatie in Europa te vergroten, leidde het ook tot het verdwijnen van veel Europese talen. Regionale en minderheidstalen werden zelden gedrukt en talen zoals het Cornish en Dalmatisch werden beperkt tot mondelinge vormen van overdracht, wat dan weer hun gebruiksbereik beperkte. Zal het Internet hetzelfde schok-effect hebben op onze talen?

---

De grote verscheidenheid aan talen in Europa is een van zijn rijkste en belangrijkste culturele bezittingen.

---

De ongeveer 80 talen van Europa zijn een van zijn rijkste en belangrijkste culturele bezittingen, en een vitaal onderdeel van Europa's unieke sociale model [3]. Hoewel talen zoals Engels en Spaans waarschijnlijk zullen overleven in de opkomende digitale marktplaats, zouden veel Europese talen irrelevant kunnen worden in een

genetwerkte maatschappij. Dit zou Europa's globale status verzwakken, en ingaan tegen het strategische doel om gelijke deelname voor iedere Europese burger te verzekeren ongeacht taal.

Volgens een UNESCO rapport over meertaligheid zijn talen een essentieel medium om fundamentele rechten uit te oefenen zoals politieke expressie, onderwijs en deelname aan de maatschappij [4].

## 2.3 TAALTECHNOLOGIE IS EEN ESSENTIËLE ONDERSTEUNENDE TECHNOLOGIE

In het verleden richtten investeringsinspanningen op het gebied van taalbehoud zich op taalonderwijs en vertaling. Volgens een schatting bedroeg de Europese markt voor vertaling, tolken, softwarelokalisatie en websiteloalisatie 8.4 miljard euro in 2008 en er wordt een groei verwacht van 10% per jaar [5]. En toch dekt dit getal slechts een klein gedeelte af van de huidige en toekomstige noden voor communicatie tussen talen. De meest overtuigende oplossing om het taalgebruik in het Europa van morgen zowel in de breedte als in de diepte te verzekeren is het gebruik van de gepaste technologie, zoals we ook technologie gebruiken om onze transport-, energie- en handicapnoden op te lossen.

Digitale taaltechnologie (die zich richt op alle vormen van geschreven tekst en gesproken uitingen) helpt mensen samen te werken, handel te drijven, kennis te delen en deel te nemen aan sociale en politieke debatten ongeacht taalbarrières en computervaardigheden. De technologie functioneert vaak onzichtbaar in complexe softwaresystemen om ons te helpen:

- informatie te vinden met een zoekmachine op het internet;
- spelling en grammatica te controleren in een tekstverwerker;

- productaanbevelingen in een online winkel te bekijken;
- de verbale instructies te horen van een navigatiesysteem in auto's;
- webpagina's te vertalen via een online dienst.

Taaltechnologie bestaat uit een aantal essentiële toepassingen die processen mogelijk maken in een groter toepassingskader. Het doel van de META-NET taalwittenboeken is om vast te stellen hoe matuur deze kerntechnologieën zijn voor iedere Europese taal.

---

Europa heeft voor alle talen robuuste en betaalbare taaltechnologie nodig.

---

Om onze positie aan de frontlinie van de globale innovatie te behouden heeft Europa taaltechnologie nodig die aangepast is aan alle Europese talen, die robuust en betaalbaar is, en nauw geïntegreerd in belangrijke softwareomgevingen. Zonder taaltechnologie zullen we niet in staat zijn een werkelijk effectieve interactieve multimedia en meertalige gebruikerservaring te bereiken in de nabije toekomst.

## 2.4 MOGELIJKHEDEN VOOR TAALTECHNOLOGIE

Op het gebied van het drukken werd de technologische doorbraak gevormd door het snelle kopiëren van een tekstbeeld (een pagina) met een daartoe uitgeruste drukpers. Mensen moesten het harde werk van het opzoeken, lezen, vertalen en samenvatten van kennis doen. We moesten wachten tot Edison om gesproken taal vast te kunnen leggen – en ook die technologie maakte niet meer dan analoge kopieën.

Digitale taaltechnologie kan nu de processen van vertaling, productie van inhoud en kennismanagement voor alle Europese talen automatiseren. Het kan intuïtieve

taal- of spraakgebaseerde interfaces mogelijk maken voor huishoudelijke elektronica, machineparken, voertuigen, computers en robots. Praktische commerciële en industriële toepassingen zijn nog in de initiële stadia van ontwikkeling, maar de resultaten van onderzoek en ontwikkeling creëren echte toegang tot nieuwe mogelijkheden. Zo is automatisch vertalen al redelijk accuraat in specifieke domeinen, en experimentele toepassingen bieden meertalige informatie- en kennismanagement evenals productie van inhoud in veel Europese talen.

