European Language Resource Coordination
Connecting Europe Facility

European Commission

TAALTECHNOLOGIE IN/VOOR DE OPENBARE SECTOR

Tom Vanallemeersch, CrossLang
8 Juli 2021

OVERZICHT

- Belang van taaltechnologie voor de openbare sector in de EU
- Voorbeelden van projecten
- Uitgelicht project: MICE (vertaaltechnologie)
- Interface voor computerondersteund vertalen

TAALTECHNOLOGIE EN DE OPENBARE SECTOR

Publieke administraties op verschillende niveaus:

- EU: Directoraten-Generaal van EC, Digital Service Infrastructures (DSIs) van EC, Europees Parlement, …
- Lidstaten: ministeries, nationale banken, culturele instanties, …
- Regionale overheden
- Gemeentelijke overheden

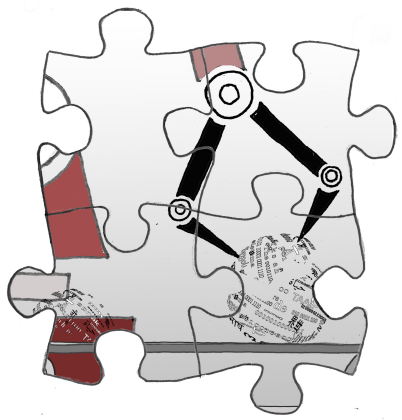
Domeinen: financieel, cultureel, juridisch, …

TAALTECHNOLOGIE EN DE OPENBARE SECTOR

Voordelen van taaltechnologie / kunstmatige intelligentie:

- Digitaal transformeren van publieke diensten
- (Kosten)efficiënter maken van processen
- Beleid uitstippelen op basis van informatieanalyse
- Verbeteren diensten voor burgers

TAALTECHNOLOGIE EN DE OPENBARE SECTOR



Rol van administraties in gebruik/ontwikkeling toepassingen:

- Aankopen bestaande toepassingen (of diensten)
- Intern verder ontwikkelen van open source software
  → Eventueel met ondersteuning door externe partners
- Subsidiëren projecten uitgevoerd door externe partners
  → Onderzoek / ontwikkeling prototype / integratie
- Samenbrengen van betrokken partijen (onderzoek, industrie, …)
  → Bv. standaardisering, catalogisering
- Verzamelen en ter beschikking stellen van data

VOORBEELDPROJECTEN (CROSSLANG)

Vertaaltechnologie:

- Vertaaldepartement van de Kanselarij
  → Bouwen van systeem voor automatische vertaling (MT) voor juridische en beleidsteksten
- NBN (Bureau voor Normalisatie), RIK (Estse registers)
  → Opzetten vertaalworkflow met MT en vertaalgeheugens (TM)

MICE
Middleware for
Customer eTranslation

VOORBEELDPROJECTEN (CROSSLANG)

CEFAT4Cities

Occam
OCR, Classification & Machine Translation

Informatie-extractie / -conversie:

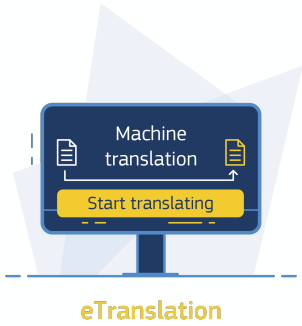
- Stedelijke overheden in Europa
  → Omzetten procedures naar meer gestructureerde vorm, toepassen MT
- DG FISMA (EU-beleid voor banken en financiën)
  → Bouwen van tool voor extractie termen, definities, rapporteringsverplichtingen
- DSIs BRIS (handelsregisters) en Europeana (cultureel erfgoed)
  →Herkennen van tekst in gescande documenten, toepassen MT en TM

MICE

Verbeteren meertalige workflows in specifieke domeinen:

- Betere resultaten produceren dan “generieke”MT
- Terminologische consistentie vergroten
- Leren uit correcties van automatische suggesties door experten

MICE



Kadert binnen Connecting Europe Facility:

- Verbinding nationale infrastructuren met centraal platform eTranslation-systeem van de EC
- Ook ontwikkeling vertaalsystemen door consortiumpartners

