

2.1.3 Definition von Recall und Precision

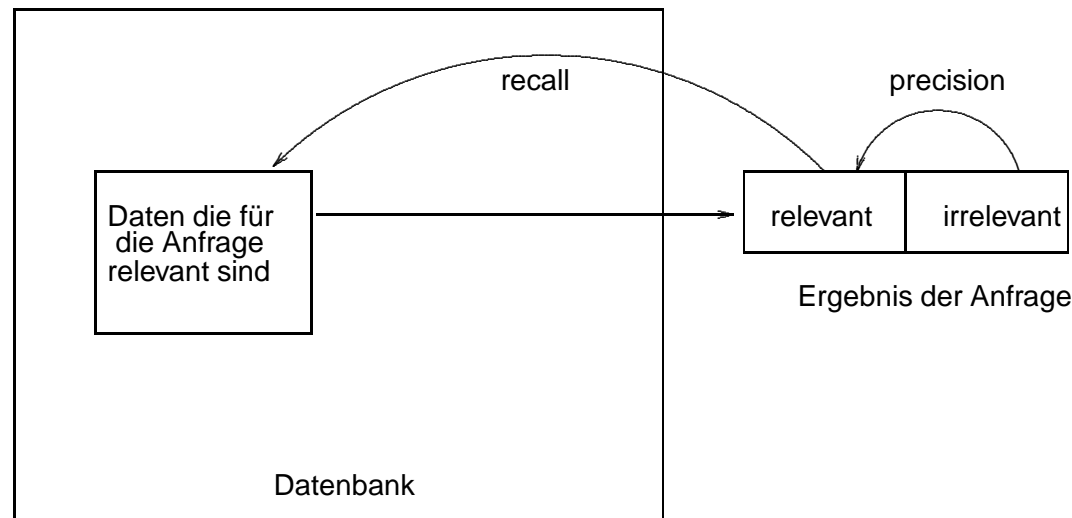
Recall mißt, wieviele der relevanten Texte im Verhältnis zum Gesamtbestand gefunden wurden.

Beispiel: Eine Datenbank aus mehreren Millionen Texten enthält 100 Texte, die für eine gegebene Fragestellung relevant sind. Wenn die Anfrage 75 Texte liefert, von denen 50 relevant und 25 irrelevant sind, dann ergibt sich ein *Recall* von $50 : 100 = 50\%$.

Precision mißt, wieviele relevante Texte im Suchergebnis enthalten sind.

Beispiel: Eine Anfrage hat 75 Texte ergeben, von denen 50 für den Benutzer relevant und 25 irrelevant sind. Dann ergibt sich eine *Precision* von $50 : 75 = 66,7\%$.

2.1.4 Korrelationen zur Bemessung von Recall und Precision



2.2 Einsatz grammatischen Wissens

2.2.1 Linguistische Methoden der Optimierung

A. Verarbeitung der Anfrage

- Automatische *query expansion*
 - (i) Die Suchwörter der Anfrage werden in ihre möglichen Flexionsformen ‘explodiert’. Dann wird die Datenbank mit allen Formen der jeweiligen Paradigmen durchsucht.
 - (ii) Über einen Thesaurus werden den Suchwörtern sämtliche Synonyme, Hypernyme und Hyponyme zugeordnet, die dann bei der Suche berücksichtigt werden – möglicherweise ebenfalls in explodierter Form.
 - (iii) Die syntaktischen Strukturen der Anfrage, z. B. **betrunkenen Fahrer**, werden automatisch in äquivalente Versionen *transformiert*. Hier z. B. in: **Fahrer, der betrunken war**.
- Interaktive Anfrageoptimierung

Vor Beginn der eigentlichen Suchprozedur wird dem Benutzer die bearbeitete Anfrage vorgelegt, um unsinnige Ergebnisse der automatischen Expansion zu eliminieren und die Fragestellung neu einzugrenzen.

B. Differenzierung der Indexierung

- **Buchstabenbasierte Indexierung:**

Diese technologische Standardmethode erlaubt es, sämtliche Vorkommen einer Buchstabenfolge in der Datenbank zu finden.

- **Morphologiebasierte Indexierung:**

Beim Einlesen wird jede Wortform im Text analysiert und ihrer Grundform zugeordnet. Mit dieser Information wird ein Index aufgebaut, der es erlaubt, zu einer bestimmten Grundform sämtliche entsprechenden Wortformen in der Datenbank zu finden.

