European Language Resource Coordination
Connecting Europe Facility

ELRC in Deutschland

Andreas Witt

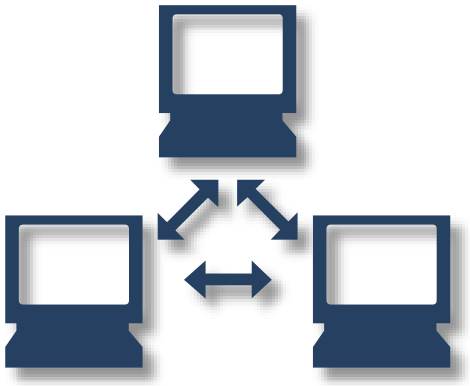
Technology
National Anchor Point

Alexandra Soska

Public Services
National Anchor Point

Die Arbeit als National Anchor Point

Fragen



Warum ist es wichtig,

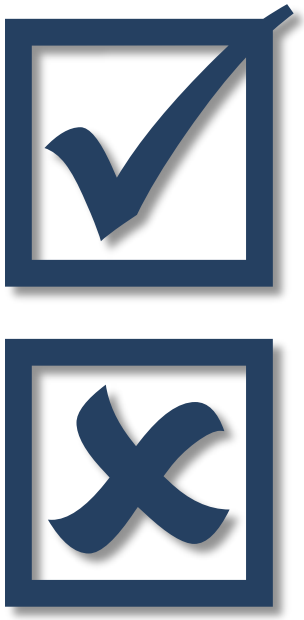
Sprachdaten auszutauschen?

European Language Resource Coordination
Connecting Europe Facility

Fragen

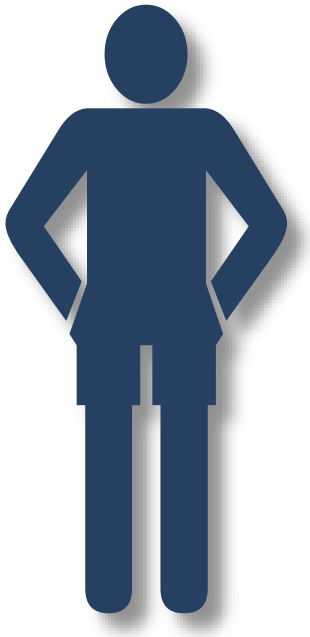
Dürfen wir Daten weitergeben?

Fragen



Wer entscheidet, welche Daten weitergegeben werden?

Fragen



Wer soll das umsetzen?

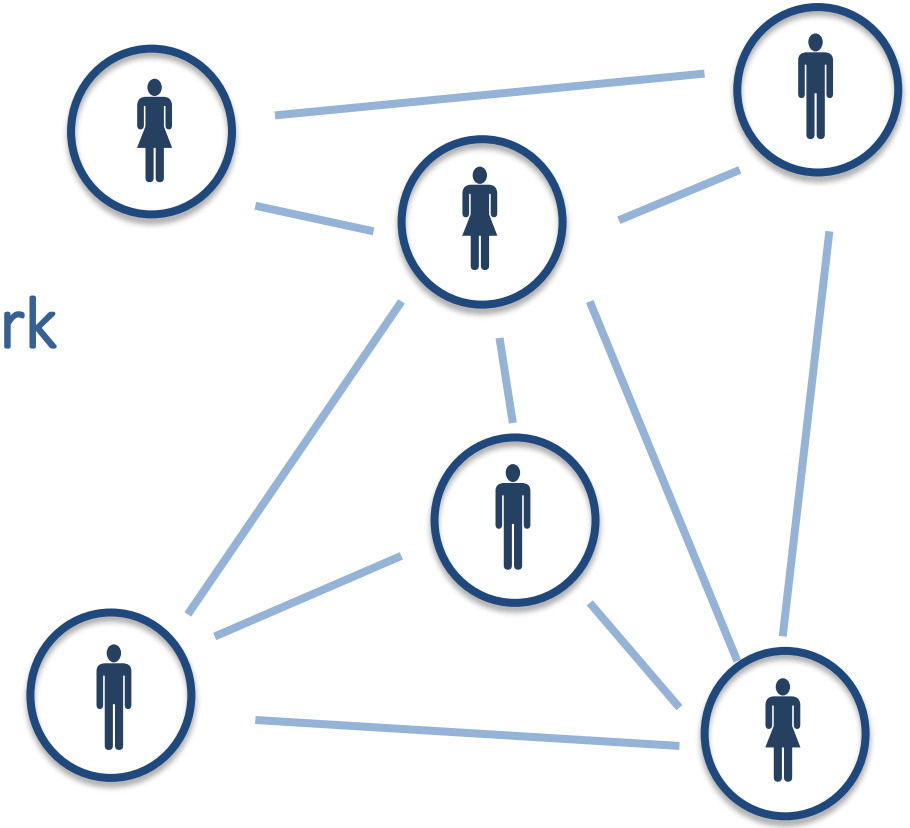
Fragen

Was ist technisch zu tun?