MICE: GEBRUIKERSCASUS NBN

NBN
BUREAU VOOR NORMALISATIE
BUREAU DE NORMALISATION

- Ontwikkeling, publicatie en verkoop van normen in België
  → Norm = overeenkomst over product, dienst, proces of methode
- Samenwerking met sectorfederaties (bv. bouw)

Nood aan vertaling:

- Engelse normen naar het Frans en (vooral) het Nederlands
- Engelse technische teksten naar het Frans en Nederlands
- Franse technische teksten naar het Nederlands en vice versa

MICE: GEBRUIKERSCASUS NBN

Opzetten omgeving voor NBN:

NBN, federaties
- Verzamelen vertaalde normen, technische teksten, glossaria, *named entities* (persoonsnamen, locaties, …)

CrossLang
- Trainen MT-systeem op basis van algemene en domeinspecifieke vertalingen
- Creëren domeinspecifiek TM (herkent bestaande / gelijkende zinnen tijdens het vertalen)

Federaties
- Vertalen documenten in interface met vertaalsuggesties

MICE: VERZAMELEN DOMEINSPECIFIEKE DATA

Opzetten omgeving voor NBN:

NBN, federaties

- Verzamelen vertaalde normen, technische teksten, glossaria, *named entities* (persoonsnamen, locaties, …)

CrossLang

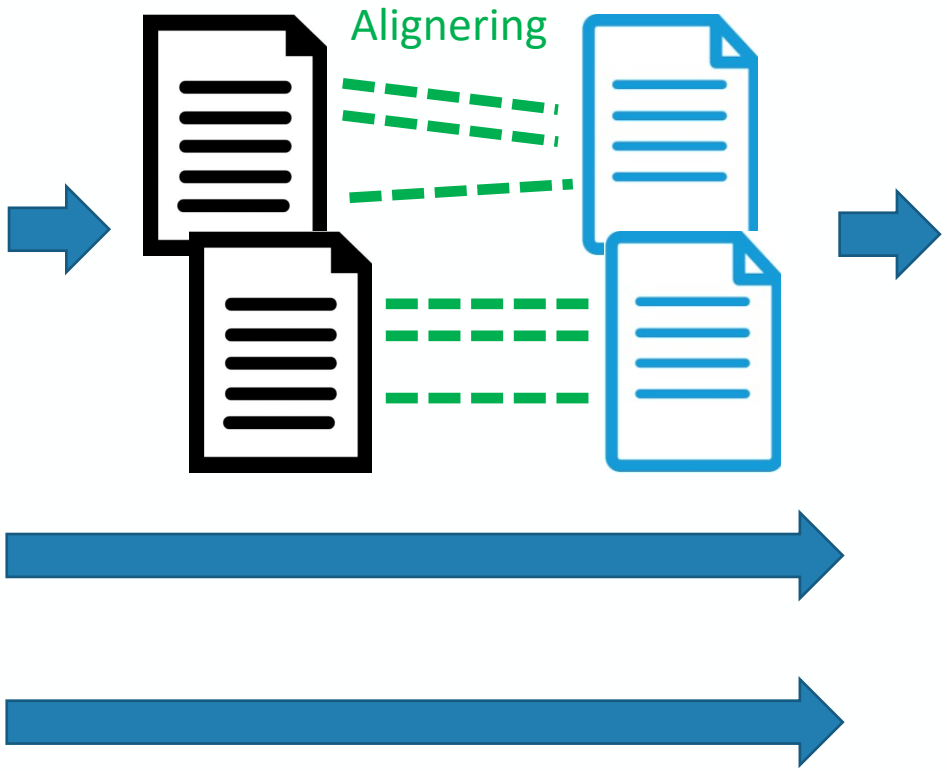
- Trainen MT-systeem op basis van algemene en domeinspecifieke vertalingen
- Creëren domeinspecifiek TM (herkent bestaande / gelijkende zinnen tijdens het vertalen)

Federaties

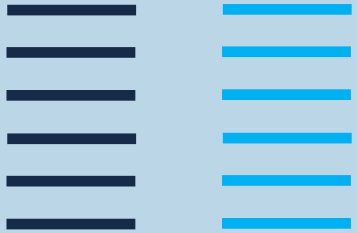
- Vertalen documenten in interface met vertaalsuggesties

