University of Zurich[UZH]

Wortalignierung & mehrsprachige Suchsysteme

CAS in Translation Technology and AI

Martin Volk & Johannes Graën

5. November 2021

Alignierung von Centauri und Arcturan

Mehrsprachige Suchsysteme

Mehrsprachige Suchsysteme (multilingual concordancers)

- Linguee – Wörterbuch mit automatisch gefundenen Belegstellen
- TradooIT – «a computer-assisted translation suite including a translation memory, a term bank and a bilingual concordancer»
- Glosbe – Online-Wörterbuch
- Reverso Context – «kontextbezogenes Wörterbuch»

Beispiele

- Suche nach «Aufmerksamkeit zollen» in Linguee (de/en)
- Suche nach «Aufmerksamkeit zollen» in Glosbe (de/en)
- Suche nach «Aufmerksamkeit zollen» in ReversoContext (de/en)

Linguee

⚠ [...] Verfassung sofern diese angenommen wird berufen, dann wird uns auch niemand außerhalb dieses Parlaments Respekt zollen. europarl.europa.eu	⚠ [...] various treaties and under the new constitution if it is adopted then no one outside this House will have respect for us. europarl.europa.eu
[...] politischer als auch auf diplomatischer Ebene, wo Botschafter Vattani eine führende Rolle gespielt hat, Respekt zollen. europarl.europa.eu	[...] by the Presidency, both at the political and at the diplomatic level, where Ambassador Vattani has played a leading role. europarl.europa.eu
⚠ [...] Innovationsleistungen und der Armutsminderung mehr Aufmerksamkeit widmen und den Menschenrechten größeren Respekt zollen. europa.eu	⚠ [...] rational use of natural resources, better innovation performance, poverty reduction, and greater respect for human rights. europa.eu
⚠ [...] möchte in diesem Zusammenhang der italienischen Präsidentschaft Dank und Respekt zollen für ihren Einsatz, den sie im Vermittlungsverfahren gezeigt hat. europarl.europa.eu	⚠ [...] trans-European transport networks and I would like to offer my thanks and respect to the Italian presidency in this connection for the efforts it made in the conciliation procedure. europarl.europa.eu
[...] die der traditionsreichen griechischen Aromatherapie Respekt zollen und nur aus natürlichen mediterranen Zutaten bestehen. tridonic.com	[...] specializes in health products for the face, body and hair, paying enormous respect to the traditions of Greek aromatherapy, with [...] tridonic.com

Multilingwis

- basiert auf automatisch berechneten Wortalignierungen
- inhärent mehrsprachig; eine Quellsprache, verschiedene Zielsprachen
- Übersicht über gefundenen Übersetzungsvarianten pro Sprache
- Wörter-in-Satz-Suche, Phrasensuche, Platzhalter
- Korpusexploration durch Rückwärtssuche
- verwaltet verschiedene Korpora mit verschiedenen Sprachen
- Software frei verfügbar, mindestens zwei Instanzen online

Gleiches Beispiel

- Suche nach «Aufmerksamkeit zollen» in Multilingwis (de/*)

Linguee ↔ Multilingwis

Linguee

- unscharfe Entsprechungen (Algorithmus/Verfahren unbekannt)
- häufig keine Entsprechung markiert
- diverse Quellen, hauptsächlich Webseiten (keine Überprüfung)
- Verknüpfung mit eigenem Wörterbuch
- Zielgruppe?

Multilingwis

- Alignierungsverfahren abhängig vom Korpus
- für FEP6/FEP9: Schnitt mehrerer Alignierer (mehr Genauigkeit)
- Ziele: Exploration von Korpus, Übersetzungsvarianten, Alignierung
- Zielgruppen: (Korpus-)Linguisten, Sprachlerner, Übersetzer (!)

Übung

...

