Aktuelle Architekturfragen in der Maschinellen Übersetzung
semantischer Transfer und Integration statistischer Information in *translate*

Kurt Eberle
Lingenio GmbH

26. Juni 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
  - Kategoriale Ambiguität von Lexemen und Grammemen
    - *Time*$_{N/V}$ *flies*$_{N/V}$ *like*$_{V/P}$ *an arrow - Zeitfliegen lieben einen Pfeil*
  - Semantische Ambiguität kategorial disambiguierter Lexeme und Grammeme
    - *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 showing the chancellor behind the bar/ pictures behind the bar showing the chancellor
  - Skopus-Ambiguität
    Few women like many men - Wenige Frauen mögen viele Männer
- 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
    **She hires the print device*
- Nicht alle Mehrdeutigkeiten müssen aufgelöst werden!
  - *printer* - *Drucker*
  - *few(x,woman,many(y,man,like(x,y))) / many(y,man,few(x,woman,like(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 swim*

  $$\boxed{\begin{array}{l} \text{s} \\ \text{s} : \text{ATT}(\text{peter}, \left\{ \ <\text{POS\_DISP}, \lambda x.rauchen(x) > \ \right\}) \end{array}}$$

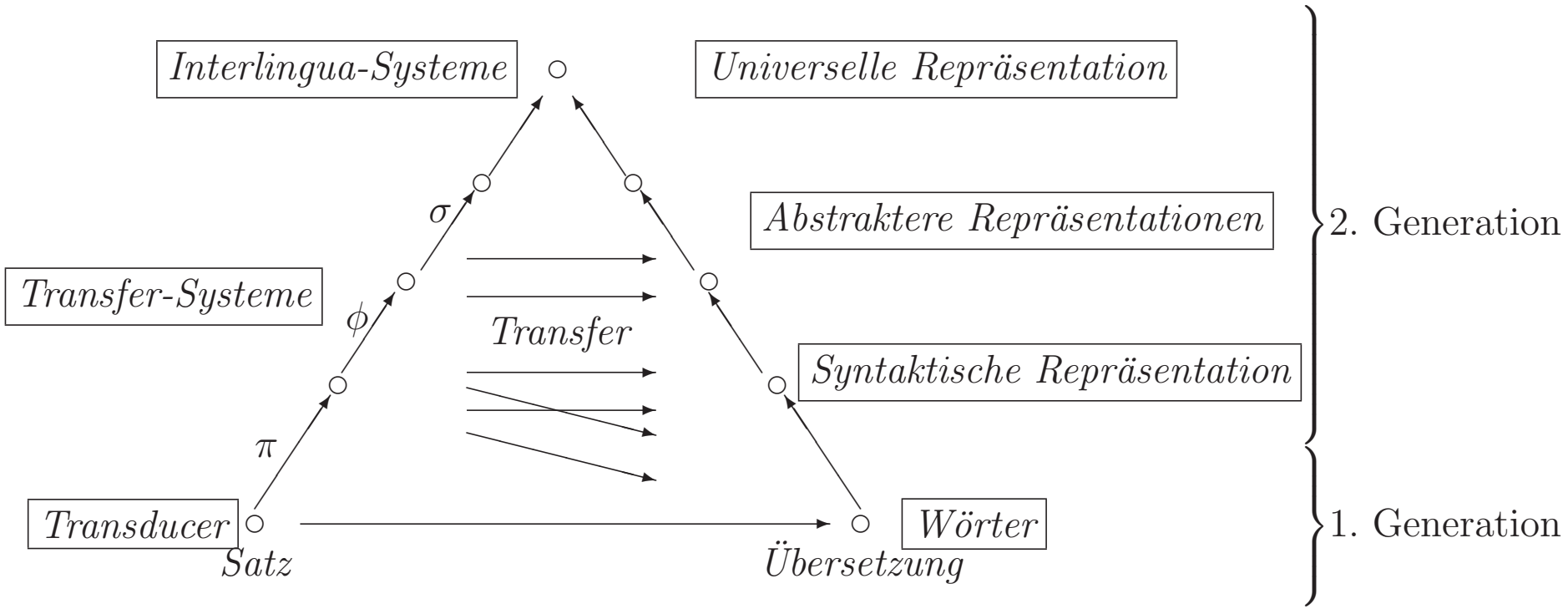
Mehrdeutigkeit, Mismatches und Analysen

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

$\rightarrow$ Optimierungsproblem

Lösungsversuche

(Vauquois 19975)



Regel-basierte Architekturen

Lösungsversuche - 3. Generation

- Statistik-basierte Systeme (SMT)
- Beispiel-basierte Systeme (EBMT)

Statistik-basierte Systeme (SMT)

(Brown et al 1990), CANDIDE

$$\hat{\mathrm{e}}_1^I = \underset{e_1^I}{\operatorname{argmax}} \{\mathrm{P}(\mathrm{e}_1^I) \times \mathrm{P}(\mathrm{f}_1^J | \mathrm{e}_1^I)\}$$

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

Lexikonmodell	*travaille* $\rightarrow$ {*bearbeitet* ($p_{l,1}$), *arbeitet* ($p_{l,2}$), ...} , *le* $\rightarrow$ {*der, die, das, ...*}	$\mathrm{P}(\mathrm{f}_1^J \mid \mathrm{e}_1^I)$
Sprachmodell	*das Holz* ($p_{s,1}$), *der Holz* ($p_{s,2}$)	$\mathrm{P}(\mathrm{e}_1^I)$
Alignment-Modell	Pos. 4 (travaille) $\rightarrow$Pos. 2 ($\mathrm{p}_{a,1}$), ...	