Stärken

Gut organisiertes Netzwerk

von Sprachendiensten



Stärken

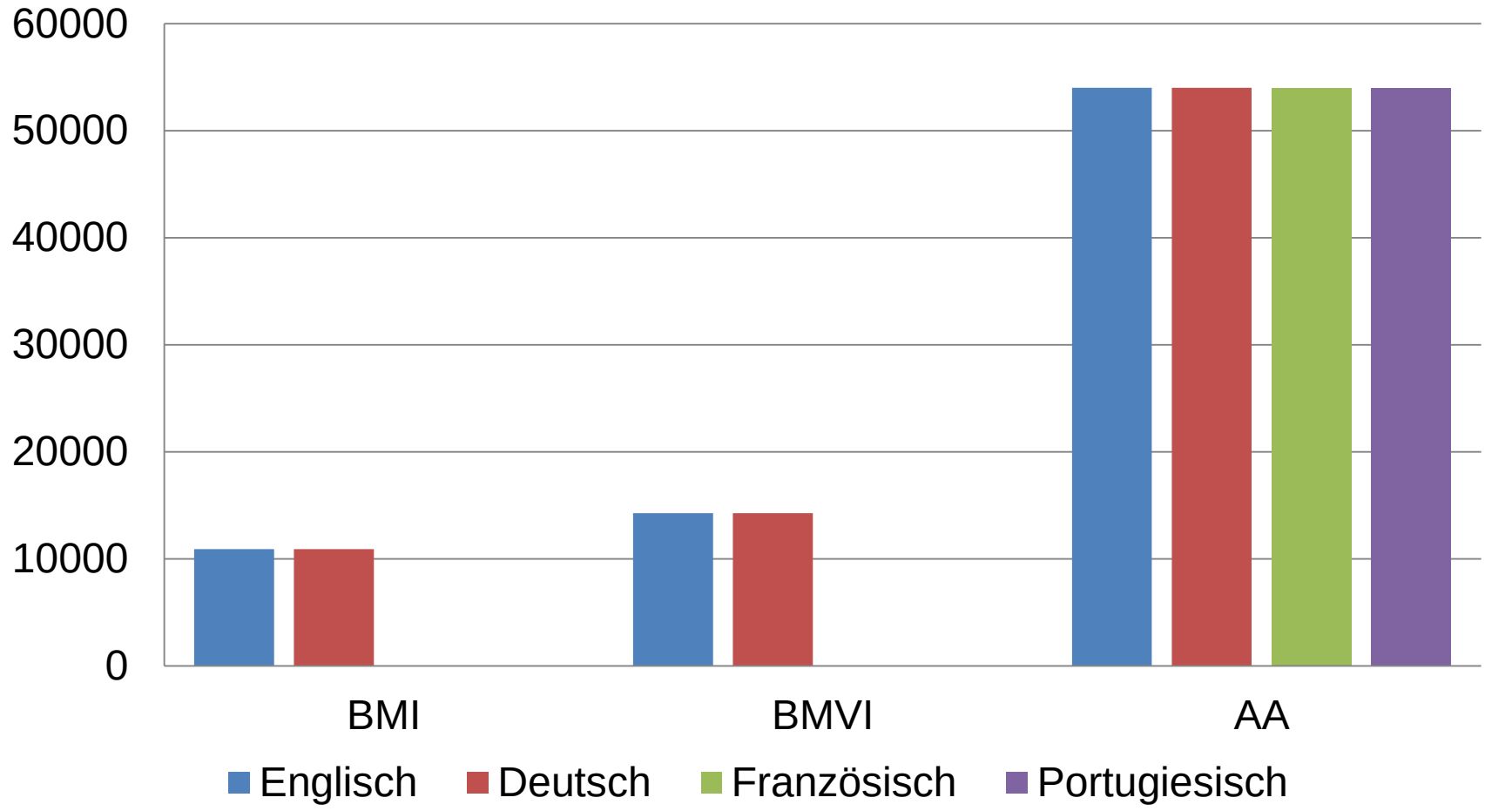
Sprachressourcen leicht auffindbar

Stärken

Gute technische Ausstattung & Kenntnisse

European Language Resource Coordination
Connecting Europe Facility

Beiträge zu ELRC-SHARE



Ansätze und Methoden der MÜ

Regelbasierte Ansätze:

