Universität Hamburg Germanisches Seminar

Maschinelle Übersetzung

Prof. Dr. Walther v.Hahn

Dr. Cristina Vertan

Seminar II im Wintersemester 2006/07

einige der Folien stammen von Cristina Vertan, Uni HH und Jaime Carbonell, Carnegie Mellon University

WWW: http://nats-www.informatik.uni-hamburg.de/view/User/WaltherVHahn

E-Mail: {vhahn,vertan}@informatik, uni-hamburg.de

Wie man uns erreichen kann (außer vor und nach dem Seminar)

Telefon (mindestens von 9.15 - 17.00 im Büro, bitte nicht zuhause anrufen)

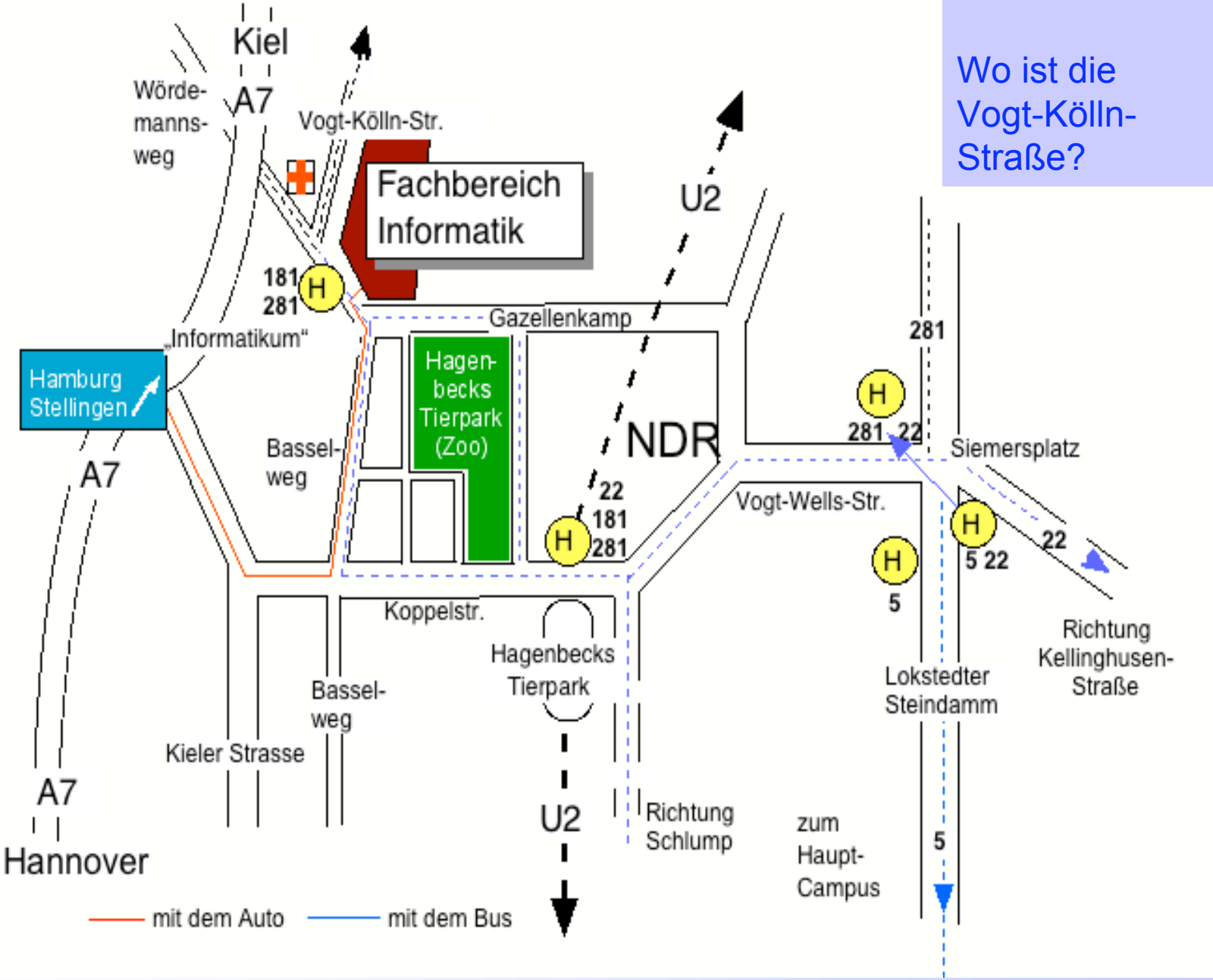
v.Hahn	428 83 2434 (Sekretariat Frau Jarck 2433)
Vertan	428 83 2519