Zoals voor de meeste technologieën geldt, zijn ook de eerste taaltoepassingen zoals stemgebaseerde gebruikersinterfaces en dialoogsystemen ontwikkeld voor zeer gespecialiseerde domeinen, en zij laten vaak beperkte performantie zien. Maar er zijn enorme marktmogelijkheden in de onderwijs- en entertainmentsectoren voor de integratie van taaltechnologieën in 'games', sites voor cultureel erfgoed, 'edutainment' pakketten, bibliotheken, simulatieomgevingen en trainingprogramma's. Mobiele informatiediensten, software voor het computerondersteund leren van talen, eLearning-omgevingen, gereedschappen voor zelfevaluatie en software voor plagiaatdetectie zijn maar enkele van de toepassingsgebieden waar taaltechnologie een belangrijke rol kan spelen. De populariteit van socialemediatoepassingen zoals Twitter en Facebook suggereren additionele noden voor gesofisticeerde taaltechnologieën die het plaatsen van berichten kunnen controleren, discussies kunnen samenvatten, trends in opinievorming kunnen suggereren, emotionele reacties kunnen detecteren, en schendingen van copyright kunnen identificeren of misbruik opsporen.

Taaltechnologie biedt de Europese Unie een enorm potentieel. Het kan ertoe bijdragen de complexe kwestie van meertaligheid in Europa aan te pakken – het feit dat verschillende talen op natuurlijke wijze naast elkaar bestaan in Europese bedrijven, organisaties en scho-

len. Maar burgers moeten kunnen communiceren over deze taalgrenzen heen dwars door de Europese Gemeenschappelijk Markt, en taaltechnologie kan helpen deze laatste barrière te overwinnen en daarmee het vrije en open gebruik van individuele talen ondersteunen.

---

Taaltechnologie draagt ertoe bij de 'handicap' van taaldiversiteit te overwinnen.

---

Als we nog verder in de toekomst kijken zal innovatieve Europese meertalige taaltechnologie een maatstaf bieden voor onze globale partners wanneer zij hun eigen meertalige gemeenschappen hiervan willen voorzien. Taaltechnologie kan gezien worden als een vorm van 'ondersteunende technologie' die de 'handicap' van taaldiversiteit helpt overwinnen en de taalgemeenschappen toegankelijker voor elkaar maakt.

Tot slot is ook het gebruik van taaltechnologie voor reddingsoperaties in rampgebieden waar succesvol functioneren een kwestie van leven of dood kan zijn een actief onderzoeksgebied: Toekomstige intelligente robots met meertalig vermogen hebben het potentieel om levens te redden.

## 2.5 UITDAGINGEN VOOR TAALTECHNOLOGIE

Hoewel taaltechnologie aanzienlijke vooruitgang geboekt heeft in de laatste paar jaar is het huidige tempo van de technologische vooruitgang en productinnovatie te langzaam.

---

Het huidige tempo van de technologische vooruitgang is te langzaam.

---

Veelgebruikte technologieën zoals programma's voor spellings- en grammaticacontrole in tekstverwerkers zijn

typisch eentalig, en zijn alleen beschikbaar voor een handjevol talen. Online diensten voor automatisch vertalen zijn nuttig om snel een redelijke benadering van de inhoud van een document te genereren maar zijn nog hoogst problematisch als het gaat om zeer accurate en volledige vertalingen.

Door de complexiteit van natuurlijke taal is het modeleren van ons taalgebruik in software en het testen ervan in de praktijk een lange en kostbare zaak die duurzame financieringstoezeggingen vereist. Europa moet daarom zijn pioniersrol behouden in het aangaan van de technologische uitdagingen voor een meertalige taalgemeenschap door nieuwe methodes uit te vinden om de ontwikkeling voor het hele gebied te versnellen. Dit zou zowel computationele innovaties als technieken zoals crowdsourcing kunnen omvatten.

## 2.6 TAALVERWERVING BIJ MENSEN EN MACHINES

Om te illustreren hoe computers met taal omgaan en waarom het moeilijk is ze te programmeren om taal te gebruiken bekijken we kort hoe mensen eerste en tweede talen verwerven, en daarna hoe taaltechnologie-systemen werken.

---

De mens maakt zich taalvaardigheden eigen op twee verschillende manieren: door te leren aan de hand van voorbeelden en door taalregels te leren.

