Grundstudiumspraktikum WiSe 04/05

Maschinelle Übersetzung

Walther v. Hahn
Cristina Vertan
{vhahn,vertan}@nats.informatik.uni-hamburg.de

Zweck des Praktikums

- Kennen lernen der Grundprinzipien maschineller Übersetzung,
- Programmierübung,
- Integration von Modulen in eine Systemarchitektur,
- Anwenden von Software- und MT-spezifischen Evaluationskriterien,
- Gruppenarbeit.

Struktur des Praktikums

- Präsentation der wesentlichen Konzepte und -methoden, soweit sie für die Implementation nötig sind:
  - Einführung in die maschinelle Übersetzung 20.10
  - Einführung in die Beispiel-basierte maschinelle Übersetzung 27.10
  - Evaluationskriterien 05.01
- Implementierungstermine (dazwischen)
- Evaluationstermine:
  - Zwischenevaluation 6.12
  - Interne Endevaluation 26.01
  - Öffentliche Präsentation 02.02

Scheinkriterien

1. Anwesenheit (maximal 2 begründete Abwesenheiten),
2. Implementierung und Evaluation eines Teils des Systems, mit Präsentation in der letzten Sitzung,
3. Integration des eigenen Moduls in die Systemarchitektur,
4. Ablieferung des entsprechenden Teils des Projektberichtes.

Punkte 2 und 4 sind von jeden Teilnehmern individuell und nicht als Gruppenarbeit zu erledigen.

Einführung in die maschinelle Übersetzung

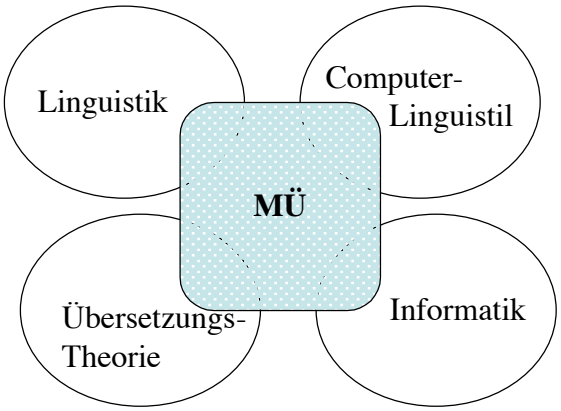
Walther v. Hahn

Häufige Abkürzungen in der englischen Literatur

- MT = Machine Translation (Maschinelle Übersetzung)
- MAT = Machine Aided Translation (Maschinell Unterstützte Übersetzung)
- MAHT = Machine Aided Human Translation
- HAMT = Human aided Machine Translation
- SL, TL = Source Language / Target Language
  (Quellsprache/Zielsprache)

Maschinelle Übersetzung als Forschungsgebiet

MÜ ist kein isoliertes Forschungsgebiet, sondern eine Anwendung von Methoden aus mehreren Bereichen:



Warum brauchen wir MÜ?

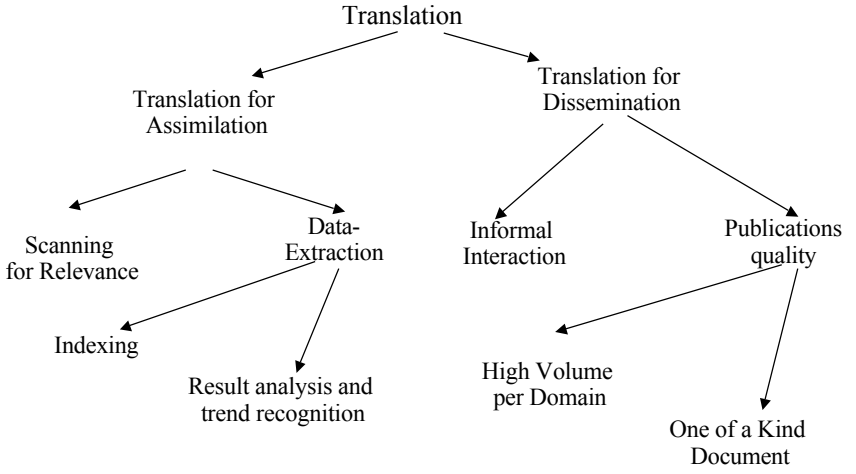
- Weltweit ist hat der Übersetzungsmarkt einen Wert (in Millionen $) von
  1989 20
  1990 500
  2003 2000
  Das durchschnittliche jährliche Wachstum ist immer noch ca 20%
- Schon 1986 waren es weltweit mehr als 500 Mio. übersetzte Seiten, davon mehr als 100 Mio. in Europa. Davon entfielen
  1% > "schöne Literatur"
  30% offizielle (Staatliche) Dokumente
  50% Industrie und Wirtschaft (meist technische Dokumentation)
- Die Zeitersparnis durch die Benutzung des MT-Systems Systran war nach Kundenauskunft ca. 75%
- Dienstverbesserung der Übersetzungsabteilung durch MAT-Systeme (nach Deutsche Airbus): 20%