Beispiel-basierte Systeme (EBMT)

(Sumita et al 1990, Maruyama Watanabe 1992)

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*}

Finde (berechne) optimale Überdeckung:

{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

  Text $\rightarrow$ *VB* $\rightarrow$ *Suche* $\{\leftarrow$ *Lexikon-, Alignment-, Sprach-Modell + Features*$\} \rightarrow$ *NB*

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

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

- Multisystem: Integriere Konkurrenzergebnisse

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*

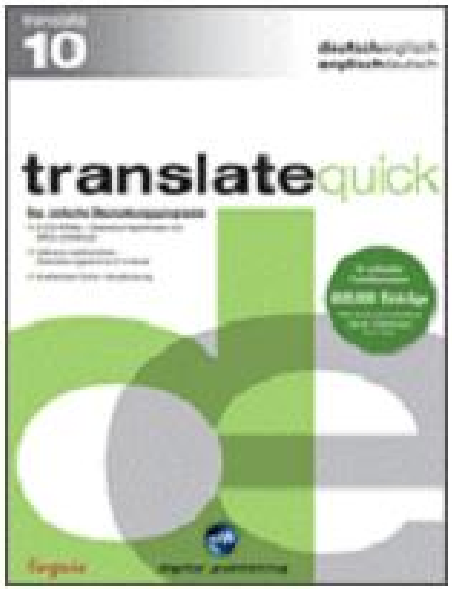
translate

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

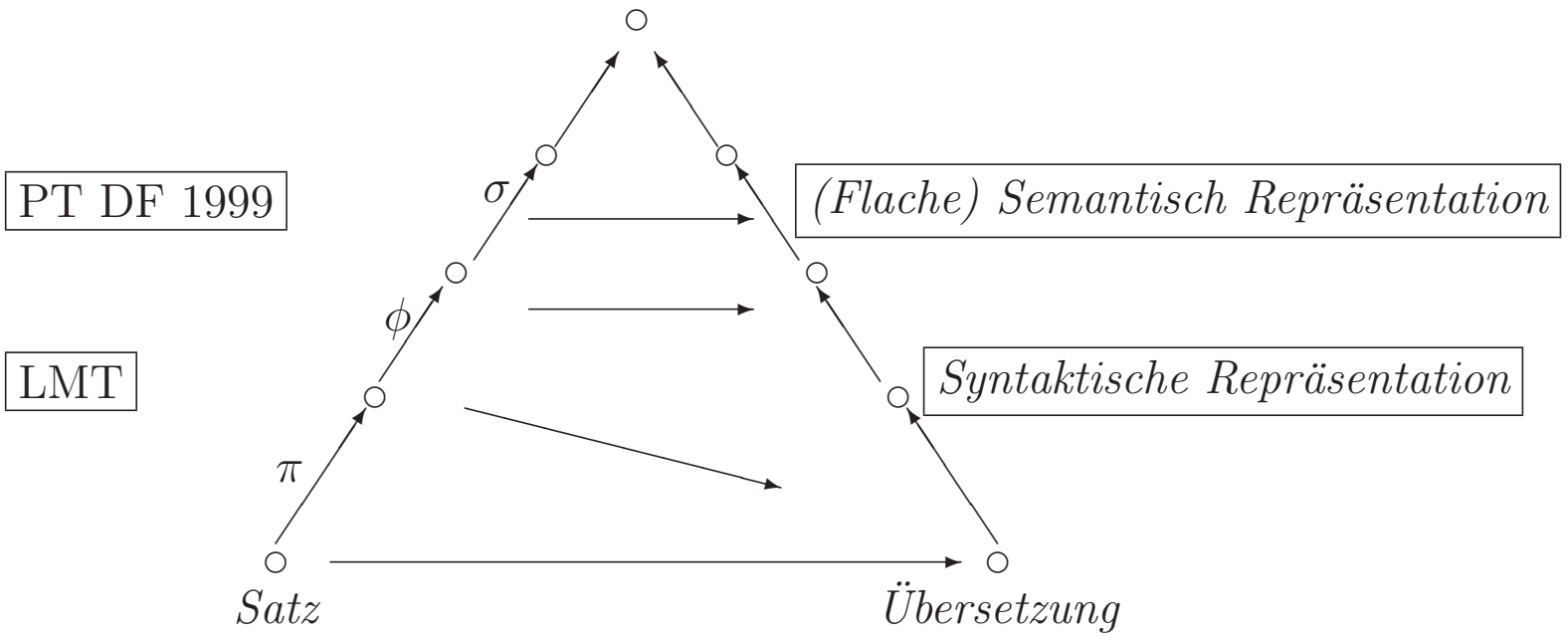
Die Oppostion hat Bundeskanzlerin Angela Merkel aufgefordert, sich bei ihrem he
Frankreichs Präsident Jacques Chirac von den umstrittenen französischen Nuklea