- **Syntaxbasierte Indexierung:**

Beim Einlesen wird der Text syntaktisch geparkt (d.h. automatisch bezüglich seiner grammatikalischen Struktur analysiert), wobei viele morphologische Ambiguitäten eliminiert werden. Aus den grammatischen Analysen wird ein Index aufgebaut, mit dem man sämtliche Vorkommen einer syntaktischen Konstruktion sowie syntaktische Variationen einer Konstruktion finden kann.

- **Konzeptbasierte Indexierung:**

Beim Einlesen wird der Text semantisch und pragmatisch analysiert. Dabei werden nicht nur semantische Ambiguitäten eliminiert und domänenabhängige Spezialverwendungen erschlossen, sondern es wird zudem über den Inhalt des Textes ein Index aufgebaut, mit dem man sämtliche Vorkommen eines bestimmten Konzepts in der Datenbank finden kann.

C. Nachbearbeitung der gefundenen Daten

- Der niedrige Precision-Wert bei einer allgemein gehaltenen Fragestellung kann durch eine automatische Nachbearbeitung der gefundenen Daten ausgeglichen werden. Da die Rohergebnisse der Suche im Vergleich zur Gesamtmenge der Texte verhältnismäßig klein sind, können sie nach dem Anfragevorgang geparkt und auf ihren Inhalt untersucht werden. Anschließend werden nur die Texte an den Benutzer ausgegeben, die sich in der Nachbearbeitung als relevant erwiesen haben.

2.3 *Smart oder Solid Solutions?*

2.3.1 Smart solutions

vermeiden schwierige, kostspielige oder theoretisch ungelöste Aspekte der anstehenden Aufgabe, wie z. B.

- Weizenbaum's Eliza Programm, das natürliche Sprache zu verstehen scheint, es aber nicht tut.
- Statistisch-basiertes Tagging, das Wortarten erraten kann.
- Direkter und Transfer-Ansatz bei der maschinellen Übersetzung, welche ein Verstehen des Quelltextes vermeiden.

2.3.2 Solid solutions

zielen auf ein vollständiges theoretisches und praktisches Verstehen der Phänomene. Bei Anwendungen wird auf fertige *off-the-shelf* Komponenten zurückgegriffen, wie z. B.

- eine automatische Wortformanalyse, die auf einem online Lexikon und einem regelbasierten Morphologieparser basiert, der alle Aspekte der Flexion, Derivation und Komposition der Sprache behandelt
- ein syntaktischer Parser für freie Texte, der die automatische Wortformanalyse verwendet
- eine semantische Interpretation der syntaktischen Analyse, die die wörtlichen Bedeutungen der Ausdrücke ableitet
- eine pragmatische Interpretation in bezug auf einen Verwendungskontext, die die Sprecherbedeutung der Äußerung ableitet.

2.3.3 Vergleich

- Solange fertige *off-the-shelf* Komponenten nicht zur Verfügung stehen, scheinen *smart solutions* zunächst billiger und schneller zu sein. *Smart solutions* sind jedoch teuer in der Wartung und ihre Leistung kann nicht wesentlich verbessert werden.
- Die Entwicklung von Grammatikkomponenten im Rahmen einer *solid solution* sind eine langfristige Investition. Die Komponenten können in einer Vielzahl von Anwendungen wiederverwendet werden, wobei Verbesserungen der Grammatik direkt zu Verbesserungen bei den Anwendungen führen.

2.3.4 Wahl zwischen *smart* oder *solid solution* hängt auch von der Anwendung ab

- Eine *smart solution*, die bei einer großen Datenbank einen 70% Recall bringt, liefert wesentlich mehr als das, was die Benutzer mit Hilfe menschlicher Arbeitskraft erreichen könnten. Außerdem wissen die Benutzer nicht, welche Texte das System übersehen hat.
- Dagegen sind die Mängel einer 70% akkuraten *smart solution* in der maschinellen Übersetzung für die Benutzer von schmerzlicher Offensichtlichkeit. Zudem gibt es die Alternative menschlicher Übersetzer, die ein wesentlich besseres Ergebnis erzielen können.

2.4 Anfänge der maschinellen Übersetzung

2.4.1 Sprachpaare

Französisch → *Englisch* und *Französisch* ← *Englisch* sind zwei verschiedene Sprachpaare.