MICE: VERZAMELEN DOMEINSPECIFIEKE DATA

Documenten en hun vertaling



Zinnen en hun vertaling (zinparen)



Termen and hun vertaling

Named entities and hun vertaling

MICE: TRAINEN MT-SYSTEEM

Opzetten omgeving voor NBN:

NBN, federaties

- Verzamelen vertaalde normen, technische teksten, glossaria, *named entities* (persoonsnamen, locaties, …)

CrossLang

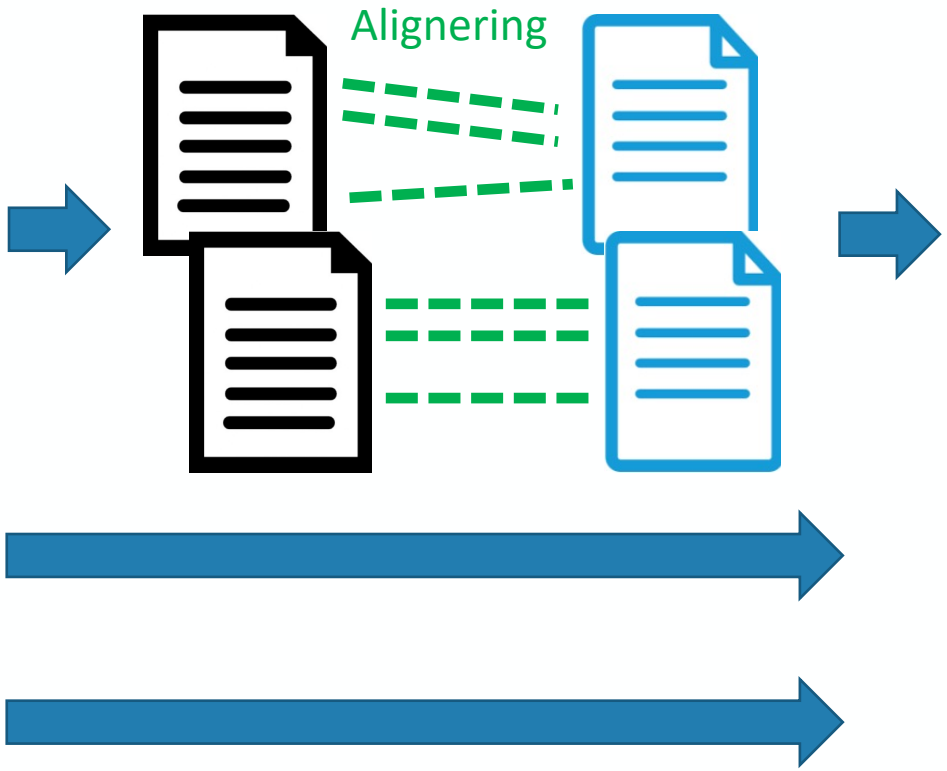
- Trainen MT-systeem op basis van algemene en domeinspecifieke vertalingen
- Creëren domeinspecifiek TM (herkent bestaande / gelijkende zinnen tijdens het vertalen)

Federaties

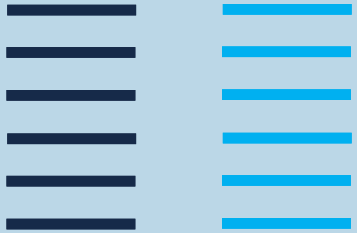
- Vertalen documenten in interface met vertaalsuggesties

MICE: TRAINEN MT-SYSTEEM

Documenten en hun vertaling



Zinnen en hun vertaling (zinparen)



Termen and hun vertaling

Named entities and hun vertaling

+ Grote hoeveelheid “generieke”, publiek beschikbare zinparen (bv. van website *Opus*)

MT-systeem

MICE: CREËREN TM

Opzetten omgeving voor NBN:

NBN, federaties

- Verzamelen vertaalde normen, technische teksten, glossaria, *named entities* (persoonsnamen, locaties, …)

