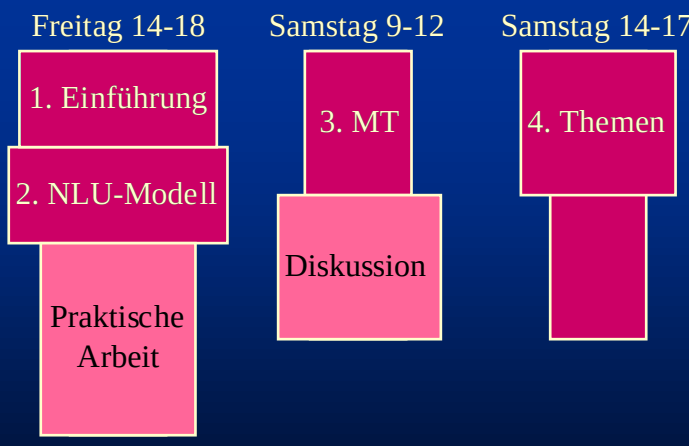
Anwendungen der Computerlinguistik

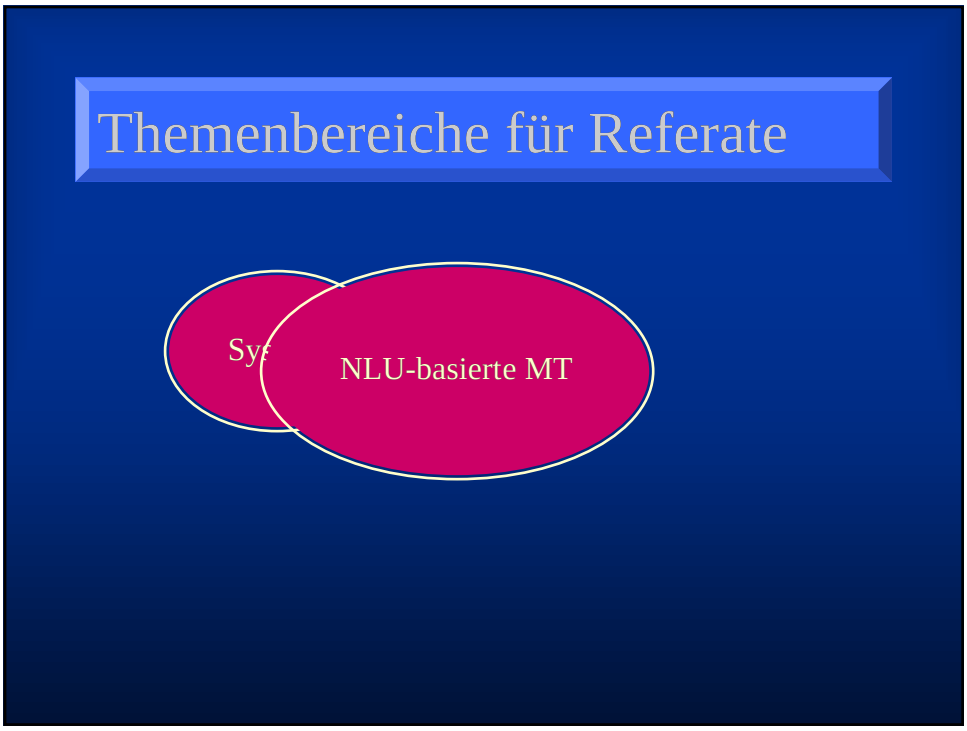
4. Übersicht der Themen

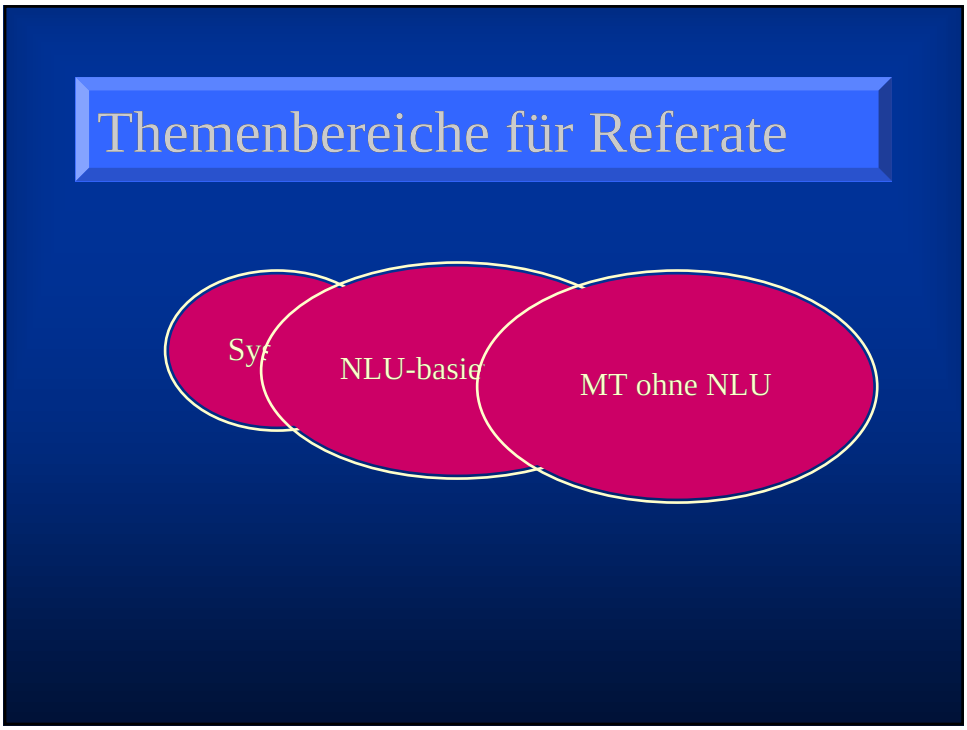
Pius ten Hacken

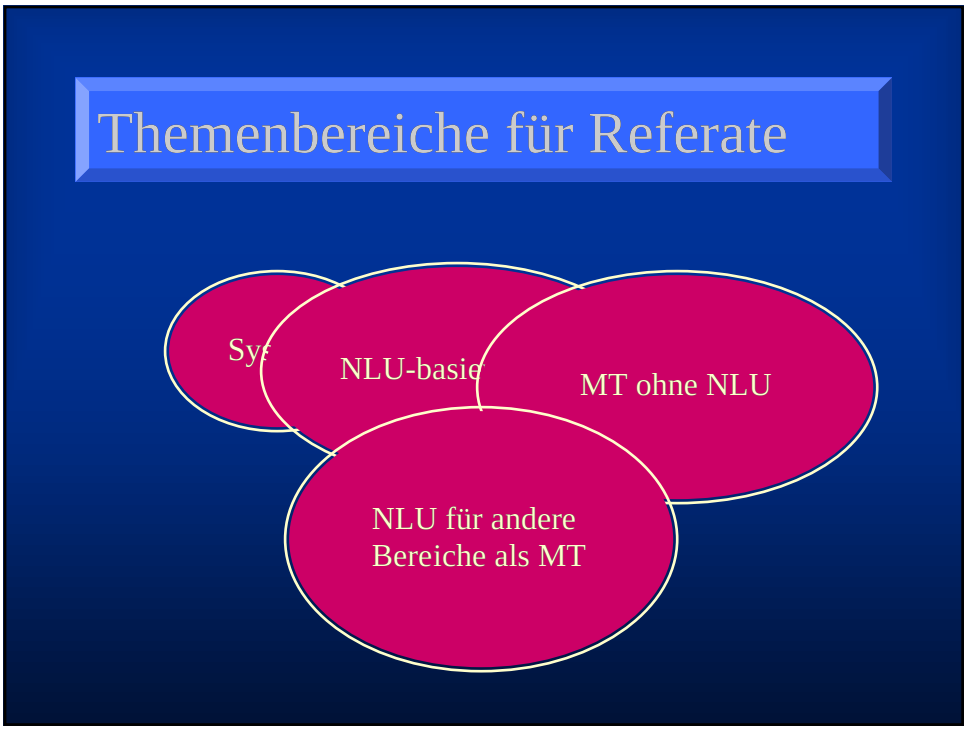
Block 1

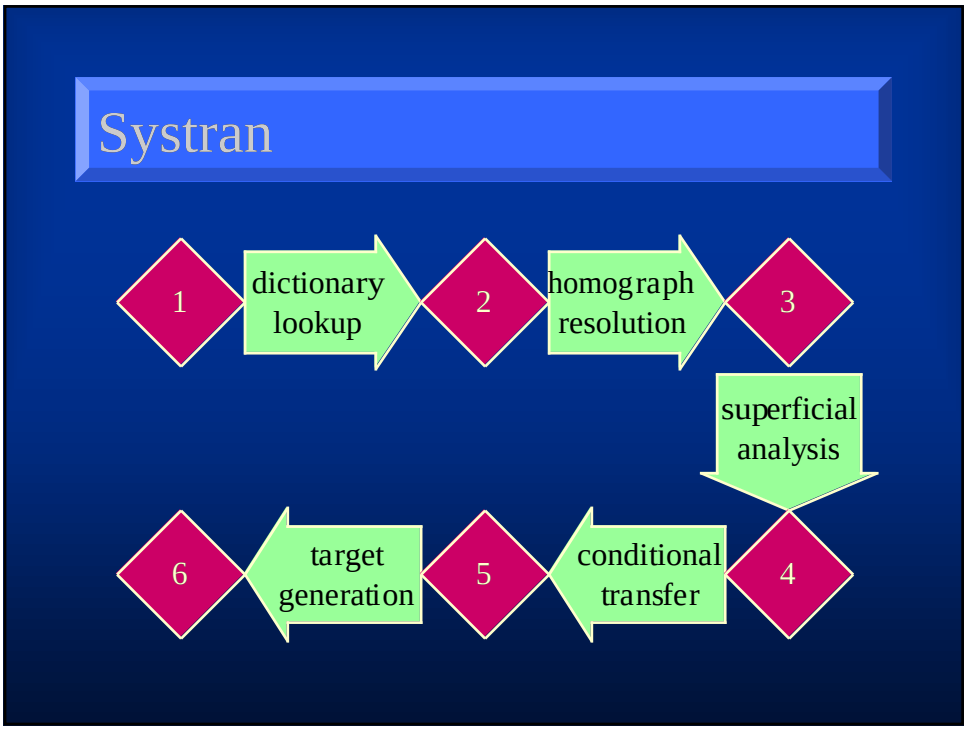


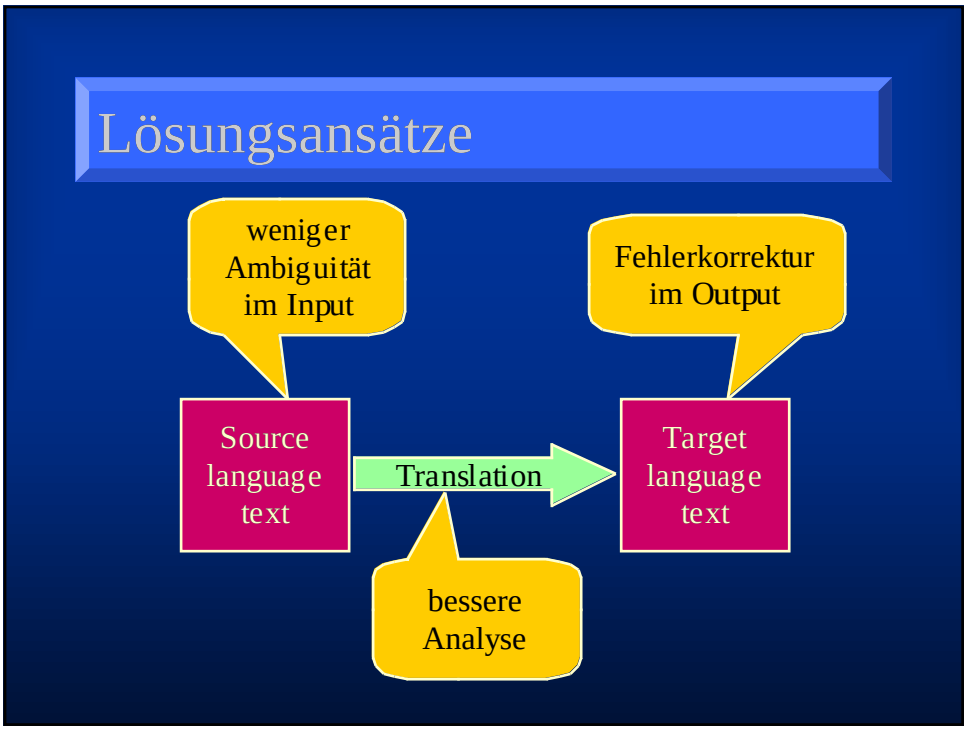

Themenbereiche für Referate
Systran

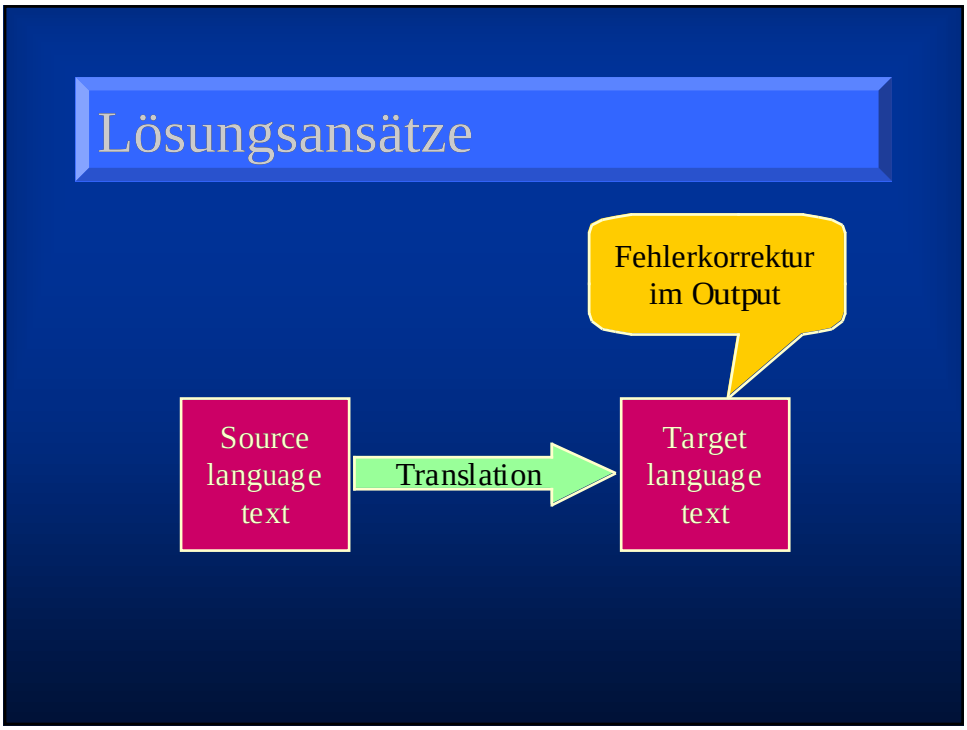