EU

- Systran hat schon 1994 140 000 Seiten an EU-Dokumenten übersetzt. 80 % der EU-Dokumente zwischen Spanisch und Französisch werden inzwischen automatisch übersetzt.
- Man kann nicht so viele ÜbersetzerInnen/DolmetscherInnen ausbilden wie gebraucht werden
- Die Situation ist nach der EU-Erweiterung noch komplizierter geworden: Zur Zeit gibt es allein 20 EU-amtliche Sprachen (mit doppelt so vielen sprachpaarspezifischen Übersetzungsdiensten.

Functional Typology of MT-Systems

©Carbonell



MÜ und Sprachtechnologie

- MÜ hat vieles mit anderen Sprachtechnologien gemeinsam:
  - Typische grundsätzliche Eigenschaften natürlicher Sprachen (s. folgende Folien)
  - Fehlende Umgebungs- und Situationsdaten/-modelle
  - Begrenzte modale Kanäle
  - Begrenzte Technologien
  - Häufiges Vorkommen unbekannter Wörter/Namen
  - Nicht-deterministische Verarbeitung

Grundsätzliche Eigenschaften natürlicher Sprachen

- Eigenschaften, die alle natürlichen Sprachen auf verschiedenen Strukturebenen haben
- Unterschiede zwischen Sprachen
  - Lexikalische Unterschiede
  - Syntaktische Unterschiede

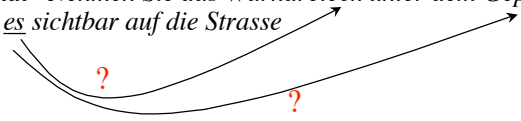
Typische Merkmale natürlicher Sprachen - 1 -

- Unklarer Analysefokus, besonders bei gesprochener Eingabe:
  Da sich mit zunehmender Geschwindigkeit die Fahrstabilität des Gespannes verringert, sollte unter ungünstigen Straßen-, Wetter- und Windverhältnissen – vor allem auf Gefällstrecken – die gesetzlich erlaubte Höchstgeschwindigkeit nicht ausgenutzt werden.
- Selbst-Referenz, metasprachliche Ausdrucksmöglichkeit *(Mit „Lampe" meinte ich das Rücklicht)*
- Valenzen, d.h. syntaktische/semantische Kookurrenzen von Kategorien.
  - *I did not take care of the oil type.*
  - *I did not take additional winter tyres.*
- Mehrwort-Lexeme und Idiome ohne regelhaft zusammengesetzte Bedeutung
  - *Give up, mit etwas gut (schlecht) fahren*
- Hierarchische Syntax in nicht-lineare Reihenfolge
  Dieser Regelvorgang macht sich durch eine pulsierende Bewegung des Bremspedals, verbunden mit Geräuschen, bemerkbar.

Typische Merkmale natürlicher Sprachen - 2 -

Ambiguitäten auf allen Ebenen:.

- Sprachsignal-Ambiguität *"Rat" vs. "Rad" "peak" vs. "peek"*
- lexikalische Ambiguität *"Fahren sie zu der nächsten Bank!"*
- Syntaktische Ambiguität *"I saw the VW-Service Center driving to the airport"*
- Pragmatische Ambiguität *"Können Sie mir bitte die Anweisungen per SMS schicken?" (Als Antwort nicht: "Ja, kann ich!")*
- Referentielle Ambiguität *"Nehmen Sie das Warndreieck unter dem Gepäck heraus und stellen Sie es sichtbar auf die Strasse*



- Ambiguität ist der Hauptunterschied zwischen formalen und natürlichen Sprachen

Typische Merkmale natürlicher Sprachen - 3 -

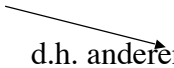
- Long distance dependencies

Eine bei allen Bedingungen einwandfrei arbeitende Abgasreinigungsanlage

- Nichtkontinuerliche Wörter (*Der Kraftstoff, den Sie getankt haben,* ***weicht*** *wahrscheinlich von der Norm* ***ab****.)*
- Ellipsen *(Hier ebenso)*
- Paraphrasen
- Kohärenz in Texten *(Daher hat die Sicherung ...)*
- Verstehen durch Weltwissen

Stellen Sie sich diese Phänomene einmal bei Programmier-sprachen vor ...