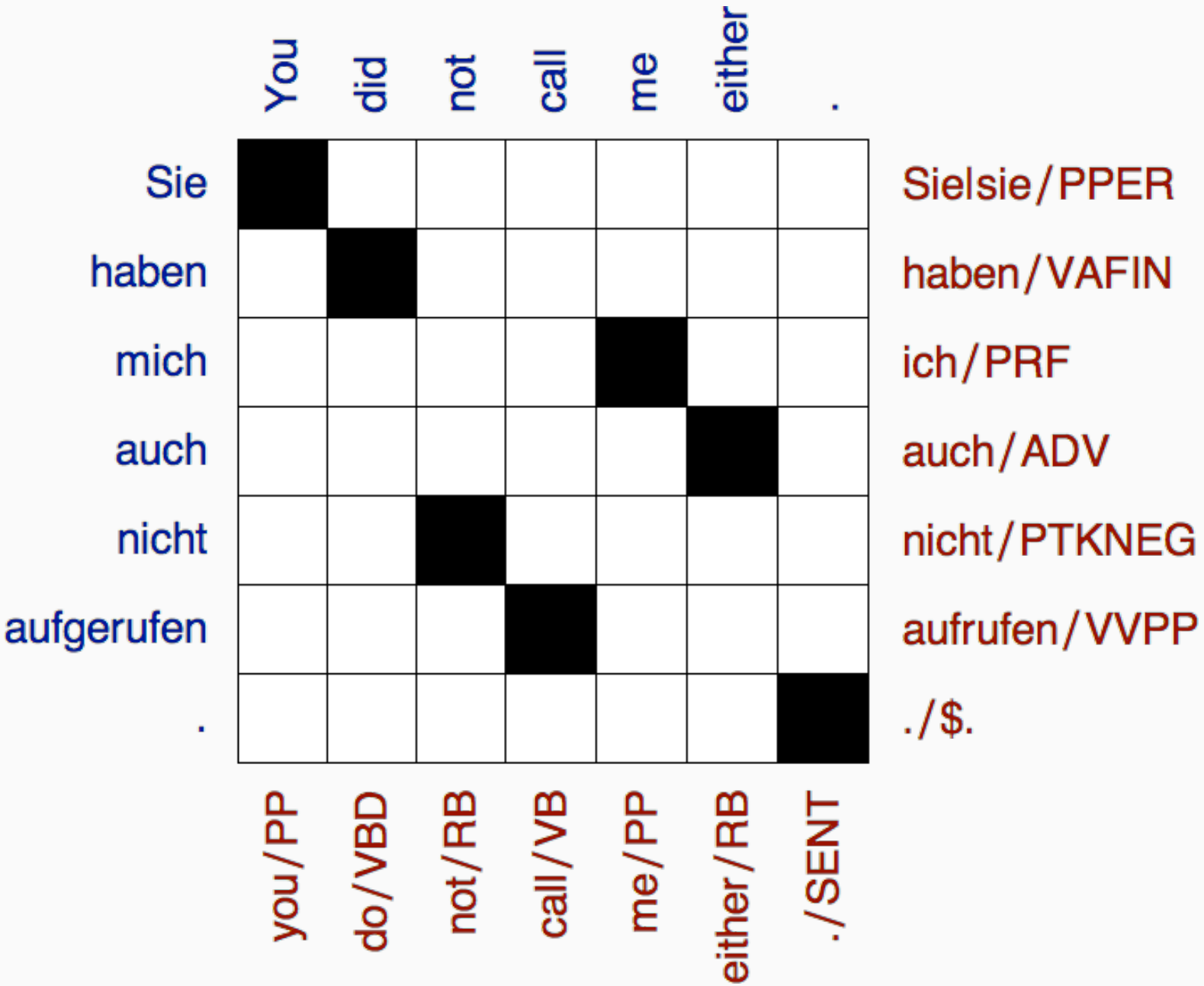
Wortalignierung

Wortalignierung

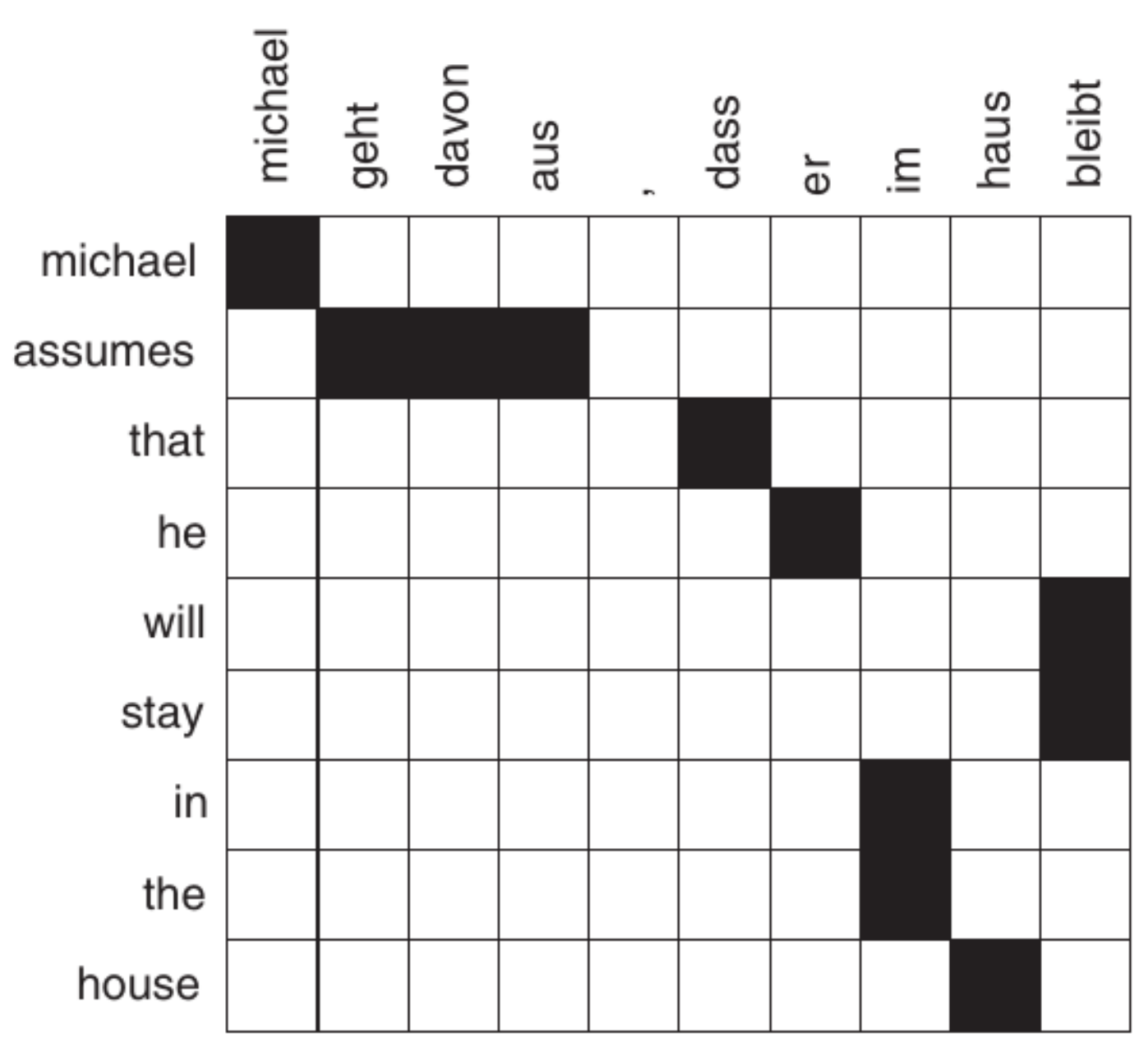
- setzt korrekt alignierte Sätze voraus
- um ein Vielfaches komplexer als Satzalignierung
  - Wortalignierung nicht monoton
  - Wortbildung sprachspezifisch, ‹Satzbildung› eher einheitlich
  - funktionale Elemente, die nicht/unterschiedlich übersetzt werden
- viele verschiedene mathematische Modelle, aber kein Standard
- Fehlerrate bei Alignierungen[1] problematisch bei der Nutzung individueller Beispiele

[1] Zum Beispiel gemessen mit der ‹alignment error rate› (AER)

Beispiel (1:1 Alignierungen)