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**

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)

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

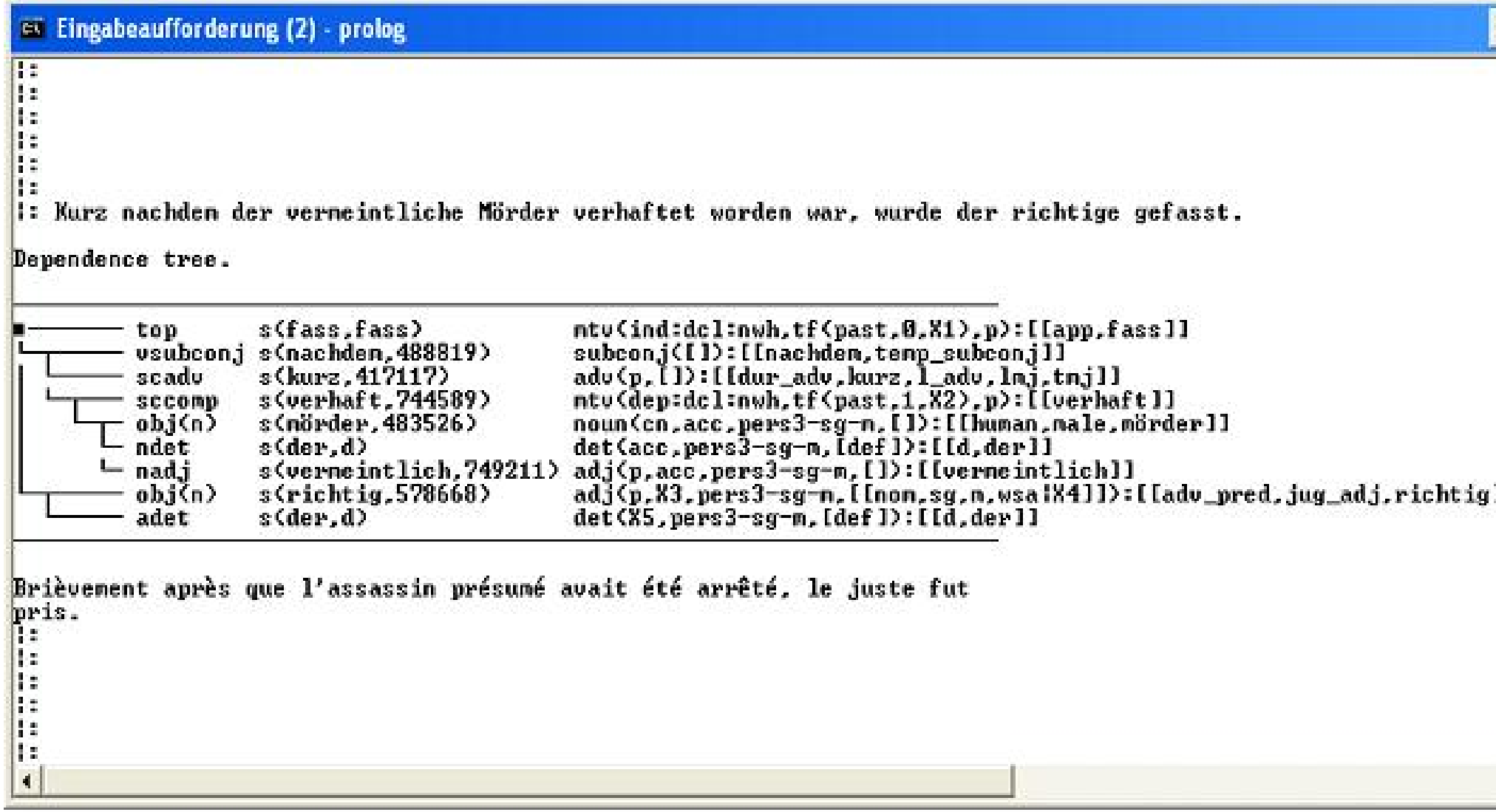
Disambiguierung nach Bedarf!