Bei Fahrten in England oder ähnlichen Ländern blendet das asymmetrische Abblendlicht den Gegenverkehr.



d.h. anderen Ländern mit Linksverkehr

Unterschiede zwischen Sprachen: Lexikon - 1 -

- Ein Wort in der Quellsprache muss durch mehrere einzelne Wörter oder Mehrwortausdrücke in der Zielsprache ersetzt (übersetzt) werden.
- Eins-zu-mehr Übersetzungen (Ein Wort in der Quellsprache hat kontextabhängig mehrere Übersetzungen)
  - SL: *Wall* (engl.) wird mit TL *Mauer* (dt.) oder *Wand* (dt.) übersetzt, abhängig davon, wo das Objekt ist (innen oder aussen). In diesem Fall müssen semantische Merkmale verglichen werden müssen.
  - Für die Übersetzung von “know” muss der grammatische Kontext muss bekannt sein:

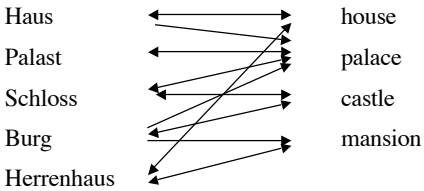
  I know him (engl.) → *Ich kenne ihn* (dt.)

  I know a solution (engl.) → *Ich weiß eine Lösung* (dt.)

Unterschiede zwischen Sprachen: Lexikon - 2 -

Mehr-zu-eins Übersetzung (Ein Wort in der Zielsprache hat kontextabhängig mehrere Bedeutungen) :

Herrenhaus (dt.) kann mit *house* (engl.) übersetzt werden, erzeugt in der TL aber die Ambiguität mit dt. Haus:



Die Gründe lexikalischer Unterschiede zwischen Sprachen sind:

- Unterschiedliche Begriffe oder Begriffsteilung
- Unterschiedliche Grammatikregeln
- Unterschiedliche stilistische Regeln

Unterschiede zwischen Sprachen: Lexikon - 3 -

- Unterschiedliche Übersetzungen abhängig von Domänen-/Weltwissen

  z.B. *library* (engl.) wird auf Deutsch mit:

  - *Bibliothek* wenn sie privat ist oder zu einer akademischer Institution gehört.
  - *Bücherei* wenn sie öffentlich ist..

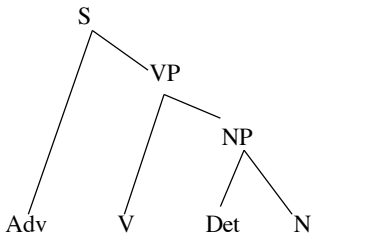
  übersetzt.

- Lexical gaps - Einzelwörter/Konzepte eine Sprache, die in der Zielsprache nur umschrieben werden können E.g. *abschleppen* (germ.) = *to take in tow* (engl.)

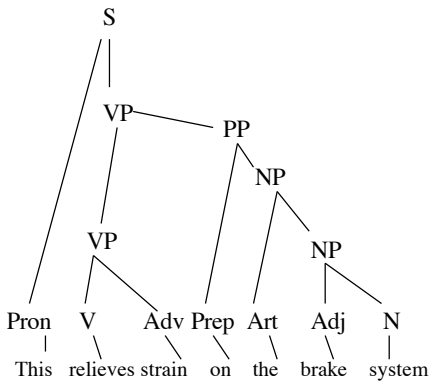
  Solche Probleme können nicht allein durch lexikalischen Transfer gelöst werden, da es z.B. im Englischen Lexikon keinen Eintrag *"to take in tow"*

  - Lexical gaps sind oft spezifisch kulturelle Konzepte (z.B. *Guten Appetit! Meldebescheinigung*), sie bleiben oft unübersetzt.

