Multilinguales Information Retrieval

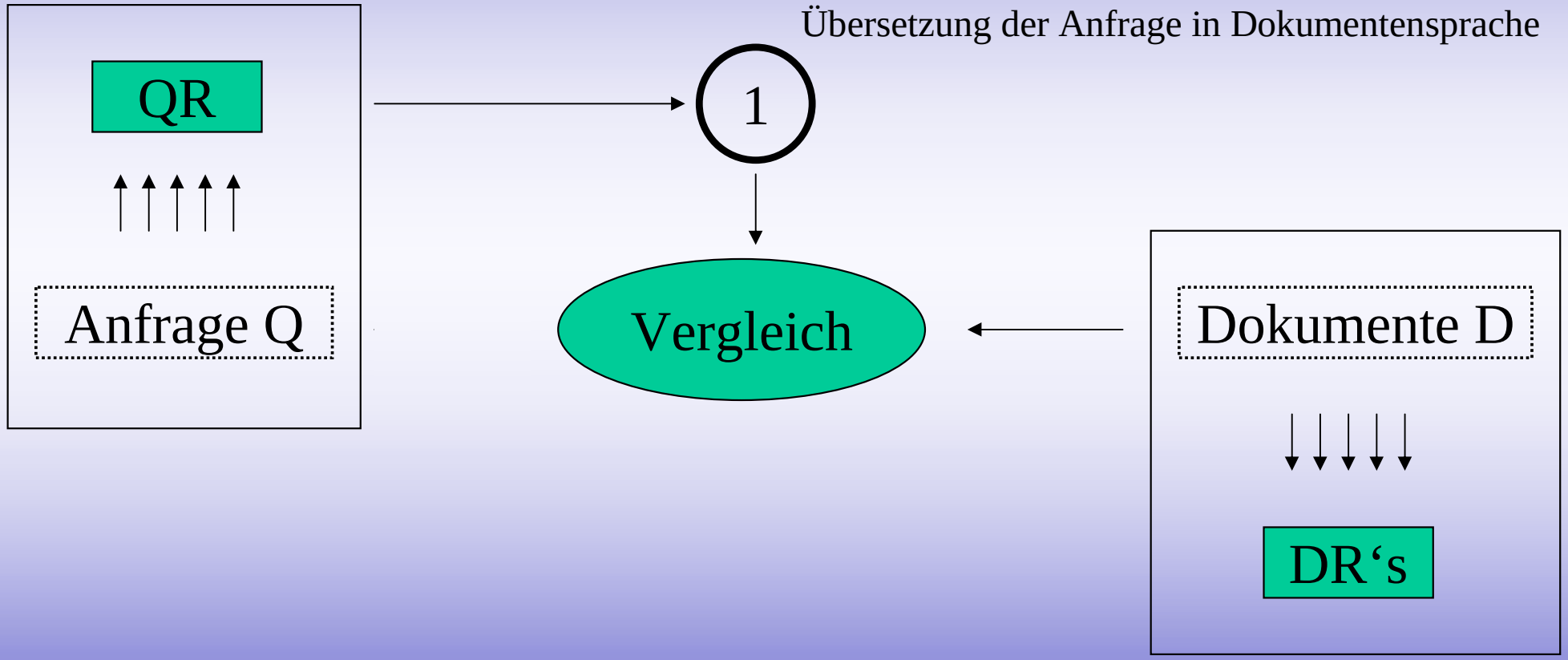
Definition

„IR in einer anderen Sprache als Englisch“
„IR auf einer einsprachigen Dokumentensammlung, die in mehreren Sprachen befragt werden kann.“
„Information Retrieval auf einer Sammlung von Dokumenten in vielen Sprachen, die in vielen Sprachen befragt werden kann.“

