Integration von Regel- und Statistik-basierten Methoden in der Maschinellen Übersetzung

Kurt Eberle
Lingenio GmbH

Köthen
4. Juli 2008

Überblick

- Das Übersetzungsproblem:
  Mehrdeutigkeit, Translation mismatches
- Lösungsvorschläge:
  1., 2., 3. Generation der MÜ
- Aktuelle Diskussion:
  Hybride Systeme
- *translate*:
  Hintergrund, Architekturen
- 'Hybridisierung' von *translate*:
  Statistische Korpus-Information für Analyse, Transfer und Generierung

Mehrdeutigkeit

- lexikalisch
  - Kategorial
    - *Time$_{N/V}$ flies$_{N/V}$ like$_{V/P}$ an arrow - Zeitfliegen lieben einen Pfeil*
  - Semantisch
    - *Sie stellt$_{PHYS/SOC}$ den Drucker ein$_{PERSON/TOOL}$ - She hires/adjusts the printer.*
- strukturell
  - Funktionale Ambiguität (Label-/Etiketten-Ambiguität)

    Ein Bild Albrecht Dürers. - A picture of/by Albrecht Dürer
  - Attachment-Ambiguität

    Bilder von der Kanzlerin hinter dem Tresen. - Pictures of the chancellor behind the bar
  - Skopus-Ambiguität
    Viele Hunde jagen eine Katze - many dogs chase a cat
- referentiell, pragmatisch
  - *They saw the Alps as they flew over Zurich.*

Mehrdeutigkeit und Übersetzung

- Filter für die Disambiguierung
  - Syntaktische Constraints
    time flies like ... - **Zeit fliegt mag/mögen ...*
  - Semantische Selektionsbeschränkungen
    Sie stellt den Drucker ein ... **She hires the print device*
- Nicht alle Mehrdeutigkeiten müssen aufgelöst werden!
  - *printer* - *Drucker*
  - *viel(x,hund,ein(y,katze,jagen(x,y)))* / *ein(y,katze,viel(x,hund,jagen(x,y)))*
- Variable Analysetiefe- Übersetzung als *negociator*
  (Kay Gawron Norvig1994)

Translation mismatches

(Kameyama Ochitani Peters 1991)

- Lexikalische Divergenz

  un novillo - ein Jungbulle - a young bull
  Rappe, Schimmel, . . .
  Boden - soil (Durrell 1988)

- Thematische Divergenz und Scrambling
  (Dorr 1990, Hutchins Somers 1992)

  Mir gefällt die Aufführung - I like the performance
  Il remet le bouquet à la femme - er überreicht der Frau den Strauß/ den Strauß der Frau

- Hinzufügen und Tilgen von Teilstrukturen
  er durchschwimmt den Fluß - Il traverse la rivière en nageant

- Strukturumkehrung (Head switching)
  Er raucht gerne- He likes to smoke
  La plupart des gens aiment le foot - Die meisten Leute mögen Fußball

Translation mismatches und Repräsentationen

- Morphosyntaktische Repräsentation

  Er schreibt an die Angestellten $\leftrightarrow$*Il écrit aux employés*

  $$[\text{Prep}\ \ [\text{DET}_{def}\ \text{N}]\ _{NP}]$$

- Funktional/Semantische Repräsentation
  Peter würde den Wein nicht mögen. $\leftrightarrow$*Peter n'aimerait pas le vin.*

$$\begin{bmatrix} \text{PRED:} & \text{"mögen}\langle(\uparrow\text{SUBJ})\ (\uparrow\text{OBJ})\rangle\text{"} \\ \text{SUBJ:} & [\text{PRED: "wein"}] \\ \text{OBJ:} & [\text{PRED: "peter"}] \\ \text{NEG:} & + \\ \text{TENSE:} & \text{COND} \end{bmatrix}$$

- Semantisch/Konzeptuelle Repräsentation

  Peter raucht gerne $\leftrightarrow$*Peter likes to smoke*

$$\text{s}$$
$$\text{s} : \text{ATT}(\text{peter}, \{\ <\text{POS\_DISP}, \lambda x.rauchen(x) >\ \})$$

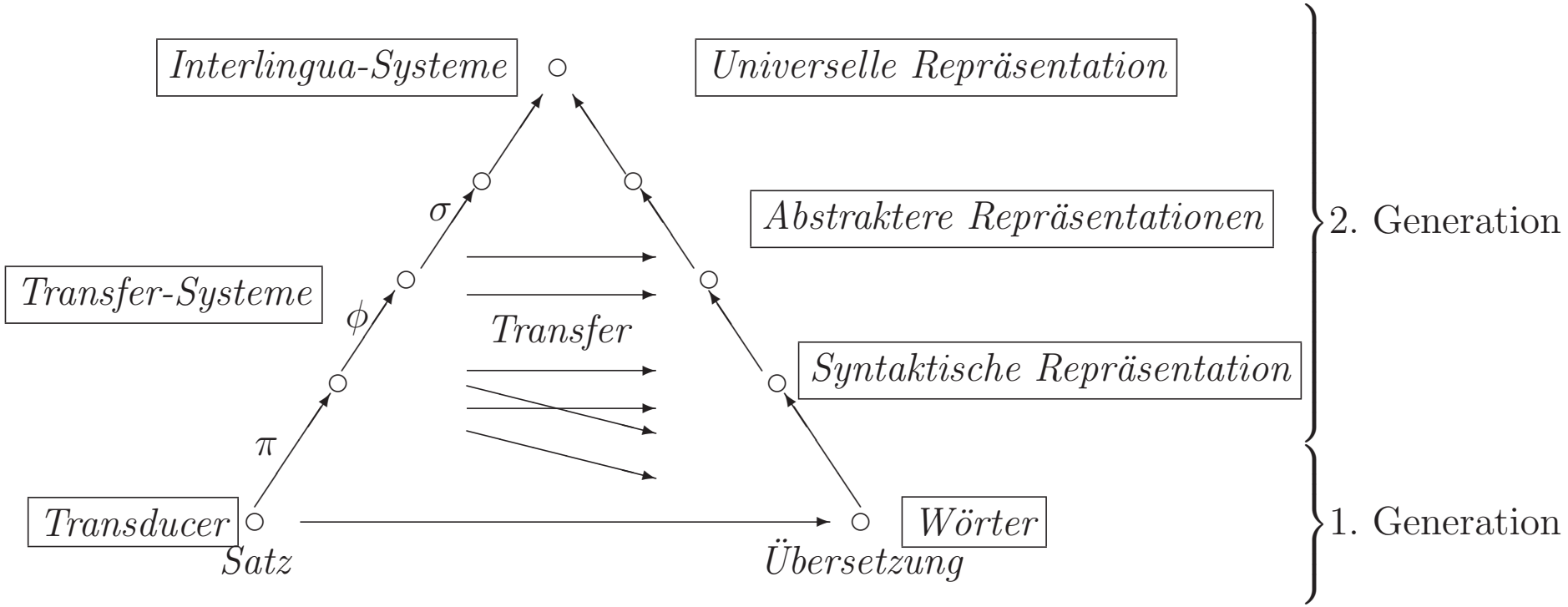
Mehrdeutigkeit, Mismatches und Analysen

- Analysen filtern!
- je abstrakter die Analyse desto geringer der Source-Target-Unterschied
- je abstrakter die Analyse desto mehr Disambiguierungsaufwand

