# Einführung in die Computerlinguistik: Maschinelle Übersetzung

WS 2010/2011

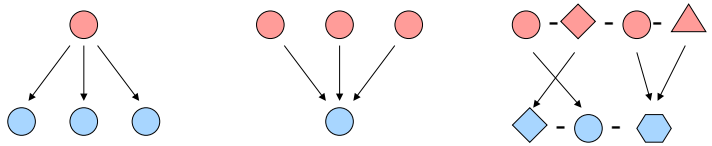
Manfred Pinkal

## Können Computer übersetzen?

Goethe
Babel Fish
Google

- Über allen Gipfeln ist Ruh. In allen Wipfeln spürest du kaum einen Hauch
- Over all summits is rest. In all treetops you do not feel breath.
- Über allen Gipfeln ist Rest. In allen Treetops glauben Sie nicht Atem.

## Übersetzungsäquivalenz: Elementare Probleme



## Können Computer übersetzen?

- Vollautomatische, qualitativ hochwertige Übersetzungen werden auf absehbare Zeit nicht möglich sein
  - insbesondere gilt das für Lyrik
- aber – approximative Übersetzung von Gebrauchstexten ist durchaus möglich und sinnvoll
- Was als Übersetzung akzeptabel ist, hängt ab
  - vom Zweck: Juristisch verbindliche Übersetzung/ Scannen auf Relevanz
  - vom Leidensdruck: Englisch vs. Japanisch
- Japan Patent Office

## Ein Beispiel: Verbmobil

- Dialogübersetzung
- Eingabe durch Mikro oder Telefon
- Domäne: Termin- und Reiseplanung
- Sprachen: Deutsch- Englisch-Japanisch
- Sprachumfang: 10000 Wörter D,E; 2500 Wörter Japanisch
- Zeitraum: 1992-2000
- Volumen: 110 Mio. DM/ ca. 60 Mio. €

## Lexikalische Mehrdeutigkeit

- Homonymie:
  - engl. *rest* → *Rest/Ruhe*
  - *dt. Warte"→ wait/control room*
- Polysemie:
  - *breath → Atem/Hauch*
  - *Termin → appointment / time slot*
- *"gehen"* in Verbmobil (6 von 15 Varianten)
  - *Gehen wir ins Theater?* – gehen_move
  - *Gehen wir essen?* – gehen_act
  - *Mir geht es gut.* – gehen_feel
  - *Es geht um einen Vertrag.* – gehen_theme
  - *Das Treffen geht von 3 bis 5.* – gehen_last
  - *Geht es bei Ihnen am Montag?* – gehen_passen

## Ambiguitätsauflösung

... durch satzinternen Kontext (Sortenbeschränkungen)

- *Wir treffen uns vor dem Frühstück*
  → *before*
- *Wir treffen uns vor dem Hotel*
  → *in front of*

Aber:

- *Wir treffen uns nach Hamburg*
  → *?*

## Ambiguitätsauflösung

*... durch den Diskurskontext*

- *Geht es bei Ihnen?*
- *Wo sollen wir uns treffen? Geht das bei Ihnen? → at your place*
- *Sollen wir uns am Fünften treffen? Geht das bei Ihnen? → for you*

## Ambiguitätsauflösung

... durch Weltwissen:

- In der Zukunft werden wir Maschinen entwickeln, die immer mehr auf ihre Umwelt reagieren und in der Lage sind, ihren Betrieb an wechselnde Bedingungen anzupassen

## Ambiguitätsauflösung

... durch Weltwissen:

- In der Zukunft werden wir Maschinen entwickeln, die immer mehr auf ihre Umwelt reagieren und in der Lage sind, ihren Betrieb an wechselnde Bedingungen anzupassen

LEO

## Jenseits von Mehrdeutigkeit

- Mehrwortausdrücke /Idioms/ Kollokationen
  - *Karten geben*
    → *to deal cards*
  - *eine Prüfung ablegen*
    → *to take an exam*
  - *eine Prüfung abnehmen*
    → *to give an exam*
  - *den Fahrschein entwerten*
    → *to validate the ticket*
- Sprachspezifische, konventionelle Kookkurenz von Wörtern, die gelernt bzw. im Lexikon explizit vorgegeben werden muss – i.d.R. keine semantische Mehrdeutigkeit