CrossLang

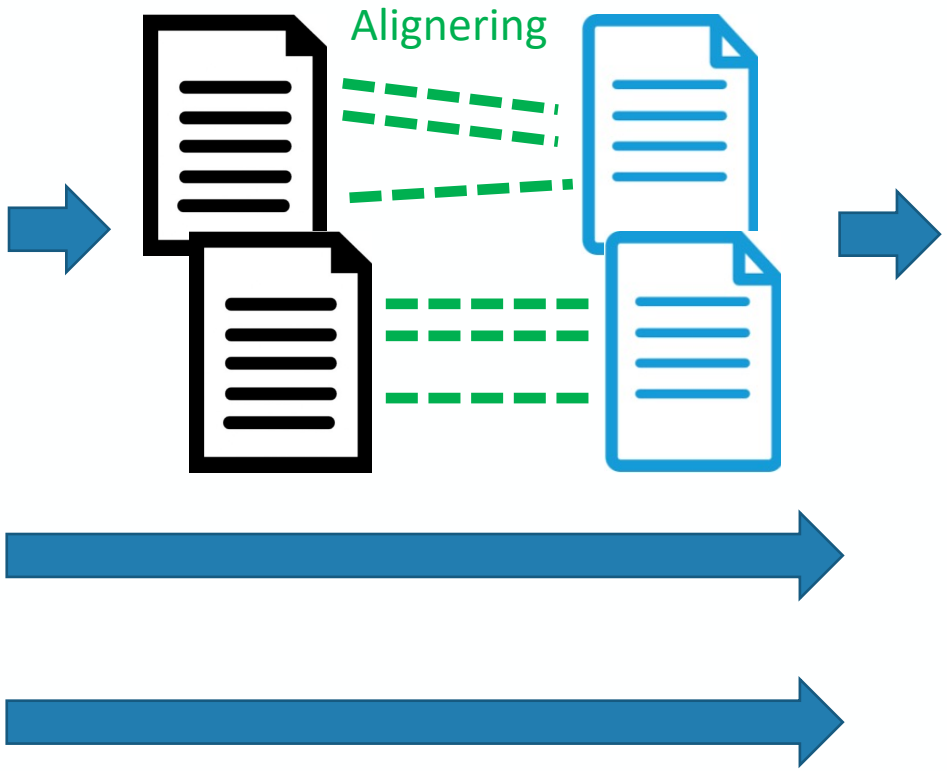
- Trainen MT-systeem op basis van algemene en domeinspecifieke vertalingen
- Creëren domeinspecifiek TM (herkent bestaande / gelijkende zinnen tijdens het vertalen)

Federaties

- Vertalen documenten in interface met vertaalsuggesties

MICE: CREËREN TM

Documenten en hun vertaling



Zinnen en hun vertaling (zinparen)

