European Language Resource Coordination
Connecting Europe Facility

ELRC Workshop in Nederland

ELRC in Nederland

Jan Odijk (Universiteit Utrecht)
Carole Tiberius (Instituut voor de Nederlandse Taal)

ELRC in Nederland: Wie is wie

Jan Odijk

Carole Tiberius

- Technical NAP
- Public Services NAP : vacature

- Hugo Keizer, Vertegenwoordiger DGT

European Language Resource Coordination
Connecting Europe Facility

Wat is ELRC?

- Het doel van ELRC is om in alle Europese landen taaldata te verzamelen om het eTranslation systeem verder mee te ontwikkelen en verbeteren.

- Typen (ééntalige en meertalige) digitale teksten:
  - interne verslagen en andere documenten
  - publicaties en andere documenten voor extern gebruik
  - websites en brochures
  - lexicons en glossaria
  - vertaalgeheugens

- 1ste workshop op 19 april 2016

Resultaten: geïdentificeerde data

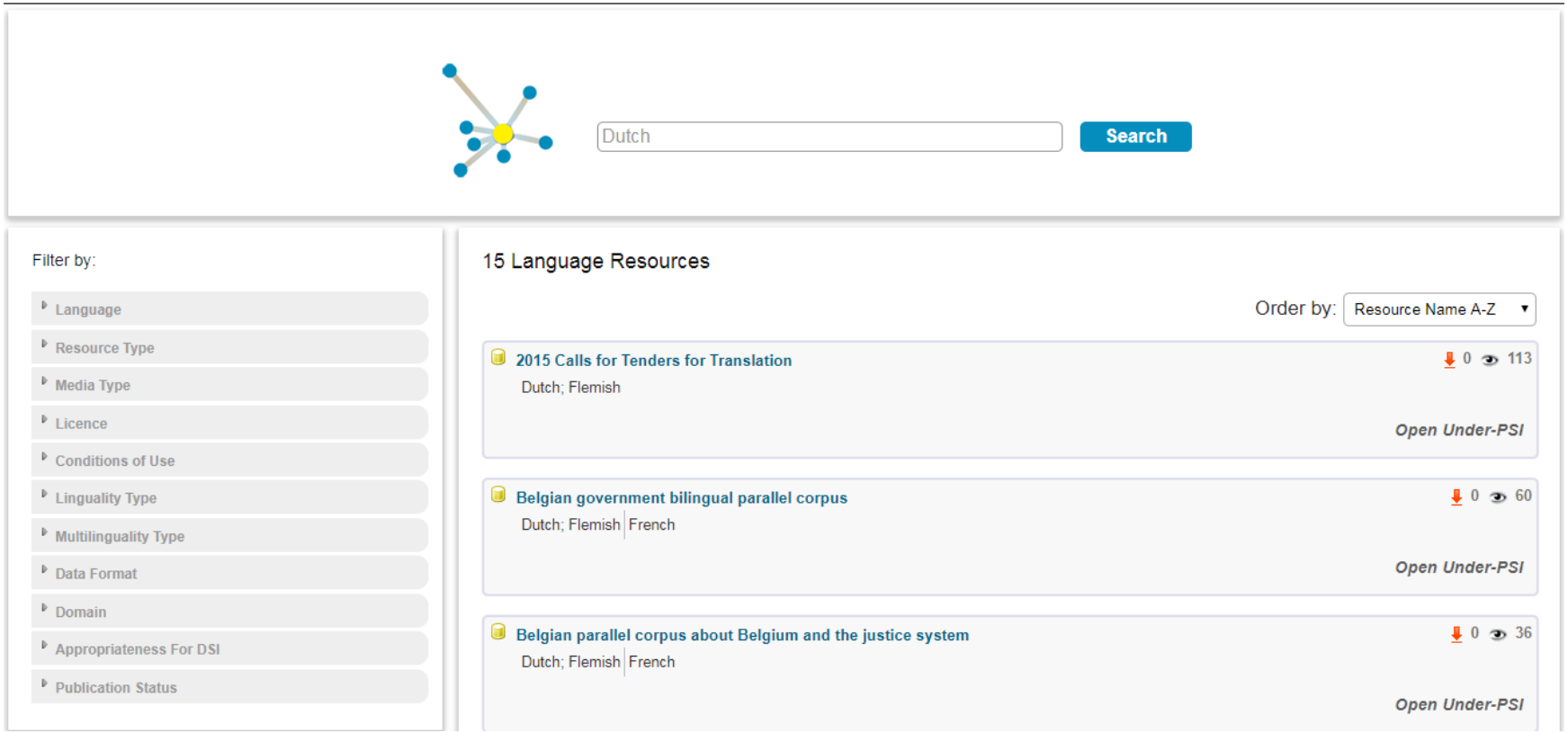
- Datasets Justitie: Aanbestedingsdocumenten Vertalen/ Tolken
  - Monolinguaal
  - 2 batches
- Dataset Sociale VerzekeringBank (SVB)
  - Parallelle Websiteteksten in meerdere talen
  - Zinstemplaten in meerdere talen
- Dataset Tweede Kamer: Teksten Nederlands Parlement
  - Monolinguaal
- Dataset Taalunie:
  - Parallel Corpus: nld, fra, eng
  - Multilingual subtitle data 2BDutch
  - Enkele monolinguale corpora (SoNaR, DAESO)
- Nederlandse teksten en parallelle corpora verzameld in België
- Nederlandse teksten van Hongaarse websites