translate - Flache Semantik

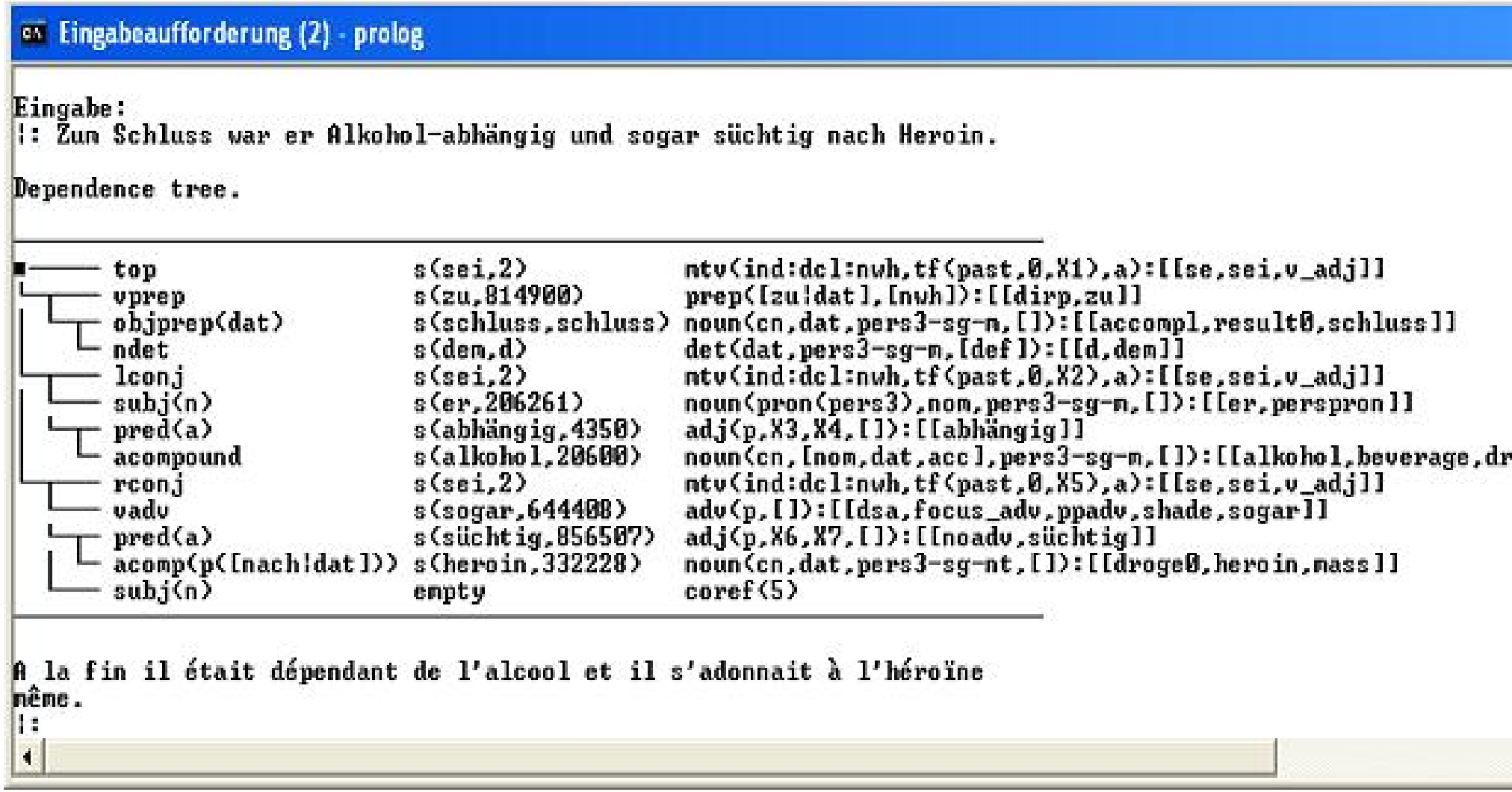
implementiert

- 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)

Dependenz-Strukturen



”Auffaltung” koordinierter Strukturen (bei Bedarf)



Implementierung im Test

Funktionale Ambiguität, Attachment-Ambiguität

Zusammenarbeit mit IMS Stuttgart, SFB 732 Projekt B3

Beispiele für die Verwendung von flachen Repräsentationen

Skopus-Ambiguität - Darstellung

In wenigstens fünf europäische Urlaubsgebiete flogen wenigstens drei Reiseveranstalter ihre Kunden in den letzten Jahren mehr als 20 Mal.

To at least five European tourist areas, at least three operators flew their clients more than 20 times within the last years.

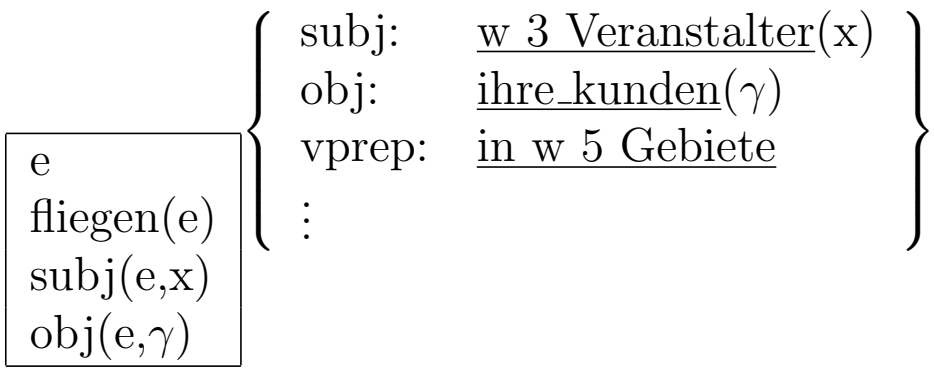
Beispiele für die Verwendung von flachen Repräsentationen

Skopus-Ambiguität - Darstellung

In wenigstens fünf europäische Urlaubsgebiete flogen wenigstens drei Reiseveranstalter ihre Kunden in den letzten Jahren mehr als 20 Mal.