Strukturelle Unterschiede zwischen Sprachen: Syntaktische Strukturen



Dadurch [wird entlastet] die Bremsanlage



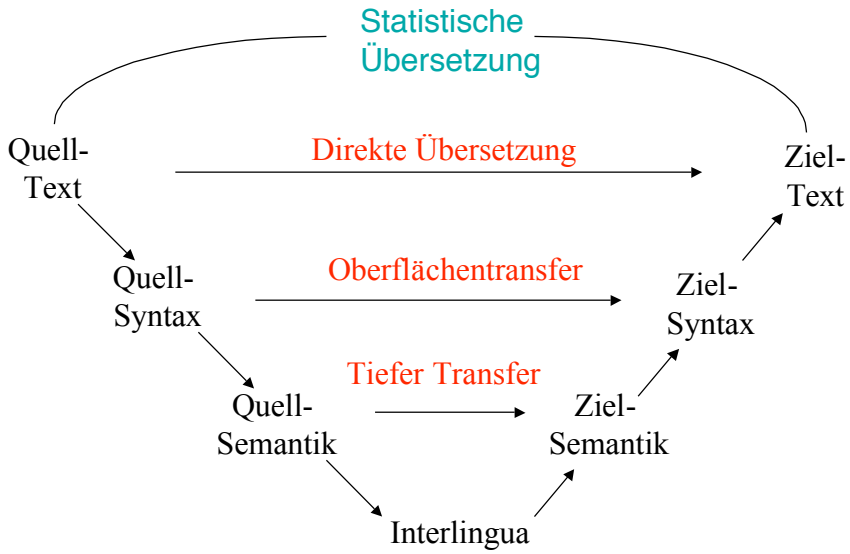
Maschinelles Dolmetschen

- Neue Forschungs- und Technologiedomäne mit Anwendungen in:
  - Konsekutivdolmetschen
  - Simultandolmetschen
  - Gesprächsdolmetschen
- Es ist hochinteressant weil es integrierte Verarbeitung zwischen
  - Signal-Ebene Phonetik und
  - Text-Ebene Linguistik

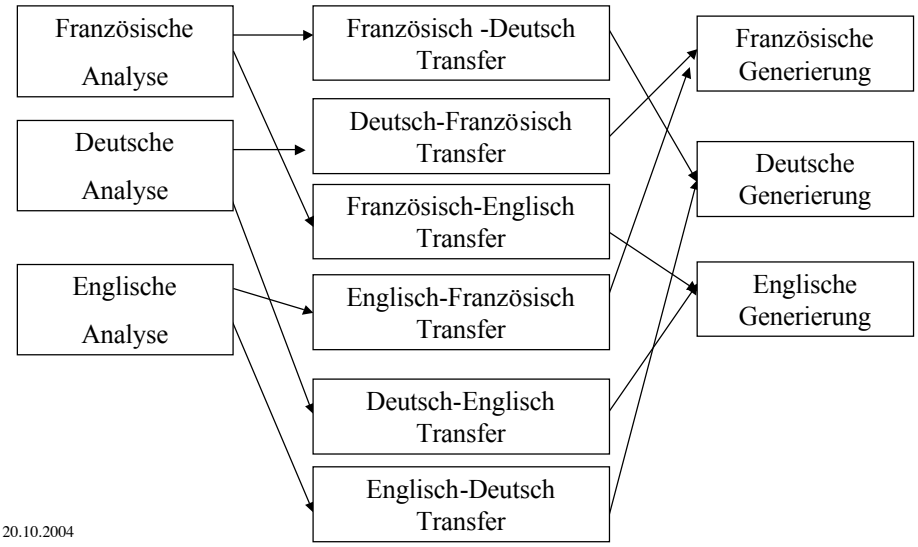
erfordert.

- Sehr relevant für die kognitiv orientierte Informatik durch
  - Dolmetschenstrategien
  - Verstehensleistungen
  - Zeitverhältnisse
  - Erschließen von Sprecher- und Sprachmerkmalen