Themenbereiche für Referate
Sys
NLU-basierte MT

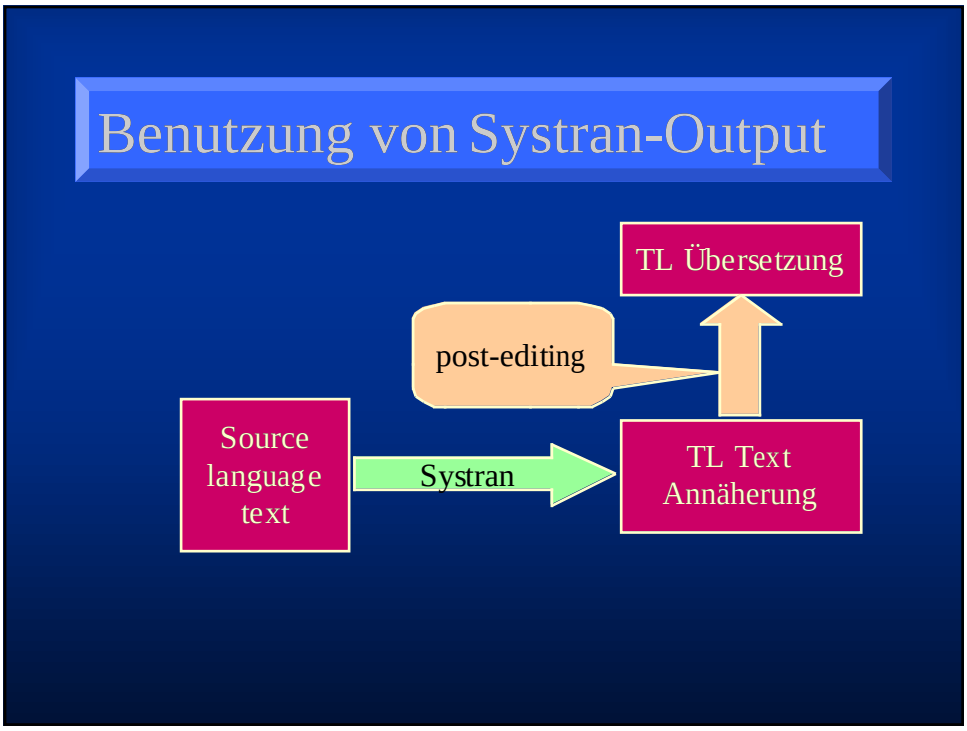

Themenbereiche für Referate
Sy
NLU-basie
MT ohne NLU


Themenbereiche für Referate
Sy
NLU-basie
MT ohne NLU
NLU für andere
Bereiche als MT


Systran
1
dictionary lookup
2
homograph resolution
3
superficial analysis
4
conditional transfer
5
target generation
6


Lösungsansätze