→ **Optimierungsproblem !**

Traditionelle Architekturen

(Vauquois 1975)



Regel-basierte Architekturen

Lösungsversuche - 3. Generation

- Statistik-basierte Systeme (SMT)
  Source-Channel-Modell (Brown et al 1990):

  $$\hat{e}_1^I = \underset{e_1^I}{\text{argmax}} \ \{P(e_1^I|f_1^J)\} = \underset{e_1^I}{\text{argmax}} \ \{P(e_1^I) \times P(f_1^J|e_1^I)\}$$

  (mit Lexikon-Sprach-Alignment-Modellen)

- Beispiel-basierte Systeme (EBMT)
  (Sumita et al 1990, Maruyama Watanabe 1992)
  Finde optimale Überdeckung aus Beispielen, verknüpfe Übersetzungen der Beispiele

  Souvent le charpentier travaille le bois - Oft bearbeitet der Zimmermann das Holz

  $\{$*le charpentier travaille*$\}$ – $\{$*der Zimmermann arbeitet*$\}$
  $\{$*travaille le bois*$\}$ – $\{$*bearbeitet das Holz*$\}$
  $\{$*le charpentier*$\}$ – $\{$*der Zimmermann*$\}$

  $\{$Souvent $\}$ $\cup$ $\{$le charpentier $\}$ $\cup$ $\{$travaille le bois$\}$

Aktuell - Hybride Systeme

- Maximum-Entropie-Modell und linguistische Features
  (Och Ney 2002)

  $$\hat{e}_1^I = \underset{e_1^I}{\text{argmax}} \{ \sum_{m=1}^{M} \lambda_m h_m(e_1^I, f_1^J) \}$$

  h_m: POS-Unterschiede, Differenz VP-Knoten, ...

- Linguistische Vor- und Nachbearbeitung:
  (Quirk Menezes Cherry 2005): Dependency treelet translation

  $$\textit{Text} \rightarrow \boxed{\textit{PRE}} \rightarrow \textit{Suche} \ \{\leftarrow \textit{Lexikon-, Alignment-, Sprach-Modell + Features}\} \rightarrow \boxed{\textit{POST}}$$

- Syntaktisch motivierte Beispiele, rekursive Beispielkomposition
  (Chiang 2006): HIERO

  $\langle (1)_{NP1}$'s $(2)_{NP2}$,DET $(2)_{NP2}$ de $(1)_{NP1} \rangle$

- Multisystem: Integriere Konkurrenzergebnisse

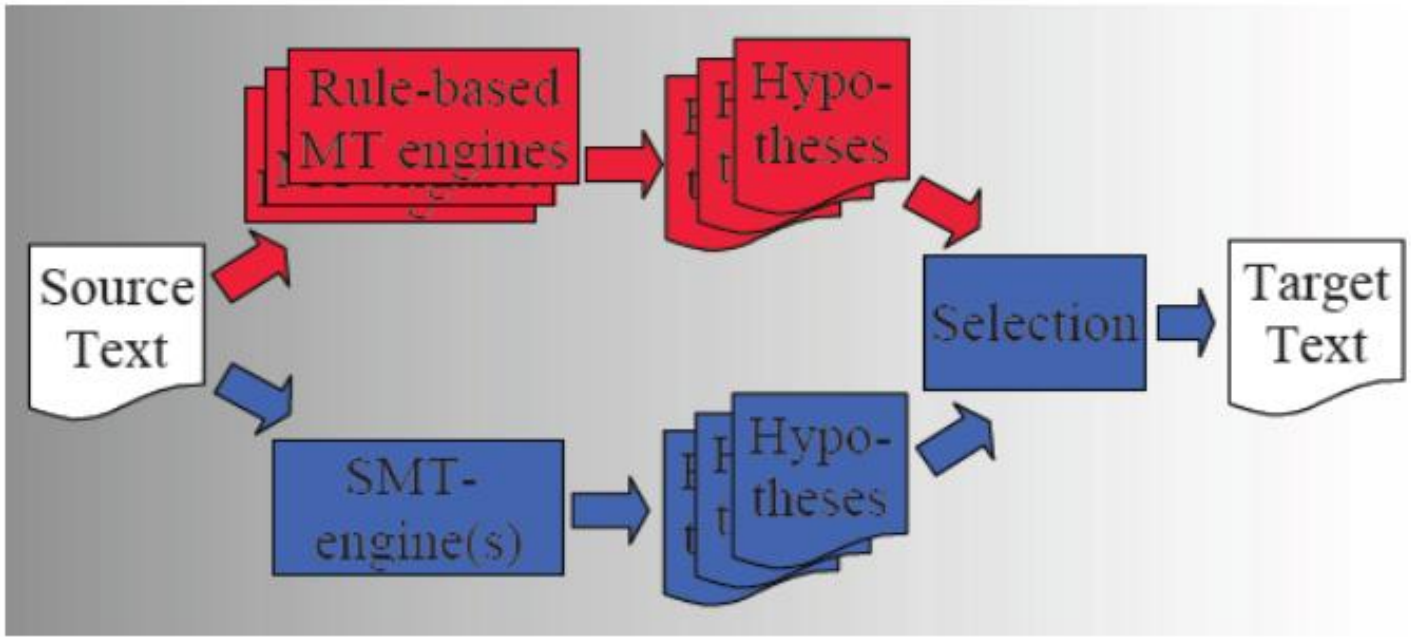
Multisysteme

- Schwache Integration
  Wähle aus Ergebnissen / nutze zum Ressourcenaufbau

- Starke Integration
  'interleaved architecture' (Reaktion auf jeweilige Teilergebnisse)

Multi-Engine MT (from Eisele, 2007)

- Pangloss (1996)
- Moore (2008)