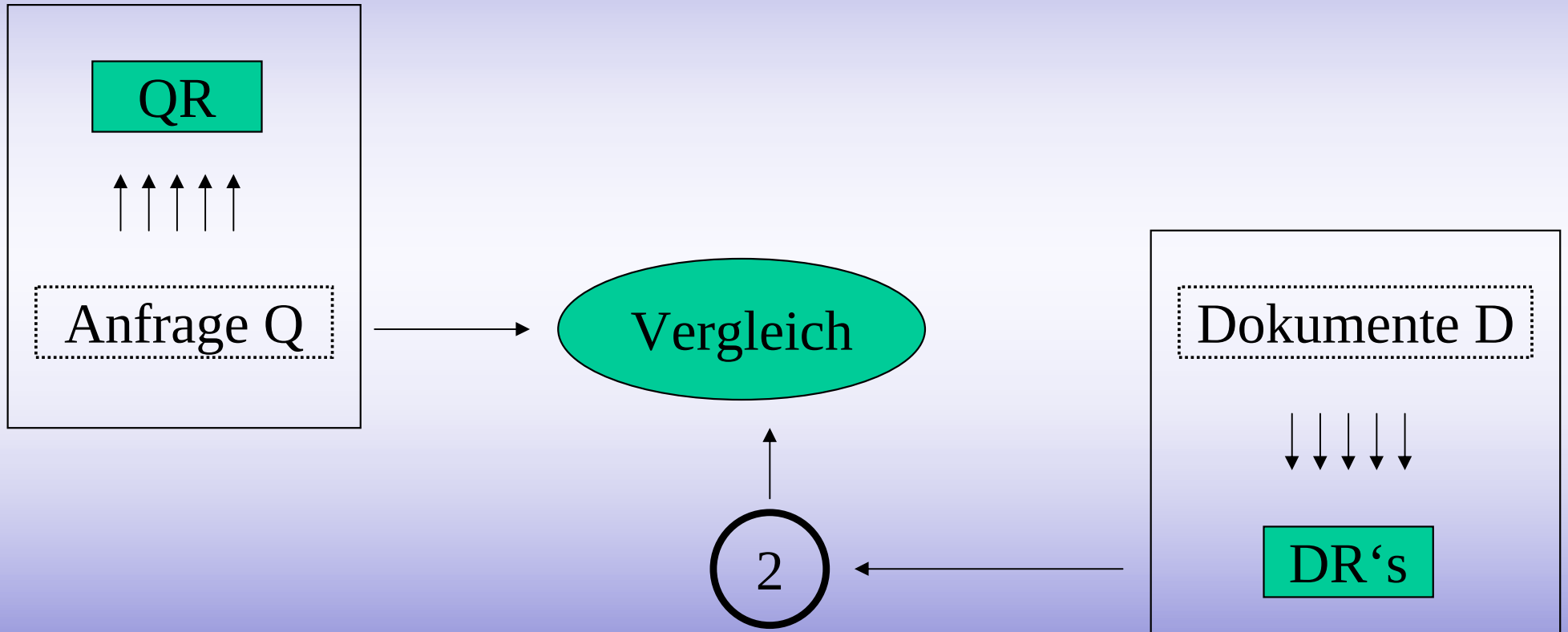
Cross Language Information Retrieval

Informationsgewinnung bei Überschreitung der Sprachgrenze.

Vorgehensweisen



Vorgehensweisen



Übersetzung der Dokumente in Anfragesprache

Übersicht, Sprachverarbeitungsmethoden

Anfrage→Dokumentensprache | Dokumente→Anfragesprache

Anfrageerweiterung, Übersetzung

Erkennung von Wortformen, -arten

Spracherkennung

Maschinelle Übersetzung

Anfrageerweiterung

Benutzer stellt Anfrage mittels Schlüsselwörtern. Sind diese nicht spezifisch genug erhält er eine unübersichtlich große Menge an Dokumenten. Um dies zu vermeiden wird die Anfrage um einige Terme erweitert.

Zwei Methoden zur Anfrageerweiterung :