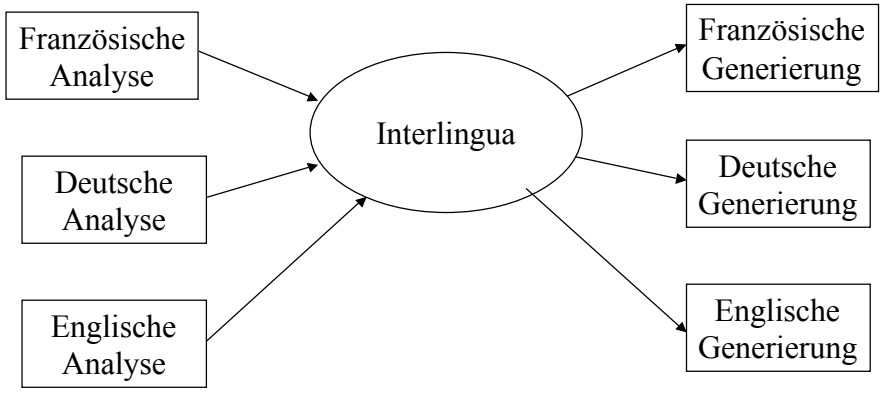
Das MÜ-Dreieck



3 Sprachiges-Transfer-System



3 Sprachiges Interlingua-System

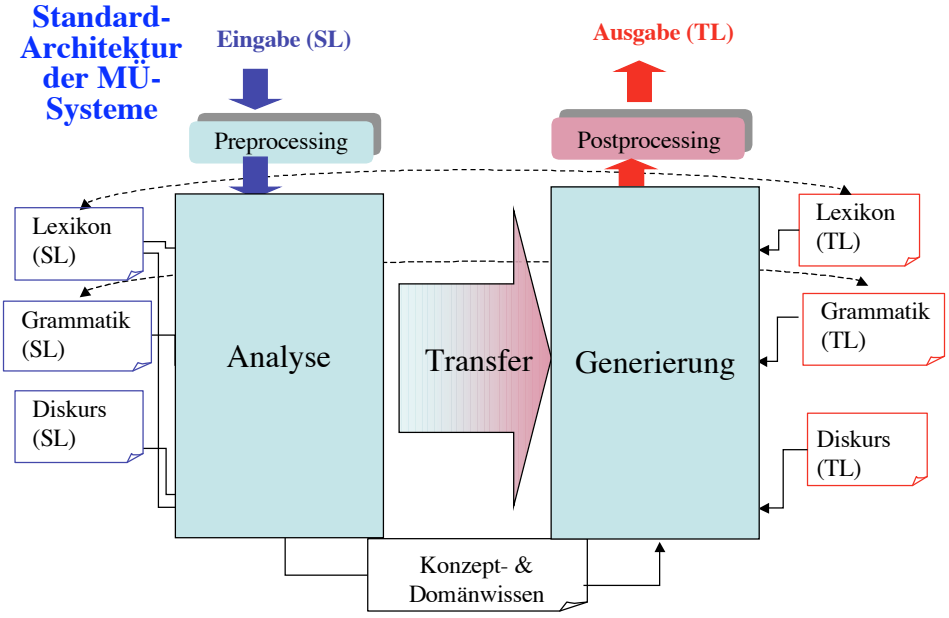


Interlingua- vs. Transfer-Systeme

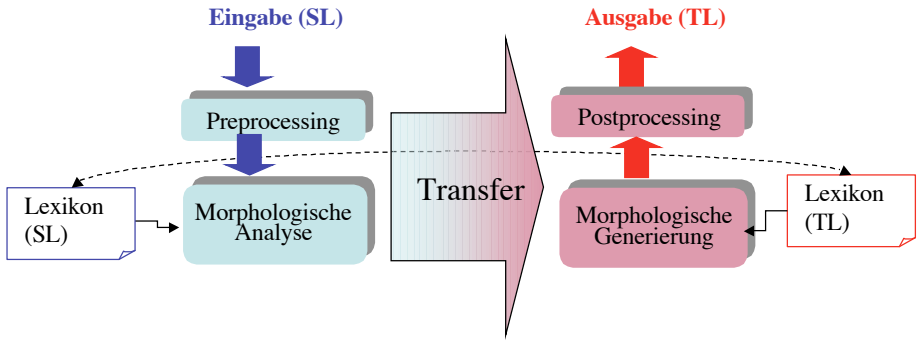
- Kein Modul ist von Analyse- oder Generierungsmodulen anderer Sprachen abhängig
- Zielsprachen haben keinen Einfluß auf dem Analyseprozeß.
- Für jede neue Sprache müssen nur 2 neue Module implementiert werden.
- „Rück-Übersetzung“ ist möglich (nützlich für Systemevaluation)
- Sehr komplizierte Repräsentation, auch für Sprachen derselben Famile)

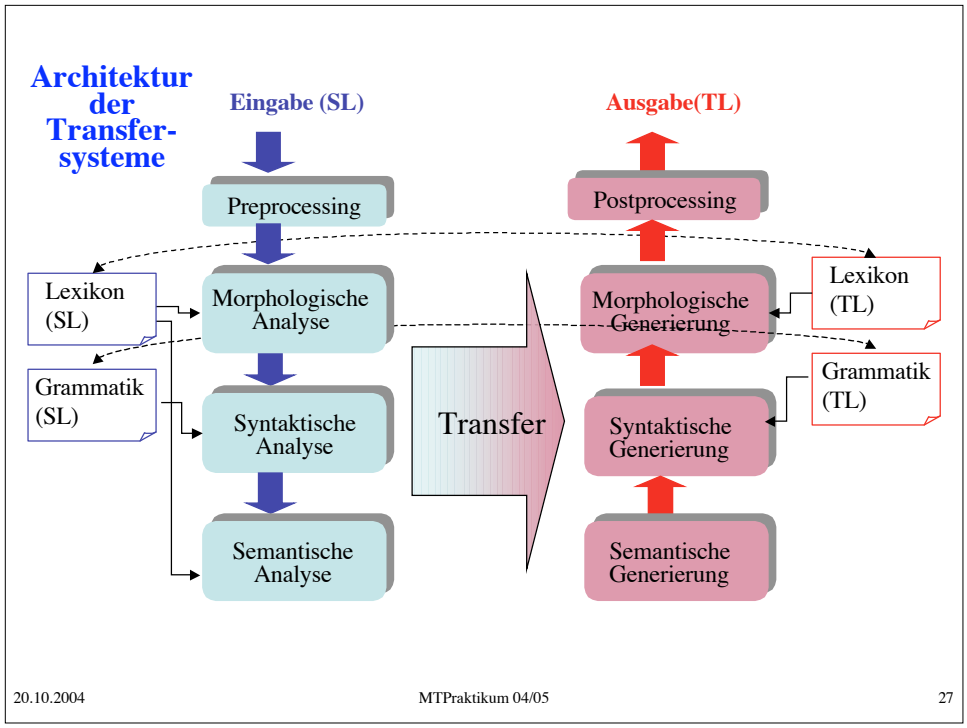
- Sind Sprachpaarabhängig
- Für jede Sprache muss eine größere Anzahl von neuen Modulen implementiert werden. (für n Sprachen: n(n-1) Modulen)
- Straight-forward Darstellung der Regeln
- Lokale Definitionen von Ähnlichkeiten zwischen Sprachen.