weniger Ambiguität im Input
Fehlerkorrektur im Output
Source language text
Translation
Target language text
bessere Analyse

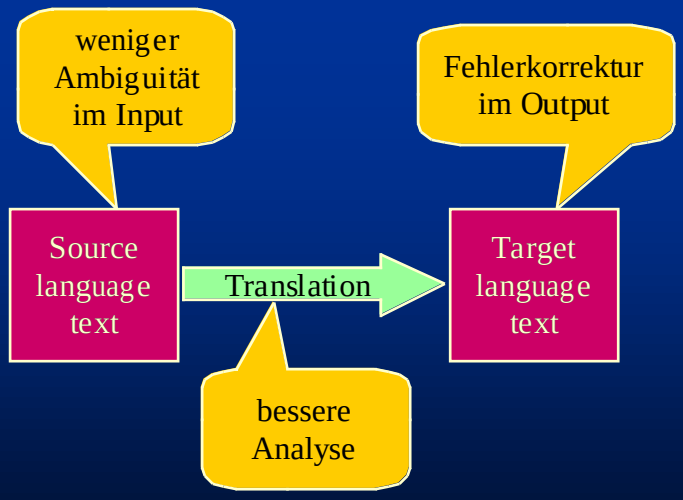

Lösungsansätze
Fehlerkorrektur im Output
Source language text
Translation
Target language text

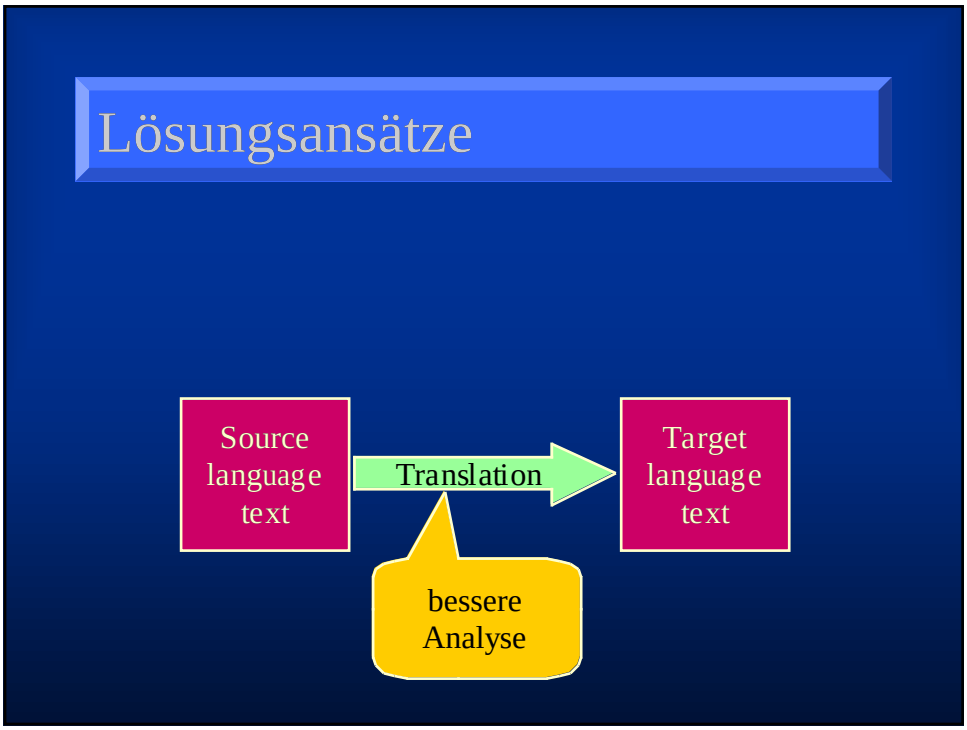

Benutzung von Systran-Output
TL Übersetzung
post-editing
Source language text
Systran
TL Text Annäherung