1. Thesaurusbenutzung
2. Korpusbenutzung

Thesaurus

Ontologie *(Wissenssammlung)*

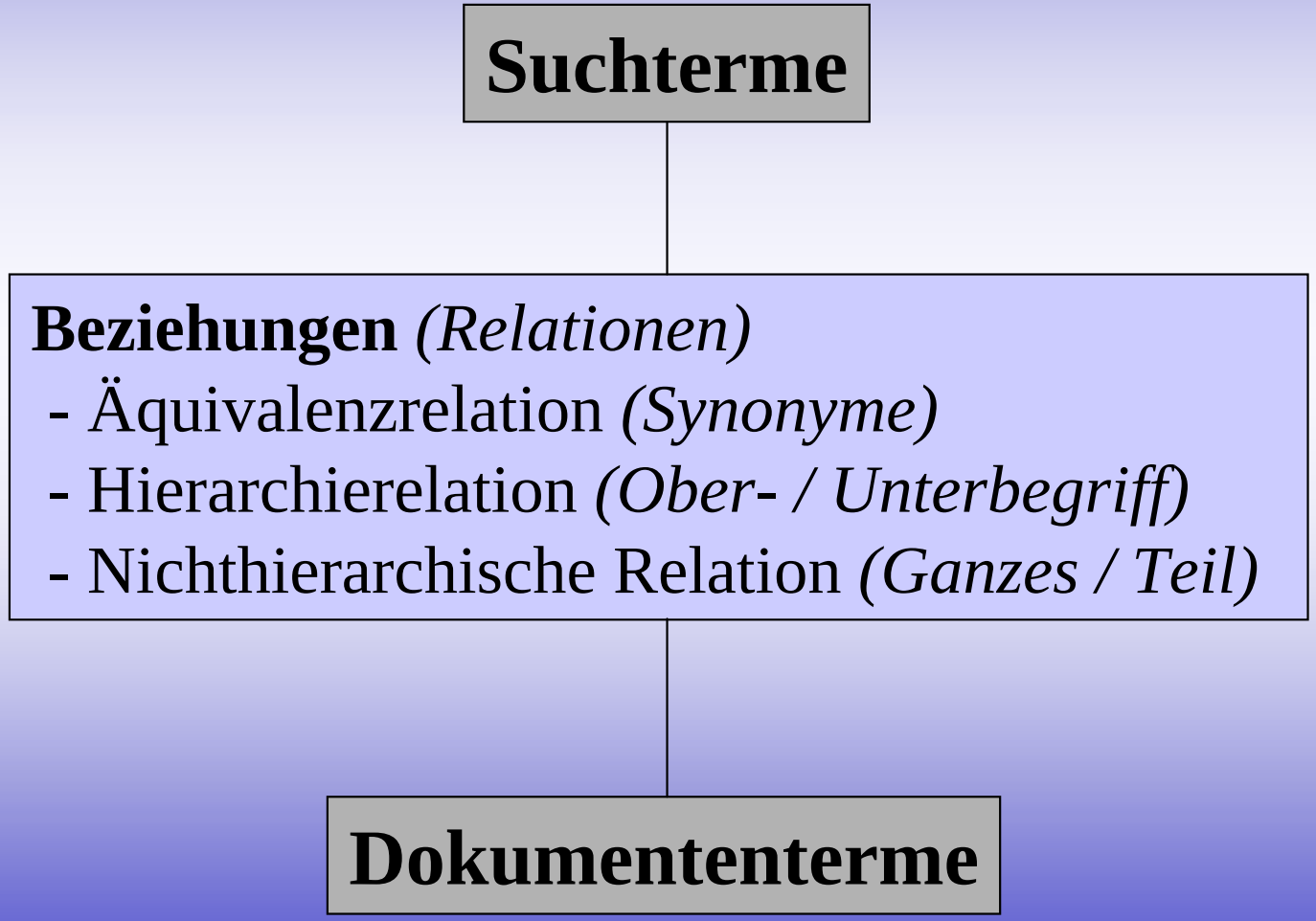
Strukturierte Konzeptliste

Deskriptor *(Bezeichner)*

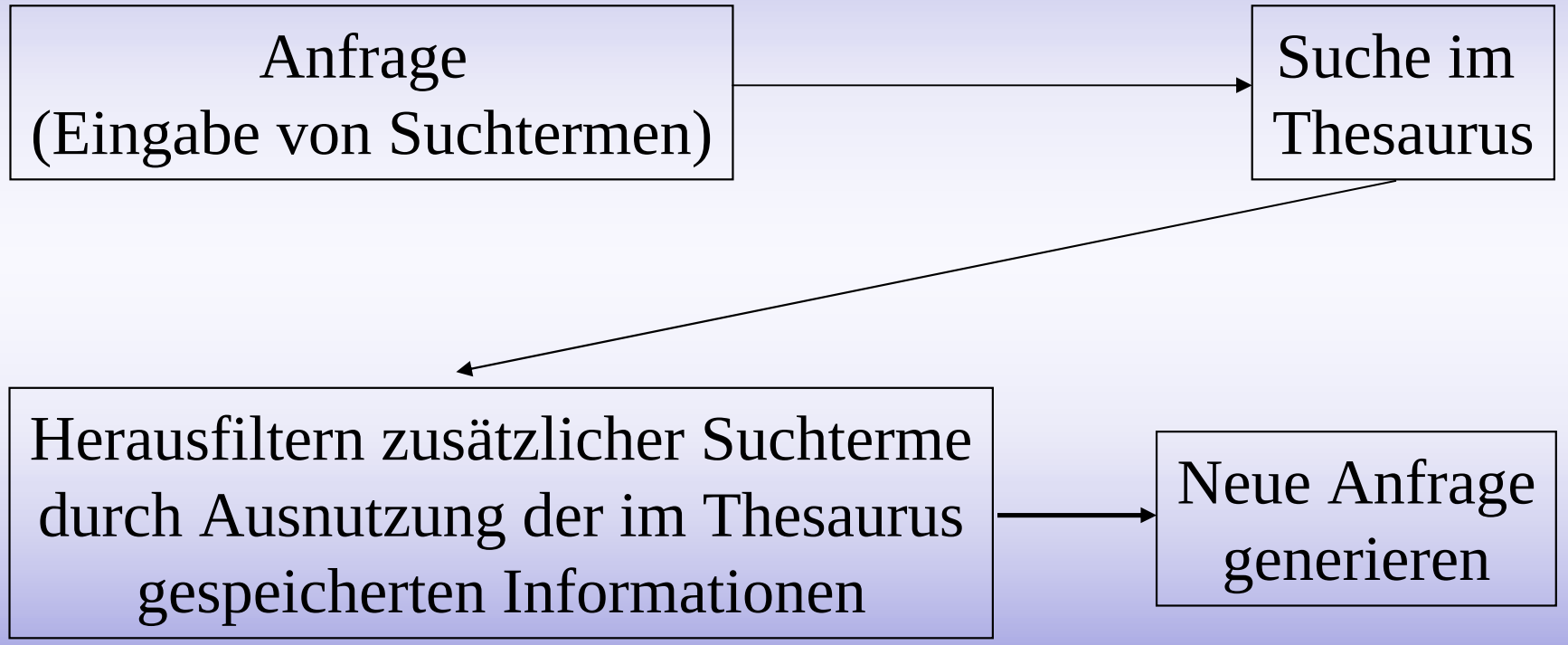
Dokumententerme *(Einträge)*

- Lexem / Wortstamm *(‚sagen‘ / ‚sag‘)*
- Phrasen *(‚ins Gras beißen‘)*
- Referenzwörter *(‚Rat(Personen)‘-‚Rat(Äußerung)‘)*
- Wortklasse *(‚verlegen(adj)‘-‚verlegen(v)‘)*