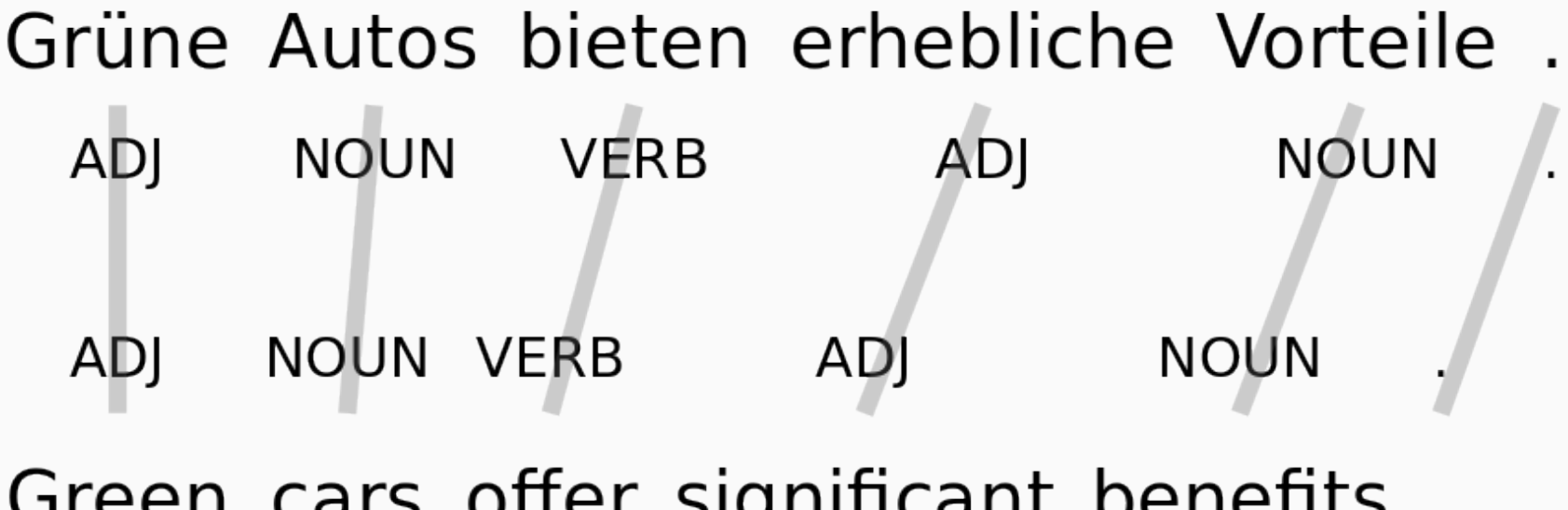
Beispiel (1:n Alignierungen)



Wortalignierungsmatrizen

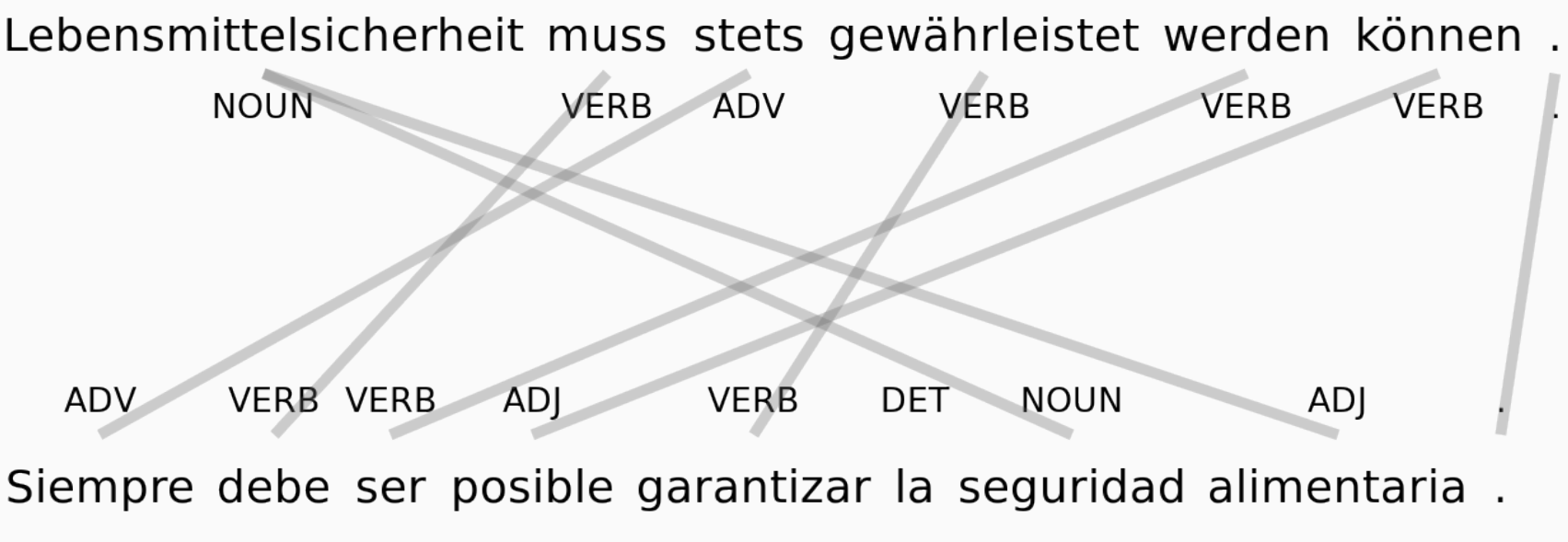
- Kombination aller Wörter aus Quell- und Zielsprache
- ausgefüllte Zellen stehen für Alignierung bzw. (partielle) Korrespondenz
- selten auch grau für ‹mögliche› Korrespondenz in Gebrauch
- andere Darstellungen: via ‹Links› oder farblicher Hervorhebung