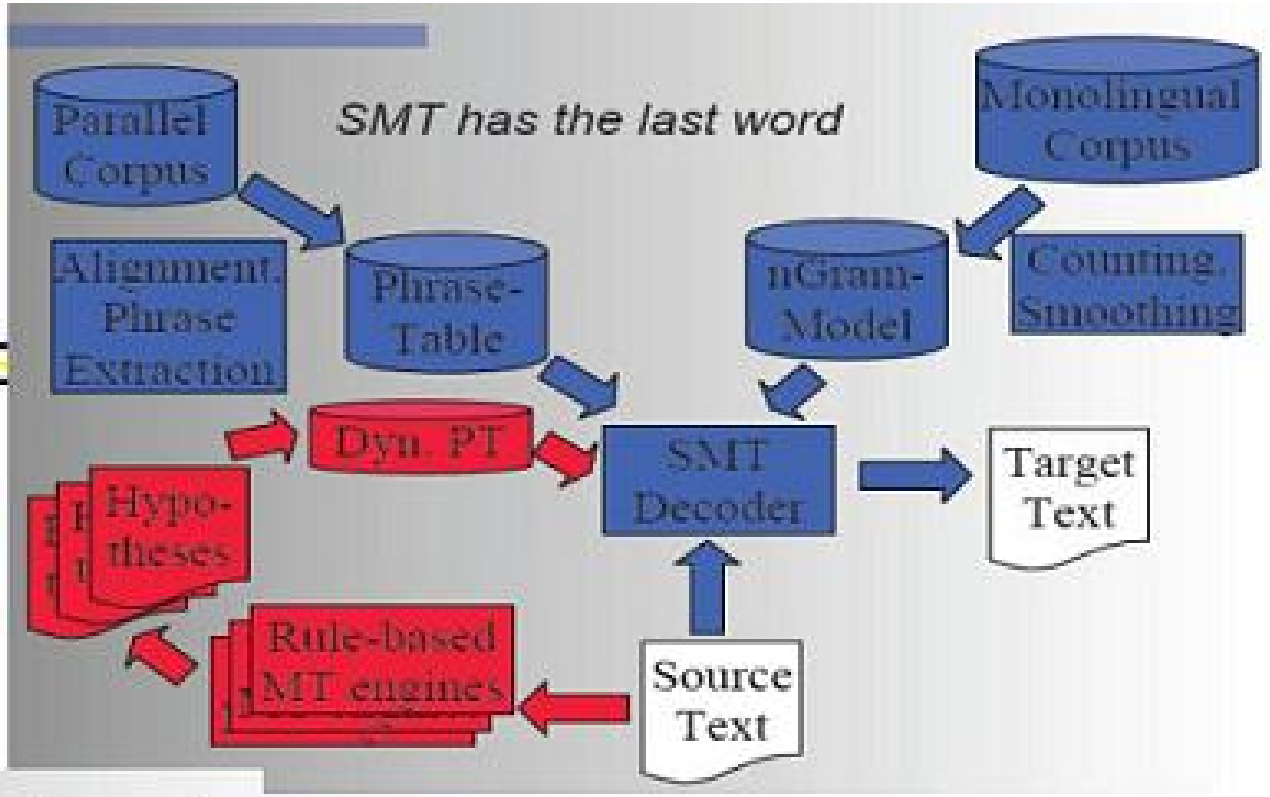


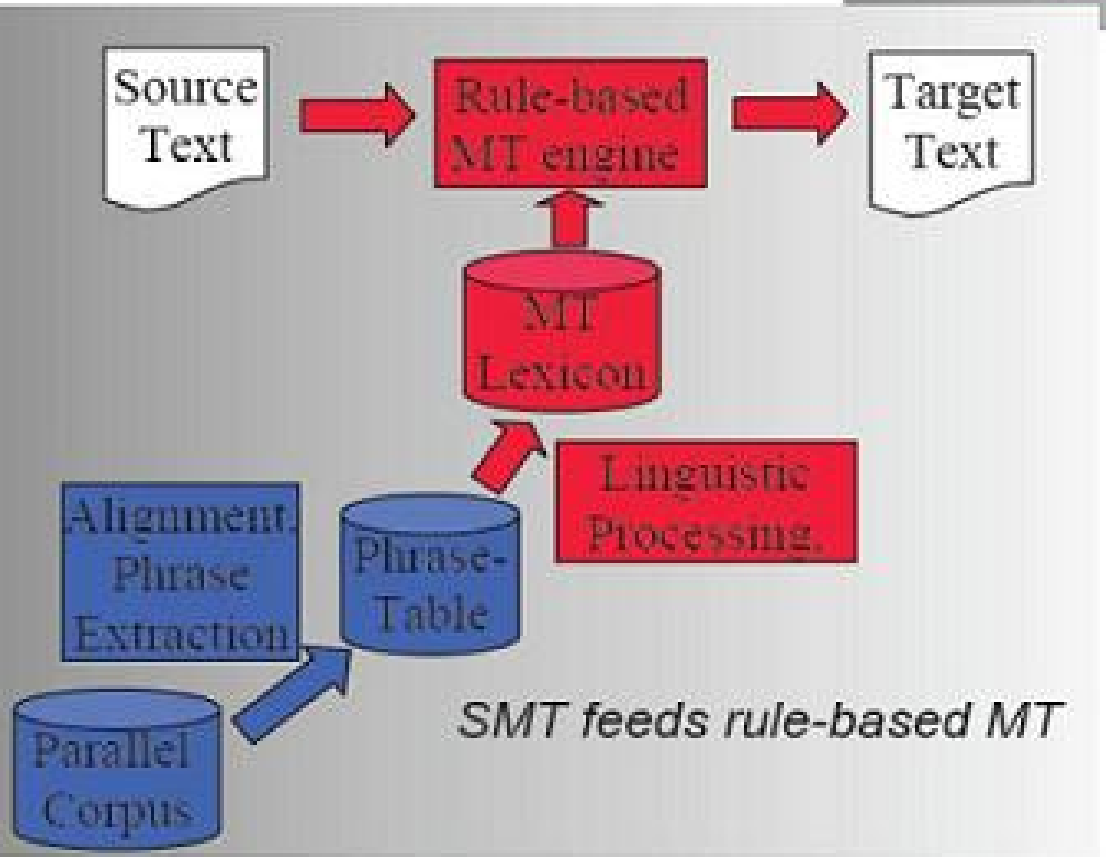
- Crucial: selection/combination
- complex
- monolingual

Architectures for weak

Integration

(Eisele 2007)





translate

Hintergrund

Lingenio

- Wissenschaftliches Zentrum der IBM
- 1999 Ausgründung unter dem Namen *linguatec* Entwicklung & Services
- Zusammenarbeit mit *linguatec* Sprachtechnologien
- seit 2004 unter dem Namen *lingenio*
- Zusammenarbeit mit *digital publishing*