Resultaten: verkregen data

- Datasets Justitie: Aanbestedingsdocumenten Vertalen/ Tolken
  - Monolinguaal
  - 2 batches
- Dataset Taalunie:
  - Dutch Parallel Corpus (DPC): nld, fra, eng
  - Multilingual subtitle data 2BDutch
  - Enkele monolinguale corpora (SoNaR, DAESO)
- Nederlandse teksten en parallelle corpora verzameld in België
- Nederlandse teksten van Hongaarse websites

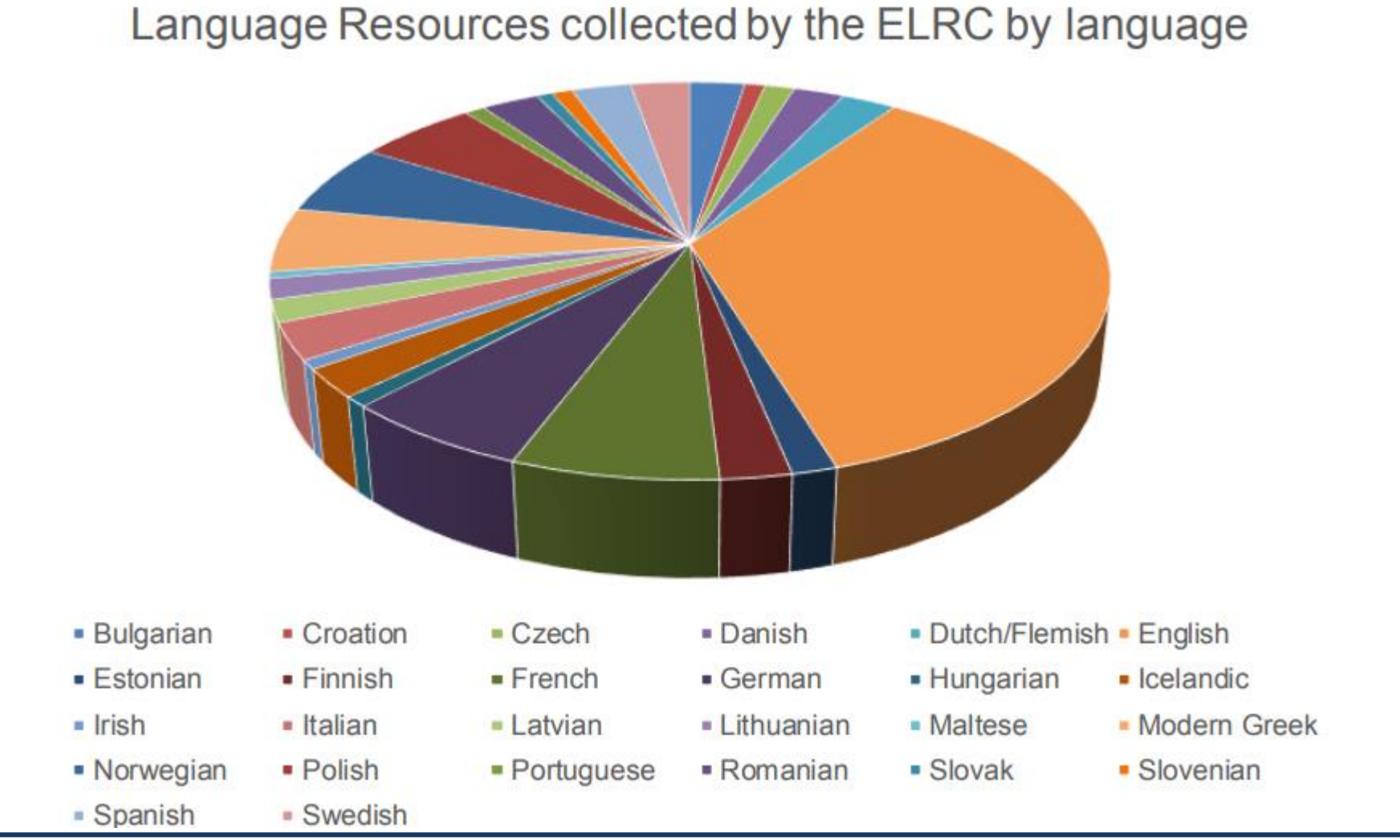
European Language Resource Coordination
Connecting Europe Facility