Alignierungslinks



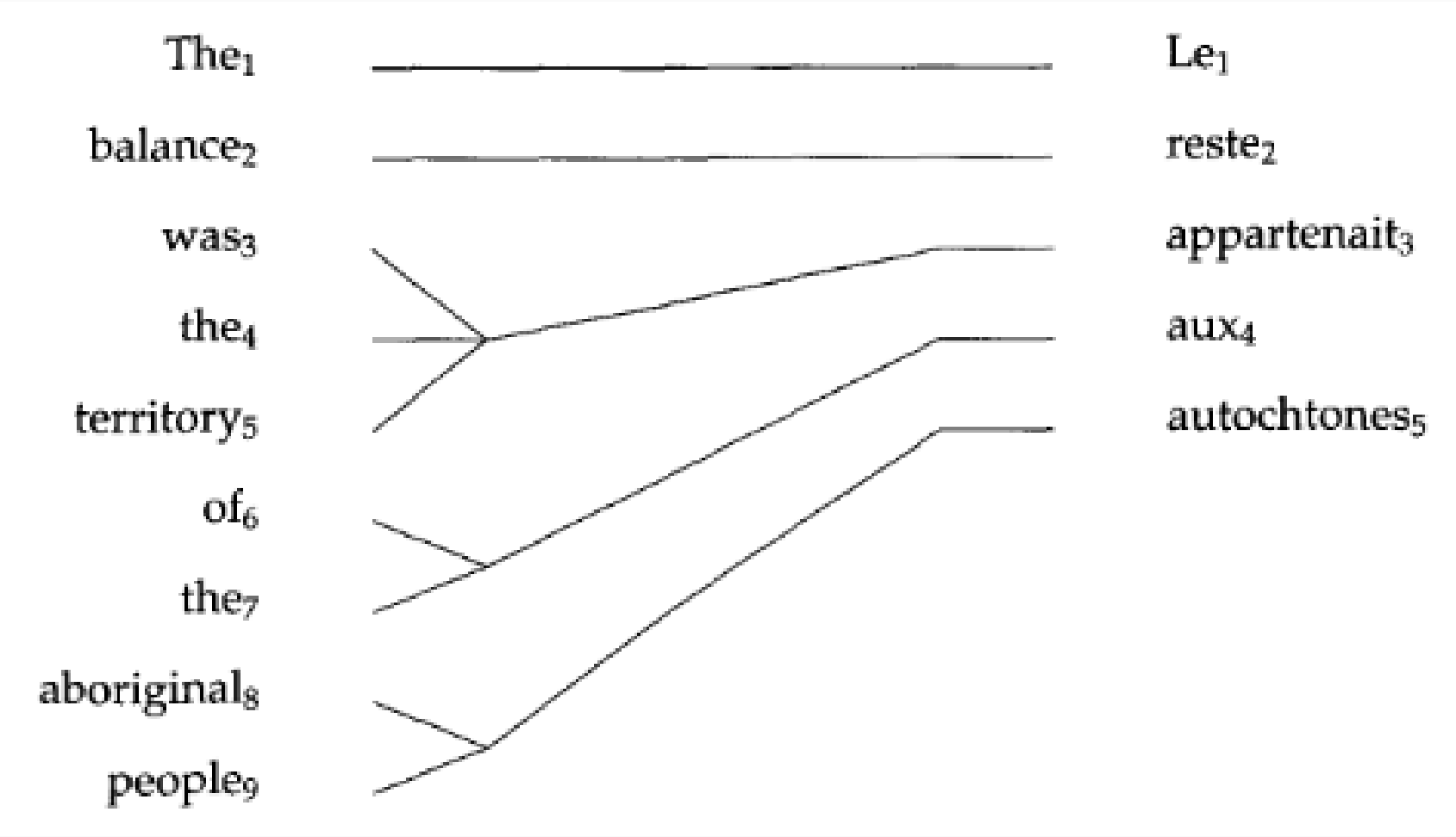
- ‹grüne› ↔ ‹green› ✓
- ‹Autos› ↔ ‹cars› ✓
- ‹bieten› ↔ ‹offer› ✓
- ‹erhebliche› ↔ ‹significant› ✓
- ‹Vorteile› ↔ ‹benefits› ✓

Alignierungslinks (2)



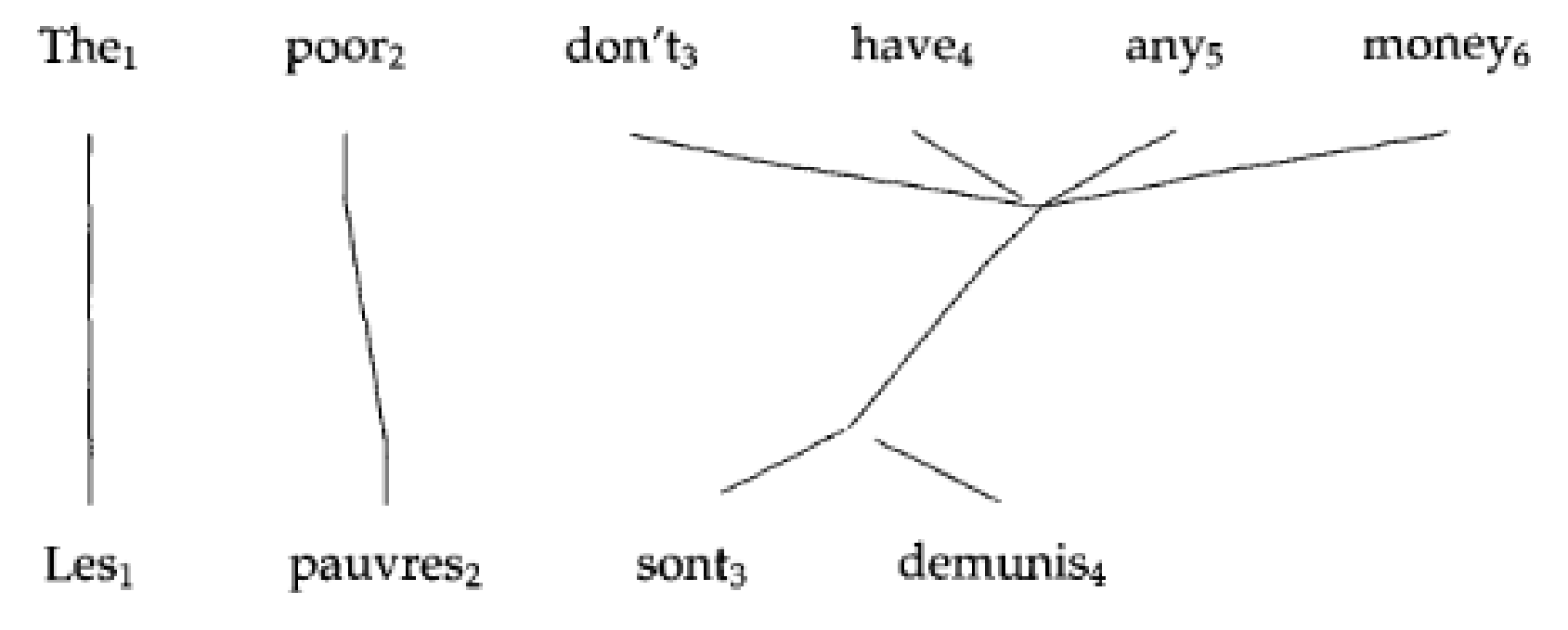
- ‹Lebensmittelsicherheit› ↔ ‹seguridad alimentaria› ✓
- ‹muss› ↔ ‹debe› ✓
- ‹stets› ↔ ‹siempre› ✓
- ‹gewährleistet› ↔ ‹garantizar› ✓
- ‹werden› ↔ ‹ser› (✓)
- ‹können› ↔ ‹posible› (✓)