Thesaurus (2)


Suchterme
Beziehungen (Relationen)
- Äquivalenzrelation (Synonyme)
- Hierarchierelation (Ober- / Unterbegriff)
- Nichthierarchische Relation (Ganzes / Teil)
Dokumententerme

Thesaurus, Anfrageerweiterung



Korpus

Ein Korpus (Textkörper) ist eine Sammlung von Dokumenten, die dazu dient, sprachliche Phänomene über statistische Analysen zu ermitteln.

Sprachliche Phänomene sind beispielweise

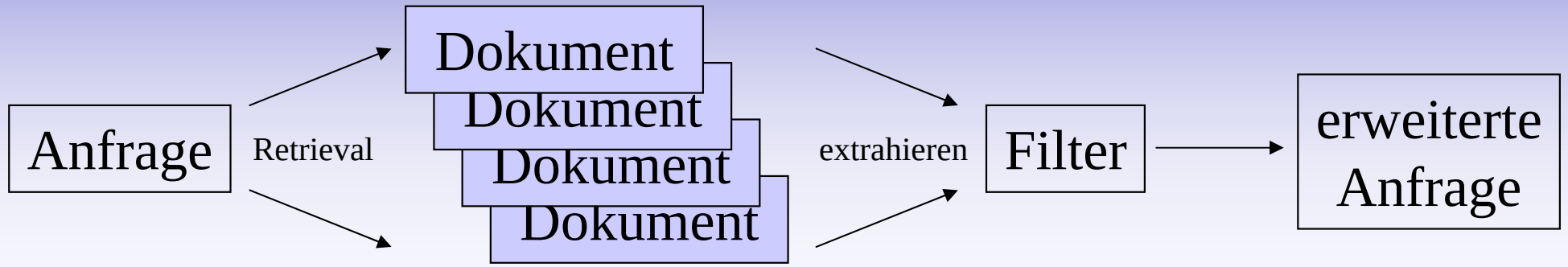
- Worthäufigkeiten
- Wortbeziehungen

Korpusbenutzung zur

→ Erzeugung einer thesaurusähnlichen Struktur

→ Anfrageerweiterung

Korpus, Anfrageerweiterung



Anfrage findet im Korpus eine Menge von Dokumenten. Durch Ermittlung der Ähnlichkeit von Dokument und Anfrage werden diese eingestuft und die besten als relevant betrachtet.

Aus diesen werden dann die Terme, die häufig auftreten, extrahiert.

Durch meist einfache Kriterien werden einige Terme ausgewählt. Es sind meist Terme die nicht zu häufig oder zu selten auftreten, da diese den Inhalt oft nicht gut beschreiben.

Korpus, Übersetzung

Übersetzungsstrategien mittels Korpusbenutzung :

Vorr.: einwandfreie Qualität
Ideal : paralleler Korpus
meist : bilingualer Korpus
Zur Übersetzung eines Wortes in der Quellsprache werden Wörter in der Zielsprache gesucht, die oft parallel dazu benutzt werden :

There‘s a dog in the garden – Da ist ein Hund im Garten
The dog is barking – Der Hund ist am bellen
The dog has a black skin – Der Hund hat ein schwarzes Fell

Korpus, Übersetzung(2)

Wörter müssen korrekten semantischen Sinn beibehalten (Auflösung der Ambiguität) → WSD (word sense disambiguation)

Adäquate Übersetzungen als Basis für WSD im Korpus meist nicht gegeben.

Zugriff auf Presseartikel :

-Ereignis am selben Ort

-Ereignis mit selbem Datum

Liefert meist gute Ergebnisse.

Hauptproblem für parallele Korpora ist die mangelnde Verfügbarkeit von Übersetzungen.

Übersicht, Sprachverarbeitungsmethoden

Anfrage→Dokumentensprache	Dokumente→Anfragesprache

Anfrageerweiterung, Übersetzung :
Thesaurus, Korpus

Erkennung von Wortformen, -arten

Spracherkennung

Maschinelle Übersetzung

Erkennung von Wortformen, -arten

In Anfragen treten Wörter meist in ihrer Grundform auf, in Texten jedoch meist in einer gebeugten Form.
Da eine Speicherung aller gebeugten Wortformen in Hinsicht auf den Platzbedarf und den Zeitbedarf bei der Suche nicht ratsam ist, wird nur der Wortstamm als Repräsentant aller Ausprägungen des Wortes aufgenommen.
Dieser wird mittels **morphologischer Analyse** erkannt.