Resultaten: Nederlands in ELRC-SHARE



European Language Resource Coordination
Connecting Europe Facility

Overzicht LR in ELRC-SHARE



Obstakels & Oplossingen(?)

- Tekst wordt meestal niet als data beschouwd ➔ anders behandeld
- We hebben de teksten vanuit de bron nodig, niet via een web-zoekinterface
  - Bewustwording hiervan creëren o.a. via deze workshop
- Fragmentatie van vertalingsprocessen en van datastromen: meerdere ingangspunten
  - een goed beeld hiervan via deze workshop en landenprofiel

Obstakels & Oplossingen(?)

- Verzoek om data is geen ‘normal business’. Wie kan / mag daarover beslissen?
- Angst voor juridische /ethische kwesties ➔licentie gewenst, soms alleen met NL partij
  - Presentatie Annemarie Beunen kan hopelijk angst wegnemen
- Vertalers delen hun vertaalgeheugens niet graag: werken voor een goedkoper tarief als ze de vertaalgeheugens niet mee hoeven te leveren
  - Opdrachtgevers zouden hier strikter de standaardclausules in contracten moeten handhaven

European Language Resource Coordination
Connecting Europe Facility

Deze Workshop:

- Focus dit keer op diensten die
  - CEF-bouwstenen integreren / gebruiken: eDelivery, eID, eTranslation, eInvoicing, eSignature), en/of
  - CEF DSI-domeinen betreffen: justitie, gezondheid, publieke aanbesteding en inkoop, handelsregisters, sociale zekerheid, cultuur, open data, veiligheid en betrouwbaarheid van internet, en/of
  - meertalige functionaliteiten gebruiken of dat willen, en/of
  - een grensoverschrijdend karakter hebben
- Landenprofiel nodig om de datastromen te begrijpen

Landenprofiel voor Nederland

- Doel: Inzicht krijgen in de LR infrastructuur bij de overheden
- Vragen:
  - Waar in de publieke sector is de noodzaak voor vertaling het grootst?
    - Welk type documenten;
    - welke taalparen;
    - hoeveel data en hoe vaak?
  - Hoe wordt nu met vertaling omgegaan in de publieke sector? Welke diensten worden gebruikt?
    - Vertaling in huis of uitbesteding?
    - Terminologiedatabanken? Vertaalgeheugens?
    - eTranslation, CAT-tools?
  - Is er uitwisseling van data op nationaal niveau?
    - Centrale terminologiebank / centrale vertaalgeheugens?
  - Is er een coördinator voor de uitwisseling van data / vertalingen?
  - Zijn data voor publieke aanbesteding open beschikbaar?
  - Waar worden CEF-bouwstenen gebruikt?