Alignierungslinks (III)



(Brown u. a. 1993)

Alignierungslinks (IV)



(Brown u. a. 1993)

Korrespondenzen auf Wortebene bei mehreren Sprachen

- Wir möchten nicht die Katze im Sack kaufen .
- Nous ne voulons pas acheter chat en poche .
- Não estamos interessados em comprar gato por lebre .
- We are not interested in buying a pig in a poke .
- Vi är inte intresserade av att köpa grisen i säcken .

Übersetzungsvarianten

Der englische Ausdruck ‹key point› kann ins Deutsche wie folgt übersetzt werden:[2]

- wichtiger Punkt
- Kernpunkt
- zentraler Punkt
- Hauptpunkt
- wesentlicher Punkt

[2] Wortwahl «largely a matter of taste» (Brown u. a. 1993)

Übersetzungsvarianten

Der englische Ausdruck ‹key point› kann ins Französische wie folgt übersetzt werden:[2]

- point clé
- point essentiel
- points-clés (Plural)
- point important

[2] Wortwahl «largely a matter of taste» (Brown u. a. 1993)

Übersetzungsvarianten

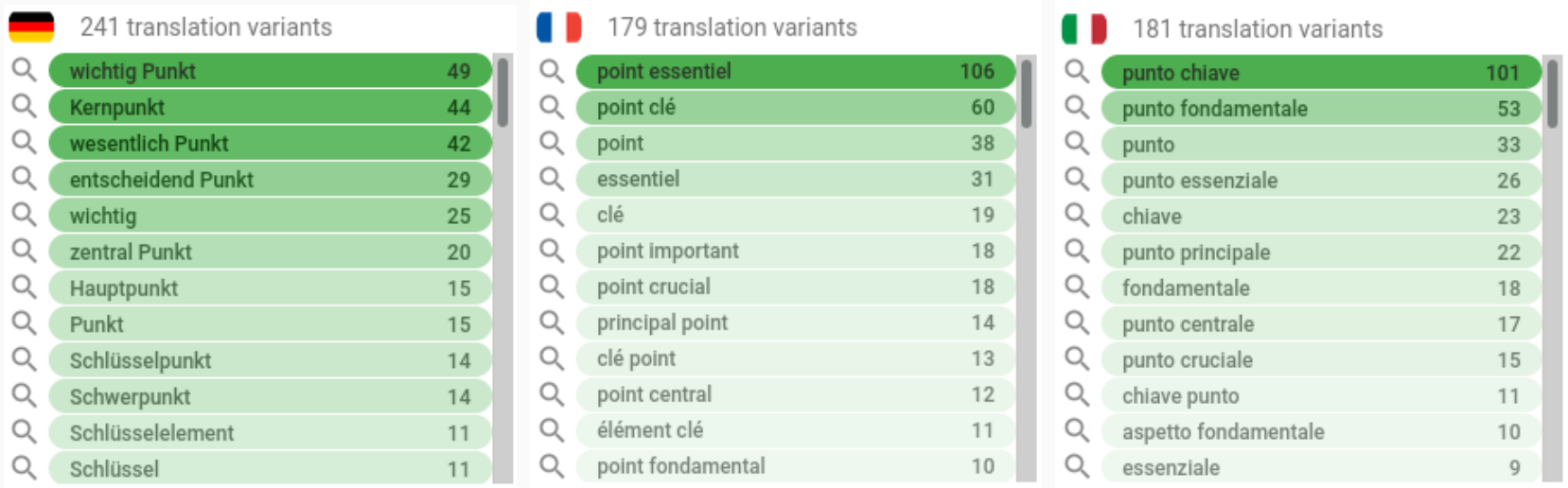
Der englische Ausdruck ‹key point› kann ins Italienische wie folgt übersetzt werden:[2]

- punto chiave
- punto principale
- punto fondamentale
- punto essenziale

[2] Wortwahl «largely a matter of taste» (Brown u. a. 1993)

Übersetzungsvarianten

- unterschiedliche Verteilungen der Übersetzungsvarianten
- Alignierungsfehler spielen bei statistischer Auswertung keine Rolle, falls sie nicht systematischer Art sind



Die IBM-Modelle

IBM-Modelle

P. F. Brown u. a. (1993). „The Mathematics of Statistical Machine Translation: Parameter Estimation“. In: *Special issue on using large corpora: II* 19.2, S. 263–311

- *„each of our models assigns a probability to each of the possible word-by-word alignments“*
- *„minimal linguistic content“*
- *„[...] it is reasonable to argue that word-by-word alignments are inherent in any sufficiently large bilingual corpus“*