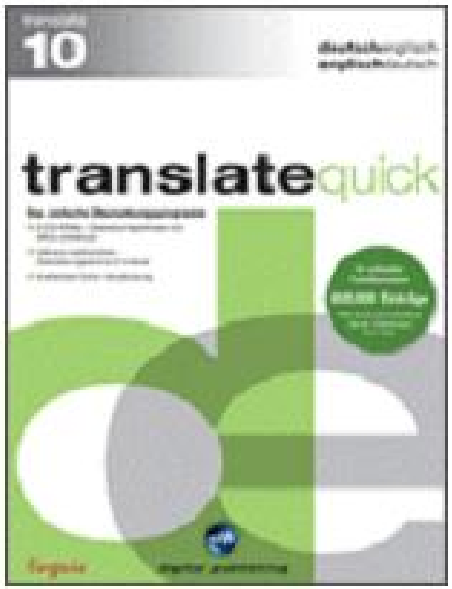
translate

Produktentwicklung

- Kommerzielles Maschinelles Übersetzungssystem

- *Logic based Machine Translation* (LMT)
  (McCord 1989), IBM
- 1996 *Personal Translator* (IBM, linguatec)
- 1999 *Personal Translator Französisch* (linguatec)
- 2004 *translate* (Lingenio)
- 2005 *office* Wörterbcher

Produkte


10
translatequick


10
translatepro


office
wörterbuch


office
wörterbuch

translate pro - Deutsch - Französisch

Datei Bearbeiten Ansicht Format Wörterbuch Übersetzen Satzarchiv Sprachausgabe Hilfe

Deutsch - Französisch

Arial 10 100%

Unbekannte Wörter

Atomdrohungen (Kompositum
h18
Oppostion
Kuhn
Atomdrohung (Kompositum)
Kraftmeierei
Blumentopf (Kompositum)
Topthemen (Kompositum)

Status

Quelltext - (unbenannt)

Topthemen

Merkel soll Pariser Atomdrohungen missbilligen
23/01/2006 08h18

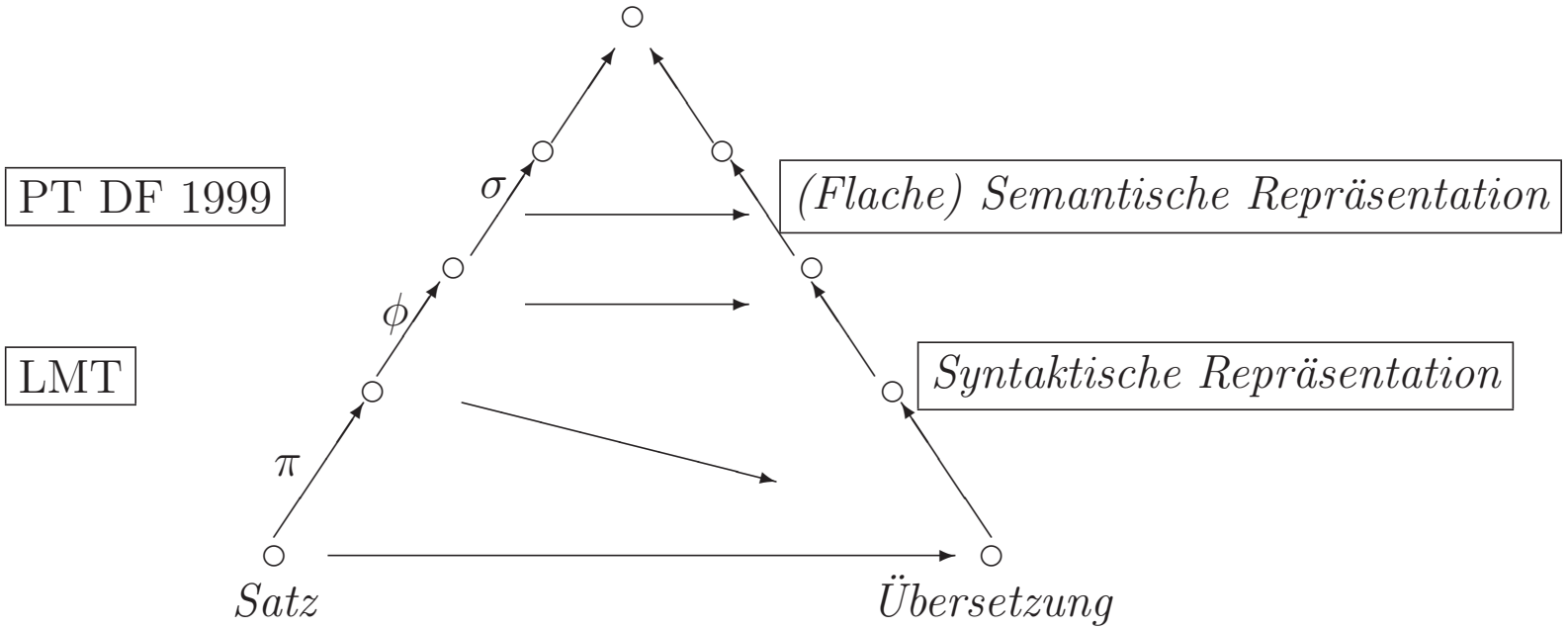
Die Oppostion hat Bundeskanzlerin Angela Merkel aufgefordert, sich bei ihrem he
Frankreichs Präsident Jacques Chirac von den umstrittenen französischen Nuklea

Zieltext - (unbenannt)

Des sujets de haut niveau

Merkel doit désapprouver des menaces d'atome parisiennes
23/01/2006 08 h18

translate-Architekturen



translate - Flache Semantik

Ziel

Flat underspecified discourse representation theory (FUDRT) (Eberle 1997)
(Erweiterung/Modifikation der UDRT (Reyle 1993))

- Lexikon:
  Semantische Repräsentationen sind Funktionen
  (die bei Bedarf schrittweise durch den Kontext ausgewertet werden)

  $\underline{\text{drucker}}(\text{x@}\neg\text{ARTEFACT}) := \text{l}_{x@PROF}$: $\boxed{\begin{array}{l}\text{x}\\ \\ \text{druck\_arbeiter(x)}\end{array}}$

  $\underline{\text{drucker}}(\text{x@}\neg\text{HUMAN}) := \text{l}_{x@EGERAET}$: $\boxed{\begin{array}{l}\text{x}\\ \\ \text{druck\_geraet(x)}\end{array}}$