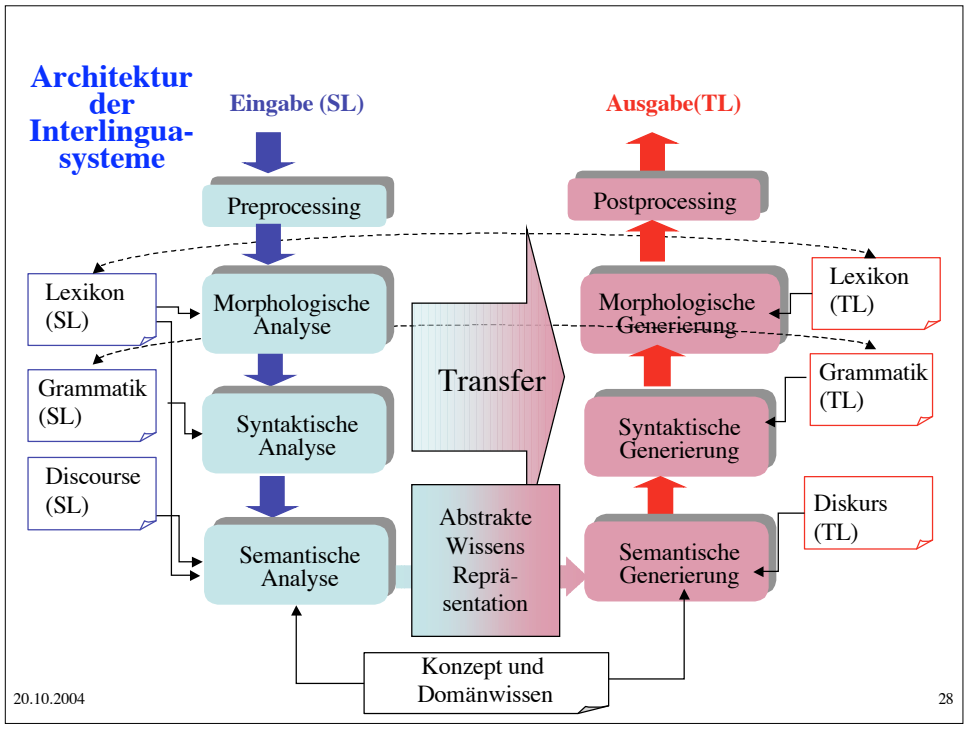
Standard-Architektur der MÜ-Systeme



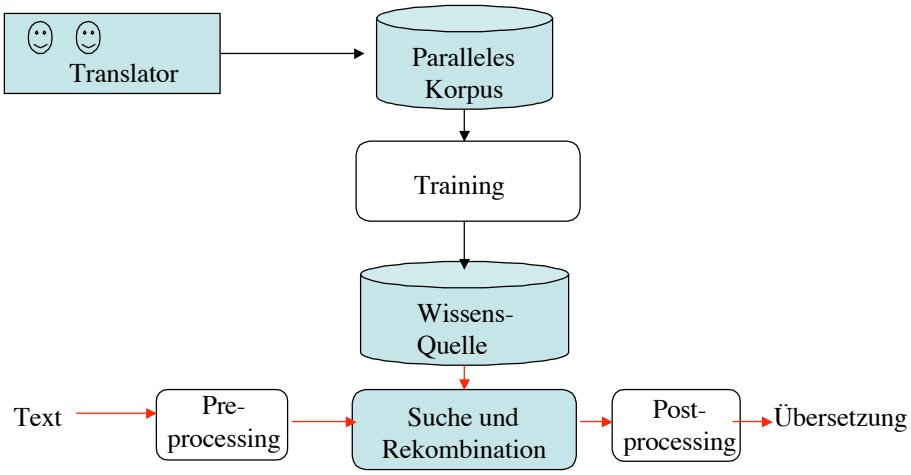
Direkt System Architekturen




Architektur der Transfer-systeme
Eingabe (SL)
Ausgabe(TL)
Preprocessing
Postprocessing
Lexikon (SL)
Morphologische Analyse
Morphologische Generierung
Lexikon (TL)
Grammatik (SL)
Syntaktische Analyse
Transfer
Syntaktische Generierung
Grammatik (TL)
Semantische Analyse
Semantische Generierung
20.10.2004
MTPraktikum 04/05
27