IBM-Modell I – «lexikalische Entsprechung»

- summierte Übersetzungs-/Entsprechungswahrscheinlichkeiten für einzelne Wörter:
  $p(\text{house}|\text{Haus}) > p(\text{house}|\text{Gebäude})$
  $p(\text{house}|\text{Haus}) \gg p(\text{house}|\text{Mas})$
- Wahrscheinlichkeiten werden aus Trainingsdaten (= Satzpaaren) gelernt (= berechnet)

IBM-Modell I (2

Beispiel

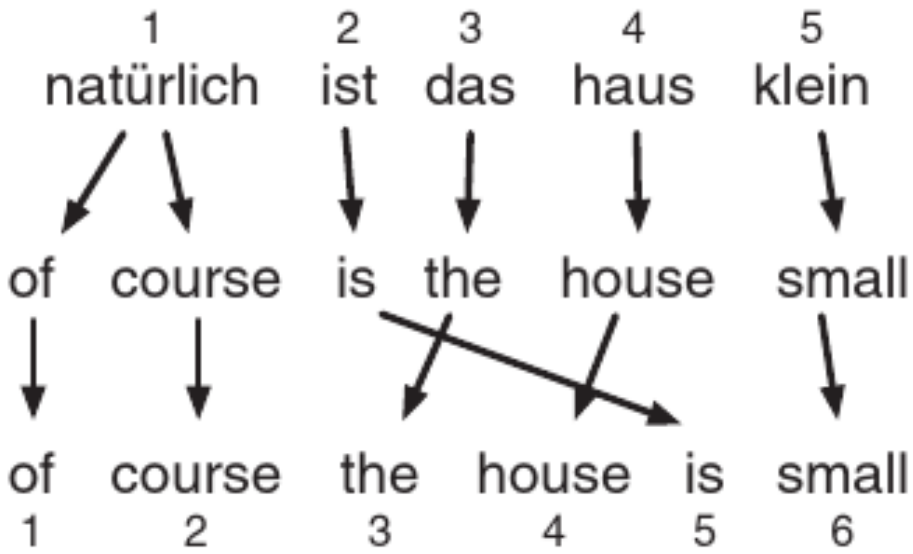
	es	regnet
it	0.9	0.2
rains	0.1	0.8

Variante	Wahrscheinlichkeit
a(it, es), a(rains, regnet)	$p(\text{it}\|\text{es}) \times p(\text{rains}\|\text{regnet}) \propto 0.72$
a(it, regnet), a(rains, es)	$p(\text{it}\|\text{regnet}) \times p(\text{rains}\|\text{es}) \propto 0.02$

IBM-Modell II – «»

- für Modell I spielt die Reihenfolge der Wörter in einem Satz keine Rolle
- Modell II ergänzt die Wahrscheinlichkeit, dass das n. Wort der einen Sprache mit dem m. Wort der anderen Sprache aligniert ist
- die Wahrscheinlichkeit der Alignierung a(it, es), a(rains, regnet) ist höher für das Satzpair «it rains» & «es regnet» als für das Paar «it rains» & «regnet es»

IBM-Modell II (2)