Sprechstunden:

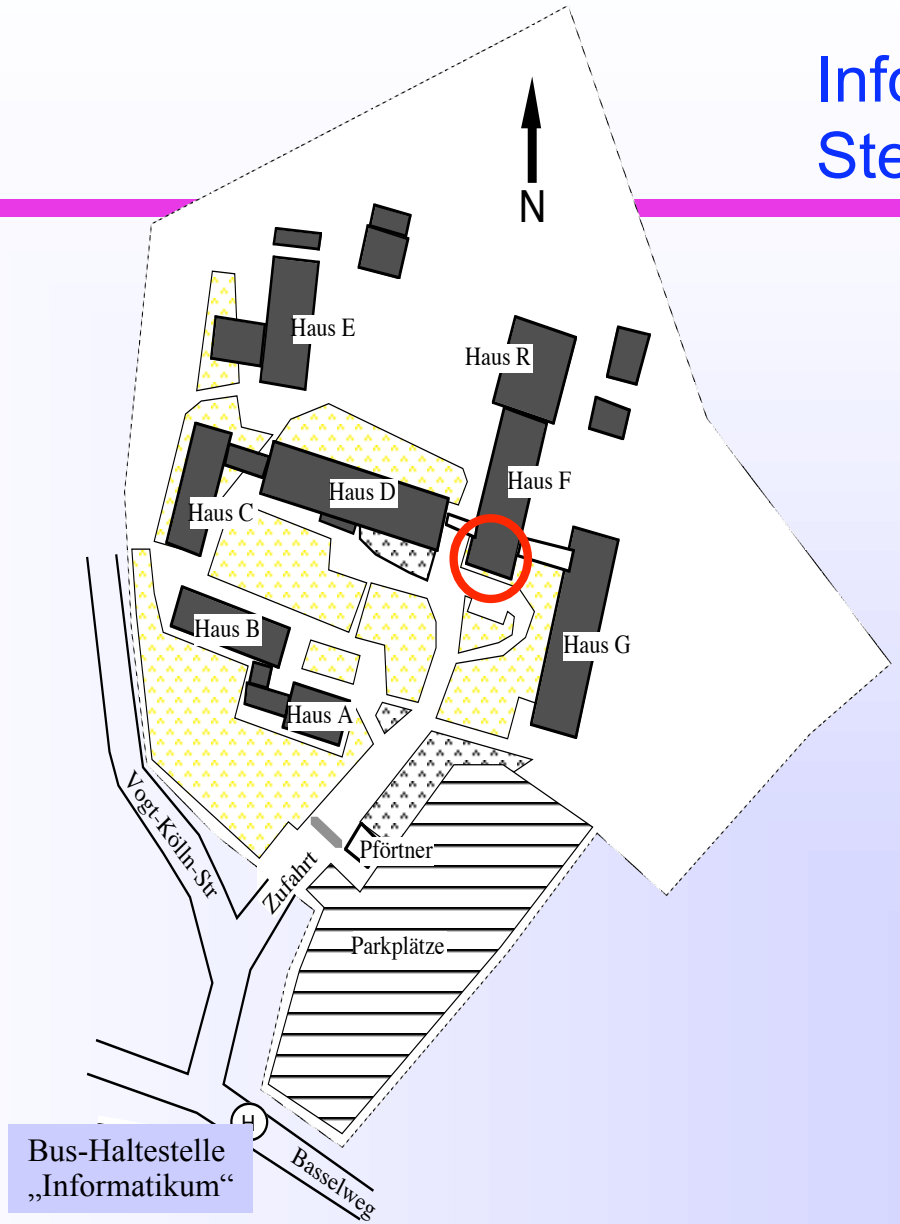
v.Hahn	Die 12:30	Phil 371
	Mo 14 - 16	Vogt-Kölln-Str. 30. - Zr. F 234
Vertan	n.V.	Vogt-Kölln-Str. 30. - Zr. F 212

Wo ist der FB Informatik? Vogt-Kölln-Str. 30 (Stellingen)

Wo ist die Vogt-Kölln-Straße?