- Satzrepräsentationen:
  Mengen von DRSen und DRS-Modifikatoren ...
  und Aussagen zu Ordnung und Art der Applikation

translate - Flache Semantik

Satzrepräsentationen

$$\underline{\text{bilder}}(\text{X})\left\{\ \text{ngen: } \underline{\text{kanzlerin}}(\text{y}),\ \text{xprep}(\text{hinter}): \underline{\text{tresen}}(\text{z})\ \right\}$$

translate - Flache Semantik

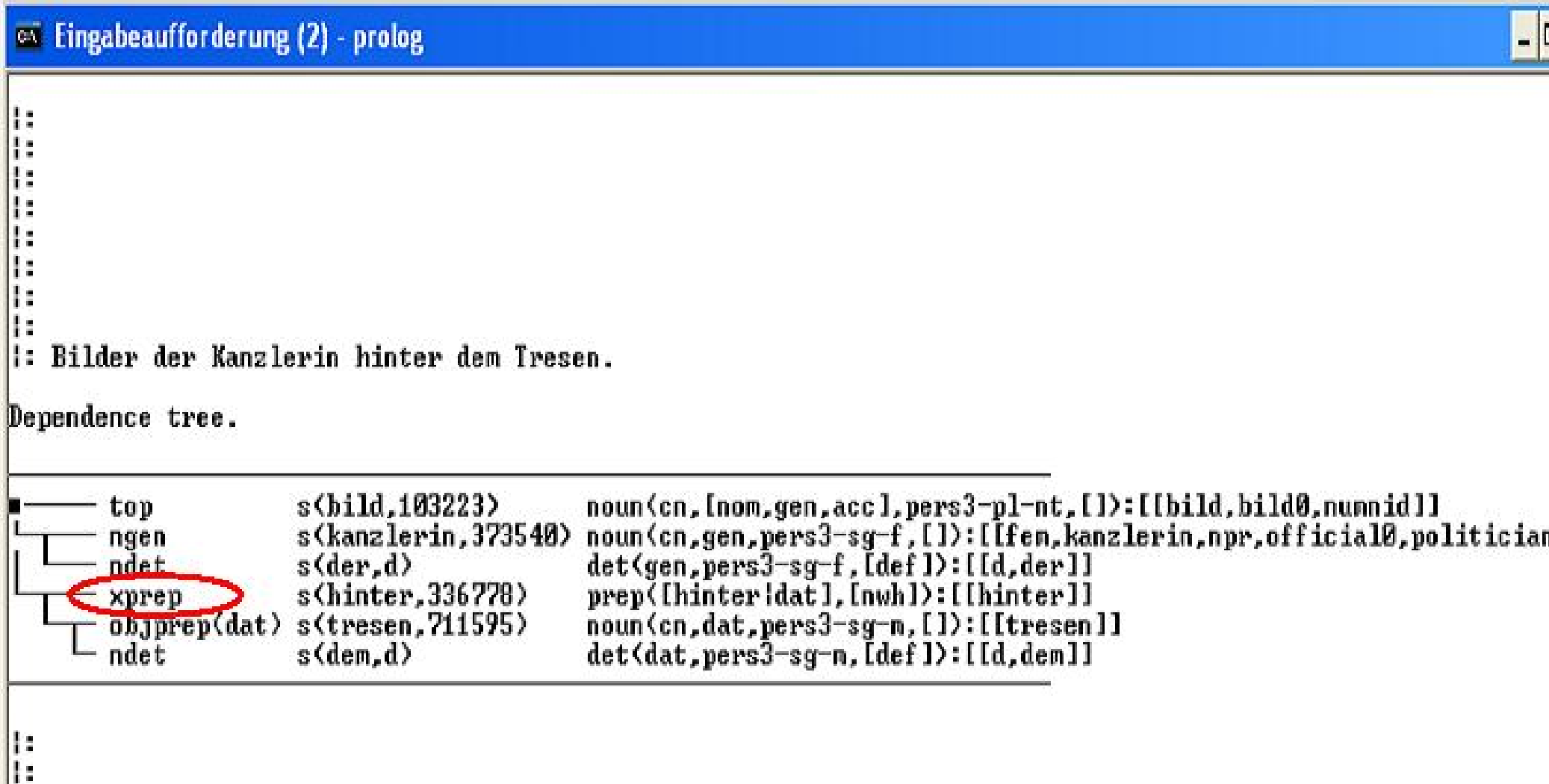
implementiert

- Dependenz-Strukturen (Prädikat-(gram.) Argument-Strukturen)
- Semantische Abstraktionen bei Koordinationen und bestimmten Angaben
- Informationsstruktur (Fokus-Hintergrundbestimmung bei Fokus-Adverbien)
- Skopusauflösung bei Bedarf
- Akzessibilitätsinformation für die Pronomenauflösung
- Transfer auf FUDRSen
- (variable Analysetiefe)

Implementierung im Test

Funktionale Ambiguität, Attachment-Ambiguität

Zusammenarbeit mit IMS Stuttgart, SFB 732 Projekt B3

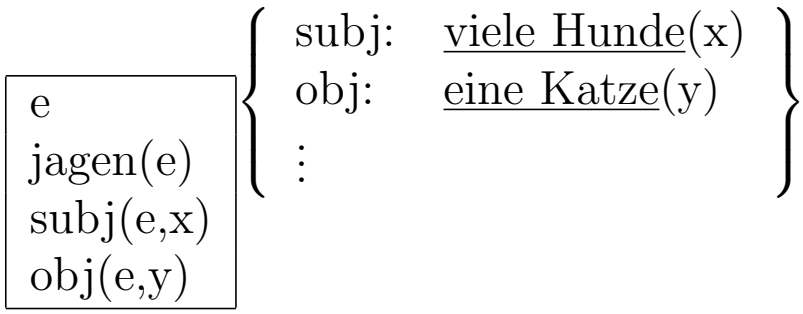


Ambiguitätserhaltende Übersetzung

Skopus-Ambiguität

Viele Hunde jagen eine Katze.

Repräsentation:



Übersetzung ohne Skopus-Disambiguierung:

Many dogs chase a cat.

Default-Transferalgorithmus

$$\tau(\underline{\text{jagen}}\left\{\begin{array}{ll}\text{subj:} & \underline{\text{viele Hunde}}(\text{x})\\ \text{obj:} & \underline{\text{eine Katze}}(\text{y})\\ \vdots & \end{array}\right\}) := \tau_n(\underline{\text{jagen}})\left\{\begin{array}{ll}\tau_s(\text{subj}): & \tau(\underline{\text{viele Hunde}})(\text{x})\\ \tau_s(\text{obj}): & \tau(\underline{\text{eine Katze}})(\text{y})\\ \vdots & \end{array}\right\}$$

$$\boxed{\begin{array}{l}\text{e}\\ \text{chase(e)}\\ \text{subj(e,x)}\\ \text{obj(e,y)}\end{array}}\left\{\begin{array}{ll}\text{subj:} & \underline{\text{many dogs}}(\text{x})\\ \text{obj:} & \underline{\text{a cat}}(\text{y})\\ \vdots & \end{array}\right\}$$

Partielle Disambiguierung

Funktionale Ambiguität

Lucie zeigte den Film ihrer Familie.

a) V dat Obj
Lucie présenta le film à sa famille.