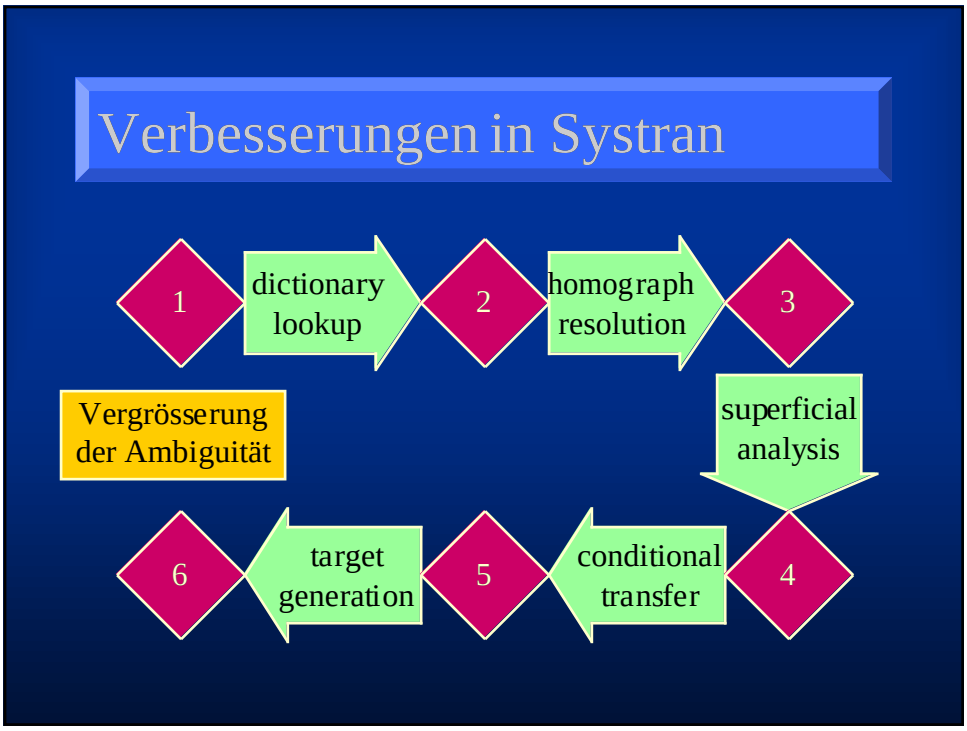
Post-editing

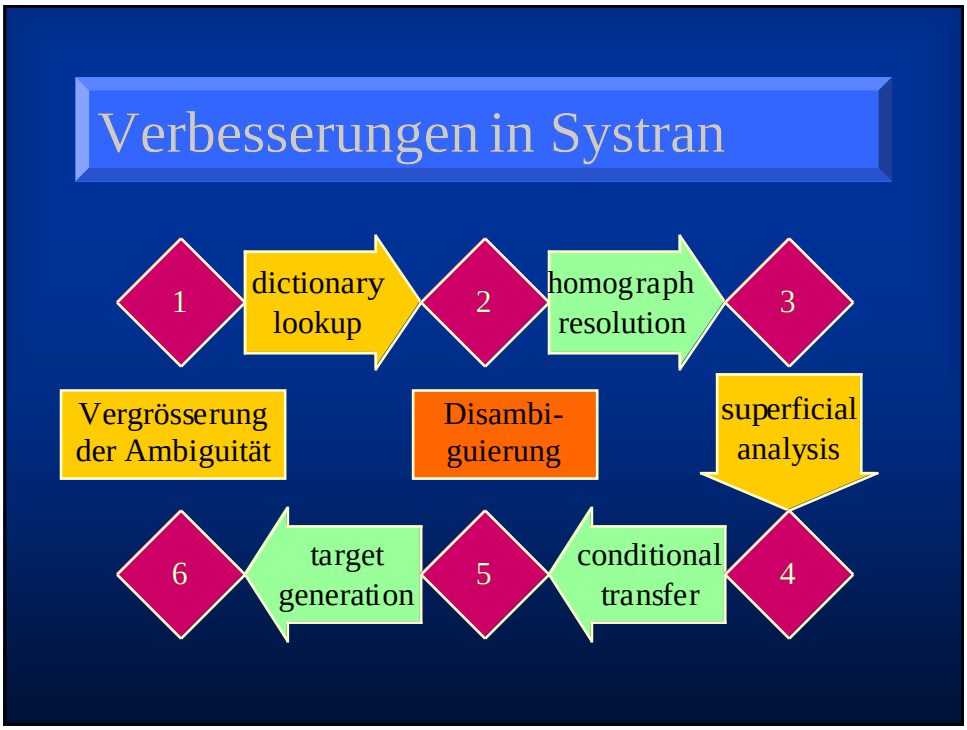
- übliche Praxis in der professionellen Übersetzung
- effizientes Post-Editing von MT Output erfordert:
  - andere Herangehensweise (minimalistisch statt perfektionistisch)
  - Unterstützung durch editing software

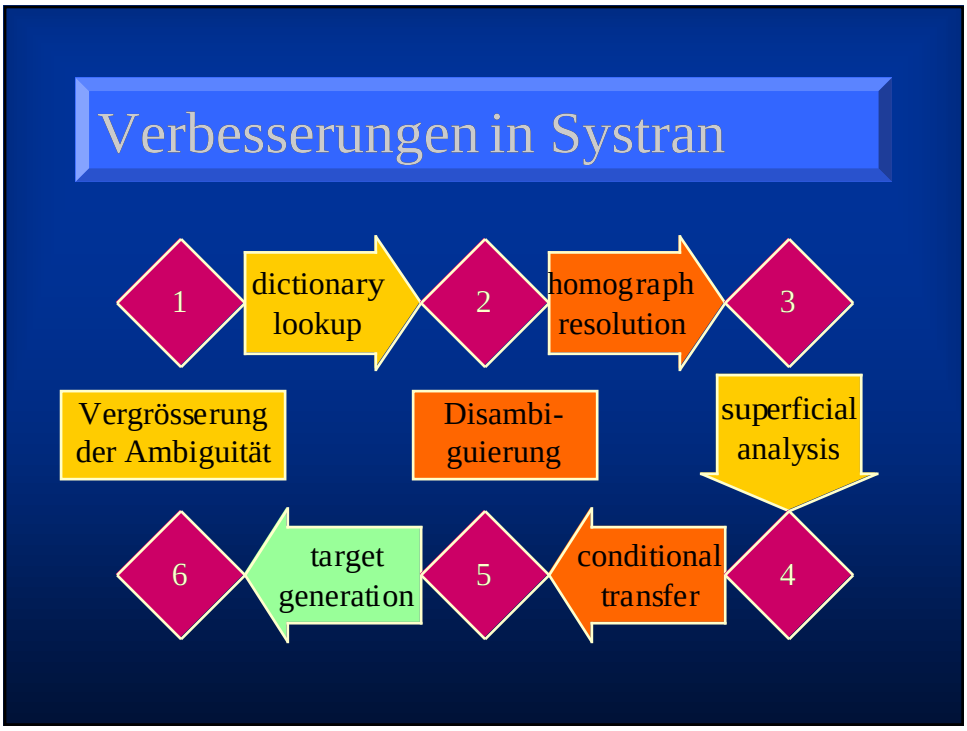
Lösungsansätze




Lösungsansätze
Source language text
Translation
Target language text
bessere Analyse


Verbesserungen in Systran
1
dictionary lookup
2
homograph resolution
3
Vergrösserung der Ambiguität
superficial analysis
6
target generation
5
conditional transfer
4


Verbesserungen in Systran
1
dictionary lookup
2
homograph resolution
3
Vergrösserung der Ambiguität
Disambi-guierung
superficial analysis
6
target generation
5
conditional transfer
4


Verbesserungen in Systran
1
dictionary lookup
2
homograph resolution
3
Vergrösserung der Ambiguität
Disambi-guierung
superficial analysis
6
target generation
5
conditional transfer
4

Reduktion der Ambiguität

- Lösung der Ambiguität innerhalb der Systran-Architektur:
  - mehr oder bessere Kontexte für homograph resolution
  - mehr oder bessere Kontexte für conditional semantics, Präpositionen, lexical routines
- Vermeidung der Ambiguität:
  - weniger Homonyme im Lexikon
  - weniger Strukturen in der Analyse

Ambiguität und Kenntnisse

- Je mehr wir über den Input wissen, desto mehr können wir die Ambiguität einschränken.
- Wissensquellen:
  - allgemeine Eigenschaften der Sprache
  - Generalisierungen über den spezifischen Input, den wir erwarten können
  - Manipulierung des Inputs

Generalisierungen

Le succès de cette méthode est évident.

The success of this method is obvious.

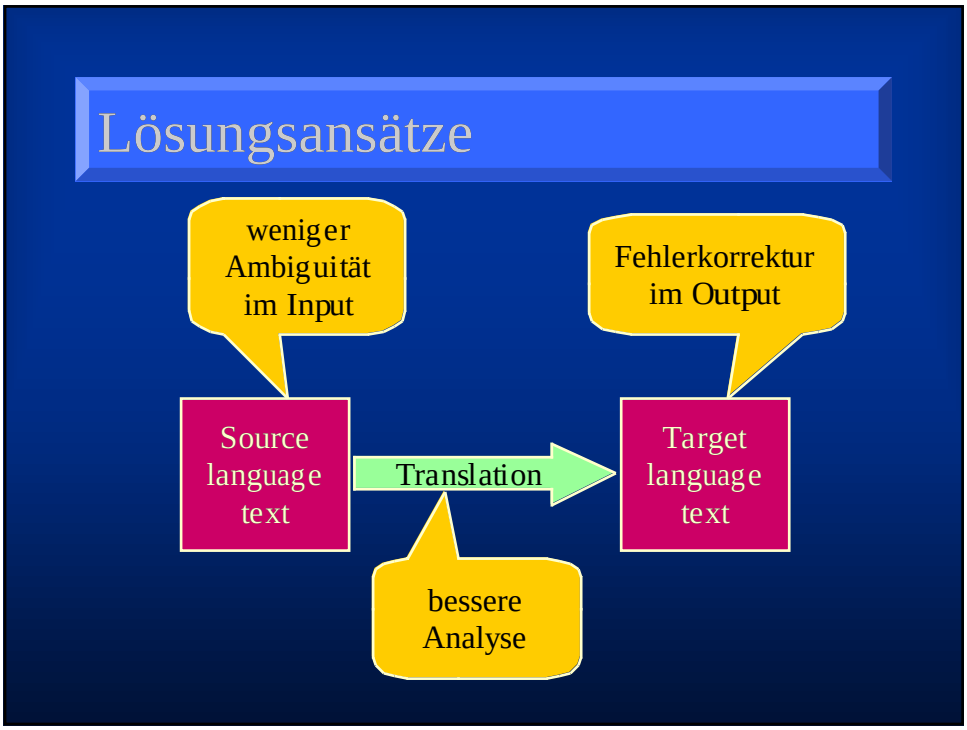
succès:	méthode:	évident:
a. success	a. method	obvious
b. bestseller	approach	evident
hit	b. manual	self-evident
c. conquest	tutor	
	primer	