Informatikum Stellingen



Anlage des Seminars

- Die Anlage des Seminars ist beispielorientiert, nur in meiner Einführung ist sie überblicksorientiert. Es sollen an einigen Publikationen/Studien wichtige Themen des Gebiets „Maschinelle Übersetzung“ aus linguistischer Sicht vorgestellt werden.
- Das Seminar ist nicht nur dazu da, ein Referat loszuwerden, sondern fachliche Fragestellungen zu diskutieren und etwas im Fach zu lernen.
- Die Teilnahme am Seminar schließt die aktive Mitarbeit durch Referat und Hausarbeit ein. Eine Teilnahme nur zum Zuhören ist nicht möglich.
- Einen (auf Wunsch benoteten) Schein erteile ich nach
  - Präsentation eines Referats
  - Abgabe der Hausarbeit bis 4 Wochen nach Semesterende
  - Besprechung der Arbeit in der Sprechstunde (des Folgesemesters)

Vorkenntnisse

- Fachliche Vorkenntnisse auf diesem speziellen Gebiet werden nicht vorausgesetzt, allerdings das
  - linguistische Grundwissen des Einführungsseminars,
  - der Umgang mit Fachliteratur und
  - die Technik des Vortrags.
- Ich erwarte nachhaltige Interesse an einer wissenschaftlichen Beschäftigung mit dem Gebiet der Sprachwissenschaft (Linguistik).
- Wenn Sie etwas nicht verstanden haben, sagen Sie das (es liegt oft nicht an Ihnen). Aber der Referent kann von sich aus oft nicht wissen, was unklar geblieben ist. Sie tun dem Referenten oder der Referentin keinen Gefallen, wenn Sie nichts fragen und nichts in Frage stellen. Ein undiskutiertes Referat ist auch für die Vortragenden wenig erfolgreich.

Ziele des Seminars

- Nichttechnische Einführung in ein neues Gebiet der Sprachwissenschaft mit hoher beruflicher Relevanz
- Einführung in ein technologisches Forschungsgebiet der Sprachwissenschaft.
- Einführung in ein interessantes Kooperationsgebiet von geisteswissenschaftlichen und technischen Fächern (Informatik)

Technisches zum Seminar

- Die Folien stehen jeweils vor der Sitzung im Netz (unter meinem Namen)
- Die Literatur ist teilweise im Germanischen Seminar vorhanden, in einem Seminarordner in Kopie in der Bibliothek oder bei mir/Cristina Vertan auszuleihen,
- Ab der dritten Sitzung werden Referate gehalten, die mit mir zuvor (telefonisch) abgesprochen werden müssen,
- Wir erwarten, daß kein Teilnehmer mehr als zweimal fehlt.
- Die Literatur ist weitgehend englisch, zur Übersetzung von Spezialtermini und anderen Sprachproblemen kann man bei uns anrufen.

Abkürzungen

MT = Machine Translation

- MÜ = Maschinelle Übersetzung
- MAT = Machine Aided Translation
- TM = Translation Memory
- MAHT = Machine Aided Human Translation
- HAMT = Human Aided Machine Translation
- FAHQT = Fully Accurate High Quality Translation
- CL = Computerlinguistik
- NLP = Natural Language Processing (ähnlich CL)
- QS, ZS= Quellsprache / Zielsprache
- SL, TL = Source Language / Target Language
- QS, ZS = Quellsprache / Zielsprache

Maschinelle Übersetzung als Disziplin

MÜ ist keine wissenschaftliche Disziplin als solche, sondern ein Anwendungsgebiet von verschiedenen Wissenschaften.