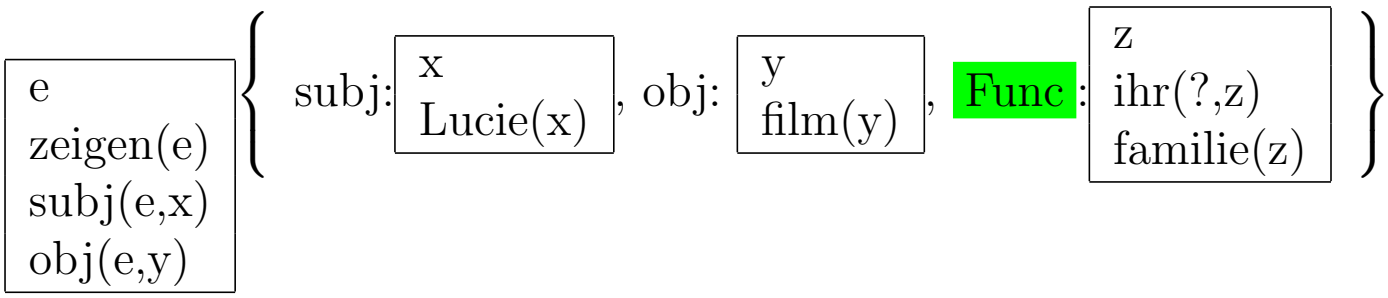
b) N gen Obj
Lucie passa le film de sa famille.

Partielle Disambiguierung

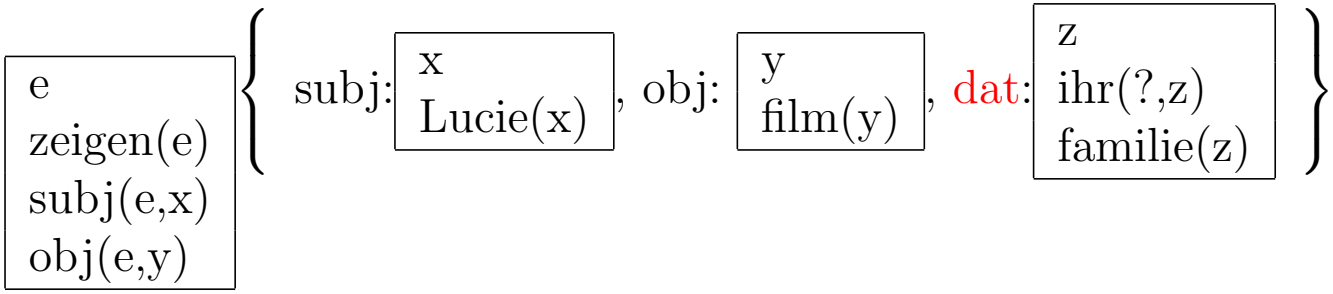
Funktionale Ambiguität

Lucie zeigte den Film ihrer Familie.