---

Mensen verwerven taalvaardigheden op twee verschillende manieren. Baby's verwerven een taal door te luisteren naar de interactie tussen de ouders, broers en zussen en andere familieleden. Vanaf een jaar of twee produceren kinderen hun eerste woorden en korte woordcombinaties. Dit is alleen mogelijk omdat mensen een genetisch bepaalde aanleg hebben om te imiteren en daarna te rationaliseren wat ze horen.

Een tweede taal leren op latere leeftijd vereist meer inspanning, vooral omdat het kind niet ondergedompeld is in een taalgemeenschap van moedertaalsprekers. Op school worden vreemde talen meestal verworven door grammaticale structuur, vocabularium en spelling te leren door drill oefeningen die taalkundige kennis beschrijven in termen van abstracte regels, tabellen en voorbeelden. Een vreemde taal leren wordt moeilijker naarmate men ouder is.

De twee hoofdtypen van taaltechnologische systemen 'verwerven' taalvaardigheden op een vergelijkbare manier. Statistische (of 'datagedreven') benaderingen verkrijgen taalkundige kennis uit gigantische collecties van concrete voorbeeldteksten. Hoewel het volstaat om tekst van een enkele taal te gebruiken om bijv. een spellingchecker te ontwikkelen, moeten parallelle teksten in twee (of meer) talen beschikbaar zijn om een automatisch vertaalsysteem te ontwikkelen. Een 'machine-learning' algoritme 'leert' dan patronen voor de vertaling van woorden, korte frases en volledige zinnen.

Deze statistische benadering kan miljoenen zinnen vereisen en de kwaliteit van de technologie neemt toe naarmate er meer tekst geanalyseerd wordt. Dit is een van de redenen waarom leveranciers van zoekmachinediensten zo graag zoveel mogelijk geschreven materiaal verzamelen. Spellingscorrectie in tekstverwerkers, en diensten zoals Google Search en Google Translate zijn allemaal gebaseerd op statistische benaderingen. Het grote voordeel van statistiek is dat de machine snel leert in continue series van trainingscycli hoewel de kwaliteit enorm kan verschillen.

De tweede benadering van taaltechnologie en automatisch vertalen in het bijzonder bestaat uit het bouwen van regelgebaseerde systemen. Experts op het gebied van taalkunde, computationele taalkunde en informatica moeten eerst grammaticale analyses (vertaalregels) inbrengen en vocabulariumlijsten (lexicons) samenstellen. Dit is zeer tijds- en arbeidsintensief. Enkele van

de regelgebaseerde automatische vertaalsystemen zijn al meer dan twintig jaar onder constante ontwikkeling. Het grote voordeel van regelgebaseerde systemen zit 'm in de gedetailleerde controle die experts hebben over de taalverwerking. Dat maakt het mogelijk om systematisch fouten in de software te corrigeren en gedetailleerde feedback te geven aan de gebruiker, vooral wanneer regelgebaseerde systemen gebruikt worden voor het leren van taal. Maar door de hoge kosten van dit werk is regelgebaseerde technologie tot nu toe alleen ontwikkeld voor de belangrijkste talen.

---

De twee hoofdtypen van taaltechnologische systemen 'verwerven' taalvaardigheden op een vergelijkbare manier.

---

Aangezien de sterktes en zwaktes van statistische en regelgebaseerde systemen complementair neigen te zijn,

richt het huidige onderzoek zich op hybride benaderingen die de twee methodologieën combineert. Tot nu toe zijn die benaderingen echter minder succesvol geweest in industriële toepassingen dan in het onderzoekslaboratorium.

Zoals we gezien hebben in dit hoofdstuk maken veel wijdverbreide toepassingen in de moderne informatiemaatschappij intensief gebruik van taaltechnologie. Vanwege de meertaligheid van de gemeenschap geldt dat in het bijzonder voor de Europese economische en informatieruimte. Hoewel taaltechnologie enorme vooruitgang geboekt heeft in de laatste paar jaar, ligt er nog een enorm potentieel om de kwaliteit van taaltechnologiesystemen te verbeteren. In de volgende secties zullen we de rol van het Nederlands in de Europese informatiemaatschappij beschrijven en de huidige toestand van taaltechnologie voor het Nederlands evalueren.

The Dutch Language in the Digital Age

Rehm, G.; Uszkoreit, H. (Eds.)

2012, VI, 79 p. 24 illus. in color., Softcover

ISBN: 978-3-642-25977-7