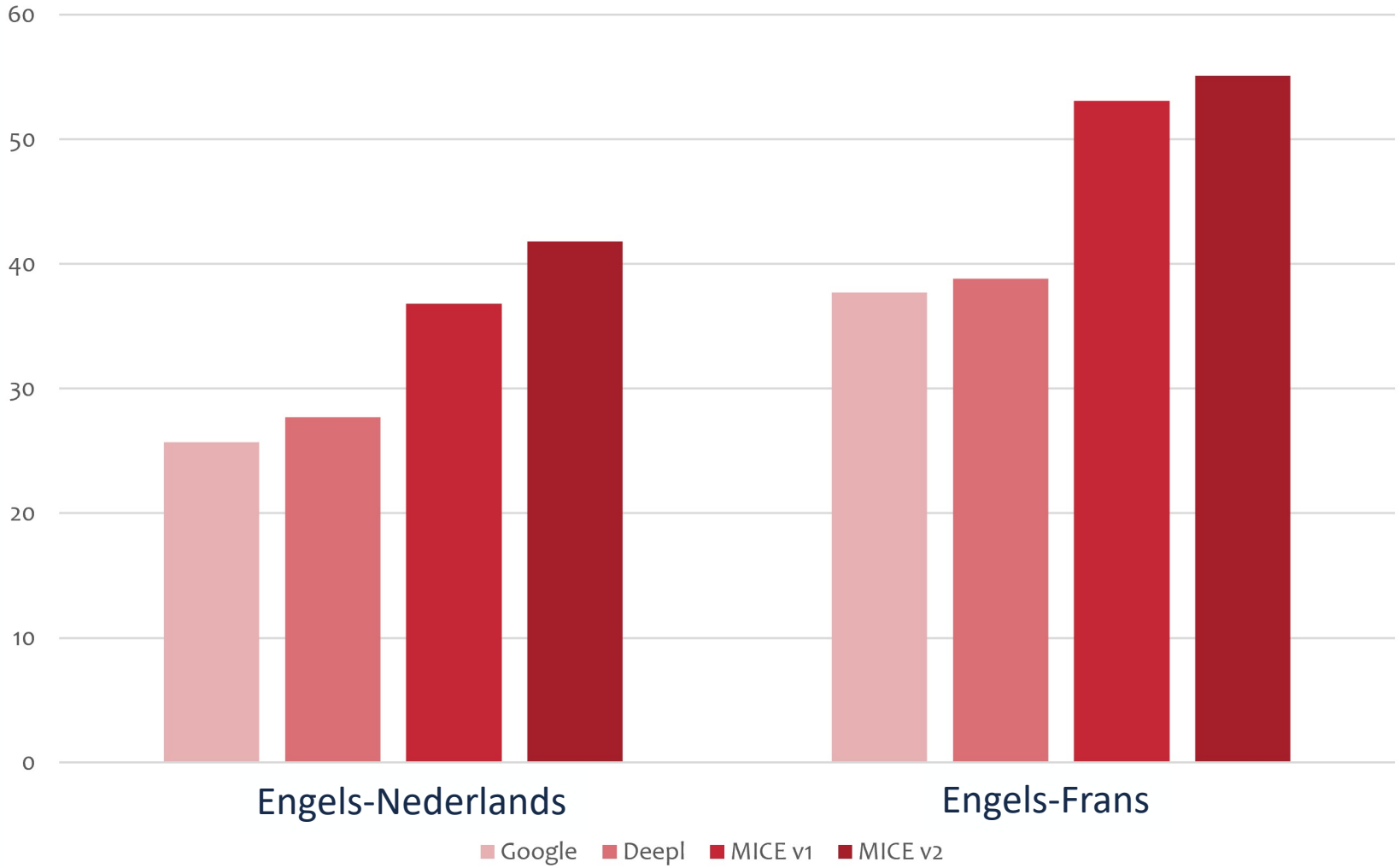
Termen and hun vertaling

Named entities and hun vertaling

Databank (TM)