## Lexikalische Granularität

- *I will go to Hamburg tomorrow.*

  *→ fahren/fliegen*
- *Ich fahre mit der Bahn nach Hamburg. In Frankfurt muss ich umsteigen.*

  *→ change trains*
- *Ich fliege nach Hamburg. In Frankfurt muss ich umsteigen.*

  *→ change planes*

## Systematische Granularitäts-Unterschiede

- Geschlechtsspezifische Personenbezeichnungen im Deutschen
  - *doctor → Arzt / Ärztin*
  - *teacher → Lehrer / Lehrerin*
- Präsens und Futur im Englischen
  - *Ich fahre nach Hamburg → I am going / I will go to Hamburg*
- Verbaspekt:
  - *Simple Present/ Progressive im Engl.*
  - *Vollendete/unvollendete Form im Russ.*

## Granularität D/E - J

Deutsch/Englisch → Japanisch

- J: Höflichkeitsformen
- J:Topikmarkierung (gegeben/ neu)

Japanisch → Deutsch/Englisch

- D: Artikel/ Definitheit (bestimmt/ unbestimmt), J hat keine Artikel
- J: „Null-Anapher“: Satzteile werden tendenziell weggelassen, wenn aus dem Kontext erschließbar ("Null-Anapher")

## Beispiel

*"Termin ausgemacht?"*

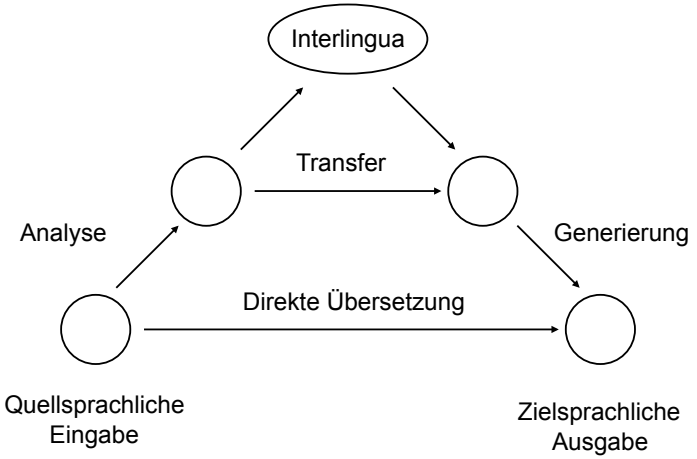
*Yotei-wa kimemashita ka. → (Er mit Ihnen)*