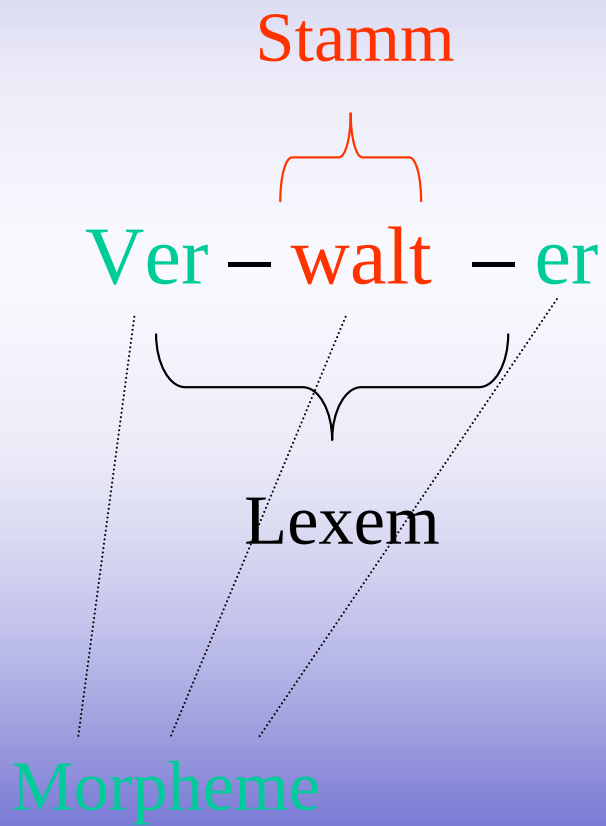
Morphologische Analyse
Begriffserklärungen

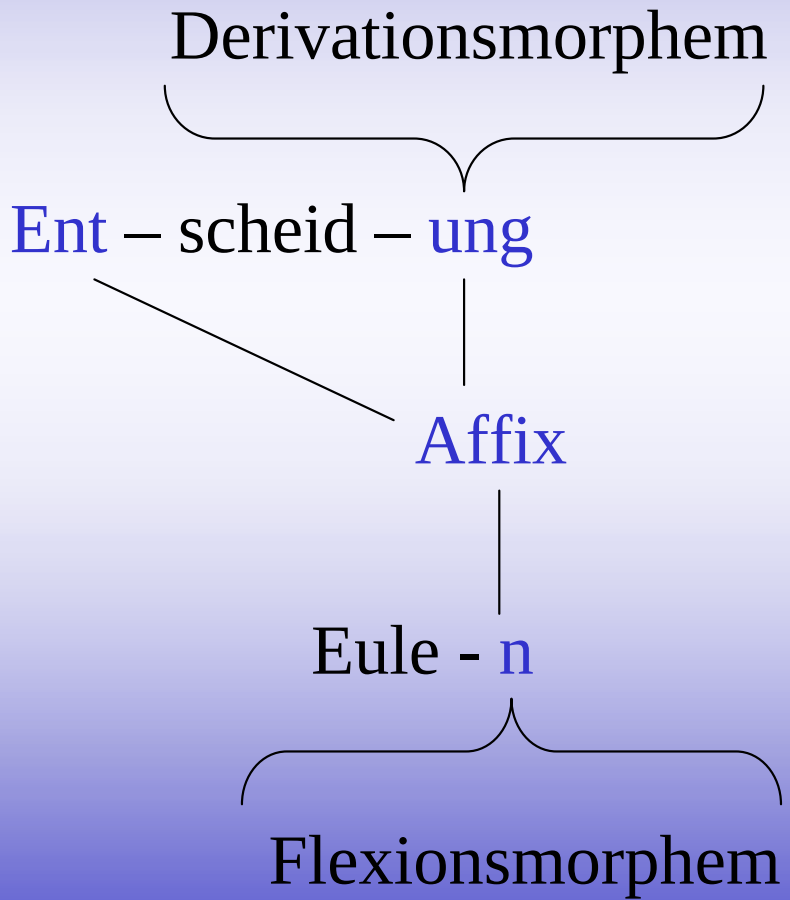
Verwalter

Lexem

Morphologische Analyse
Begriffserklärungen

Morphologische Analyse
Begriffserklärungen





Morphologische Analyse Vorgehensweise

Transformation der gegebenen Wortform in Stammform oder Wortform mit Stammqualitäten.

- Flexionsmorpheme entfernen
- Derivationsaffixe entfernen
- bei Verben Infinitiv bilden

→ Stemming Verfahren

Morphologische Analyse Stemming

Verfahren zur morphologischen Analyse einer Wortform. Wird heute standardmäßig zur Bildung von Dokumenten-repräsentationen eingesetzt.