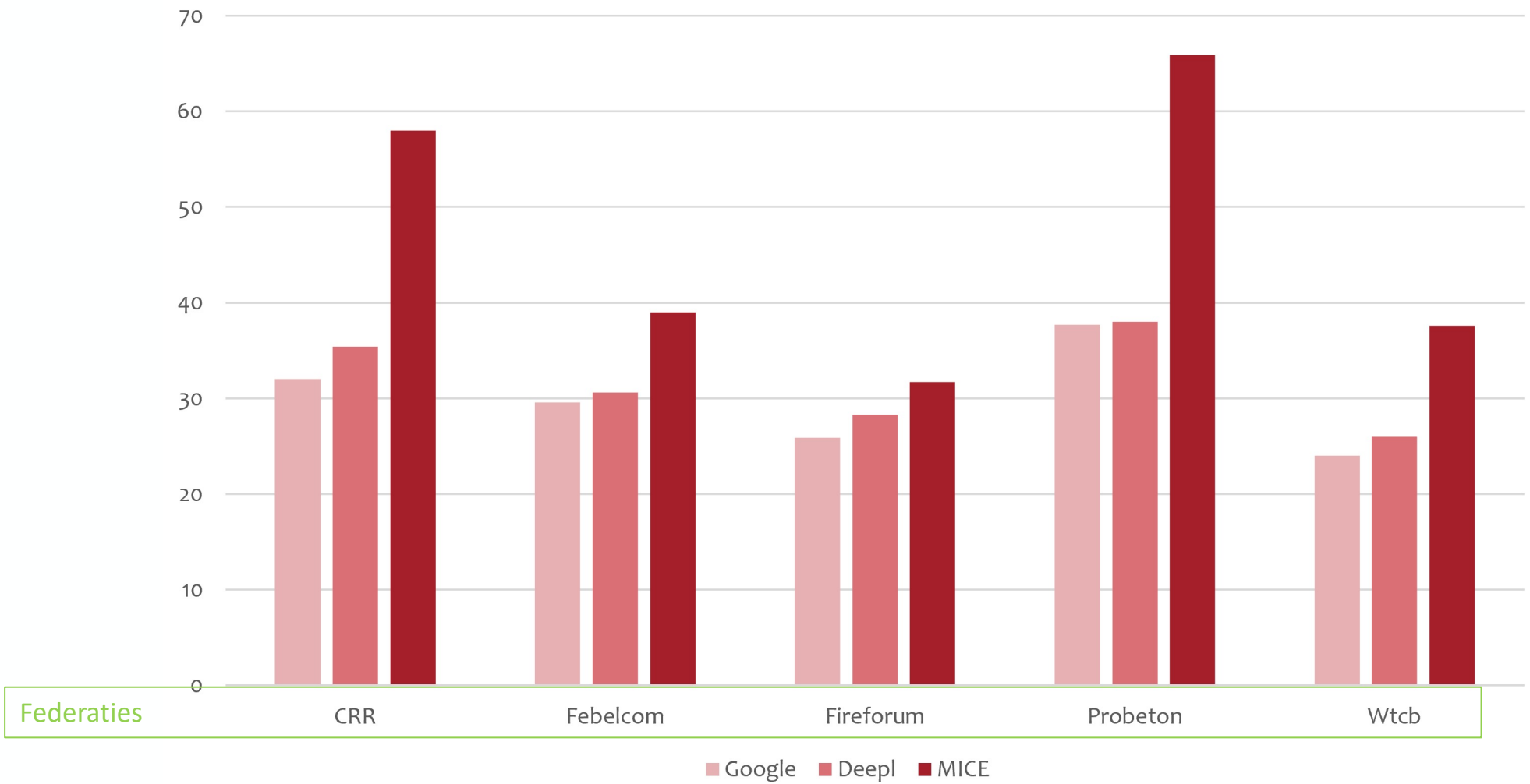
MICE: INSCHATTING KWALITEIT MT

BLEU-scores (vergelijking MT-output en bestaande vertaling, maximumscore is 100)

