TAALDATA ALS SLEUTEL TOT DE ONTWIKKELING VAN TAALTECHNOLOGIE

TOM VANALLEMEERSCH
CROSSLANG (POWERLING-GROEP)

LDS-WORKSHOP VOOR BELGIË EN NEDERLAND
BREDA, 23 JANUARI 2025

CrossLang: profiel

Meertalige, op AI gebaseerde software en diensten

Bedrijfsklanten

Door de EC gesubsidieerde projecten

Opgericht in 2002

Deel van Franse groep sinds 2024

powerling

Verlener van taaldiensten (top 75 wereldwijd)

DOCUMENTATIEPROCESSEN

Domeinen

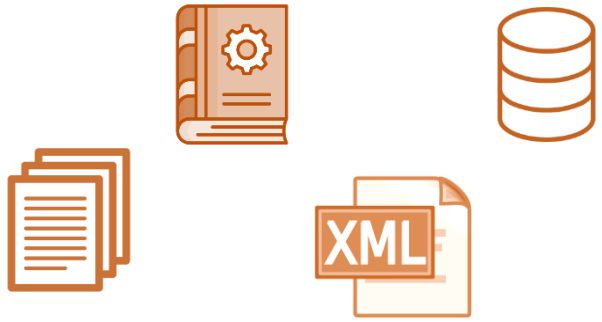
- Autoproducenten: onderhouds- en gebruikshandleidingen
- E-commerce: beschrijvingen produkten
- Financiële dienstverlening: documenten rond regelgeving
- Juridisch domein: wetten, verordeningen
- Publieke administraties: beleidsdocumenten
- Academische wereld: wetenschappelijke artikels, blogs, …
- …

Functies

- Documenten voor intern gebruik
- Publicatiewaardige documenten
- Samenvattingen
- Documentatie voor meertalig publiek
- Documenten voor laaggeletterd publiek
- …

Formaten

- Documenten in bewerkbare formaten (vrije tekst)
- Ingescande documenten
- Afbeeldingen met tekst
- Databanken / bestanden met gestructureerde tekst
- Geluidsopnames
- …



ONDERSTEUNING MET TAALTECHNOLOGIE

(Semi-)automatische systemen

- Vertalen
- Corrigeren
- Samenvatten
- Bevragen (vraag-antwoordsystemen)
- Doorzoekbaar maken data
  - Ook geluidsopnames, gescande documenten

Verlagen marktintroductietijd

Reduceren van kosten

Verhogen consistentie
(bv. terminologie)

Data voor trainen systemen

- Publiek beschikbare data
  - ➢ Teksten in allerlei domeinen, zinparen (zinnen + vertaling)
- Interne data: organisatiespecifiek
  - ➢ Documenten, vertaalgeheugens (databanken met professionele vertalingen), glossaria
- Domeinspecifieke data vanuit andere organisaties
  - ➢ Voorwaarden: gebruik, financiële compensatie

<u>Voorbeelden:</u>

OPUS-website (documenten met vertaling)
CommonCrawl

MICE-project (zie verder)

Translations and Open Science
Language Data Space
LLMs4EU (begin: 2025)

SCENARIO: AUTOMATISCH VERTALEN

Soorten vertaalsystemen

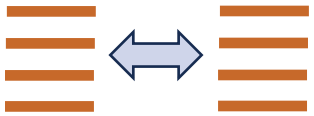
- Voor breed gebruik:
  - Commercieel: Google Translate, DeepL, …(gratis of betalend, zeer krachtig)
  - Niet-commercieel: eTranslation (Europese Commissie, focus op officiële EU-talen)
  - Trainingdata: niet vermeld

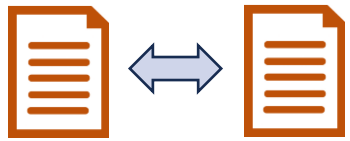
- Op maat (niet publiek toegankelijk):
  - Trainingdata: publiek beschikbare + organisatie-/domeinspecifieke teksten
  - Hogere performantie voor specifieke domeinen
  - Nood aan taaltechnologische expertise
  - Garantie op confidentiële behandeling van te vertalen documenten

Bouwen vertaalsysteem

Verzamelen data
(zinparen, termen)

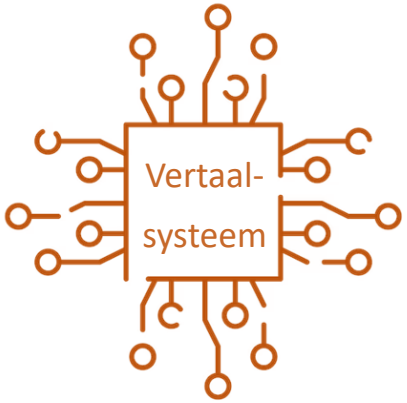
Publiek beschikbare data,
vertaalgeheugens, glossaria