- meist nur Suffixbehandlung
- schrittweise Entfernung von Endungen
- Ausnutzung von Regeln zur Ersetzung von Derivationssuffixen (Reich-tüm-er→Reich-tum)
- Abgleich mit evtl. vorhandenem Wörterbuch
- Achtung : Schick-sal →schick

Tagging - Verfahren

Mittels Tagging werden Informationen über inhaltliche Beziehungen / Semantik von Wortarten in einem Text aufrecht erhalten.
Tagging ist gebunden an die Benutzung natürlicher Sprache (besonders bei Anfragen wichtig).
Es wird die Wortart (POS - Part of Speech) eines Terms innerhalb eines Satzes mit einer entsprechenden Etikette (engl. tag) markiert.

Tagging – Verfahren Vorgehensweise

Wortklassen Tags *(Auszug)* :
[NN] – Substantiv
[JJ] – Adjektiv
[VB] – Verb
[VBZ] – Hilfsverb
[WRB] – Interrogativpronomen
[DT] – Artikel
[IN] – Präposition

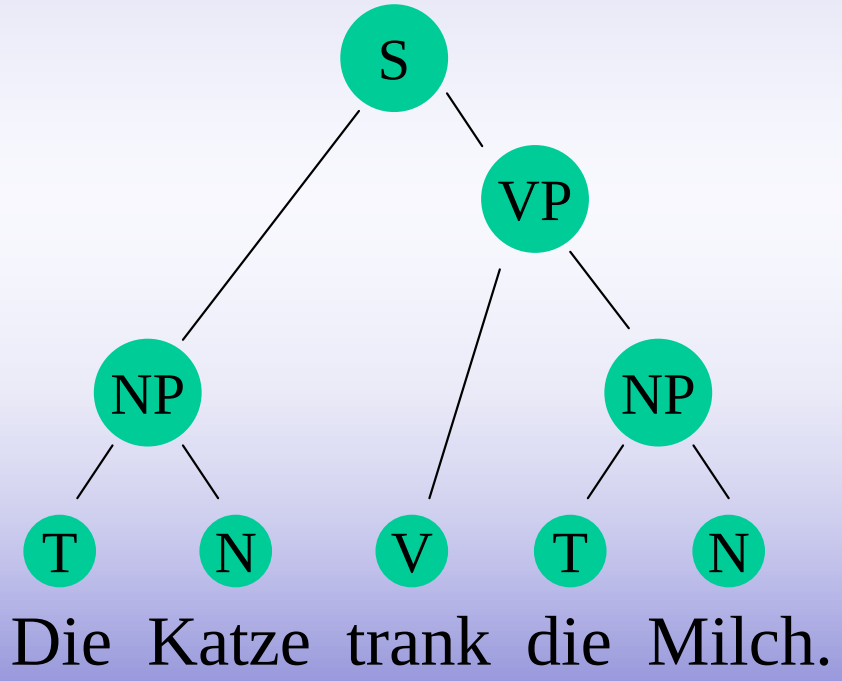
Bestimmung der Wortklasse mittels Regeln oder stochastischen Analysen :

How [WRB] has [VBZ] the [DT]
threat [NN] of [IN] swine [NN]
fever [NN] affected [VB]
international [JJ] trade [NN] ?

Wort und POS-Tag ergeben Token, welches in Thesaurus oder Wörterbuch gesucht werden kann.

Phrasenstrukturgrammatik

Syntaktische Analyse (und Synthese) von Sprachen.



T – Artikel
N – Substantiv
V - Verb

NP –Nominalphrase
VP - Verbalphrase

S -Satz

<u>Grammatik</u>

$S \rightarrow NP + VP$

$NP \rightarrow T + N$

$VP \rightarrow V + NP$

$T \rightarrow$ ‚die‘

$N \rightarrow$ ‚Katze‘ | ‚Milch‘

$V \rightarrow$ ‚trank‘

Übersicht, Sprachverarbeitungsmethoden

Anfrage→Dokumentensprache | Dokumente→Anfragesprache

Anfrageerweiterung, Übersetzung :
Thesaurus, Korpus

Erkennung von Wortformen, -arten :
morphologische Analyse, Tagging, Phrasenstrukturgrammatik

Spracherkennung

Maschinelle Übersetzung

Spracherkennung

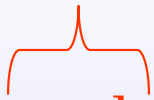
Linguistische Methoden arbeiten bei Kenntnis der Sprache Effektiver, da sie explizites Wissen über die jeweilige Sprache anwenden können.

1. Kodierung erkennen
   - ISO-LATIN-1, JIS
2. Spracherkennung
   - n-Gramm Statistiken, Stoppwortlisten

Spracherkennung
n-Gramm Statistiken

n-Gramm : beliebige Teilzeichenkette der Länge n aus einem Wort

Trigramm (3-Gramm)



M a s c h e

Lange Kombinationen sind eindeutiger einer Sprache zuzuordnen. Durch meist einmalige Silbenstruktur erzielt man schon mit Trigrammen gute Ergebnisse.

Spracherkennung Stoppwortlisten

Diese Listen bestehen meist aus kleinen Worten wie Artikel oder Präpositionen.

Für jedes Land existiert eine länderspezifische Stoppwortliste.

Das Auftreten eines Stoppwortes im Dokument wird gezählt.