*Go-yotei wa okimeni narimashita ka. → (Sie mit ihm)*

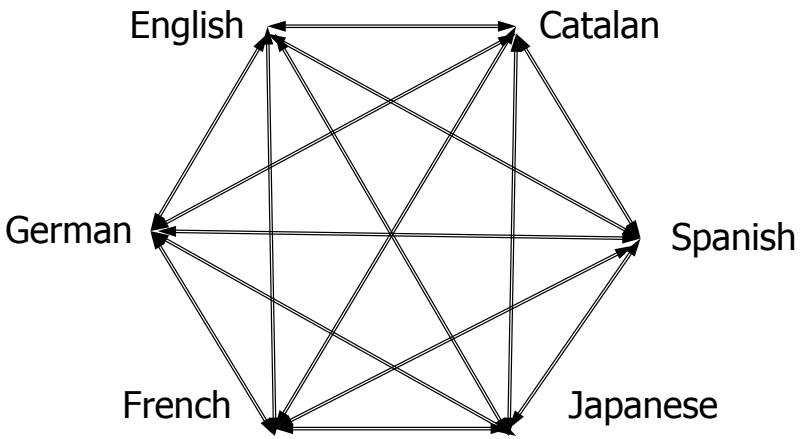
## Frameworks für die MÜ

- Wissensbasierte MÜ
- Statistische MÜ

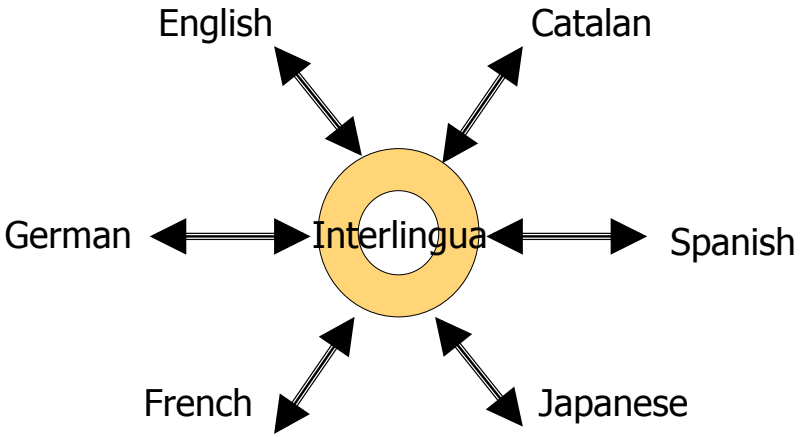
## Das "Vauquois-Dreieck"




## Das Transfer-Modell




## Das Interlingua-Modell



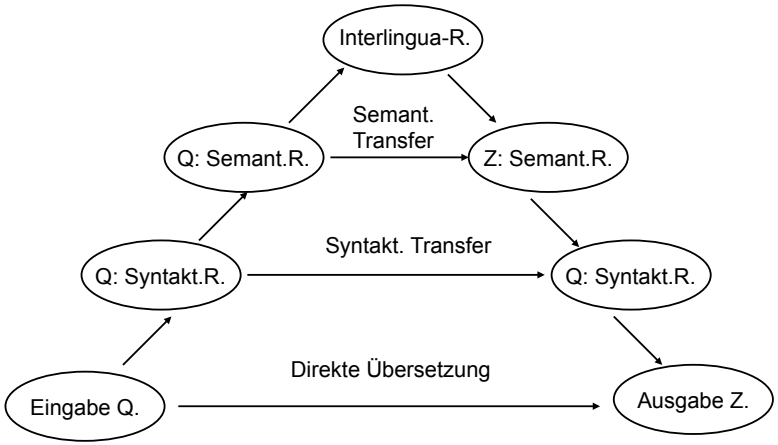
## Interlingua und Transfer

- Die Übersetzung in die / aus der Interlingua muss für jede neue Sprache nur (je) einmal bereitgestellt werden. – Wenn im Transfermodell zu n Sprachen eine neue hinzukommt, müssen 2n neue Übersetzungsrichtungen bereitgestellt werden.
  - Beispiel: Durch die letzten EU-Erweiterungen wachsen die offiziellen EU-Sprachen von 11 auf 23 an. Statt 110 Übersetzungspaaren benötigt man 506.
- Interlingua muss extrem feingranular sein, da alle Unterschiede in allen Sprachen darstellbar sein müssen. Das erfordert bei der Übersetzung einen immer gleich hohen und für viele, insbesondere eng verwandte Sprachpaare unnötigen Übersetzungsaufwand.
  - Beispiel: Übersetzung D-E benötigt keine detaillierte Bestimmung von Höflichkeitsinformation

## Interlingua und Transfer

- Der syntaktische Transfer ist hoch komplex: Unterschiedliche Wortstellung, unterschiedliche Konstruktionen ("Head Switching"-Problem)
  - *Ich schwimme gern*
  - *I like to swim*
- Kompromiss zwischen Interlingua und syntaktischem Transfer ist semantisches Transfer-Modell