2.4.2 Formel zur Berechnung der Zahl von Sprachpaaren

$n \cdot (n - 1)$, wobei $n = \text{Zahl der verschiedenen Einzelsprachen}$

Zum Beispiel muß eine EU mit derzeit $11 \cdot 10 = 110$ Sprachpaaren fertig werden.

2.4.3 Sprachpaare bei der Übersetzung eines französischen Dokuments in der EU

Französisch → Englisch

Französisch → Deutsch

Französisch → Italienisch

Französisch → Holländisch

Französisch → Finnisch

Französisch → Spanisch

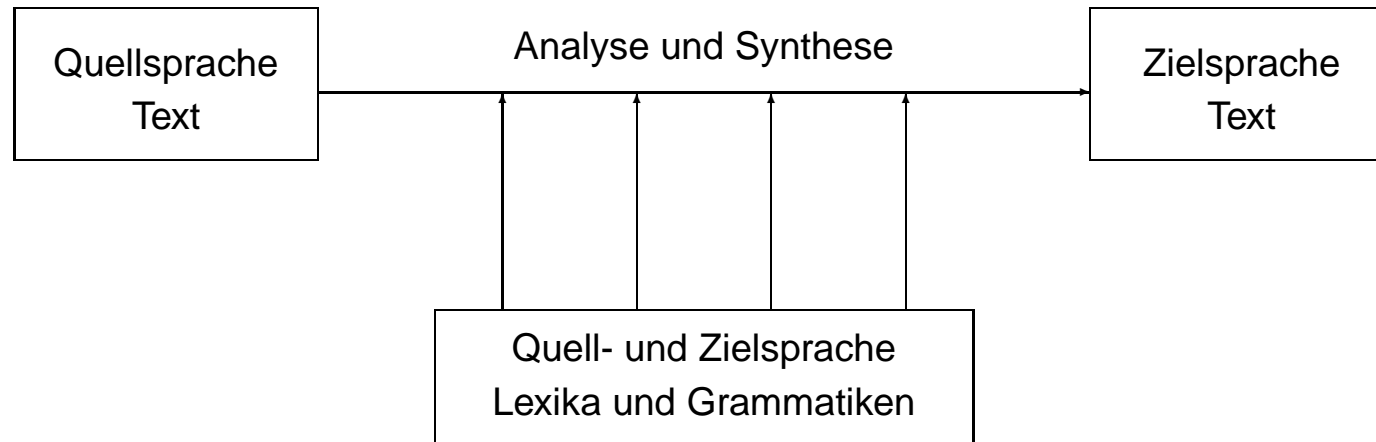
Französisch → Portugiesisch

Französisch → Griechisch

Französisch → Dänisch

Französisch → Schwedisch

2.4.4 Schema der direkten Übersetzung



2.4.5 Was ist FAHQT?

FULLY AUTOMATIC HIGH QUALITY TRANSLATION

2.4.6 Beispiele von automatischen Fehlübersetzungen

Out of sight, out of mind. \Rightarrow *Invisible idiot.*

The spirit is willing, but the flesh is weak. \Rightarrow *The whiskey is alright, but the meat is rotten.*