- direkte Übersetzung
  - Wort-zu-Wort-Übersetzung mittels Wörterbuch
- Transfer
  - satzbasierte Übersetzung
  - drei Phasen: Analyse, Transfer und Generierung
- Interlingua
  - Analyse und Generierung
  - Interlingua: sprachpaarunabhängig Zwischenrepräsentation
  - keine ideale Repräsentationssprache vorhanden

Ansätze und Methoden der MÜ:
korpusbasierte MÜ

- beruht auf Informationen aus Textkorpora (zwei- oder auch mehrsprachige Korpora)
- Alignierung: manuelle, automatische oder halbautomatische Auszeichnungen zwischen den Korpora auf Wort-, Phrasen-, Satzebene, etc. zur Verbesserung der Qualität der Übersetzung
- manuelle Alignierung:
  - zuverlässige Ergebnisse, aber nicht effizient bei großen Textmengen
  - Verwendung für Evaluierung oder für statistische Verfahren
- automatische Alignierung:
  - Einteilung in Übersetzungseinheiten
  - Zuordnung durch linguistische oder statistische Algorithmen
  - Effizient, aber fehleranfällig
- halbautomatischen Alignierung:
  - Automatischer Prozess und
  - manuelle Überwachung

Ansätze und Methoden der MÜ: Reinstatistikbasierte Übersetzung

- meist werden korpusbasierte Lernverfahren verwendet
- Ebenen: Wörter, Phrasen, ganze Sätze
- Wahrscheinlichkeiten für die Übersetzung eines Terms in einen zielsprachlichen Term werden berechnet
- beeinflusst durch Worthäufigkeit, Position des Wortes im Satz, Satzlänge etc., d.h. nicht durch Wörterbücher oder explizites linguistisches Wissen
- erlerntes Übersetzungsmodell:
  - Übersetzungsmöglichkeiten für ausgangssprachliche Textsequenzen und die
  - dazugehörige Wahrscheinlichkeit
- Sprachmodelle: einsprachigen Korpora
- Übersetzungsvorgang:
  - Decodierung via Übersetzungsmodells (Ermittlung möglicher Übersetzungen)
  - Decodierung via Sprachmodell (Einbeziehung kontextueller Übersetzungsalternativen)

Ansätze und Methoden der MÜ:
Reinstatistikbasierte Übersetzung

- **Wortbasierte statistische MÜ**:
  - Übersetzungsmodell für Wort-für-Wort-Übersetzungen
  - Annahme, dass die einzelnen Wortübersetzungen unabhängig voneinander existieren
  - Problem: ein Wort kann mehreren Wörtern übersetzt werden und umgekehrt
  - Kontextinformationen werden nicht berücksichtigt

- **Phrasenbasierte statistische MÜ:**
  - Übersetzungsmodell aus Mehrwortsequenzen
  - Neuanordnung der Segmente mithilfe eines Sprachmodells
  - Behandlung von Ambiguitäten durch Einbeziehung des Kontextes

- **Neuronale MÜ:**
  - seit wenigen Jahren verfügbar
  - hohe Qualität
  - es sind nicht für alle Sprachpaare zweisprachige Korpora nötig

Ansätze und Methoden der MÜ:
Reinstatistikbasierte Übersetzung

Vorteile:

- effizient, da geringe Investitionskosten und hohe Funktionssicherheit
- Verzicht auf Grammatikregeln und Ausnahmen
- schnelle Realisierung und Erweiterung um weitere Sprachen

Nachteil:

- benötigt großen Mengen der aufbereiteten Textkorpora
- Textmaterial ist domänenspezifisch auszuwählen

Folgen für ELRC:

- Trainingsdaten aus dem Bereich Behördensprache werden benötigt, denn:
- qualitativ hochwertige mehrsprachige Datensätze werden die maschinelle Übersetzung erheblich verbessern

European Language Resource Coordination
Connecting Europe Facility

Fragen?

European Language Resource Coordination
Connecting Europe Facility

Danke!

ELRC in Deutschland

Andreas Witt

E-Mail:
witt@ids-mannheim.de

Alexandra Soska

E-Mail:
Alexandra.Soska@bmi.bund.de

Email: info@lr-coordination.eu
Website: www.lr-coordination.eu