lexical translation step

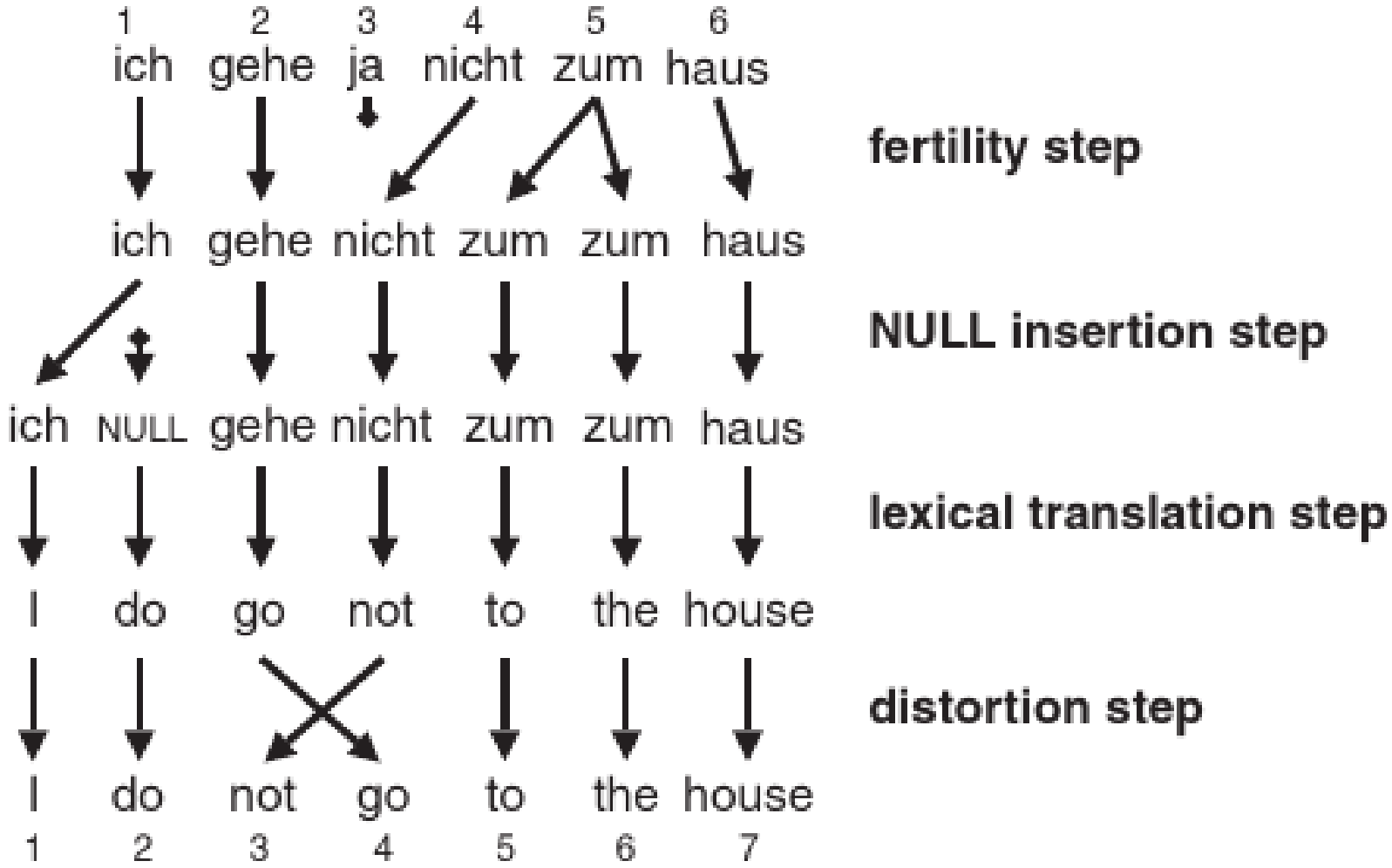
alignment step

(Koehn 2010)

IBM-Modell III – «Fertilität»

- die vorherigen beiden Modelle gehen von 1:1 Alignierungen aus
- Modell III fügt Wahrscheinlichkeiten für die sogenannte Fertilität hinzu, d.i. eine Wahrscheinlichkeitsverteilung für die Anzahl Wörter, mit denen jedes Wort aligniert ist
- diese Verteilung umfasst die Wahrscheinlichkeit für 0 alignierte Tokens (eine Null-Alignierung)
- in diesem Schritt wird auch die Wahrscheinlich für das Hinzufügen von Wörtern via Alignierung modelliert, z.B. «do» für Alignierungen mit englischen Verben

IBM-Modell III (2)



(Koehn 2010)

IBM-Modelle IV und V

- Modell IV berücksichtigt, dass Wörter der einen Sprache nicht wahllos einer bestimmten Position in der anderen Sprache entsprechen, sondern mehrere aufeinanderfolgende Wörter beim Übersetzen häufig ebenfalls nah beieinander zu finden sind
- Modell V bringt eine Optimierung ins Spiel, die verhindert, dass Satzalignierungen erzeugt werden, die technisch unmöglich sind, da mehrere Wörter sich an der gleichen Stelle im Satz befinden müssten

Weitere Ansätze/Verbesserungen

- Symmetrisierung
- symmetrische Modelle (vermeiden sogenannte ‹garbage collector words›, 1:n Alignierungen mit sehr grossem n)
- Sampling
- Triangulation in multiparallelen Korpora

Wortalignierungswerkzeuge

- GIZA++ (Och und Ney 2003)
  implementiert die IBM-Modelle (asymmetrisch)
- BerkeleyAligner (Liang u. a. 2006)
  filtert sehr unwahrscheinlich Alignierungen, inhärente Symmetrisierung
- fastalign (Dyer u. a. 2013)
  schnell, aber weniger gute Ergebnisse
- efmaral/eflomal (Östling und Tiedemann 2016)
  komplexerer mathematischer Ansatz; Modelle können auch gespeichert und geladen werden
- SimAlign (Sabet u. a. 2020)
  basiert auf bilingualen ‹word embeddings›
- AWESOME (Dou und Neubig 2021)
  basiert ebenfalls auf bilingualen ‹word embeddings›

Wortalignierungswerkzeuge

- die meisten Alignierungswerkzeuge generieren unidirektionale Alignierungen (1:n)
- deshalb braucht es zwei Modelle, eines für jede Alignierungsrichtung
- ... mit anschliessender Symmetrisierung
- mehrere Alignierer zu kombinieren erhöht die Genauigkeit

Wortalignierungen bewerten

- Genauigkeit (precision) und Trefferquote (recall)
- für die Berechnung von Genauigkeit und Trefferquote braucht es in jedem Fall Gold-Daten, d.i. händisch alignierte Satzpaare, die als ‹korrekt› angenommen werden
- F_1-Wert: $F_1 = \frac{2 \times P \times R}{P+R}$
- Alignment error rate (AER) entspricht inversem F_1-Wert

Alignierungen berechnen

Vorgehen

Parallelkorpus ‹Rumantsch Grischun› aus der ‹Parallel Corpus Collection› (PaCoCo) herunterladen

Satzpaare Deutsch/Rumantsch (de/rm) extrahieren

gemeinsame Vorkommen von Wortpaaren de/rm und Einzelvorkommen zählen

für jedes Wortpaar die ‹pointwise mutual information› berechnen und nach den Werten sortieren