## Das "Vauquois-Dreieck", erweitert




## Wissensbasierte MÜ

- Techniken: Stemmer/Morphologien, Grammatiken, Lexika für Quell- und Zielsprache, Transferregeln, sprachunabhängige Ontologien, Weltwissen, Inferenzregeln
- Probleme
  - Abdeckung: Ungeheure Vielfalt von syntaktischen und semantischen Phänomenen und Übersetzungsäquivalenten
  - Präzision: Ambiguität und Granularitätsunterschiede
- Klassisches (und noch immer aktuelles) Beispiel:
  - SYSTRAN (Babel Fish)

# Statistische MÜ

- Gesucht: Die wahrscheinlichste zielsprachliche Wortfolge, gegeben ein quellsprachlicher Satz:

$$\max_Z P(Z \mid Q)$$

- Das erinnert an das Problem der Spracherkennung: Gesucht ist die wahrscheinlichste Wortfolge, gegeben eine Folge akustischer Merkmale:

$$\max_W P(W \mid O)$$

# Wie bestimmen wir *P(Z|Q)*?

- Quellsprachlicher Satz: „Symptom“
- Zielsprachlicher Satz : „Ursache“ – Welchen Satz in der Zielsprache hat der Sprecher/ Schreiber „eigentlich“ ausdrücken wollen
- Bayes-Regel :

$$P(W \mid O) = \frac{P(O \mid W) \cdot P(W)}{P(O)} \qquad P(Z \mid Q) = \frac{P(Q \mid Z) \cdot P(Z)}{P(Q)}$$

$$\max_W P(W \mid O) = \max_W \frac{P(O \mid W) \cdot P(W)}{P(O)} \qquad \max_Z P(Z \mid Q) = \max_Z \frac{P(Q \mid Z) \cdot P(Z)}{P(Q)}$$

$$= \max_W P(O \mid W) \cdot P(W) \qquad = \max_Z P(Q \mid Z) \cdot P(Z)$$

# Sprachmodell: n-Gramm-Technik

$$\max_Z P(Z \mid Q) = \max_Z P(Q \mid Z) \cdot P(Z)$$

- Wie berechnen wir *P(Z) = P($w_1 w_2 \ldots w_n$) ?*
- Kettenregel erlaubt die Reduktion von *P($w_1 w_2 \ldots w_n$)* auf bedingte Wahrscheinlichkeiten:

  *P($w_1 w_2 \ldots w_n$)*

  *= P($w_1$)*P($w_2|w_1$)*P($w_3|w_1 w_2$)* ... *P($w_n|w_1 w_2 \ldots w_{n-1}$)*

  Sparse-Data-Problem!
- Beispiel Bigramm-Approximation:
  - *P($w_n|w_1 w_2 \ldots w_{n-1}$)≈ P($w_n|w_{n-1}$)*

    *P($w_1 w_2 \ldots w_n$) ≈ P($w_1$)*P($w_2|w_1$)*P($w_3|w_2$)* ... P($w_n|w_{n-1}$)*

# Übersetzungsmodell

$$\max_Z P(Z \mid Q) = \max_Z P(Q \mid Z) \cdot P(Z)$$

- Problem: Zuordnung von quellsprachlichen und zielsprachlichen Wörtern: Tilgungen, Einfügungen, Mehr-zu-Eins-Entsprechungen, Wortstellung
- Wortalignierung auf der Grundlage eines großen Lexikons, das alle Übersetzungsalternativen enthält, mithilfe eines Alignierungsalgorithmus (z.B. Levenshtein-Distanz)
- Alignierung großer Parallelkorpora (z.B. Europarl: Akten des Europäischen Parlaments)
- Training von Übersetzungsmodellen auf dem alignierten Parallelkorpus: Frequenzermittlung quellsprachlicher Wortentsprechungen.

# Statistische MÜ

- Liefert im Allgemeinen Resultate, die den besten wissensbasierten Systemen vergleichbar sind.
- Systeme lassen sich vergleichsweise schnell trainieren und auf neue Sprachen/ Domänen adaptieren.
- Für bestimmte Anwendungen sehr hochwertige Übersetzungen, weil Muster aus Parallelkorpora komplett übernommen werden können.
- Beispiel: Google Translate