Die Sprache der Liste, deren Elemente am häufigsten in dem Dokument vorkamen, wird gewählt und das Dokument mit dem passenden Sprachbezeichner markiert.

Übersicht, Sprachverarbeitungsmethoden

Anfrage→Dokumentensprache	Dokumente→Anfragesprache

Anfrageerweiterung,Übersetzung :
Thesaurus, Korpus

Erkennung von Wortformen, -arten :
morphologische Analyse, Tagging, Phrasenstrukturgrammatik

Spracherkennung :
n-Gramm Statistiken, Stoppworterkennung

Maschinelle Übersetzung

Maschinelle Übersetzung

Maschinelle Übersetzung hat zum Ziel, jeden Text aus einer Sprache in jede beliebige andere Sprache übersetzen zu können.

Dies erfordert einen sehr großen Aufwand und zeigt gerade an wichtigen Stellen Schwächen.

Maschinelle Übersetzung Fehler

„This drives me nuts“

- „Dies fährt mich verrückt“
- „Dieses fährt mich Nüsse“

„John took Mary for a drive“

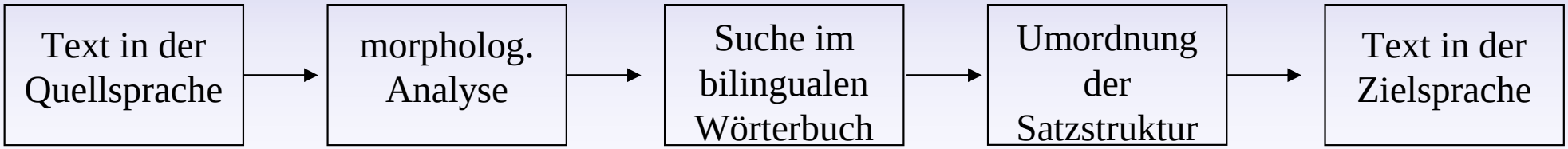
- „John nahm Mary für einen Elan“
- „John hielt Mary für eine Fahrt“

„Tell me yor name !“

- „Erzählen Sie mir Ihren Namen !“

Quelle : c‘t

Maschinelle Übersetzung direkte MÜ-Systeme



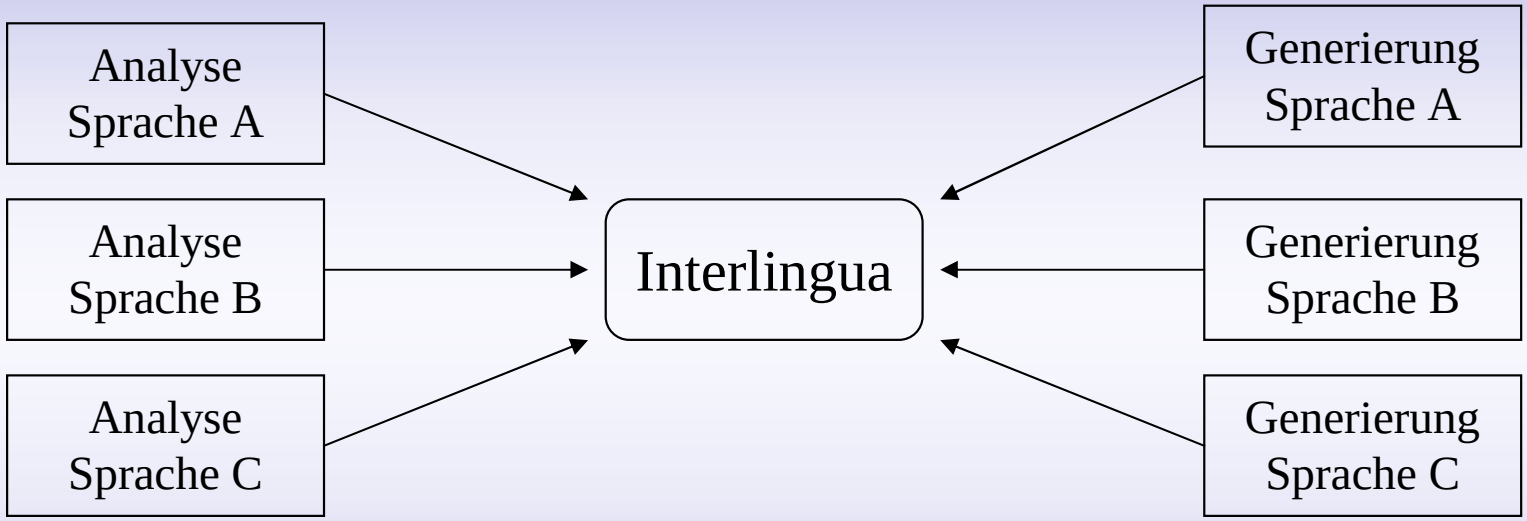
Ermittlung der Grundform

Übersetzung durch eindeutige Wort zu Wort Beziehung

Sehr grobe Umstrukturierung, keinerlei Rücksicht auf semantische Bedeutung oder syntaktische Beziehungen

Übersetzung ist meistens irreversibel.

