European Language Resource Coordination
Connecting Europe Facility

European Commission

# Die Bedeutung von Daten für die Entwicklung von Sprachtechnologien

HANNES PIRKER
AUSTRIAN CENTRE FOR DIGITAL HUMANITIES AND CULTURAL HERITAGE (ACDH-CH)
ÖSTERREICHISCHE AKADEMIE DER WISSENSCHAFTEN (ÖAW)

# DIE WICHTIGKEIT VON DATEN

- “Daten sind der Lebenssaft (*lifeblood*) der ökonomischen Entwicklung
  - sind Basis für viele neue Produkte und Dienstleistungen
  - treiben die Steigerung der Produktivität und Ressourceneffizienz in allen Wirtschaftsbereichen an
  - ermöglichen personalisierte Produkte und Dienstleistungen
  - ermöglichen eine bessere Politikgestaltung und Verbesserung der staatlichen Dienste.“

(Quelle: https://ec.europa.eu/info/sites/info/files/communication-european-strategy-data-19feb2020_en.pdf)

# DATENGETRIEBENE ANWENDUNGEN…

- … sind tatsächlich das erfolgreichste Paradigma in der Sprachtechnologie
  - Stichwort: Maschinelles Lernen und Künstliche Intelligenz
  - Lernen Strukturen aus Daten mit *bekannten* Zusammenhängen (*Trainingsdaten / ground truth / Gold-Standard*), um diese später auf unbekannte Zusammenhänge anzuwenden.

# WELCHE DATEN EIGENTLICH?

- *Beispiel 1:* Gesichtserkennung / Lokalisation
  - Trainingsdaten:(viele!) Fotos von menschlichen Gesichtern + Fotos von anderen Objekten
- *Beispiel 2:* Gesichtserkennung / Personen-Identifikation
  - Trainingsdaten: Sammlung von unterschiedlichen Fotos *einer* Person

# VERZERRUNGEN IN DEN TRAININGSDATEN?

z.B. der bekannter – oft kritisierter – *racial bias*

- System (nur) mit Bildern von weißen Erwachsenen trainiert wird für Europäer besser funktionieren als etwa für Asiaten


# FAZIT BISHER:

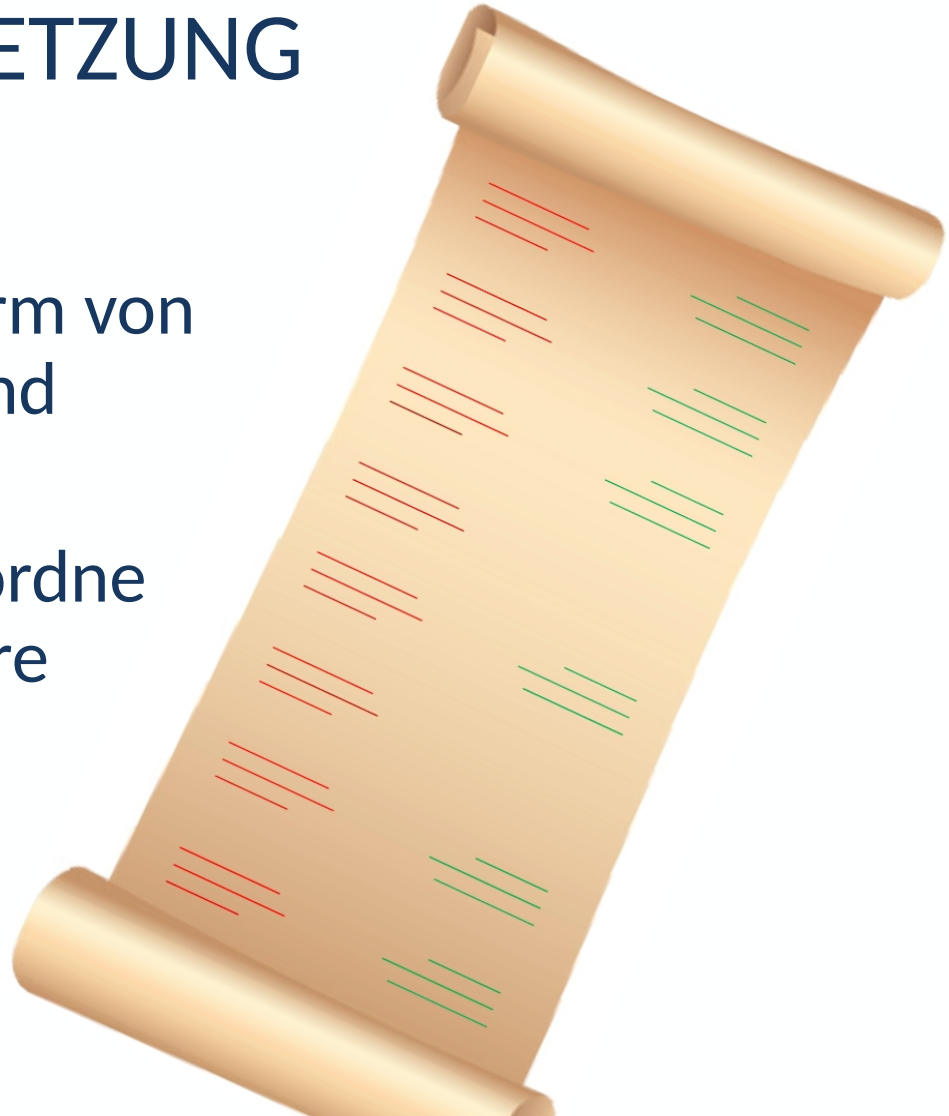
- Daten müssen zur Lernaufgabe passen
- Verzerrungen/*bias* vermeiden