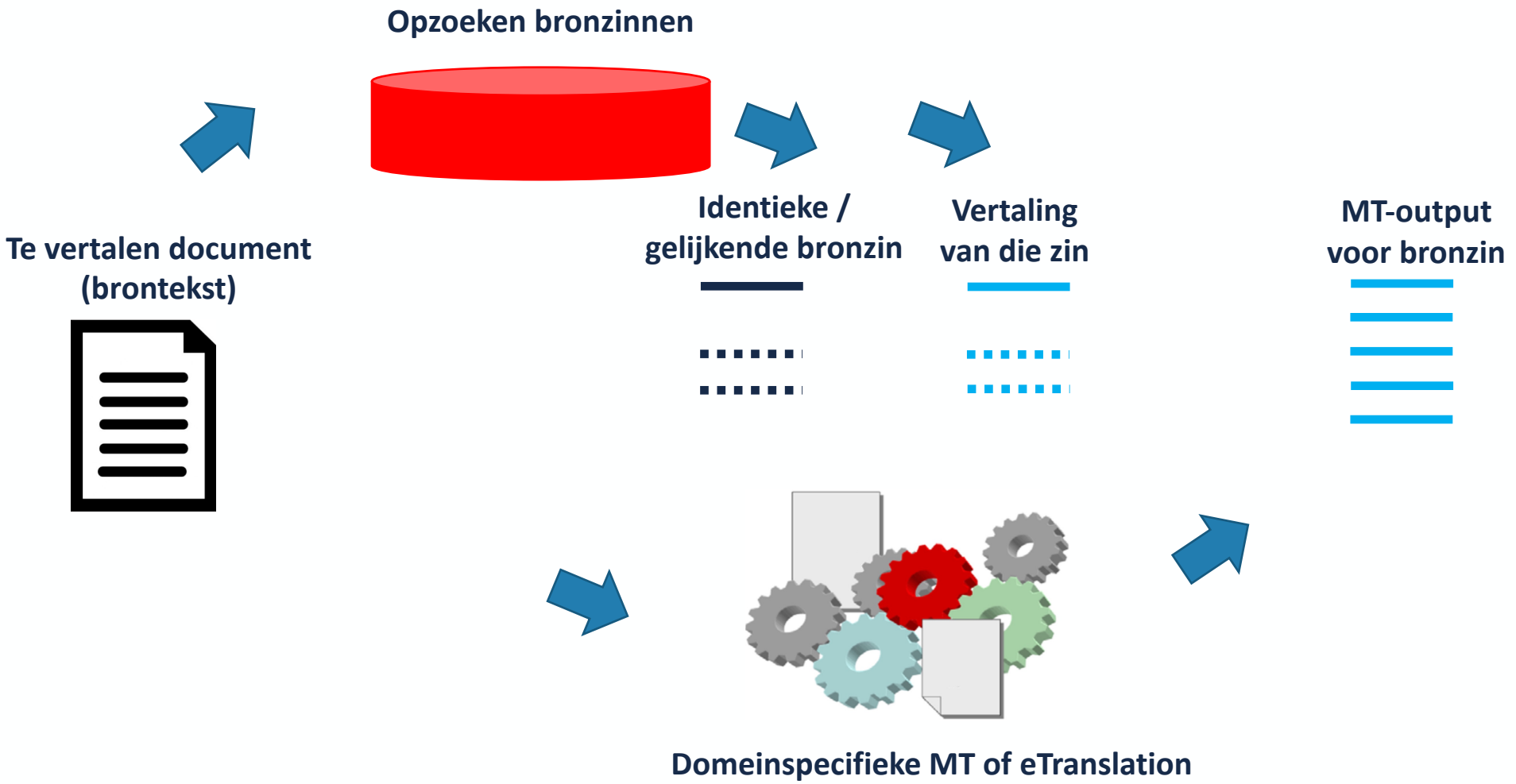


MICE: INSCHATTING KWALITEIT MT

Vertaling Nederlands naar Frans



MICE: VERTALEN DOCUMENT MET MT



Bestaande vertalingen in domein

Vertaalgeheugen

Te vertalen norm

Voorlopige / finale vertaalde norm

Subject Matter Experts

opladen

opzoeken, toevoegen

Tekst en layout scheiden / samennemen

corrigeren

MICE-software

Gebruikers-interface

opzoeken, toevoegen

Systeem voor automatische vertaling

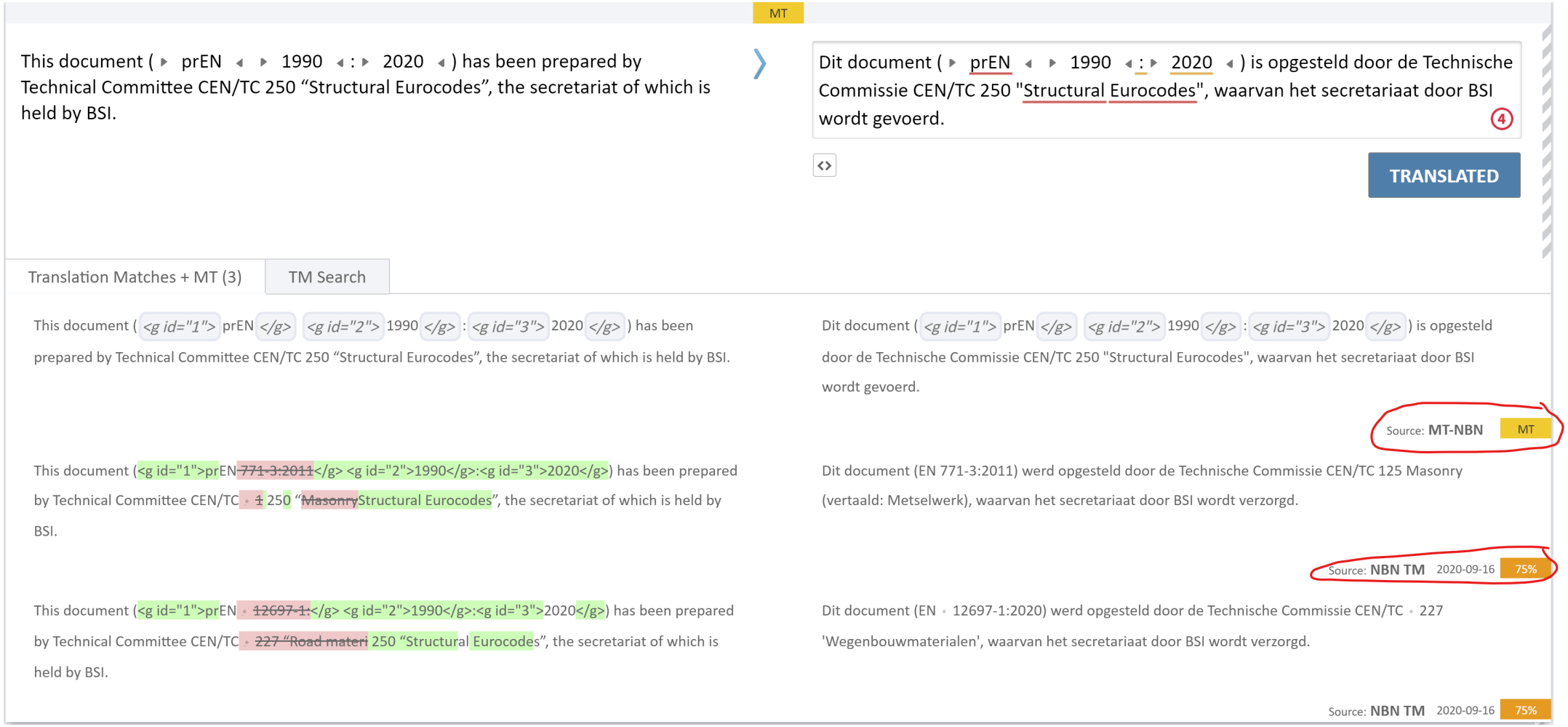
Generieke zinparen:

Engels naar Nederlands
Engels naar Frans
Nederlands naar Frans
Frans naar Nederlands

MICE: VERTAALINTERFACE

- Aangepaste versie van open source software MateCat
- Mogelijkheid tot integreren specifiek systeem voor vertaalgeheugens (bv. NEC TM, ontwikkeld in EC-project)
- Aankoppeling MT-systeem
- Functionaliteit voor termherkenning (glossaria)
- Focus op expert reviewer (niet op professionele vertaler)

MICE: DEMO VERTAALINTERFACE



MICE

Graduele vergroting vertaalefficiëntie:

- Continu aanvullen TM door reviewer
  - →Gecorrigeerde MT-vertalingen
  - →Aangepaste TM-vertalingen
  - →Van nul vertaalde zinnen

  Gevolg: kans op vinden identieke / gelijkende zinnen vergroot