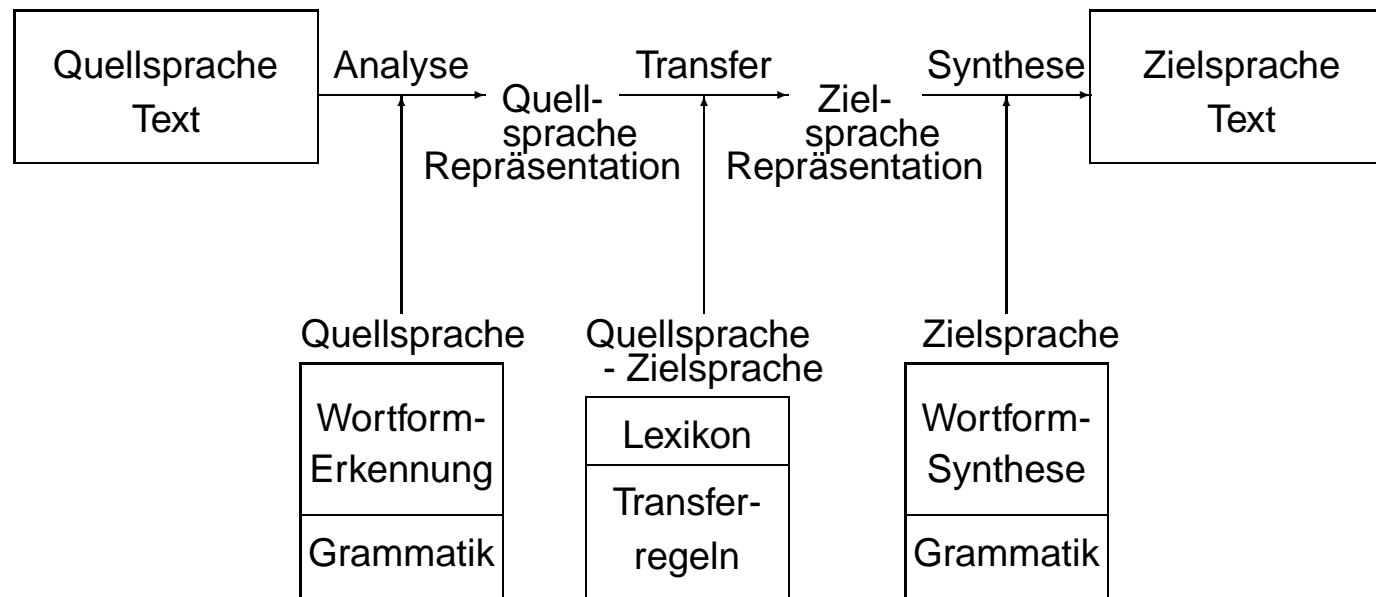
La Cour de Justice considère la création d'un sixième poste d'avocat général. \Rightarrow *The Court of Justice is considering the creation of a sixth avocado station.*

2.4.7 Transfer-Ansatz

versucht eine modulare Trennung

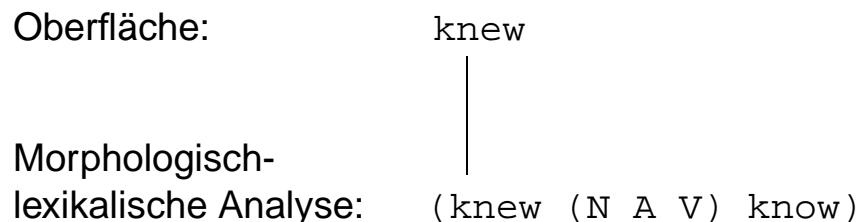
- von Quellsprach-Analyse und Zielsprach-Synthese,
- von linguistischen Daten und Verarbeitungsprozeduren und
- von Lexika für die Quellsprach-Analyse, den Quellsprach-Zielsprach-Transfer und die Zielsprach-Synthese.

2.4.8 Schema des Transfer-Ansatzes



2.4.9 Drei Phasen eines Wortform-Transfers *Englisch-Deutsch*

1. Quellsprach-Analyse der englischen Wortform knew:



Die Quellsprach-Analyse der unanalysierten Oberfläche **knew** ist ein geordnetes Tripel, bestehend aus der Oberfläche, der syntaktischen Kategorie und der Grundform **know**.

2. Quell-Zielsprach-Transfer:

Aus der Grundform der Quellsprach-Analyse werden über ein Lexikon die Grundformen korrespondierender Zielsprach-Wörter erzeugt:

know \implies wissen
kennen

3. Zielsprach-Synthese:

Über die Kategorie der Quellsprach-Form und die Grundformen der korrespondierenden Zielsprach-Wörter werden mittels Zielsprach-Morphologie und -Lexikon die gewünschten Zielsprach-Wortformen erzeugt:

wußte	wüßte	kannte	kennte
wußtest	wüßtest	kanntest	kenntest
wußten	wüßten	kannten	kennten
wußtet	wüßtet	kanntet	kenntet

2.4.10 Nachteile des direkten und des Transfer-Ansatzes

- Jedes Sprachpaar erfordert eine eigene Quell-Zielsprachkomponente.
- Analyse und Synthese beschränken sich auf einzelne Sätze.
- Die Ansätze vermeiden eine semantisch-pragmatische Analyse und versuchen ohne ein Verstehen der Quellsprach-Texte auszukommen.