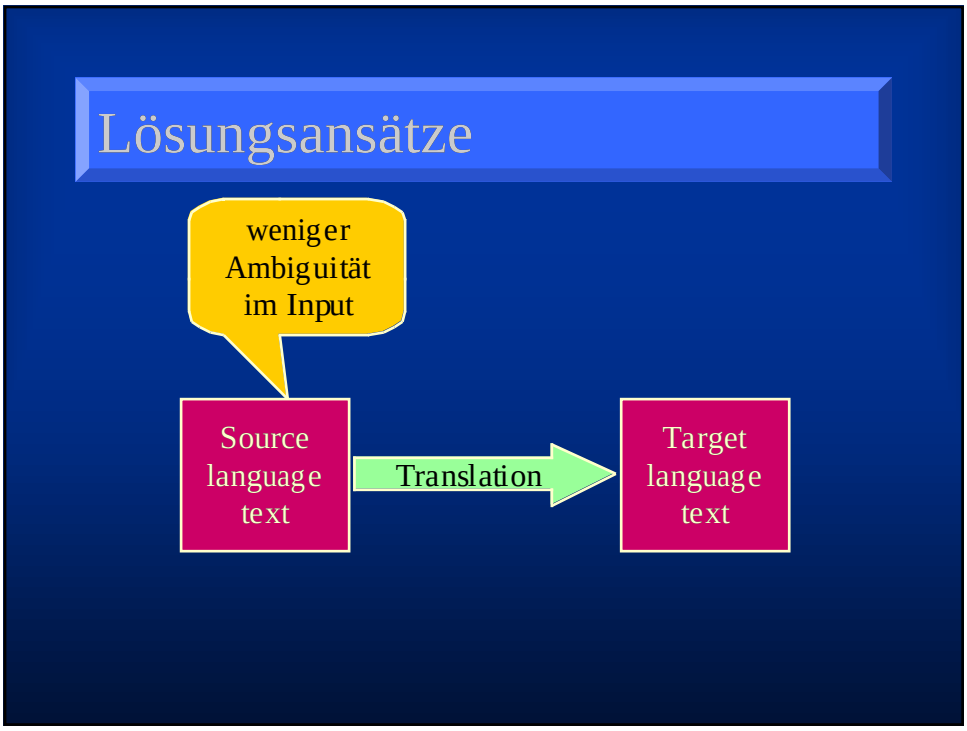
Customization

- Entfernung von irrelevanten Bedeutungen im Lexikon
- Generalisierungen für die verbleibenden Bedeutungen
- Fehleranalyse im Post-Editing

Software-Hersteller

EndbenutzerIn

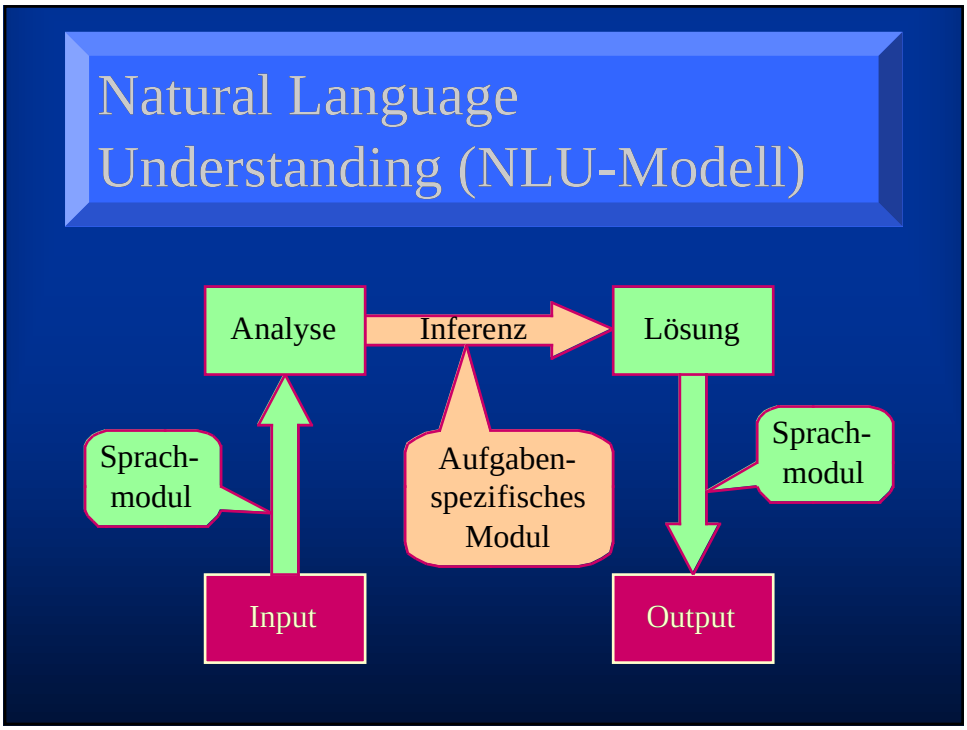

Lösungsansätze
weniger Ambiguität im Input
Fehlerkorrektur im Output
Source language text
Translation
Target language text
bessere Analyse

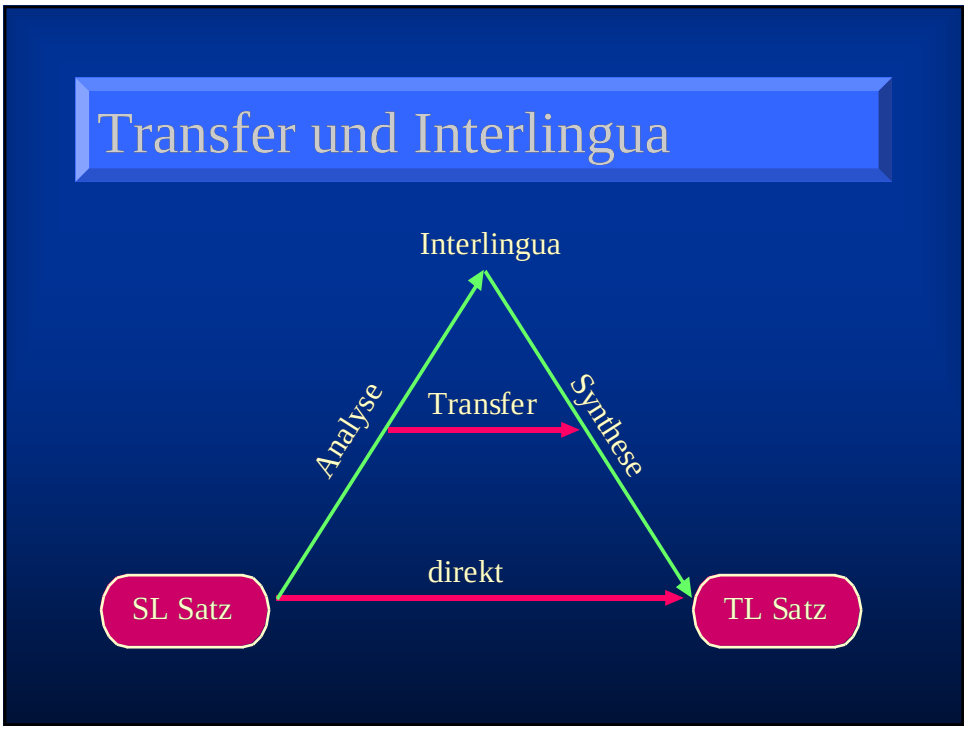

Lösungsansätze
weniger Ambiguität im Input
Source language text
Translation
Target language text

Controlled language

- selbst-erstellte Texte
- beschränkte Domäne
- genügende Textmenge
- Ausschluss von Ambiguität im Input
- Probleme:
  - genügend Ausdrucksmöglichkeiten
  - Unterstützung bei der Einhaltung der Einschränkungen

Themenbereiche für Referate


Natural Language Understanding (NLU-Modell)
Analyse
Inferenz
Lösung
Sprach-modul
Aufgaben-spezifisches Modul
Sprach-modul
Input
Output