MÜ basiert auf

- **Linguistik**, ist *aber keine "programmierte Theorie der Mehrsprachigkeit"*
- **Computerlinguistik**, *ist aber keine "programmierte Grammatik"*
- **Übersetzungstheorie**, *aber kein Simulationsmodell des Übersetzers*

und

- **Informatik**, *aber kein algorithmisches Anwenderprogramm wie Gehaltsabrechnungen*

Warum wir maschinelle Übersetzung brauchen

Weltweiter Übersetzungsmarkt hatte einen Wert (in Millionen $) von

1989	20
1990	500
2003	2000

Geschätztes Wachstum liegt bei 20% je Jahr

1986 weltweit über 500 Mio Seiten Übersetzungen, mehr als 100 Mio in Europa.

1% > "schöne Literatur"

30% staatliche Stellen

50% Industrie und Handel (überwiegend technische Dokumentationen)

Zeitersparnis beim Übersetzungsvorgang lt Systran: 75%

Verbesserung des Dienstes durch computergestütztes Übersetzen (nach MBB): 20%

Systran übersetzte im Jahr 1994 140 000 Seiten EU-Dokumente. 80 % der EU-Dokumente zwischen Spanisch und Französisch werden maschinell übersetzt.

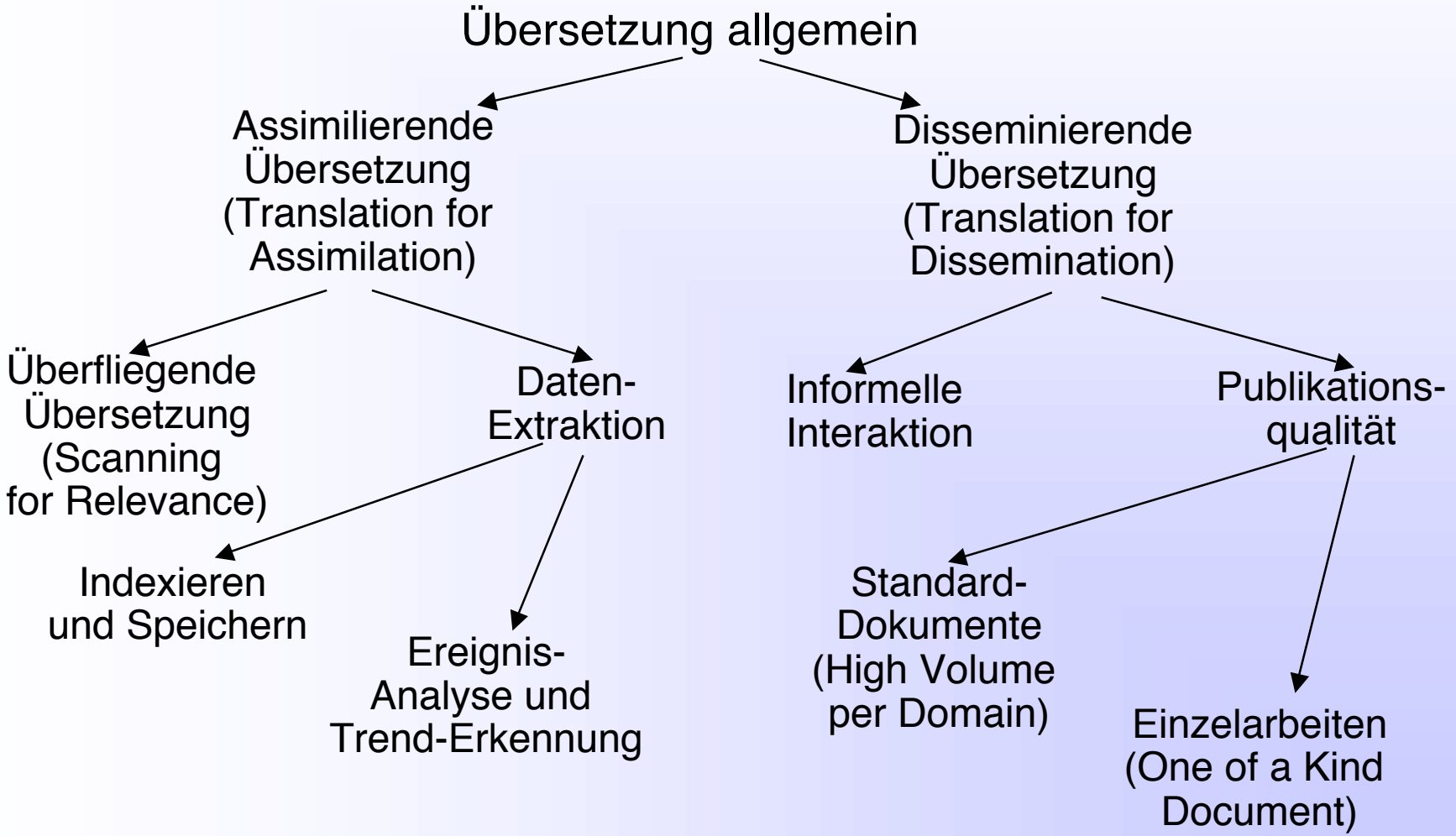
Wir können schon nicht mehr so viele Übersetzer ausbilden, wie wir Übersetzungsbedarf haben.