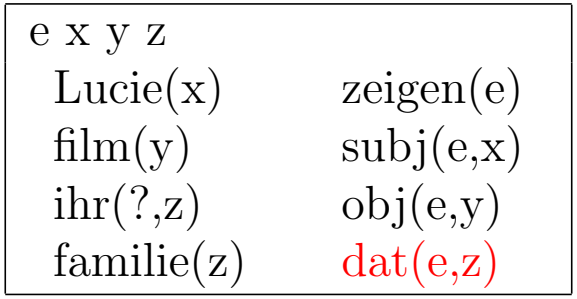
Ausgangssituation



a) Interpretiere Func als Dativ-Rolle der Verbrepräsentation

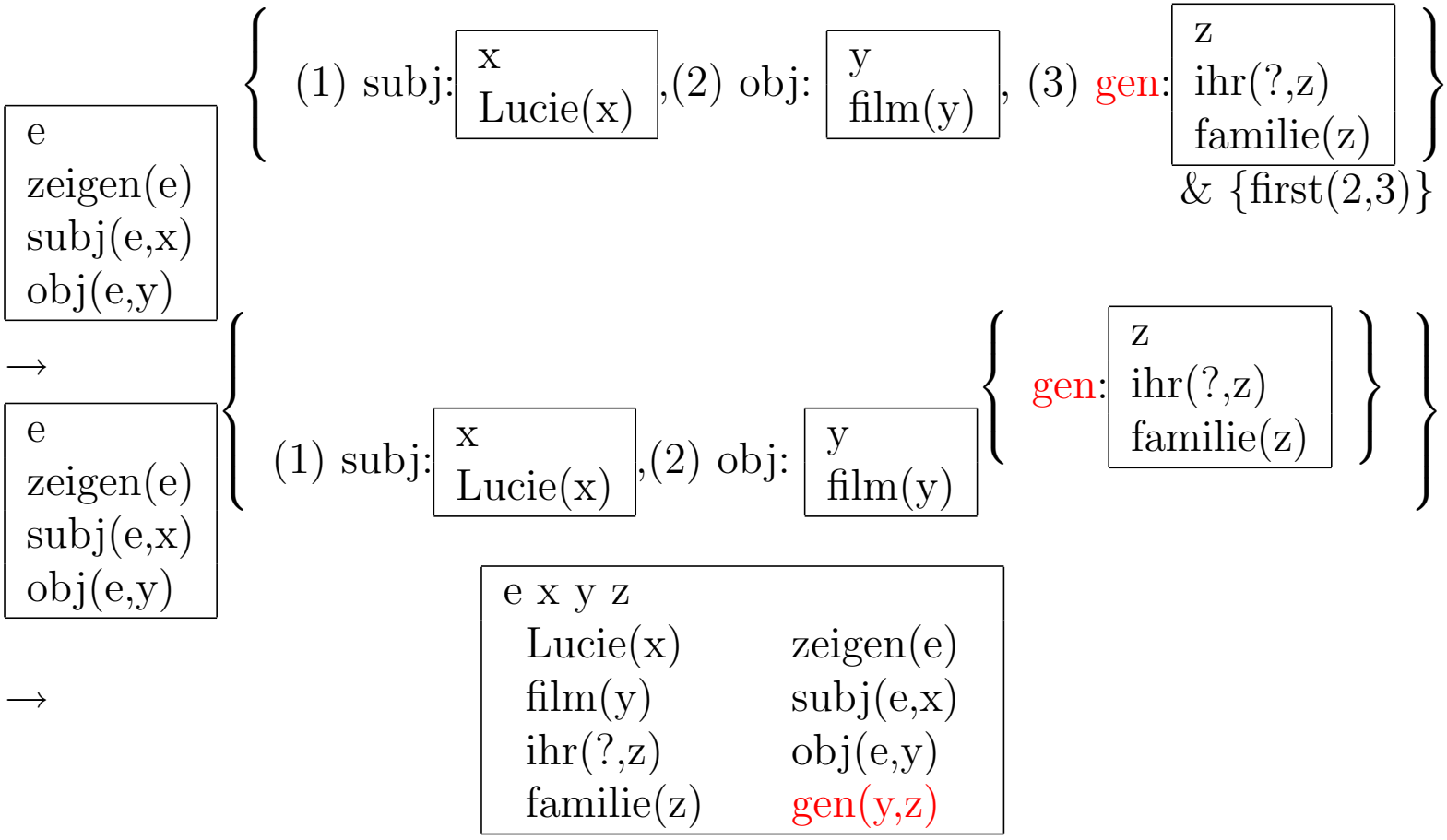


→



Lucie présente le film à sa famille.

b) Interpretiere Func als Genitiv-Rolle der Objektrepräsentation



Lucie passe le film de sa famille.

Partielle Disambiguierung

Wodurch werden tiefere Auswertungen ausgelöst?

- Durch den Transfer
- nach Maßgabe im entsprechenden Eintrag im Lexikon

zeigen(e):	Cond:	$C \vdash_D$ filled(obj,FILM), empty(dat)	τ: passer
zeigen(e):	Cond:	$C \vdash_D$ filled(obj,FILM), filled(dat)	τ: présenter

→Funktionale Ambiguität wird reduziert !

Partielle Disambiguierung

Wodurch werden tiefere Auswertungen ausgelöst?

- Durch Propagieren von Information im Kontext !

Beispiel: Pronomenresolution

Kürzlich erst hatte sie den Drucker eingestellt.
a) Jetzt kündigte er schon wieder.
b) Jetzt war er schon wieder defekt.
It was only recently that she had hired/adjusted the printer.
a) Now he already dismissed.
b) Now it already was defective again.

Pronomenauflösung und Antezedent-Übersetzung interagieren!

. . . den Drucker$_u$. . . . Er$_x$ war defekt.

s s: defekt(x)	$\vdash_D$	MASCHINE(x)

u drucker_HUMAN(u)	+ x=u :	Widerspruch!

versus

u drucker_MASCHINE(u)	+ x=u :	kein Widerspruch!

Verb- und Nomeninterpretation interagieren qua semantischer Selektionsbeschränkung!

einstellen(e,x,y@HUMAN) := $l_{e@SOC\_ev}$:

e
einstellen_hire(e) agens(e) = x thema(e) = y

einstellen(e,x,y@MASCHINE) := $l_{e@PHYS\_ev}$:

e
einstellen_adjust(e) agens(e) = x thema(e) = y

...

- *einstellen* [subj(n),obj(n)]

C: $\vdash_D$ *SOC_ev*
τ: *hire*

C: $\vdash_D$ *PHYS_ev*
τ: *adjust*

...

Partielle Disambiguierung

- Propagieren von Information
  - Selektionsbeschränkungen
  - Pronomen und Kennzeichnungen
  - ...
- Transfer-Notwendigkeit
  →Legitimation??

Integration regelbasierter und statistischer Methoden

- regelbasierter Kern
  - ökonomische Repräsentation (flach)
    vermeide Entscheidungen die nicht durch ('harte') Regeln begründet sind
  - wenig Weltwissen (Typhierarchie . . .)
  - variable Analysetiefe
    Auswertung nur per *constraint propagation* und bei *Transfer-Bedarf*
  - mächtiger Lexikonformalismus
  - keine Transfergrammatik
  - modulares Design
    unterscheide *'harte'* Regeln von *'flexibler'* Bewertung
- statistikbasierte Peripherie
  - Entscheidungshilfe

'Hybridisierung' von *translate*

Statistische Korpus-Information für

- Analyse
  - Lexem-Disambiguierung
  - strukturelle Disambiguierung
- Transfer
  - Lernen von Übersetzungsbeziehungen
  - Bewertung von Transfer-Äquivalenten
- Generierung
  - Wortstellung

Statistische Information für die Analyse

Lexem-Disambiguierung

> *Little John was looking for his toy box. Finally, he found it.*
> *The box was in the pen. John was very happy.*
> (Bar-Hillel 1959)

Weltwissen	$\rightarrow$	(pen = playpen/Laufstall)
oder	$\rightarrow$	*Word sense disambiguation* (WSD)

pen $\rightarrow$pen(x)$_{\mid C}$	C $\vdash_D$ sit_x $\in$BABY_FRAME	$\rightarrow$Laufstall(x)
	C $\vdash_D$ sit_x $\in$WRITING_FRAME	$\rightarrow$Stift(x)

- verfeinere *automatische Sachgebietserkennung*
- lerne signifikante Kontexte

Statistische Information für die Analyse

Funktional-/Attachment-Ambiguität

Er zeigt den Film der Familie (im Urlaub, . . .)

Bootstrapping (verwende/verbessere Analyse)

- Flache Analyse hat folgende Charakteristik:

$$\underline{\text{V}}(\text{e @ e\_TYP}) \left\{ \begin{array}{l} \text{subj:}\underline{\text{NP}}, \text{ obj:}\underline{\text{NP}}(\text{y @ obj\_TYP}), \\ \text{Func:}\underline{\text{NP}}(\text{z @ func\_TYP}), \text{ xprep(in):}\underline{\text{NP}}(\text{p @ prep\_TYP}) \end{array} \right\}$$

- Korpus-Suche mit entsprechender Query
  (Zusammenarbeit SFB 732, Projekt B3, Prof. Heid)

- Evaluation
  signifikante Muster für verschiedene Lesarten

- (Hindle/Rooth): V NP P NP - Strukturen, . . .