Transfer und Interlingua
Interlingua
Analyse
Transfer
Synthese
direkt
SL Satz
TL Satz

Eurotra

- Transfersystem
- neun Sprachen, dezentrale Entwicklung
- Minimierung des Transfers durch Angleichung der IS
- linguistisch informierte Analyse
  - eklektische Behandlung der Theorie
  - keine absolute Vereinheitlichung des Analyseprozesses

Rosetta

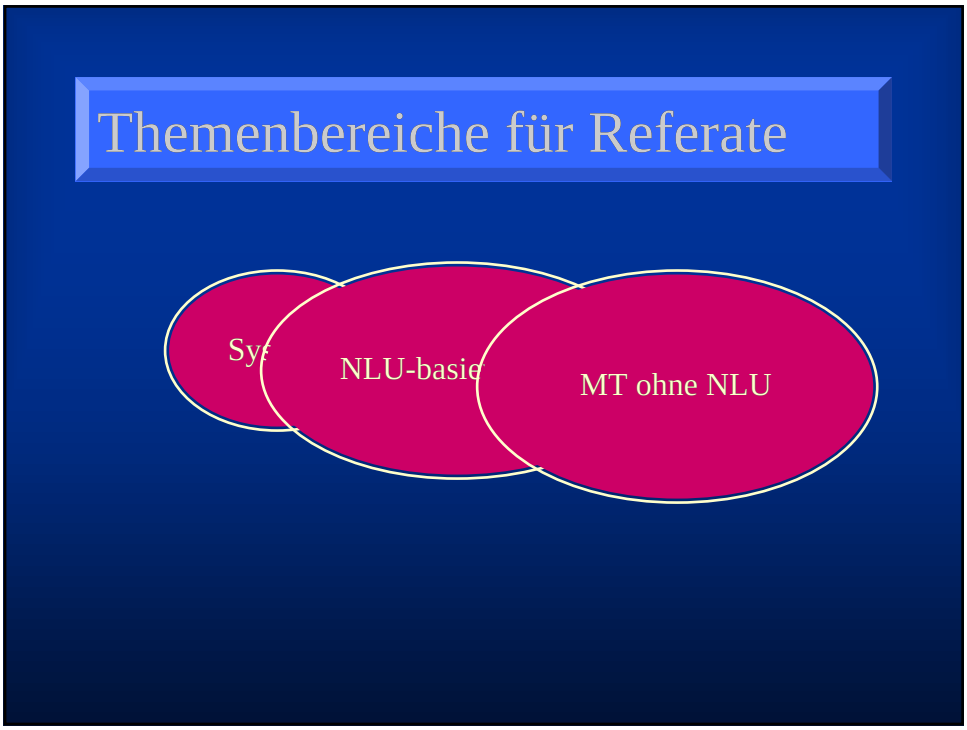
- Interlingua-System
- drei Sprachen, zentrale Entwicklung
- Gesamtarchitektur basiert auf Logik:
  - Kompositionalität
  - “Menge der möglichen Übersetzungen”
- eklektische Einpassung von Analysen einzelner Phänomene

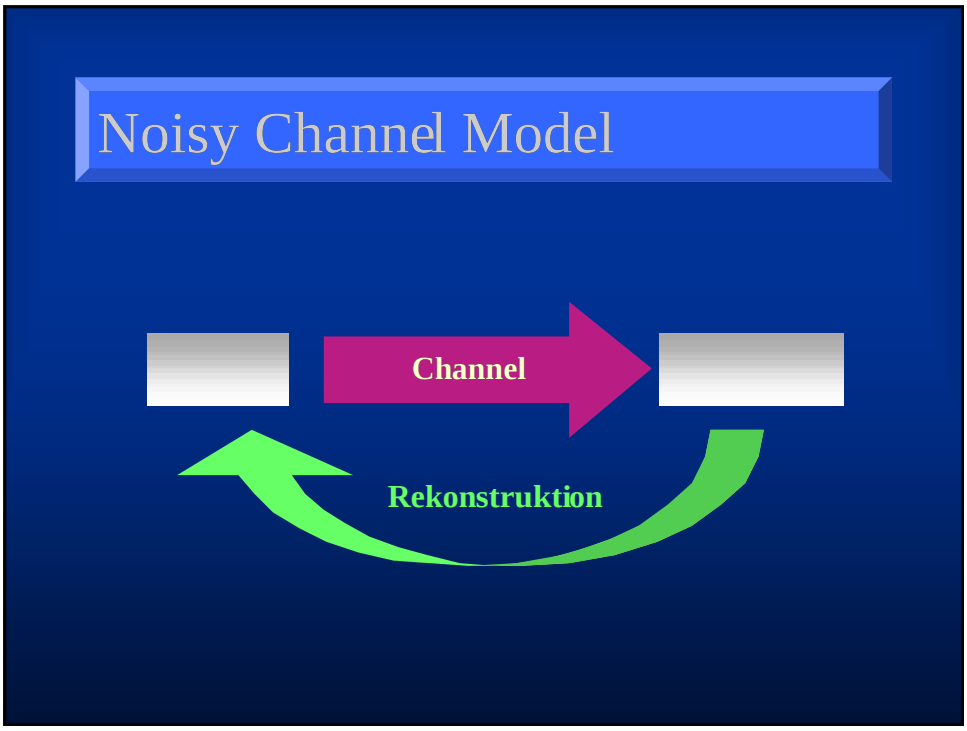
Unitran

- Transfersystem
- one-woman project
- Linguistisch basierte Analyse
  - Syntax gemäss Chomsky
  - Semantik gemäss Jackendoff
- Klassifikation und Lösung von Divergenzen in der Übersetzung

Knowledge-Based MT

- Interlingua-Ansatz
- Künstliche Intelligenz als Basis
- Sprachtheorie als Hilfsmittel
- Repräsentation einer Domäne mit sprachunabhängigen Konzepten


