# Einführung in die Computerlinguistik: Maschinelle Übersetzung

WS 2008/2009

Manfred Pinkal

## Gliederung

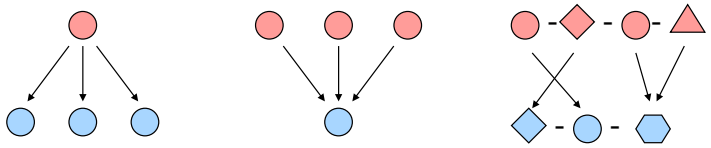
- Können Computer übersetzen?
- Was sind die Probleme?
- Was sind die möglichen Lösungen?
- Mensch-Computer-Interaktion bei der Übersetzung

## Können Computer übersetzen?

Goethe
Babel Fish
Google

- Über allen Gipfeln ist Ruh. In allen Wipfeln spürest du keinen 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 nicht für Dichtung und rechtlich relevante Dokumente).

## Gliederung

- Können Computer Übersetzen?
- Was sind die Probleme?
- Wie sehen Lösungen aus?
- Mensch-Computer-Interaktion bei der Übersetzung

## 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 (2)

- 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 (2)

- 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

## Globaler Kontext

- *Geht es bei Ihnen?*
- *Wo sollen wir uns treffen? Geht das bei Ihnen? → at your place*
- *Sollen wir uns am 5. treffen? Geht das bei Ihnen? → for you*

## 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*
- *Simple Present/ Prgressive im Engl.*
- *Vollendete/unvollendete Form im Russ. (Aspekt)*

## Granularität D/E - J

Deutsch/Englisch → Japanisch

- Höflichkeitsformen (unter anderem)

Japanisch → Deutsch/Englisch

- Japanisch hat keine Artikel
- Satzteile (Subjekt, Objekte) werden tendenziell weggelassen, wenn aus dem Kontext erschließbar ("Null-Anapher")

## Beispiel

*"Termin ausgemacht?"*

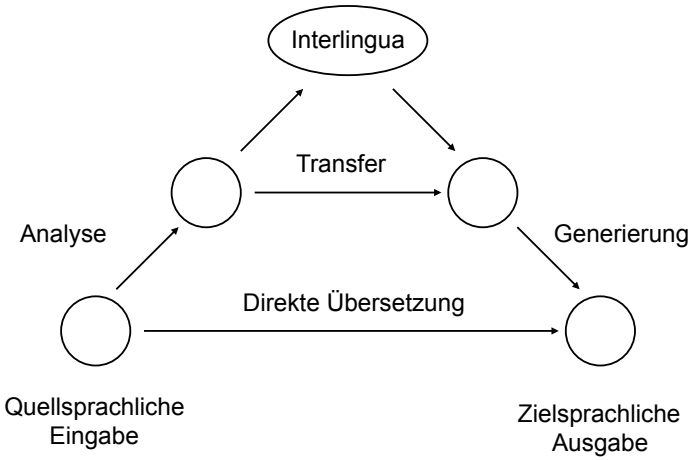
*Yotei-wo kimemashita ka. → (Er mit Ihnen)*

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

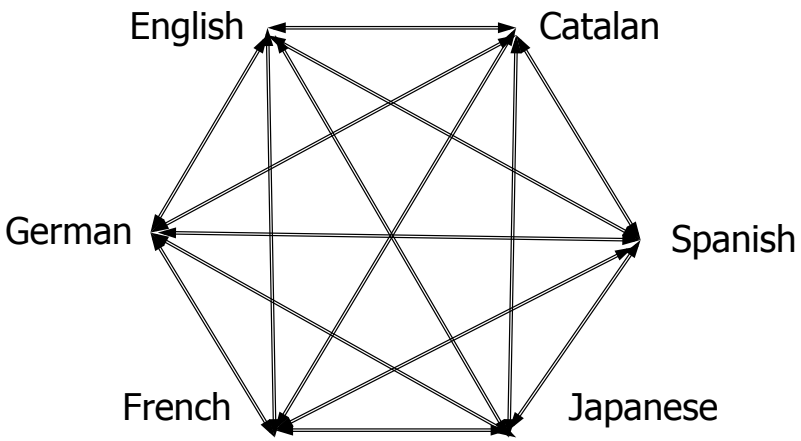
## Gliederung

- Können Computer übersetzen?
- Was sind die Probleme?
- Wie sehen Lösungen aus?
- Mensch-Computer-Interaktion bei der Übersetzung
- Was gibt es sonst?