Statistische Information für den Transfer

Bewertung von Transfer-Äquivalenten

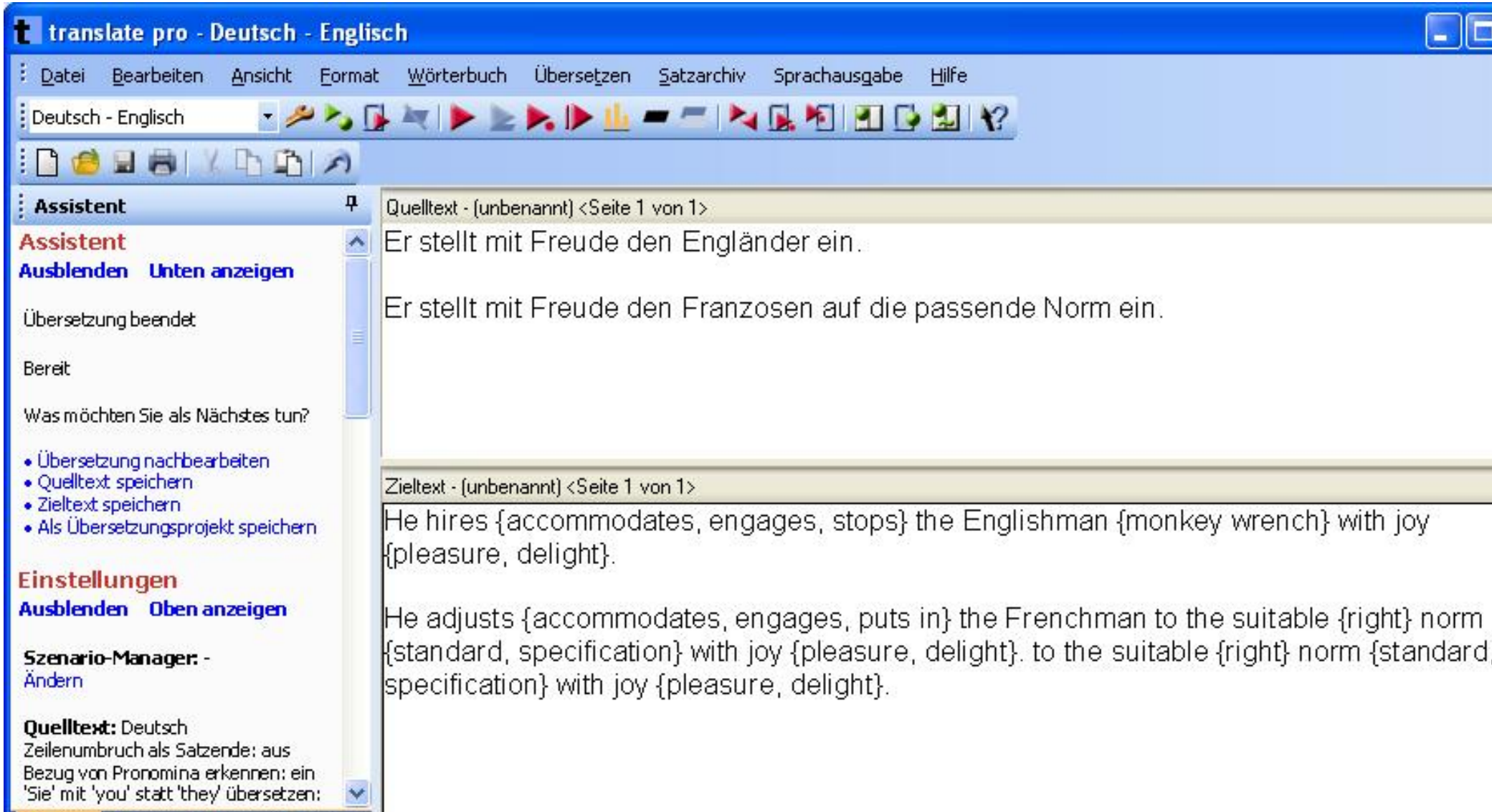
- *translate 11*:
  Bootstrapping (Analyse und bilinguales Lexikonwissen)
  - Transfer-Vorschläge des Lexikons
  - Korpus-Test (Europarl)
  - signifikante Verteilungen

  → Kontextbewertungen für Übersetzungsalternativen:

  Er stellt das ein - He adjusts {*adjust 0.35, stop 0.3, suspend 0.15*} *it.*

Statistische Information für den Transfer

Bewertung von Transfer-Äquivalenten



Statistische/EBMT- Information für den Transfer

Lernen von (analytischen) Übersetzungsbeziehungen

> *Aus bestimmten Gründen stellten die beiden Fraktionen ihre Feindseligkeiten vorübergehend durch einen Waffenstillstand ein und vereinbarten ...*
> *For some reason, a temporary cease-fire in the hostilities between the two factions was established and ....* (Datei ep-96-09-18.al, Zeile 1318)

Bootstrapping (Lexikonwissen, Source-Targetanalysen)

- Source-Analyse
- Target-Analyse
- Zuordnung von Teilen mittels bilingualen *Cognates* aus dem Lexikon
- Generalisierung von Bedingungen
- Signifikanz-Test

(Konvens 2008)

Aus bestimmten Gründen stellten die beiden Fraktionen ihre Feindseligkeiten vorübergehend durch einen Waffenstillstand ein und vereinbarten . . .
For some reason, a temporary cease-fire in the hostilities between the two factions was established and . . . . (Datei ep-96-09-18.al, Zeile 1318)

- l_0:einstellen [subj(n),obj(n)]

C: d(vadv):l_1:vorübergehend & d(subj):l_2:fraktion
& d(obj):l_3:feindseligkeit & d(prep(durch)):l_4:waffenstillstand

τ: establish [$\emptyset$,obj(n):$\tau(l_4)$]
& τ(d-l_1)=τ(l_0)-d(obj)-d(nadj)
& τ(d-l_3)=τ(l_0)-d(obj)-d(prep(in))
& τ(d-l_2)=τ(l_0)-d(obj)-d(prep(in))-d(prep(between))

Verallgemeinerungen . . .

jmd./ stellt STATE/ durch EVENT/ ein - s.o./ establishes EVENT/ in STATE

Statistische Information für die Generierung

Lernen von Wortstellungsregeln

- Wortstellung

  Poirot remet la lettre à la femme →*den Brief der Frau*/ *der Frau den Brief*

  →Frequenzanalysen

    Flache Analysen zu Korpussätzen mit Typinformationen für NPs
    (Zusammenarbeit SFB 732, Projekt D2, Prof. Rohrer)

Zusammenfassung

- Aktuell: 'Hybridisierung'
- Gängig: Integration linguistischen Wissens in SMT
- *translate*: Integration von SMT- und EBMT-Methoden in RBMT
  - Flache Repräsentationen
  - variable Analysetiefe, Auswertungstrigger
  - Modularisierung:
    Harte Regeln (deklarative Grammatiken) - *Weiche* Entscheidungskriterien (Disambiguierung)
  - bei Analyse, Transfer, Generierung
- Signifikante Qualitätsverbesserung
- ökonomisch, effizient
- autonome, Human-lesbare Wörterbücher