Linken van zinnen tussen documenten
en (kwaliteitsvolle) webpagina's

“Voeden” van data aan
zelflerende software
(*deep learning*)



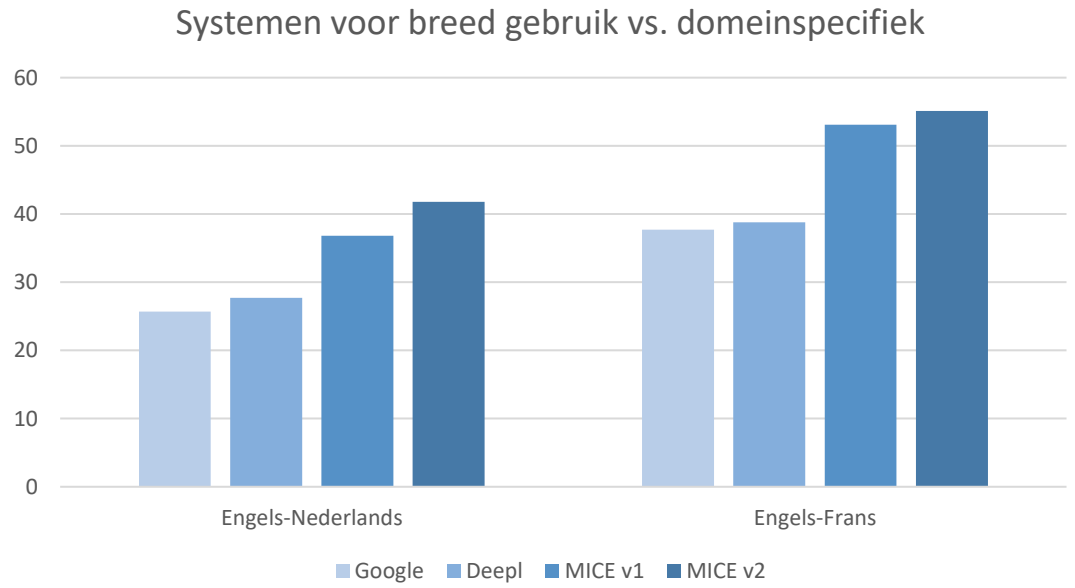
Casus: publieke instelling

Samenwerking CrossLang met NBN (Belgisch Bureau voor Normalisatie) en sectorfederaties in MICE-project

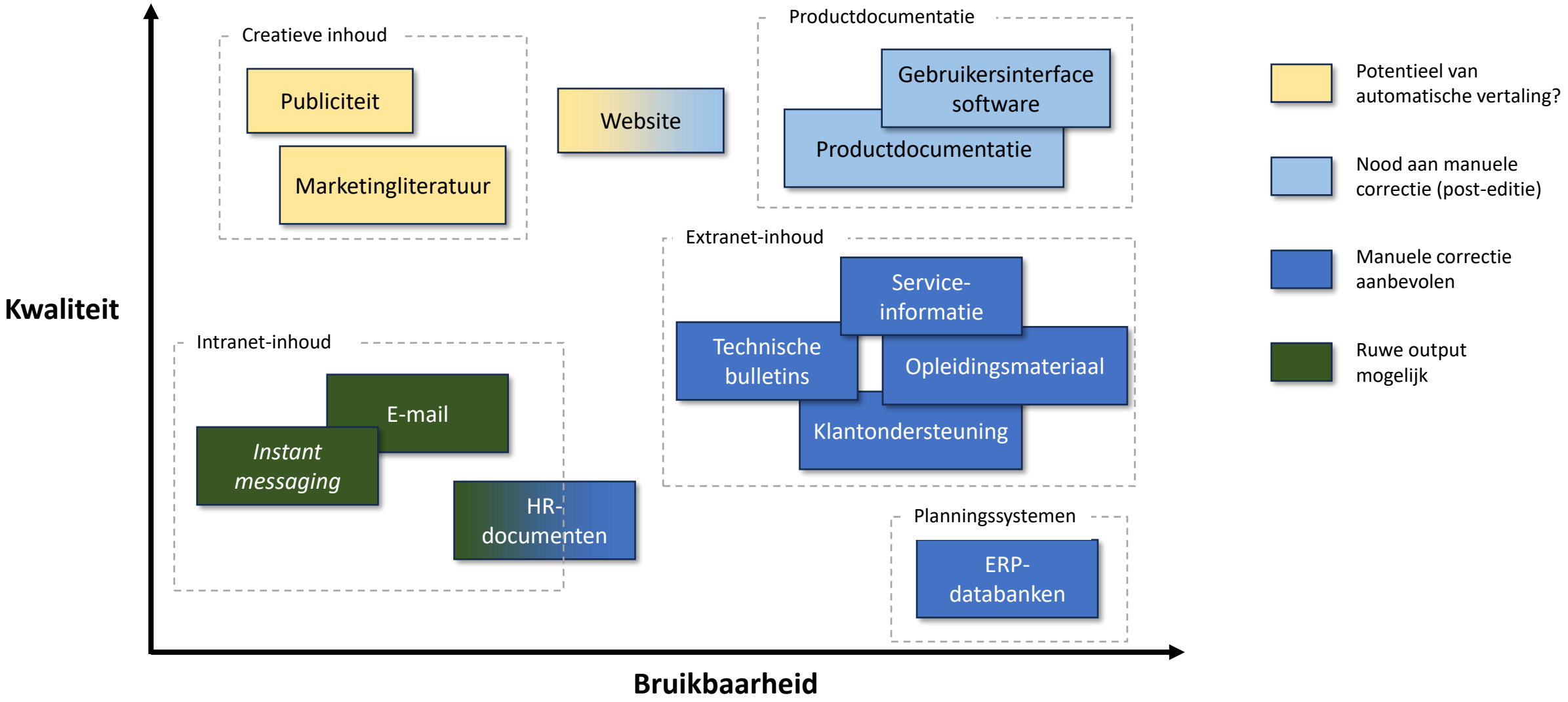
- Vertaling van normatieve documenten (technische teksten over bv. produkttype)
- Doel: verbetering concurrentiepositie Nederlandstalige KMO's
- Trainingdata voor systeem: publieke/interne documenten, glossaria
- Post-editie van output door domeinexperts