- Regelmatig updaten van trainingsmateriaal MT-systeem op basis van die aanvullingen

  Gevolg: hogere MT-kwaliteit, minder correcties nodig door reviewer

KLASSIEKE AANPAK NBN VS MICE

Klassiek (beroep op vertaalbureau)	MICE
Jaarlijks budget voor vertaling laat verwerking van klein aantal nomen toe	Zelfde budget, besteed aan computerinfrastructuur en inspanning domeinexperten (review) laat verwerking van meer normen toe
Afwerking vertaling norm duurt lang	Vertaling is veel sneller beschikbaar, eventueel in voorlopige versie; TM laat toe de vertaalde zinnen uit die versie eenvoudig te hergebruiken voor finale versie
Controle over vertaalproces in handen van vertaalbureau	Controle over vertaalproces bij eigenaar documenten
Bijkomende expertreview vertalingen nodig	Expertreview is deel van proces; voor publicatiedoeleinden dient er wel nog een taalkundige review plaats te vinden
Nederlandstalige KMO's hebben structurele achterstand op vlak van kennis van normen t.o.v. moedertaalsprekers Engels en (ten dele) Frans	Concurrentiepositie Nederlandstalige KMO's verbetert

European Language Resource Coordination
Connecting Europe Facility

European Commission

DANK VOOR UW AANDACHT !

Contact: tom.vanallemeersch@crosslang.com