Maschinelle Übersetzung Interlingua - Systeme



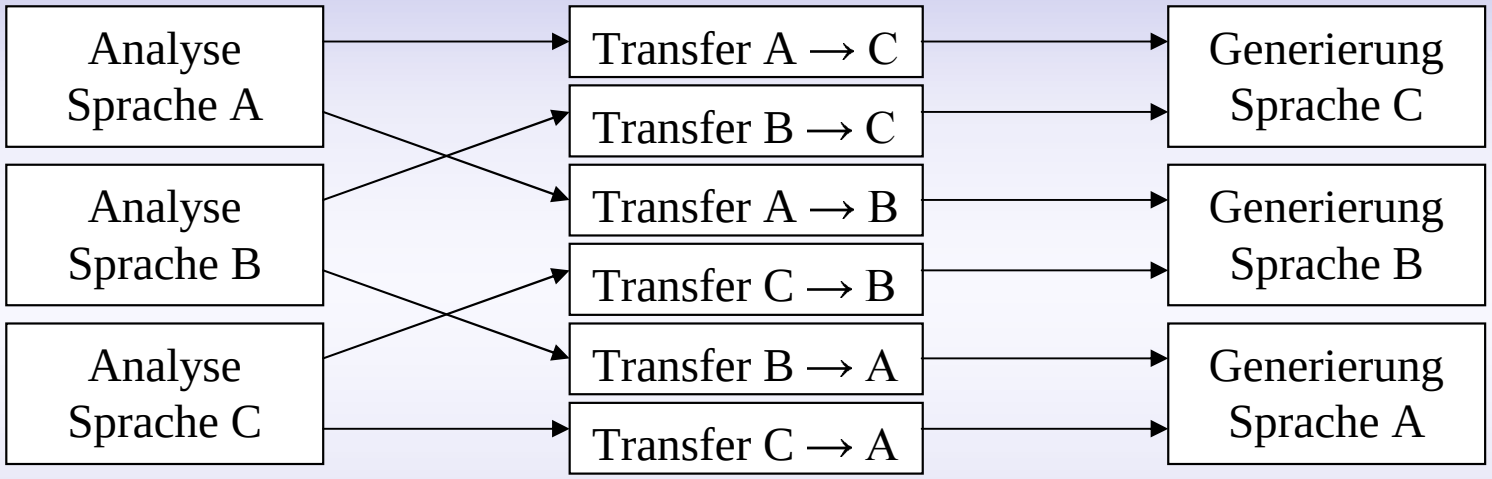
Man kann von jeder Sprache in jede beliebige andere Sprache übersetzen, wenn es ein Analysemodul für die Quellsprache und ein Generierungsmodul für die Zielsprache gibt.

Maschinelle Übersetzung Interlingua – Systeme (2)

Alle Wörter der Quellsprache werden mit Hilfe von einfachen Konzepten aus dem Interlingua Lexikon soweit wie möglich vereinfacht. (Seher → ‚Person, sehen‘) .
Ein Satz wird so in eine Interlingua Formel gebracht, die auch alle semantischen und syntaktischen Informationen enthält.
Aus dieser Formel können dann alle Übersetzungen, für die ein Generierungsmodul vorhanden ist, erzeugt werden.

Problem : Zwischensprache für die Formel

Maschinelle Übersetzung Transfer Systeme



Zwischen Analyse der Quellsprache und Generierung der Zielsprache ist eine Einheit geschaltet, die sogenannte Transfereinheit, welche die Quellsprache genau auf die Zielsprache abbildet.

Maschinelle Übersetzung Transfer Systeme (2)

Bei der Analyse wird eine Zwischenrepräsentation (ZP) des Textes erzeugt.

Die Transfereinheit erhält mit dieser ZP alle morphologischen, semantischen und syntaktischen Informationen und erstellt daraus (mit Hilfe von Grammatikregeln, bilingualem Wörterbuch, etc.) eine neue ZP in der Zielsprache.

Im Generierungsmodul wird aus der neu gewonnenen ZP der Text in der Zielsprache erzeugt.

Hauptarbeit ist die Transformation der syntaktischen Strukturen.

(‚gangsters on the run' – ‚to run a business')

Maschinelle Übersetzung
Fazit

Aufgrund des komplexen Zusammenspiels von Morphologie, Syntax und Semantik ist der Aufwand an Ressourcen und Arbeitszeit bei maschineller Übersetzung momentan extrem hoch und macht sie unattraktiv für MLIR.

Übersicht, Sprachverarbeitungsmethoden

Anfrage→Dokumentensprache	Dokumente→Anfragesprache

Anfrageerweiterung, Übersetzung :
Thesaurus, Korpus

Erkennung von Wortformen, -arten :
morphologische Analyse, Tagging, Phrasenstrukturgrammatik

Spracherkennung :
n-Gramm Statistiken, Stoppworterkennung

Maschinelle Übersetzung :
direkte MÜ-, Interlingua und Transfersysteme

Vielen Dank für die Aufmerksamkeit !