Themenbereiche für Referate
Sy
NLU-basie
MT ohne NLU

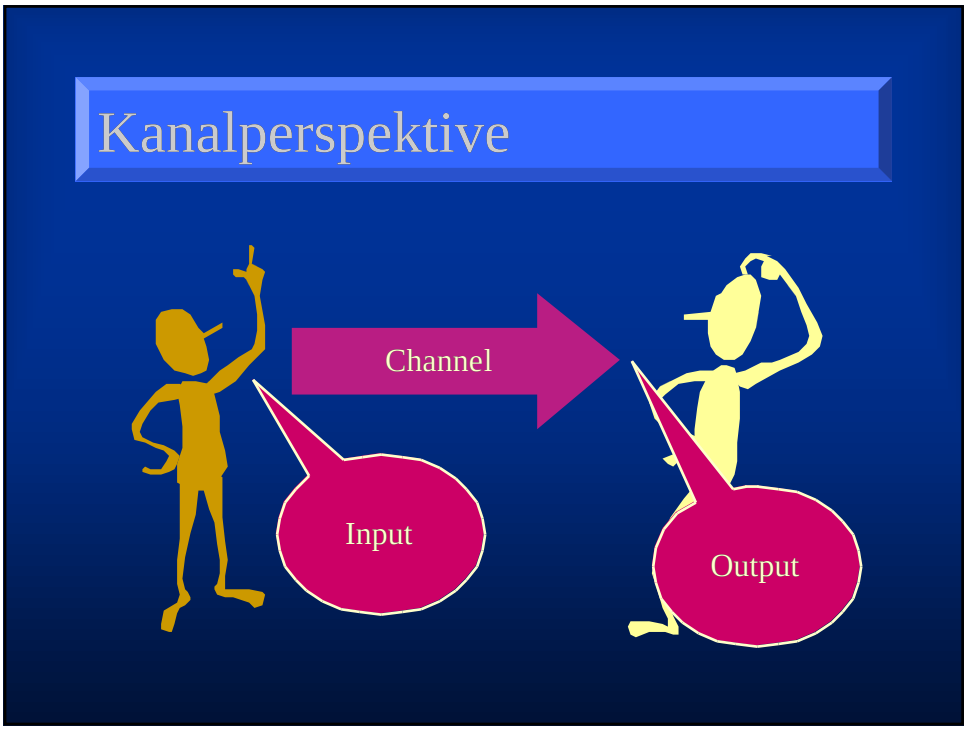

Noisy Channel Model
Channel
Rekonstruktion

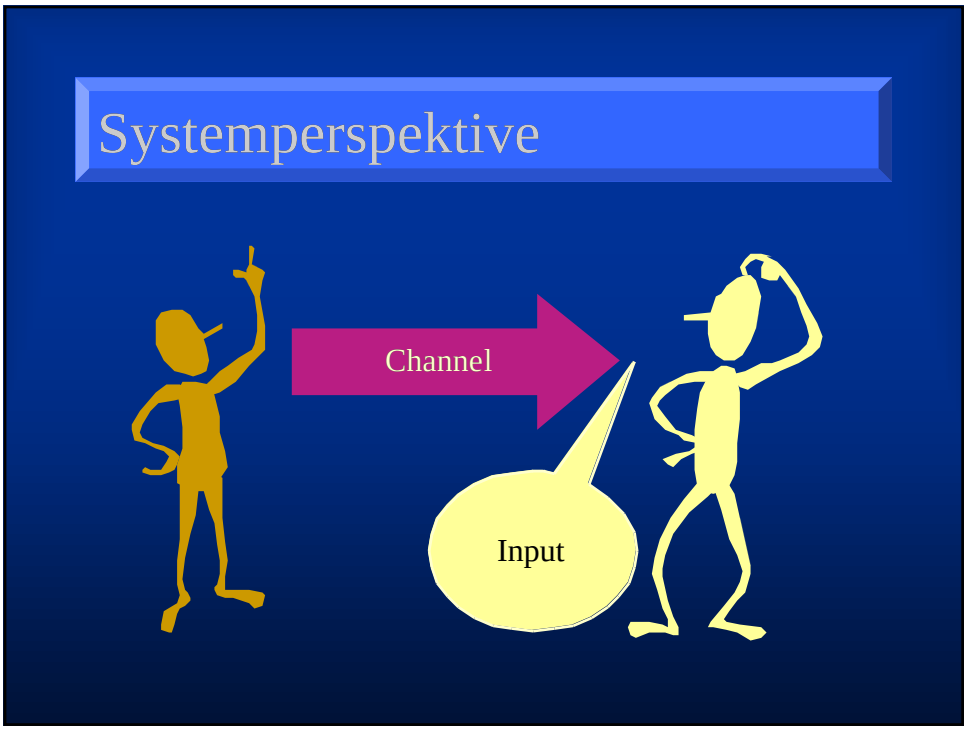
Information Theory

- Claude Shannon,
  - Bell Laboratories, 1948
- Information als Problem für Engineering
- Rekonstruktion auf der Basis von probabilistischen Verfahren

Form und Bedeutung

- "Frequently the messages have *meaning*; that is they refer to or are correlated according to some system with certain physical or conceptual entities.
- These semantic aspects of communi-cation are irrelevant to the engineering problem." [Shannon (1948)]

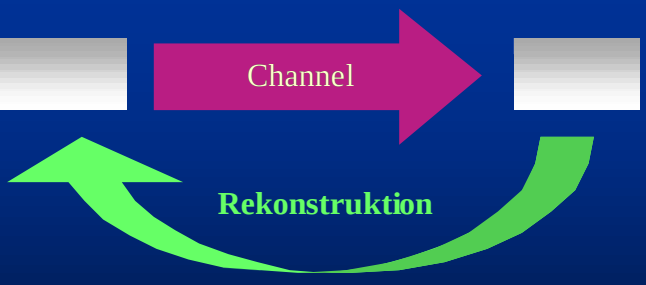

Kanalperspektive
Channel
Input
Output


Systemperspektive
Channel
Input

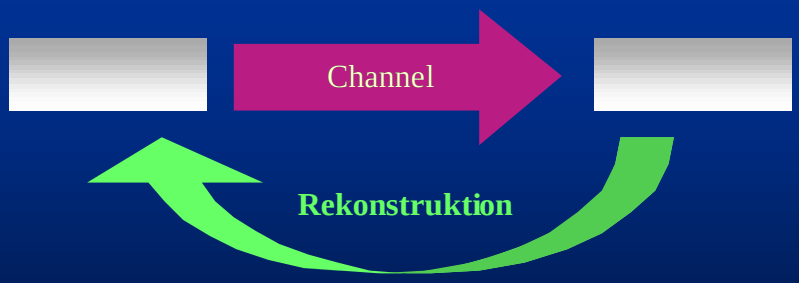
Rekonstruktion

- Der Output des Kanals wird bestimmt von:
  - Eigenschaften des Inputs
  - Eigenschaften des Kanals

Tagging im Noisy Channel Model



MT im Noisy Channel Model



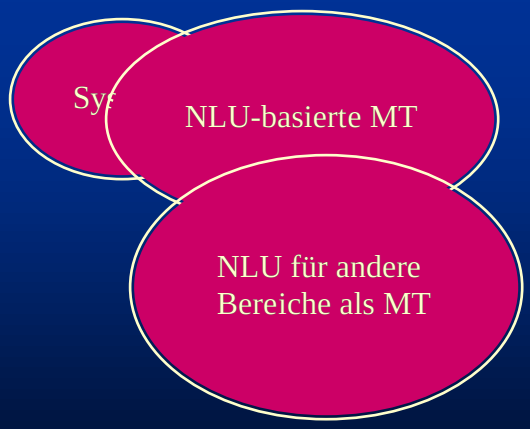
Schritte in der Entwicklung

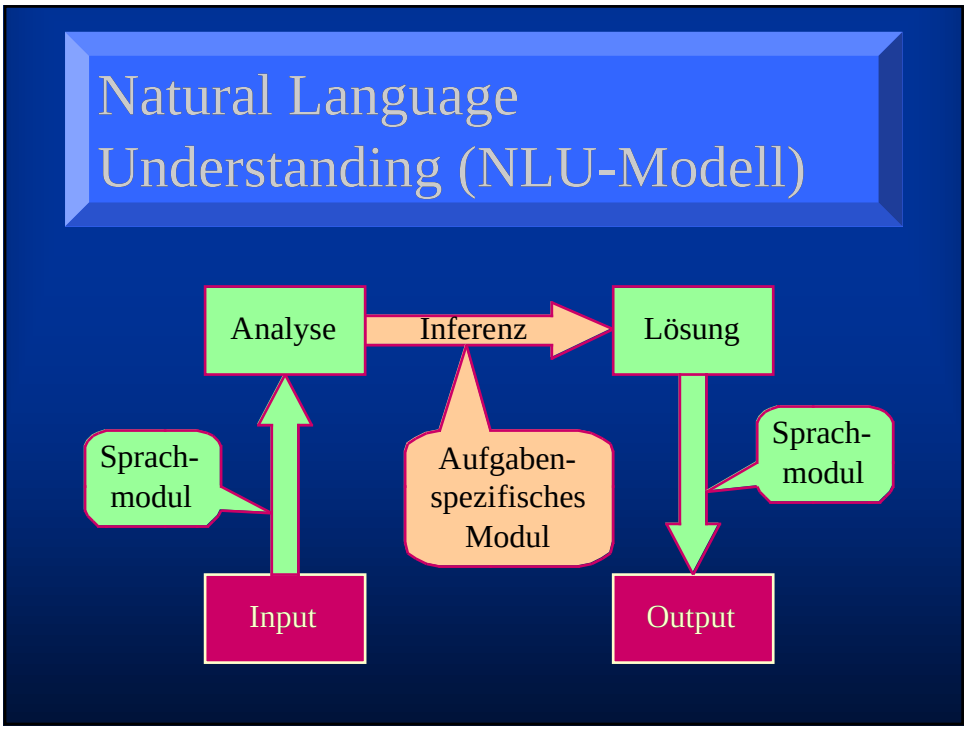
- Corpussammlung
- Alignment
- Evaluation der Corpora
  - Korrespondenz der Wörter aus SL mit TL
  - Abfolge der Wörter in der TL
- Maximierungsalgorithmus