hogere testscore =

meer overeenkomst tussen
automatische en gewenste vertaling



Casus: automobielsector

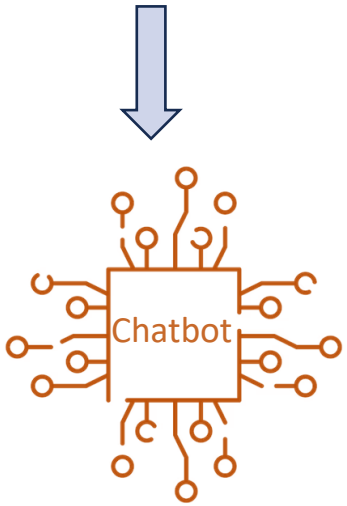


SCENARIO: AUTOMATISCH CORRIGEREN

Casus: corrigeren vertaling met glossarium

Opstellen instructie (“prompt”) voor corrigeren (automatisch) vertaalde zin

De Engelse zin *The Court must be notified of any change to those details (date and time of removal, address etc.).*
werd vertaald naar het Frans als *Toute modification de ces données (date et heure du déménagement, adresse, etc.) doit être notifiée à la …*
Corrigeer de vertaling aan de hand van het volgende glossarium:
…
removal → renvoi
…



Toute modification de ces informations (date et heure du renvoi, adresse, etc.) doit être notifiée à la Cour dans les plus brefs délais.

Zie: Kostikova et al. (2023). *Term Translation: Convert or Converse?* Proceedings of Translating and the Computer 45.

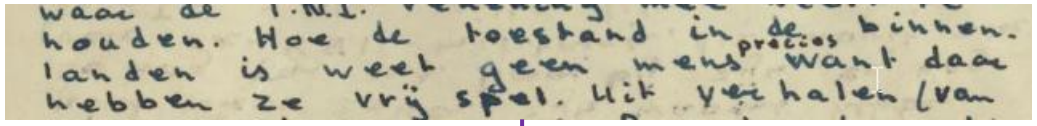
SCENARIO: DOORZOEKBAAR MAKEN

Casus: cultureel erfgoed

AI4Culture-project: software, data en handleidingen voor toepassen AI door experts in het veld

➢ Ondersteuning voor een betere doorzoekbaarheid van tekst in gescande documenten

https://ai4culture.eu



Optische karakterherkenning

Hoe de toestand **d. binnen. landen** is weet geen mens **imprecios** want daar hebben ze vrij spel.

Correctie met chatbot

Hoe de toestand in **de binnenlanden** is weet geen mens **precies** want daar hebben ze vrij spel.

Automatische vertaling

How the situation **d. inside. countries**, no one knows **imprecios** because they have free rein there.

No one knows **exactly** what the situation is like **in the interior**, because they have free rein there.

SAMENVATTEND

SLEUTELFACTOREN IN TAALTECHNOLOGIE

- Gebruik van taaldata
- Relevantie van data voor een specifieke omgeving en toepassing
- Hoge kwaliteit van data

- Menselijke interventie:
  - Selecteren van data
  - Opzetten procedures voor ontwikkelen systemen (verwerken data, instructies voor chatbots, …)
  - Evalueren van systeemoutput

Vragen?

tom.vanallemeersch@crosslang.com