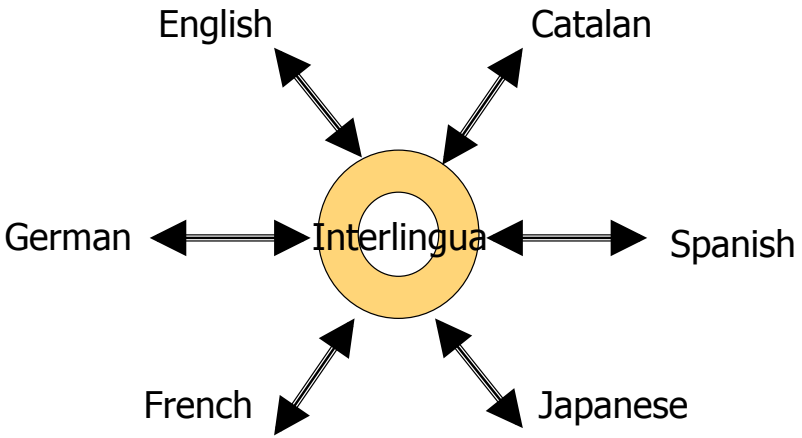
## Das "Vauquois-Dreieck"



## Das Transfer-Modell




## Das Interlingua-Modell



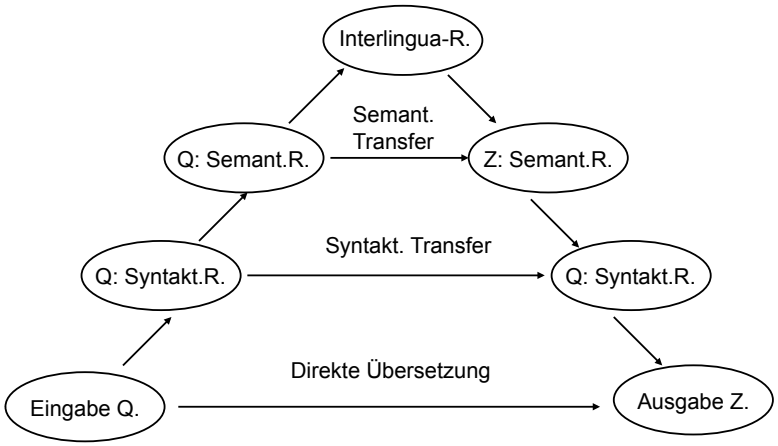

## Interlingua und Transfer

- 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 letzte EU-Erweiterung wachsen die offiziellen EU-Sprachen von 11 auf 20 an. Statt 110 Übersetzungspaaren benötigt man 380.
- 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




## Können Computer übersetzen?

- Keine qualitativ hochwertige Vollübersetzung,
- aber – approximative Übersetzung von Gebrauchstexten ist durchaus möglich und sinnvoll

## Alternative Zugänge zur MÜ

- Wissens- und regelbasierte Verfahren
  - Techniken: Stemmer/Morphologien, Grammatiken, Lexika für Quell- und Zielsprache, Transferregeln, sprachunabhängige Ontologien, Weltwissen, Inferenzregeln
  - Problem: Es ist extrem schwierig und teuer, gute Abdeckung und damit einheitliche und akzeptable Übersetzungsqualität zu erreichen.

## Alternative Zugänge zur MÜ

- Statistische Verfahren:
  - Suche nach der zielsprachlichen Formulierung (T), die Treue zum quellsprachlichen Original (S) und linguistische/stilistische Güte in bestmöglicher Weise verbindet.
  - Beide qualitativen Maße werden durch korpusbasierte Wahrscheinlichkeiten modelliert:

# Alternative Zugänge zur MÜ

- Beide qualitativen Maße werden durch korpusbasierte Wahrscheinlichkeiten modelliert:
  - Ein Satz T der Zielsprache ist umso besser, je höher die Wahrscheinlichkeit ist, dass er tatsächlich genutzt wird:

    P(T), berechnet auf der Grundlage von Korpusfrequenz in der Zielsprache.
  - Ein Satz T der Zielsprache umso originalgetreuer, je größer die Wahrscheinlichkeit ist, dass T genau in den quellsprachlichen Satz S überführt wird.

    P(S|T) wird berechnet auf der Grundlage von Korpusfrequenzen in bilingualen, alignierten Korpora.
  - Die beste Übersetzung von S in T ist diejenige, die

    P(S|T)*P(T) maximiert.

# Alternative Zugänge zur MÜ

- Wissens- und regelbasierte Verfahren
- Statistische Verfahren
- Übersetzung am Beispiel ("Translation by Example")/ Web-basierte Übersetzung: Im Web stehen immer häufiger Dokumente in verschiedenen Sprachen zur Verfügung. Solche bilingualen Dokumentpaare können identifiziert und auf Satz- und soweit möglich auf Wortebene aligniert und tragen so zu einem riesigen bilingualen Korpus bei. Zugriffheuristiken finden für einen gegebenen quellsprachlichen Satz das Gegenstück mit der größten Übereinstimmung (fuzzy match). Der zielsprachliche Satz wir teilweise übernommen und wo nötig modifiziert.