Landenprofiel voor Nederland: Data

- Dataportaal van de Nederlandse overheid: min. BZK: https://data.overheid.nl/
  - vooral niet-tekstuele data; (tekstuele data < 2%; incl onduidelijk max. 14%)
  - vaak links naar websites, links naar websites die zelf weer een link bevatten naar een dataset
  - we moeten de tekstuele data aan de bron zien te krijgen.
    - Bijv. Basiswettenbestand (XML)
  - Taal: Nederlands 13193 Engels 25 Nederlands/Engels 1
  - Hayo Schreijer
- Open data portaal Tweede Kamer: https://opendata.tweedekamer.nl/
- Open Data Nederland: https://opendatanederland.org/
  - 97 tekstuele datasets
- Nationaal Archief: https://www.nationaalarchief.nl/
- ……..

Landenprofiel voor Nederland: CEF-Bouwstenen

- Relevante organisaties (websites):
  - PIANOo:
    - Engelstalige deel
    - Tenderned
  - RINIS
    - Engelstalig deel
  - ICTU
    - eDelivery
  - Logius

CEF-bouwstenen:

- eHealth
- eID
- eInvoicing
- eSignature
- eDelivery

Landenprofiel voor Nederland: Vertaling

- Ministerie BuZa:
  - Directie Vertalingen (bron)
    - Wolfram Metz
    - Terminologie: Michel Verhagen
    - Nieuwe technologie: Marie Hendrikse

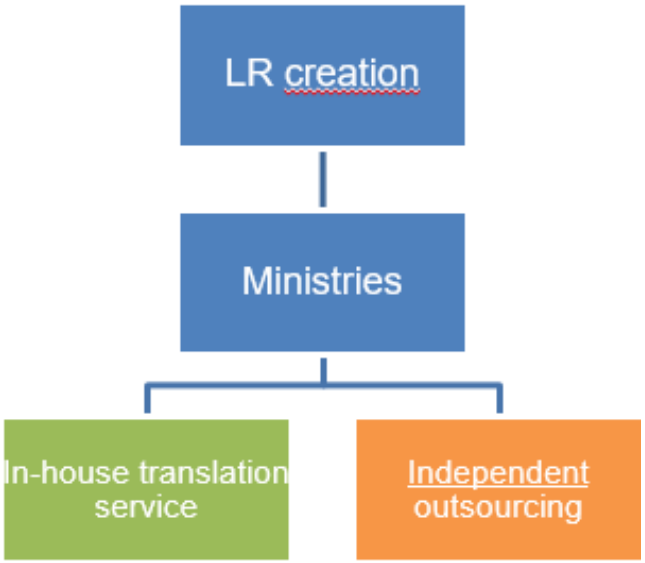
- Ministerie Justitie & Veiligheid:
  - Tolk- en Vertaaldiensten
    - Edwin de Koning, Categoriemanager Tolk- en Vertaaldiensten

- Ministerie van Defensie
  - Talencentrum Defensie (TCD)
  - Jan-Carel Annevelink

- Elders?

Landenprofiel voor Nederland

status quo



- Dataportaal van de Nederlandse Overheid
- Open Data Portaal Tweede Kamer
- Open Data Nederland
- Nationaal Archief
- ????

European Language Resource Coordination
Connecting Europe Facility

Samenvatting

- Veel data geïdentificeerd sinds 2016
- Groot aantal ervan ook verkregen
- Maar veel meer is nodig!
- Deze workshop:
  - Bewustwording creëren: u zit op een schat aan tekstuele data!
  - Goed beeld nodig van de productie, datastromen, en vertaling van tekstuele data, en de rol van CEF-bouwstenen (landenprofiel)
  - Levert hopelijk meer inzicht in deze kwesties
  - Levert hopelijk grote hoeveelheden nieuwe tekstuele data

European Language Resource Coordination
Connecting Europe Facility

Dank voor uw aandacht!

Email: j.odijk@uu.nl; carole.tiberius@ivdnt.org

Website: http://lr-coordination.eu/nl/l2netherlands