To at least five European tourist areas, at least three operators flew their clients more than 20 times within the last years.

Repräsentation:



Default-Transferalgorithmus

$$\tau(\underline{\text{fliegen}}\left\{\begin{array}{ll}\text{subj:} & \underline{\text{w 3 Veranstalter}}(\text{x})\\ \text{obj:} & \underline{\text{ihre\_kunden}}(\gamma)\\ \text{vprep:} & \underline{\text{in w 5 Gebiete}}\\ \vdots & \end{array}\right\}) := \tau_n(\underline{\text{fliegen}})\left\{\begin{array}{ll}\tau_s(\text{subj}): & \tau(\underline{\text{w 3 Veranstalter}})(\text{x})\\ \tau_s(\text{obj}): & \tau(\underline{\text{ihre\_kunden}})(\gamma)\\ \tau_s(\text{vprep}): & \tau(\underline{\text{in w 5 Gebiete}})\\ \vdots & \end{array}\right\}$$

$$\boxed{\begin{array}{l}\text{e}\\ \text{fly(e)}\\ \text{subj(e,x)}\\ \text{obj(e,}\gamma\text{)}\end{array}}\left\{\begin{array}{ll}\text{subj:} & \underline{\text{al 3 operators}}(\text{x})\\ \text{obj:} & \underline{\text{their clients}}(\gamma)\\ \text{vprep:} & \underline{\text{to al 5 areas}}\\ \vdots & \end{array}\right\}$$

Partielle Disambiguierung

Funktionale Ambiguität

Lucie zeigte den Film ihrer Familie.

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

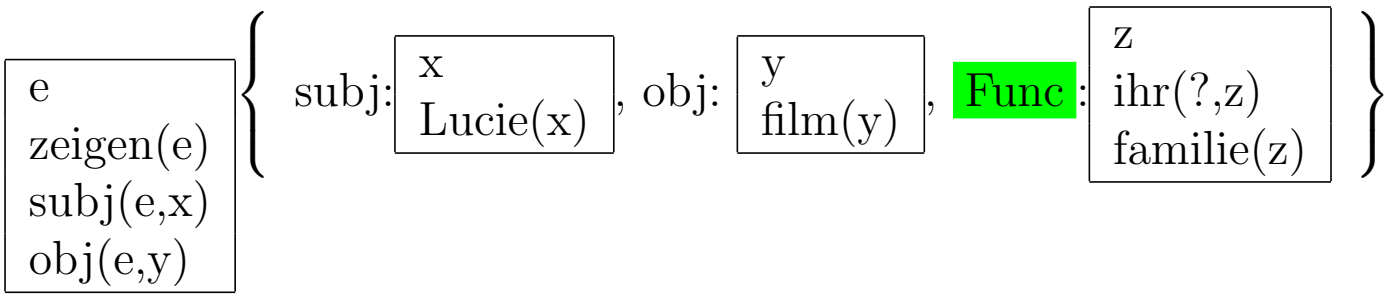
b) N gen Obj
Lucie passa le film de sa famille.

Partielle Disambiguierung

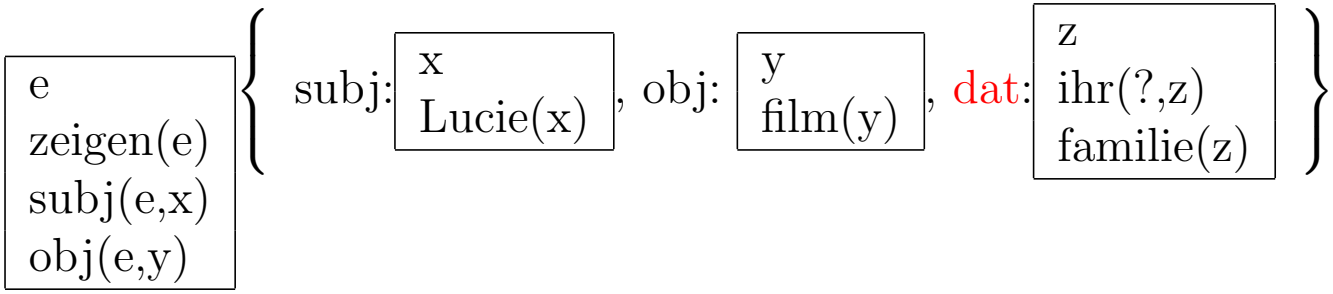
Funktionale Ambiguität

Lucie zeigte den Film ihrer Familie.