Qualitätsmerkmale

Eigenschafte, die wir mindestens von MÜ erwarten müssen:

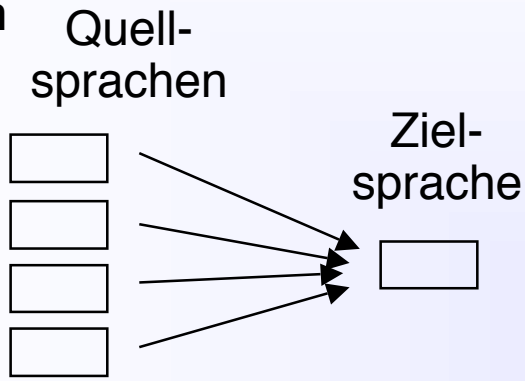
- Semantische Angemessenheit
- Stilistische und pragmatische Angemessenheit
- Niedrigere Kosten gegenüber Humanübersetzern
- Höhere Geschwindigkeit
- Konsistenz im Text und zwischen Texten

Funktionale Typologie von MÜ-Systemen



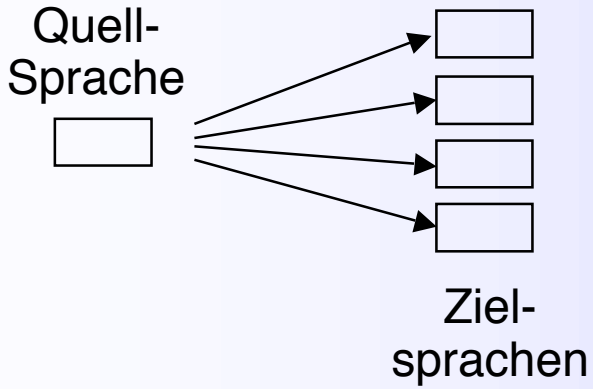
Assimilierendes und Disseminierendes Übersetzen

Assimilation



- Jede Sprache
- Jede Stilebene
- Fast jedes Thema
- Allzweck-Übersetzugen
- Wenig semantische Analyse
- Verlangt Nach-Editierung

Disseminierung



- Eine Quellsprache
- Definierter Sti
- Ein Themaoder Fach
- Spezialübersetzungen
- Volle Semantische Analyse
- Kein Nach-Editieren

Bar Hillels Argumente von 1960

1. Ein Text muß (wenigstens teilweise) verstanden worden sein, bevor man eine sinnvolle Übersetzung anfangen kann
2. Computer-Verstehen von Texten ist zu komplex
3. Deshalb ist Automatische Übersetzung unmöglich

Heutiger Stand:

Prämisse 1 ist richtig

Prämisse 2 war nur 1960 richtig

Conclusion stimmt heute nicht mehr

Historische Entwicklung

Jahr	U.S.	Europa	Japan
1950er	Start von größeren MÜ- projekten		Frühe MÜ-Forschung
1960er	ALPAC Ende der MÜ	Start der MÜ	
1970er	(SYSTRAN, METAL NLP Grundlagen-Forschung	GETA EUROTRA	
1980er	Erneuter Start in MÜ-Forschung (SYSTRAN)	EUROTRA (METAL SYSTRAN)	MU-System MÜ Boom in der Industrie
1990er	Offizielle MÜ-Forschung (SYSTRAN) Mehrsprachige Systeme	Ende von EUROTRA NLP Grundlagen-forschung, VERBMOBIL	MÜ-Produkte Grundlagen-Forschung CICC, EDR, …

Größere kommerzielle Systeme

Sprachen	LOGOS E, G, S, F	SYSTRAN (FTD) E, R	SYSTRAN (mehrspr.) E, F, S, etc	METAL E, G
Typ	Transfer, Etwas Semantik	direkt	Transfer, Etwas Semantik	Transfer, Kasus-Semantik
Themen	Handbücher und Allg.	Technische Texte	Handbücher und Allg.	Handbücher und Allg.
Status	im Handel	US-gefördert	Im Handel	Im Handel
Firma	Logos Corp.	LATSEC Inc.	SYSTRAN	(SNI)

Übersetzung und Verstehen

Beispiel: Englisch ➜ Spanisch

While driving down route 72, John swerved and hit a tree

Problem:

Im Spanischen kann man “hit” übersetzen mit:

1.	*pegar*	mit der Absicht, zu versetzen	
2.	*chocar*	zufällig, durch ein bewegtes Objekt	
3.	*acertar*	ins Schwarze treffen	
4.	*golpear*	(andere Variante)	etc.

Wie kann ein MÜ-System die korrekte Wahl treffen?

Nur mit lexikalischen Äquivalenten, ohne (wenigstens lexikalische) Semantik aussichtslos.