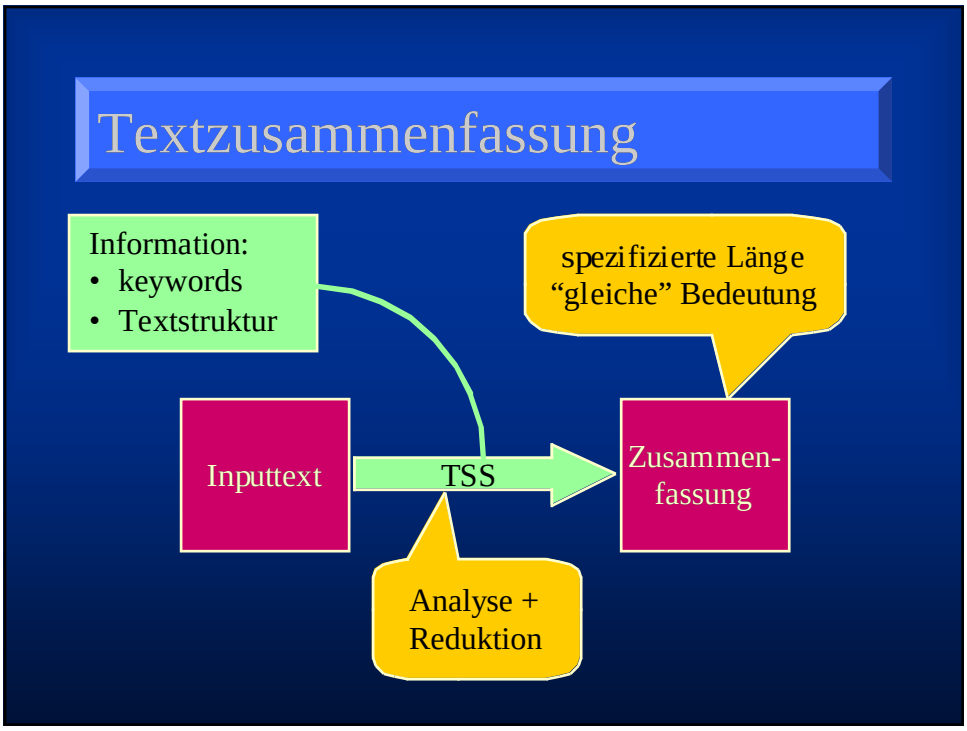
Example-Based MT

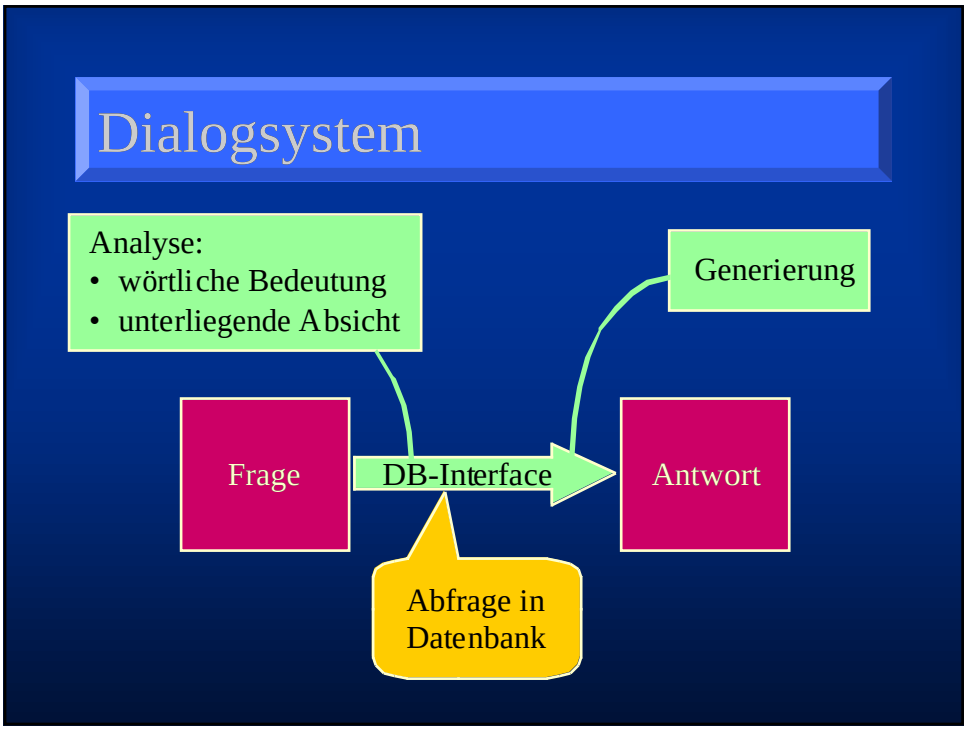
- Alternative zur Anwendung von Bayes's Theorem
- Datenbank von Beispielen mit korrekten Übersetzungen
- Algorithmus zur Erkennung der richtigen Beispiele

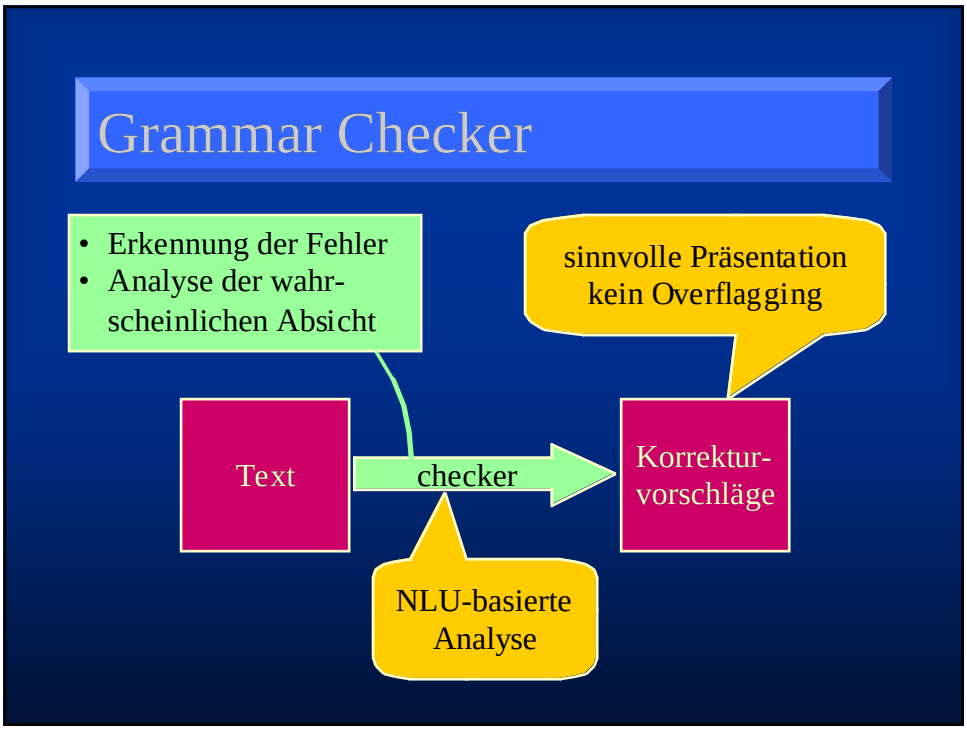
Themenbereiche für Referate




Natural Language Understanding (NLU-Modell)
Analyse
Inferenz
Lösung
Sprach-modul
Aufgaben-spezifisches Modul
Sprach-modul
Input
Output


Textzusammenfassung
Information:
• keywords
• Textstruktur
spezifizierte Länge
“gleiche” Bedeutung
Inputtext
TSS
Zusammen-fassung
Analyse + Reduktion


Dialogsystem
Analyse:
• wörtliche Bedeutung
• unterliegende Absicht
Generierung
Frage
DB-Interface
Antwort
Abfrage in Datenbank


Grammar Checker
• Erkennung der Fehler
• Analyse der wahr-scheinlichen Absicht
sinnvolle Präsentation kein Overflagging
Text
checker
Korrektur-vorschläge
NLU-basierte Analyse

Übersicht der Themen

- Controlled Language
- Eurotra
- Rosetta
- Unitran
- KBMT
- Statistisches MT
- EBMT
- Text Summarization
- Dialogsysteme
- Grammar Checkers

Block 1