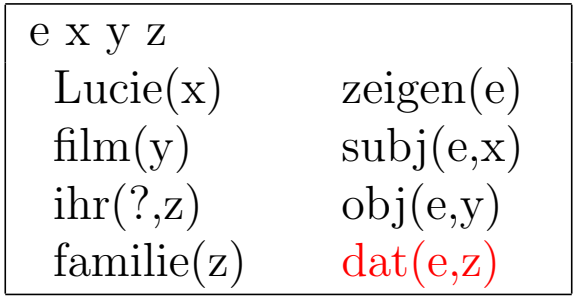
Ausgangssituation



a) Interpretiere Func als Dativ-Rolle der Verbrepräsentation

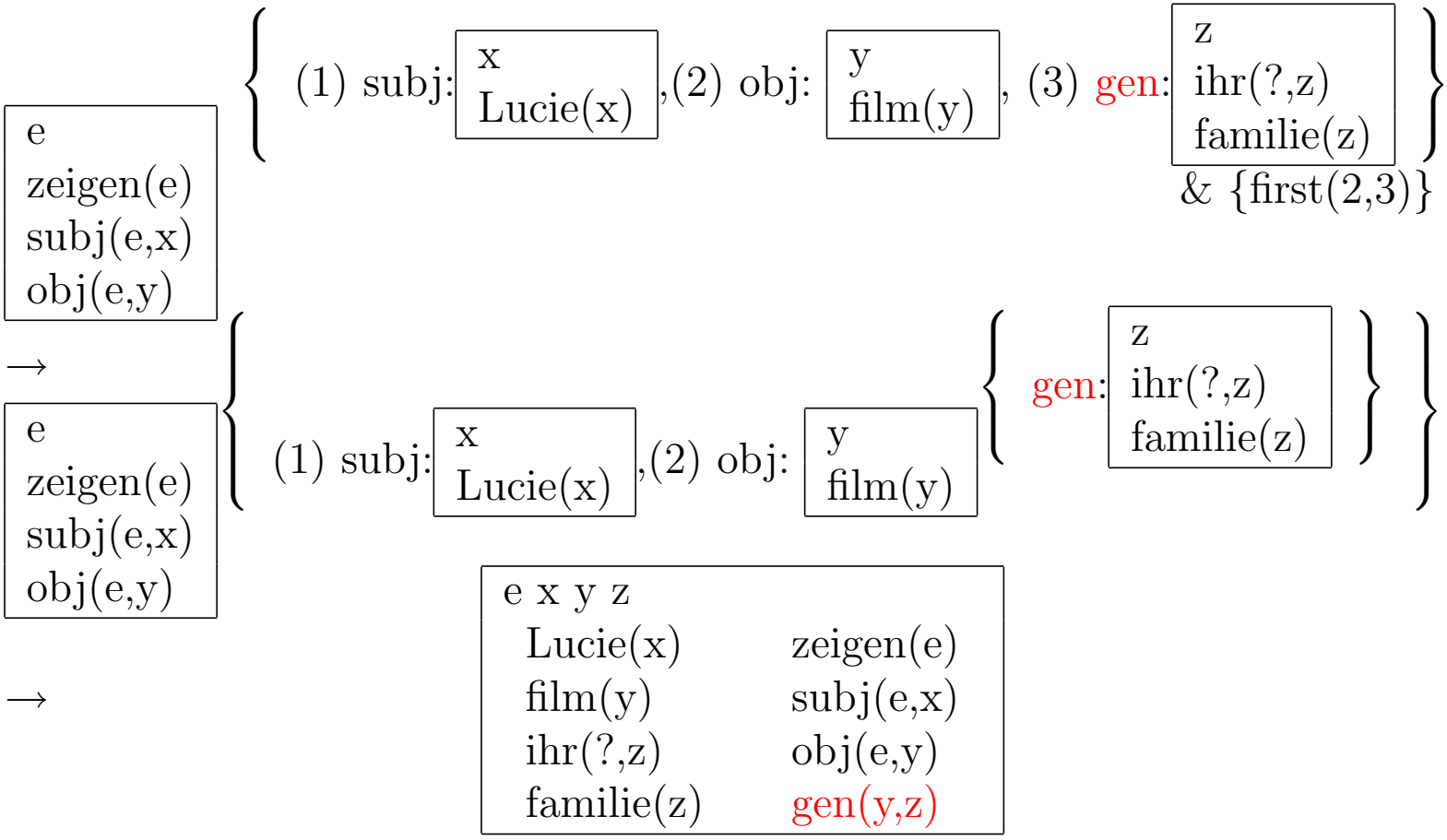


→



Lucie présente le film à sa famille.

b) Interpretiere Func als Genitiv-Rolle der Objektrepräsentation



Lucie passe le film de sa famille.

Partielle Disambiguierung

Funktionale Ambiguität

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

Interaktionen

Pronomen - Nomenambiguität

Pronomen haben nicht immer dieselbe Übersetzung:

elle	→	sie:	elle est intélligente	←	la femme
	→	er:	elle est visible	←	la lune
	→	es:	elle est bien élevée	←	la jeune fille

Der Bezug entscheidet!

Koreferenz-Entscheidungen haben komplexe Konsequenzen:

Enfin il trouva la balle. Elle était dégonflée.

elle → la balle → die Kugel → sie
der Ball → er

Pronomenauflösung und Antezedent-Übersetzung interagieren!