Architektur der Interlingua-systeme
Eingabe (SL)
Ausgabe(TL)
Preprocessing
Postprocessing
Lexikon (SL)
Morphologische Analyse
Morphologische Generierung
Lexikon (TL)
Grammatik (SL)
Transfer
Grammatik (TL)
Syntaktische Analyse
Syntaktische Generierung
Discourse (SL)
Abstrakte Wissens Reprä-sentation
Diskurs (TL)
Semantische Analyse
Semantische Generierung
Konzept und Domänwissen
20.10.2004
28

Architektur eines Korpus-basierten MT-Systems

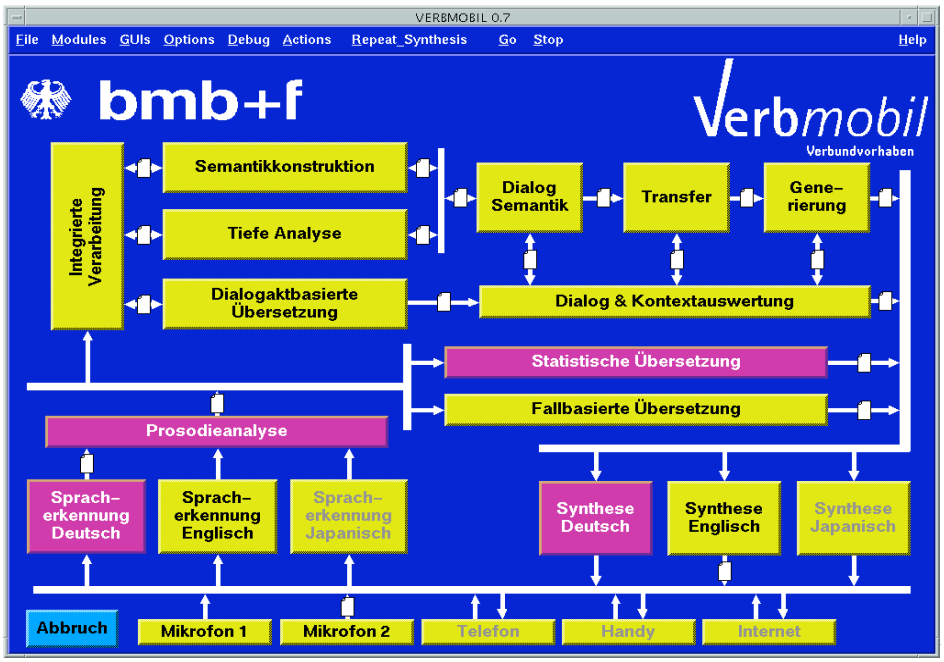


Unterschiedliche MÜ-Methoden

- Regelbasierte MÜ
- Wissensbasierte MÜ
- Statistikbasierte MÜ
- Beispielbasierte MÜ

Methoden für computerunterstützte Übersetzung

- Maschinell unterstütze Übersetzung
- Translation Memories (Workbenches)



MÜ-spezifisches Pre-editing

- In den Quelltexten werden bekannte schwierige Verarbeitungsprobleme aufgesucht und wenn möglich ersetzt.
- Beispiele für solche Operationen:
  - Identifizierung von Eigennamen
  - Markierung von grammatischen Kategorien bei Homographen
  - Markierung von eingebettete Äusserungen
  - Markierung von koordinierten Strukturen
  - Ersetzung von unbekannten Wörtern
  - Extreme Form: Neuformulierung in einer kontrollierten Sprache (manchmal in technischer Dokumentation)

Post-Editing

- Korrektur der Ausgabe nach einem vorher vereinbarten Standard.
  - Minimal bei Übersetzung für Assimilation
  - Sehr stark für Dissemination
- Häufige Operationen:
  - Ersetzung von Wörtern durch geeignete “Synonyme”
  - Ersetzung von Einzelwörtern durch Idiome
  - Syntaxkorrektur.