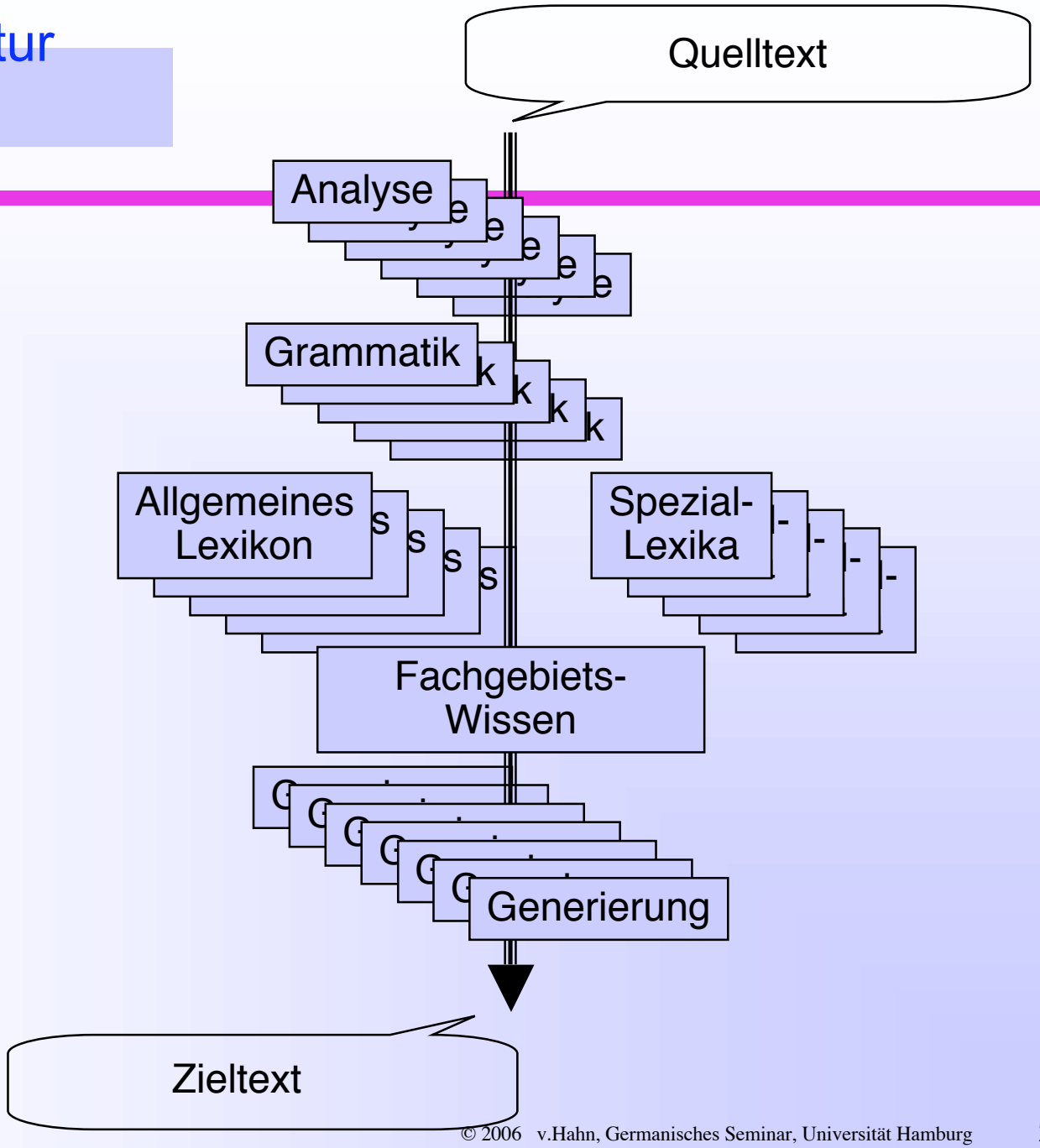
Ambiguität in der Quellsprache - 1 -

- Syntaktische Ambiguität:
  - *I saw the Grand Canyon flying to New York*
  - *Ich sah die Expedition auf dem Berg mit dem Fernrohr*
- Lexikalische Abiguität
  - *The man went to the bank to get some cash*
  - *The man went to the bank and jumped in*

Ambiguität in der Quellsprache - 2 -

- Kasusrollen-Ambiguität
  - *He ran the mile in four minutes*
  - *He ran the mile in the Olympics*
- Referentielle Ambiguität
  - *I took the cake from the table and cleaned it*
  - *I took the cake from the table and ate it*
- Pragmatische Ambiguität
  - *Can you open the door?*
  - *Haben Sie eine Uhr?*

Standard-Architektur für MÜ-Systeme



Maschinelles Dolmetschen

- Neues Forschungs- unt Technologiegebiet mit Anwendung im
  - Konsekutivdolmetschen
  - Simultandolmetschen
  - Gesprächsdolmetschen
- Interessant wegen der Verbindung von
  - Signalebene Phonetik und
  - Textebene Linguistik
- Hohe Relevanz für kognitive Linguistik
  - Dolmetschstrategien
  - Verstehen
  - Zeitverhalten
  - Abbildung von Sprecher- und Spracheigenschaften

Forschungsbedarf

Linguistische Theorie

Theorie der ComputerlinguistikCL

“Linguistic Engineering”, so etwas wie Linguistische Technologie

Industrielle Forschung und Entwicklung für Sprachprodukte

Benutzerforschung

Korpusforschung und -sammlung

Eine neue Forschungs- und Entwicklungsebene: Language Engineering

In anderen Fächern gibt es jeweils neben der wissenschaftlichen Ebene die Ingenieur-Ebene:

Chemie-Ingenieurwesen, Bau-Ingenieurwesen, Fahrzeug-, Flugzeug- ...

Auch in der Sprachverarbeitung entsteht ein eigener Bedarf an Forschung und Entwicklung zwischen der akademischen Forschung und der Herstellung.

Dort sollte bearbeitet werden:

- Funktionale und technische Evaluation von Produkten,
- Testverfahren und Fehlerstatistiken,
- Technologie der Module und wiederverwendbaren Komponenten,
- Marktsegmente und Kosten,
- Funktionale Spezifikationen,
- Standardtechniken und deren Zuverlässigkeit,
- Benutzertypologie und Bedarfsanalyse.