. . . la balle$_u$. . . . Elle$_x$ était dégonflée.

s s: dégonflé(x)	$\vdash_D$	hohlkörper(x) norm_mit_gas_gefüllt(x)

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

versus

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

Schlussfolgerungen

Enfin il trouva la balle. Elle était dégonflée.

a) Der Bezug *balle* - *elle* ist möglich.

b) Wenn der Bezug gewählt wird, muss *balle* im Sinne von *Ball* verstanden werden!

c) Dann ist *elle* mit *er* zu übersetzen.

und ...
Die Koreferenz-Information kann weitere Effekte in immer größeren Kontexten zeitigen!

Il tire la balle. Puis, tout d'un coup, elle se dégonfle.

- Achte auf sinnvolle Abfolge von Übersetzungs- und Auswertungsschritten

Interaktionen

Pronomen - Strukturambiguität

Pronomenbezug und Teilrepräsentationen interagieren!

Der Polizist zeigte jedes der sichergestellten Videos einer Verdächtigen.
Er versuchte, sie dadurch in Widersprüche zu verwickeln.

- Die Videos zeigen alle dieselbe Verdächtige (N gen)

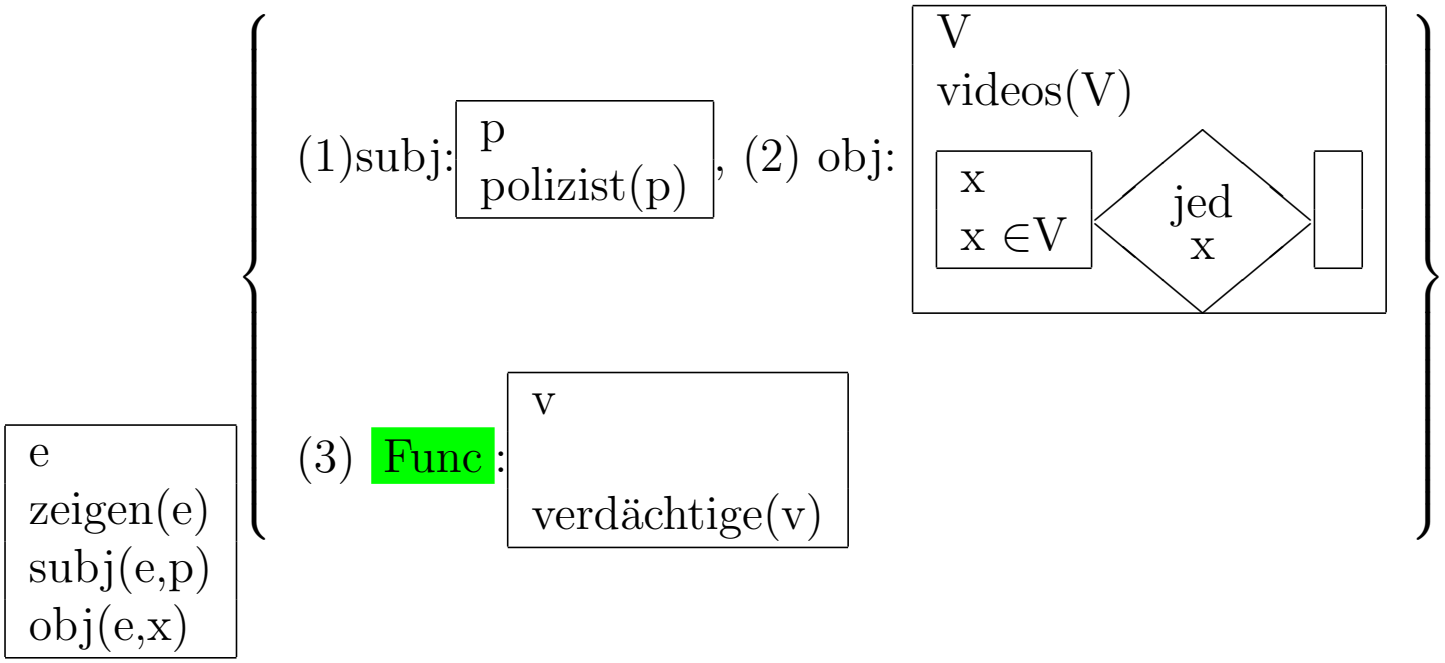
  sie = diese eine Verdächtige $\rightarrow_{\tau}$ elle
  – *Il passait chacune des vidéos d'une suspecte. . . . elle*

- Er zeigte jedes Video einer anderen Verdächtigen (V dat nach acc)

  sie = diese vielen Verdächtigen
  – *Il présentait chacune des vidéos à une suspecte (différente). . . . elles*

- . . .

Interdependenz der Übersetzungen und der Bezüge der Pronomen und der Teilrepräsentationen!



Zusammenhänge

$$sie = \text{Pl\_Pro} \rightarrow_\tau elles \quad \Rightarrow \quad \underset{zeigen}{(2)} > \underset{\rightarrow_\tau \; présenter\ à}{(3)} \wedge \text{Func=dat}$$

$$sie = \text{Sg\_Pro} \rightarrow_\tau elle \quad \Rightarrow \left(\underset{présenter\ à}{(3)} > (2) \wedge \text{Func=dat} \right) \vee \left(\underset{passer}{\text{first}(2,3)} \wedge \text{Func=gen} \right)$$

Pronomenresolution in *translate*:

- morpho-syntaktischer Filter (Lappin McCord 90)
- semantische Bedingungen des gegebenen Kontexts
- pragmatische Bewertungen (Lappin Leass 94)
- Entscheidung !