Visualisierung von Übersetzungsvarianten

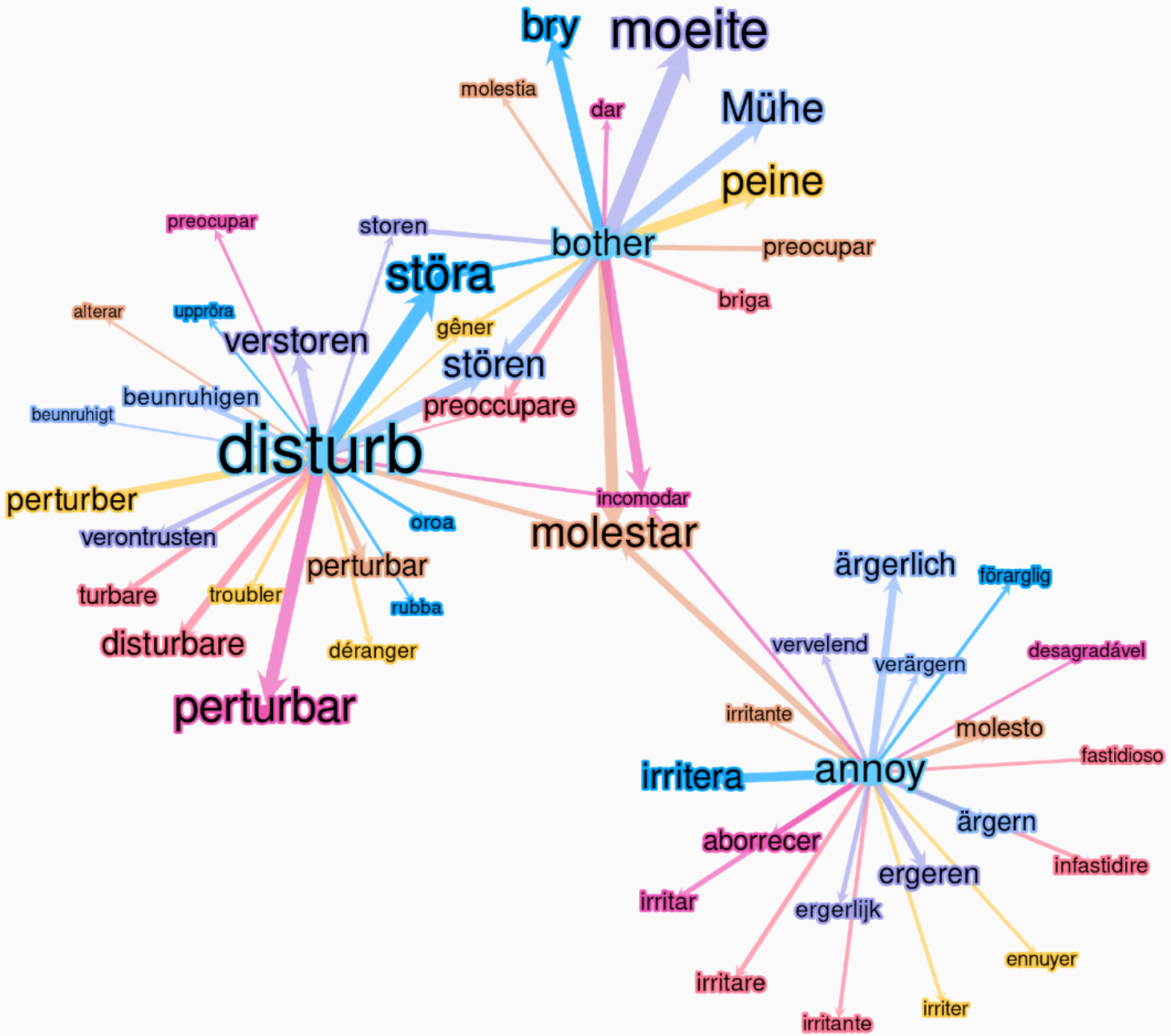
Überschneidung von Alignierungswahrscheinlichkeiten

- bedingte Alignierungswahrscheinlichkeit aus Korpus extrahiert
- Verbindung zwischen zwei Wörtern (Lemmata) mittels gemeinsamer Übersetzungsvarianten[3]

absolute Häufigkeit		relative Häufigkeit	
f_a(en: cow|es: vaca)	= 305	p_a(en: cow|es: vaca)	$\approx 82\,\%$
f_a(en: cattle|es: vaca)	= 43	p_a(en: cattle|es: vaca)	$\approx 12\,\%$
f_a(en: beef|es: vaca)	= 4	p_a(en: beef|es: vaca)	$\approx 1\,\%$

[3] mit einem Filter, um Fehlalignierungen möglichst nicht mitzuzählen

Beispiel →


bry
moeite
molestia
dar
Mühe
peine
preocupar
storen
bother
preocupar
störa
briga
alterar
upprôra
gêner
verstoren
stören
beunruhigen
preoccupare
beunruhigt
disturb
perturber
incomodar
oroa
molestar
verontrusten
perturbar
turbare
troubler
rubba
disturbare
déranger
perturbar
ärgerlich
förarglig
vervelend
verärgern
desagradável
irritante
molesto
fastidioso
irritera
annoy
ärgern
aborrecer
infastidire
ergeren
irritar
ergerlijk
ennuyer
irritare
irriter
irritante

Anwendungsfälle

- False Friends (verschiedene Sprachen)
  - (es) entender & (fr) entendre
  - (en) deception & (es) decepción
  - (en) assist & (es) asistir
- Quasi-Synonyme (gleiche Sprache)
  - (es) solicitar & (es) pedir & (es) rogar
  - (de) steigen & (de) ansteigen
  - (en) disturb & (en) bother & (en) annoy
- Übersetzungsfehler
  - (en) July & (es) julio

Fehlübersetzungen

Fehlübersetzungen

in Multilingwis

Fragen bis hierhin?


Danke für Ihre Aufmerksamkeit.


Thank you for your attention.


Gracias por su atención.


Kiitoksia mielenkiinnostanne.


Merci pour votre attention.


Grazie dell’ attenzione.


Dziękuję państwu za uwagę.

Literatur

- Brown, P. F., V. J. Della Pietra, S. A. Della Pietra und R. L. Mercer (1993). „The Mathematics of Statistical Machine Translation: Parameter Estimation“. In: *Special issue on using large corpora: II* 19.2, S. 263–311.
- Dou, Z.-Y. und G. Neubig (2021). „Word Alignment by Fine-tuning Embeddings on Parallel Corpora“. In: *Conference of the European Chapter of the Association for Computational Linguistics (EACL)*.

References ii

- Dyer, C., V. Chahuneau und N. A. Smith (2013). „A Simple, Fast, and Effective Reparameterization of IBM Model 2“. In: *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*. Association for Computational Linguistics (ACL), S. 644–649.
- Koehn, P. (2010). *Statistical Machine Translation*. Cambridge University Press.
- Liang, P., B. Taskar und D. Klein (2006). „Alignment by Agreement“. In: *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*. Association for Computational Linguistics (ACL), S. 104–111.

References iii

- Och, F. J. und H. Ney (2003). „A Systematic Comparison of Various Statistical Alignment Models“. In: *Computational Linguistics* 29.1, S. 19–51.
- Östling, R. und J. Tiedemann (2016). „Efficient word alignment with Markov Chain Monte Carlo“. In: *Prague Bulletin of Mathematical Linguistics* 106, S. 125–146.
- Sabet, M. J., P. Dufter, F. Yvon und H. Schütze (2020). „Simalign: High quality word alignments without parallel training data using static and contextualized embeddings“. In: *arXiv preprint arXiv:2004.08728*.