European Language Resource Coordination
Connecting Europe Facility

European Commission

# UMGELEGT AUF MASCHINELLE ÜBERSETZUNG

- Die verwendeten Daten (Sprachressourcen) haben die Form von *Satzpaaren* aus der Ursprungs- und Zielsprache
- Verwende übersetzte Texte und ordne die korrespondierenden  Satzpaare einander zu



European Language Resource Coordination
Connecting Europe Facility

European Commission

# UMGELEGT AUF MASCHINELLE ÜBERSETZUNG

- Übersetzungsaufgaben sind oft domänen- und genre-spezifisch
- Schlüsselfragen bei der Datenauswahl:
  - Was ist die Übersetzungsaufgabe?
  - Übersetzungstool für „alles“? Für Rechtstexte? Für Twitter-Meldungen?
  - ein System mit EuroParl trainiert, wird z.B. für Nachrichtentexte oder Chats schlechter funktionieren.

European Language Resource Coordination
Connecting Europe Facility

European Commission

# UNGENÜGENDE RESULTATE TROTZ „GUTER“ DATEN?

- Datenmenge nicht ausreichend?
- je unspezifischer die Übersetzungsaufgaben, desto mehr Daten sind nötig

Ein gutes „generelles“ Übersetzungssytem für DE-EN benötigt Millionen von Satzpaaren

# SPRACHRESSOURCEN…

- müssen zur *Aufgabe passen*
- Quantität zählt – *„je mehr, desto besser“*

# SPRACHTECHNOLOGIE IST MEHR ALS MASCHINELLE ÜBERSETZUNG

viele weitere Anwendungen, mit anderen, tw. *„einfacheren“* Ansprüchen an die Sprachressourcen, z.B.

- Spracherkennung
- Textklassifikation
- Trainieren von Worteinbettungen (*word embeddings*) – eine Möglichkeit um z.B. bedeutungsverwandte Wörter automatisch zu erkennen

# ABER WOHER DIESE DATENMENGEN NEHMEN?

- *„Gold Standard“-Daten …*
  - … werden nicht nur für das Trainieren sondern auch für das Evaluieren benötigt!
  - … können daher bereits in kleineren Mengen „Goldes wert“ sein!
- Kollaborative Netzwerke nützen!
  - ELRC / CLARIN / ELDA-ELRA / LDC / …

European Language Resource Coordination
Connecting Europe Facility

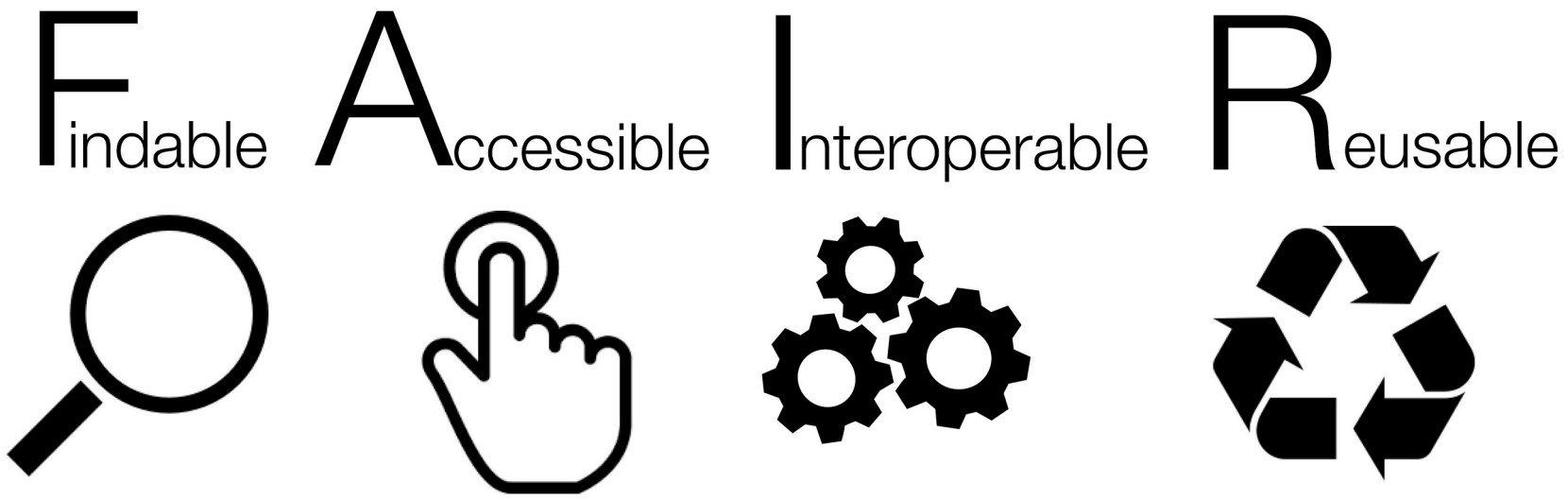
European Commission

# GEMEINSAM GEGEN DEN „ZU-WENIG-Ö-DEUTSCH“-BIAS!

- Deutsch in Österreich ist mehr als *Topfen* und *Paradeiser*
  - („Protokoll 10-Begriffe“)
- Amts- und Verwaltungsbereich:
  - *Akt*
  - *Exekutionsrecht (AT) vs. Zwangsvollstreckungsrecht (DE)*
  - …

# GUTE RESSOURCEN SIND ÜBRIGENS *FAIR*

- nicht nur *passend* bzgl. Inhalt und Umfang
- sondern auch *FAIR*



European Language Resource Coordination
Connecting Europe Facility

European Commission

# DANKE FÜR IHRE AUFMERKSAMKEIT!

Website: https://www.oeaw.ac.at/acdh
Email: hannes.pirker@oeaw.ac.at