→ neue semantische Bedingungen

→abschnittsweise Übersetzung (aktuelle Entwicklung)

Schwierige Mismatches

Head switching

(Sadler Thompson 1991, Butt 1994)

Peter schwimmt gerne.
Peter likes to swim.

Strukturelle Umkehrung:

gerne (Adjunkt) → like (syntaktischer Kopf)
schwimmen (Kopf) → swim (Komplement)

$$\tau((1)\ \text{schwimmen}\left\{\begin{array}{ll}\text{subj:} & (2)\ \text{Peter}\\ \text{vadv:} & (3)\ \text{gerne}\end{array}\right\}) = \text{like}\left\{\begin{array}{ll}\text{subj:} & \tau(2)\\ \text{comp:} & \tau(1')\end{array}\right\} \quad ?$$

Lösung

- vermeide Zirkularität
- durch partielle Skopusfestlegung
- Anweisung im Lexikon!

  gerne
  $>$ *adv* []
  $>$ τ*:*
  (u-cat(verb) $\rightarrow$ aimer [item(subj(X'),τ(u-d(subj(X)))),item(obj(binf),τ(u$^{\not\geq id}$)))]

. . . stößt eventuell weitere Disambiguierungen an. . .

Worauf bezieht sich *gerne* ?

Wie Franzi würde Peter gerne schwimmen.

a)

gerne [*wie Franzi schwimmen*] (können)

←*wie Franzi* befindet sich im Skopus von *gerne*

Pierre aimerait (savoir) nager comme (/de la façon de) Franzi.

b)

wie Franzi gerne [*schwimmen*] (gehen)

←*wie Franzi* befindet sich nicht im Skopus von *gerne*

Comme Franzi, Pierre aimerait (aller) nager.

- ein semantisches Skopusphänomen!
- andere Lösungsvorschläge:
  Restriktionsoperator (Kaplan Wedekind 1993) , nicht ko-deskriptiv (Verbmobil)

Partielle Skopusfestlegung

Wie Franzi würde Peter gerne schwimmen.

$$\boxed{\begin{array}{l}\text{e}\\ \text{schwimmen(e)}\\ \text{subj(e,x)}\end{array}}\left\{\text{subj:}\boxed{\begin{array}{l}\text{x}\\ \text{Peter(x)}\end{array}},\ \text{adv:gerne, pp: wiePetra}\right\}$$

(A)

$$\left(\text{(1)}\ \left(\boxed{\begin{array}{l}\text{e}\\ \text{schwimmen(e)}\\ \text{subj(e,x)}\end{array}}\left\{\text{pp: (4)wieFranzi}\right\}\right)\left\{\text{adv:(3)gerne}\right\}\right)\left\{\text{(2) subj:}\boxed{\begin{array}{l}\text{x}\\ \text{Peter(x)}\end{array}}\right\}$$

gerne [schwimmen wie Franzi] $\rightarrow_\tau$ aimer savoir nager comme Franzi

(B)

$$\Big(\Big(\,(1)\boxed{\begin{array}{l}\text{e}\\ \text{schwimmen(e)}\\ \text{subj(e,x)}\end{array}}\Big\{\text{adv: (3)gerne}\Big\}\Big)\Big\{\text{pp: (4)wieFranzi}\Big\}\Big)\Big\{\text{subj:(2)}\boxed{\begin{array}{l}\text{x}\\ \text{Peter(x)}\end{array}}\Big\}$$

wie Franzi gerne [schwimmen] $\rightarrow_\tau$ comme Franzi, aimer aller nager

Repräsentationen, Formalismen - Fazit

- ökonomische Repräsentation (flach)
- Auswertung nach Bedarf
- mächtiger Lexikonformalismus
- keine Transfergrammatik
- modulares Design

'Hybridisierung' von *translate*

Statistische Korpus-Information für Analyse, Transfer und Generierung

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 → (pen = playpen/Laufstall)
  oder → *Word sense disambiguation* (WSD)

  pen(x) →pen(x)$_{|C}$ C $\vdash_d$ sit_x ∈BABY_FRAME →Laufstall(x)
  C $\vdash_D$ sit_x ∈WRITING_FRAME →Stift(x)

- Funktional-/Attachment-Ambiguität

  Beispiel: V NP Prep NP (Hindle/Rooth)
  →Bootstrapping: flache Analyse, Evaluation

'Hybridisierung' von *translate*

Transfer

- *translate 11*:

  Bootstrapping: Analyse und bilinguales Lexikonwissen

  →Kontextbewertungen für Übersetzungsalternativen:

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

- Lernen von (analytischen) Übersetzungsbeziehungen
  Bootstrapping: Lexikonwissen, Source-Targetanalysen

'Hybridisierung' von *translate*

Generierung

- 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