Evaluation von MT-Systemen

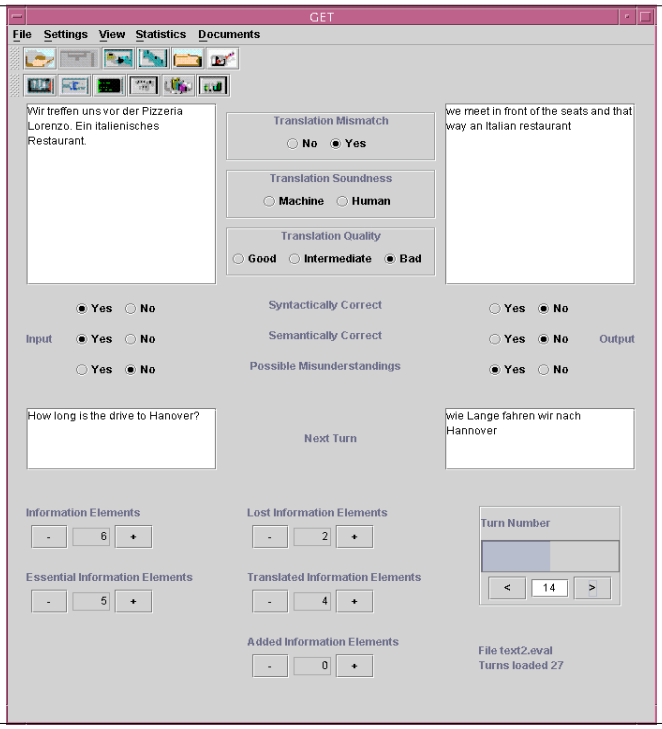
- Im Gegensatz zu anderen Softwarewerkzeugen gibt es bei Übersetzungen keine “Musterlösung”, mit der das Ergebnis verglichen werden kann
- Für eine Eingabe gibt es mehrere korrekte Übersetzungen
- Die Evaluation eines MT-Systems ist von seinen Aufgaben und den Anforderungen seiner möglichen Nutzer abhängig.

Evaluationsstrategien

Test Suite vs. Test corpus

- Carefully constructed set of examples, each testing a particular linguistic or translation problem (e.g. different lexical and structural differences)
- Problem: it is assumed that the behaviour of a system can be projected from carefully constructed examples to real texts
- Test suite evaluations are difficult to compare

- An adequate corpus (for the domain of the system) is used as input
- Problem: it does not test systematically all possible sources of incorrect translations, but considers the most frequent constructions
- It is difficult to estimate the behaviour of the system for other types of text



Was implementieren wir im Praktikum?

- Ein Deutsch-English-Deutsches beispielbasiertes Übersetzungssystem.
- Domäne: Autoreparatur
- Eingabetyp: simulierte Spracheingaben für Fragen, sowie Antworten (Telefongespräch mit einem VW-Service-Center):
  - *Ich habe Winterdiesel. Mein Auto springt trotzdem nicht an.*
  - *Ich verliere Bremsflüssigkeit. Was soll ich machen?*

Vorhandenes Material

- Englische und Deutsche Versionen des Volkswagen-Handbuchs (als paralleles Korpus)
- Die konzeptuelle Struktur eines deutsch-englischen Lexikons
- Test-Korpus (Beispieleingaben)

Themen, Termine und Zwischentermine

• 20.10. Einführung MÜ	
• 27.10. Einführung Beispiel-basierte MÜ	
• 03.11. Systemarchitektur, Gruppenaufteilung	
• 10.11. Implementierung	Lexikon Vers. 0.1 fertig
• 17.11. Implementierung	Schnittstellendefinition Vers 0.1 fertig
• 24.11. Implementierung	Beispieldatenbank Vers. 0.1 fertig
• 01.12. Implementierung	Analyse-Modul Vers. 0.1 fertig
• 08.12. Implementierung	Rekombination-Modul Vers. 0.1 fertig
• 15.12. Zwischenevaluation	System Vers. 0.1
• 05.01. Evaluationskriterien / Implementierung	Lexikon Vers. 0.2
• 12.01. Implementierung	Beispieldatenbank Vers. 0.2
• 19.01. Implementierung	Analyse- und Rekombinationsmodul Vers. 0.2
26.01. Interne Evaluation	System fast fertig, Vers 0.2 Beta
• 02.02. Öffentliche End-Präsentation	System fertig Vers. 0.9

Kontakt

- Webseite des Praktikums

http://nats-www.informatik.uni-hamburg.de/view/MachineTranslation04/WebHome

- Walther v. Hahn:
  - F-234
  - E-mail: vhahn@informatik.uni-hamburg.de
  - Sprechstunden: Montag: 14-16
- Cristina Vertan:
  - F-211
  - E-mail cri@nats.informatik.uni-hamburg.de
  - Sprechstunden: Dienstag: 10-12
- Sekretariat: Karin Jarck
  - F-206
  - E-mail: jarck@nats.informatik.uni-hamburg.de
  - Besetzungszeit: jede Tag 9-13