2.5 Maschinelle Übersetzung heute

2.5.1 Warum Sprachverstehen für die Übersetzung unverzichtbar ist

- Syntaktische Ambiguität der Quellsprache
 1. Nicht kopierfähige Folien schmelzen und verkleben die Maschine.
Folien (schmelzen und verkleben die Maschine)
(Folien schmelzen) und (verkleben die Maschine)
 2. Susanne beobachtete die Yacht mit dem Fernglas.
Susanne beobachtete den Mann mit dem Bart.
 3. The mixture gives off dangerous cyanide and chlorine fumes.
(dangerous cyanide) and chlorine fumes
dangerous (cyanide and chlorine) fumes
- Lexikalische Quell-Zielsprach-Unterschiede
 1. Die Männer ermordeten die Frauen. Sie wurden drei Tage später gefaßt.
Die Männer ermordeten die Frauen. Sie wurden drei Tage später begraben.
 2. know – wissen – savoir
kennen – connaître
 3. The watch included two new recruits that night.

- Syntaktische Quell-Zielsprach-Unterschiede

- Deutsch:

- Auf dem Hof sahen wir einen kleinen Jungen, der einem Ferkel nachlief.
Dem Jungen folgte ein großer Hund.

- Englisch:

- In the yard we saw a small boy running after a piglet.
A large dog followed the boy.
The boy was followed by a large dog.

- Kollokation und Idiom

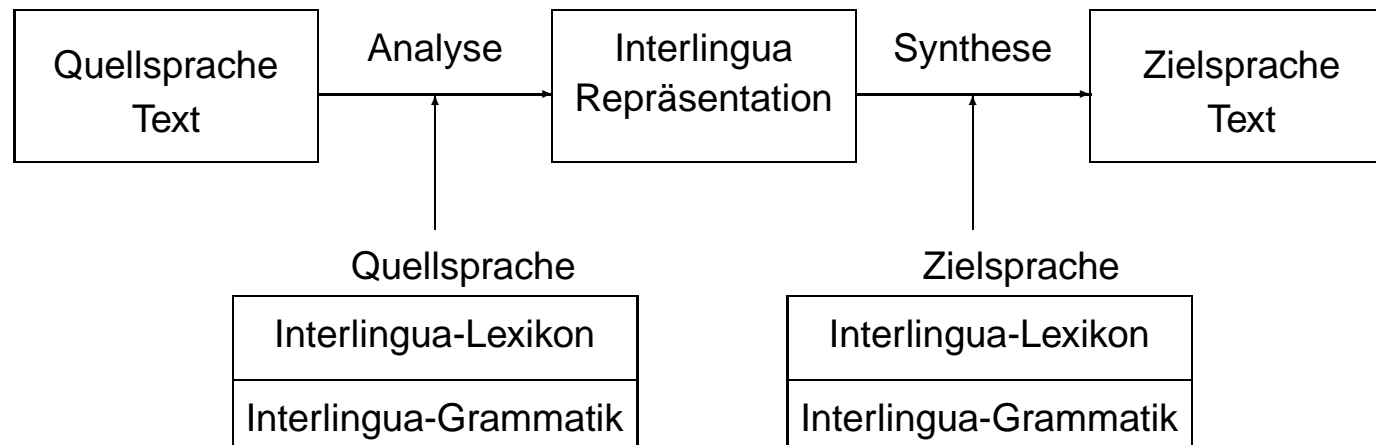
strong current | high voltage (but: *high current | *strong voltage)

bite the dust | ins Gras beißen (but: *bite the grass | *in den Staub beißen)

2.5.2 Praktische Teillösungen in der MÜ

1. **MAT** (*machine aided translation*), also menschliche Übersetzung, aber mit Hilfe von elektronischen Hilfsmitteln wie Online-Lexika, Textverarbeitung, morphologischer Analyse etc.
2. **Rohübersetzungen**, etwa im Rahmen des Transfer-Ansatzes, die stark von Übersetzern nachbearbeitet werden müssen.
3. **Kontrollierte Sprache**, d.h. zwar vollautomatische Übersetzung, aber nur von Texten, deren lexikalische und syntaktische Struktur kanonisch beschränkt ist.

2.5.3 Schema des Interlingua-Ansatzes



2.5.4 Vorteil des Interlingua-Ansatzes

Die Grundstruktur des Interlingua-Ansatzes hat zur Folge, daß für $n(n - 1)$ Sprachpaare nur $2n$ interlinguale Teilsysteme benötigt werden (nämlich n Analyse- und n Synthese-Module).

2.5.5 Kandidaten, die als Interlingua vorgeschlagen worden sind

- eine künstliche Logiksprache,
- eine Mischsprache wie Esperanto, die zwar konstruiert ist, aber wie eine natürliche Sprache funktioniert,
- eine Menge semantischer Urausdrücke, die allen natürlichen Sprachen gemeinsam ist, eine Art universales Vokabular.