Kurt Eberle

Integration von regel- und statistikbasierten Methoden in der Maschinellen Übersetzung

1 Einführung

Warren Weavers Appell an die akademische Welt, zu untersuchen inwieweit es möglich ist, Texte automatisch zu übersetzen, wird gemeinhin als Beginn der *Maschinellen Übersetzung* verstanden (Weaver (2003); Hutchins (1995)). Seither sind rund 60 Jahre vergangen und das Problem der automatischen Übersetzung von Texten ist keineswegs gelöst, steht aber aktuell im Fokus der computerlinguistischen Forschung wie kaum ein anderes.

Zu Beginn der Forschung standen eher Rechnerprobleme im Vordergrund und architektonisch die sogenannte *direkte Übersetzungsarchitektur*, die schlagwortartig auch als Wort-zu-Wort-Übersetzung gekennzeichnet wird. Danach, in der zweiten Generation der Maschinellen Übersetzung, standen die sogenannten *regelbasierten* Übersetzungssysteme im Zentrum, deren gemeinsames Grundprinzip, bei aller Vielfalt, die im Lauf der Jahre entstanden ist, gekennzeichnet ist durch die Idee, Sätzen abstrakte strukturelle Analysen zuzuweisen und auf dieser Basis zu übersetzen. (Diese Systeme werden zusammengefasst unter der Bezeichnung *RBMT* für *Rule Based Machine Translation*). In der dritten Generation stehen statistische Modelle im Vordergrund (diese sind Instanzen der sog. *SMT* für *Statistics based Machine Translation*). Ohne noch eine echte vierte Generation zu begründen, stehen heute Forschungen im Zentrum, die versuchen, möglichst viel Wissen aus Sprachdaten abzuleiten und dabei Methoden verschiedener Übersetzungstraditionen möglichst effizient in sogenannten *hybriden* Ansätzen zu verbinden.

Eines der größten Probleme für die Maschinelle Übersetzung, vermutlich das zentrale Problem überhaupt, war und ist die Mehrdeutigkeit. Diese Eigenschaft erlaubt es den natürlichen Sprachen, mit einer möglichst geringen Anzahl von Zeichen und Zeichenkombinationen eine maximale Ausdruckskraft zu erzielen. Verwirrung wird dabei vermieden, indem Kontextwissen äußerst effizient ausgenutzt wird, um die richtige Bedeutung hervorzuheben und die falschen Interpretationen auszufiltern. Dies aber ist das größte Hindernis für den Erfolg einfacher Übersetzungskonzeptionen. Wegen der Mehrdeutigkeit genügt es nicht, Übersetzungsregeln als isolierte ein-eindeutige Wortbeziehungen anzulegen, sondern sie müssen als kontextsensitive n:m-Beziehungen definiert werden, wobei die qualitativ wirklich gute Übersetzung bedeutet, dass zum Schluss der ganze Text und der Zweck des Texts in den Blick genommen werden muss, um die kontextuellen Einschränkungen vollständig zu erfassen.

Das ist die Herausforderung, mit der Maschinelle Übersetzung konfrontiert ist.

Wir werden im Folgenden die hauptsächlichen Arten von Mehrdeutigkeit skizzieren, die die Maschinelle Übersetzung potenziell auflösen können muss und die Lösungsansätze, die dazu von den verschiedenen MÜ-Generationen vorgeschlagen wurden.

Nach diesem ersten, eher historisch orientierten und grundlagenbezogenen Teil werden die Hauptlinien hybrider Lösungsansätze vorgestellt, wie sie aktuell in der Literatur diskutiert werden.

Im dritten und letzten Teil wird gezeigt, welche Möglichkeiten bestehen und nahe liegen, ein regelbasiertes MÜ-System semi-automatisch mit Wissen aus Sprachdaten zu vervollständigen. In der Debatte um Übersetzungsarchitekturen wird dabei die Position des linguistisch orientierten Vorgehens eingenommen, statt, etwas zugespitzt formuliert, linguistisches reguläres Wissen aus Sprachdaten erst abzuleiten. Motiviert und skizziert werden die Vorschläge anhand des kommerziell verfügbaren Übersetzungssystems *translate*.

2 Mehrdeutigkeit und *translation mismatches*

2.1 Arten von Mehrdeutigkeit

Alle Arten von sprachlicher Mehrdeutigkeit können Auswirkungen auf die Übersetzung haben, von der Formenlehre von Wörtern bis zu satzübergreifenden pragmatischen Phänomenen. Im folgenden seien einige Beispiele für verschiedene Klassen von Mehrdeutigkeit genannt, ohne dabei vollständig zu sein.

2.1.1 Lexikalische Mehrdeutigkeiten

(1) *a.* *Time*$_{N/V}$ *flies*$_{N/V}$ *like*$_{V/P}$ *an arrow.*
(Die) Zeit fliegt wie ein Pfeil.
Zeitfliegen lieben einen Pfeil.
...
b. *Er vertreibt Mäuse.*
He expels mice.
He sells mice.

(1.a) greift Zenons Paradoxon auf und ist ein bekanntes Beispiel für *kategoriale* Mehrdeutigkeit, wobei die Subskripte die diversen kategorialen Lesarten anzeigen. Entsprechend gibt es neben der (gemeinten) Lesart, bei der die Zeit mit einem fliegenden Pfeil verglichen wird, noch eine Reihe anderer Lesarten. Das '*Die*' in Klammern illustriert, dass es neben den kategorialen Mehrdeutigkeiten hier noch andere Übersetzungsprobleme gibt, die in dem Fall mit unterschiedlichen Konventionen der Sprachen bei der Wiedergabe von Determinationsinformation zu tun haben. Wichtig bei diesem Beispiel ist auch die

Tatsache, dass nicht alle kategorialen Mehrdeutigkeiten greifen können. Grammatikregeln sorgen dafür, dass beispielsweise Lesarten mit *flies*$_V$ *like*$_V$ ausgefiltert werden. Die Filterwirkung von strukturellen Analysen ist das Hauptargument für die Verwendung von entsprechenden Komponenten in Übersetzungssystemen.

Die Mehrdeutigkeit von *vertreiben* in (1.b) ist rein *semantisch* und nicht abhängig von einer kategorialen Mehrdeutigkeit. Auch bei diesen lexikalisch-semantischen Mehrdeutigkeiten gilt, dass reguläres syntagmatisches Wissen isoliert gegebene Lesarten ausfiltern kann: *vertreiben* in der Bedeutung *expel* setzt voraus, dass es sich bei dem direkten Objekt um eine Instanz des semantischen Typs *ANIMAL* handelt. Das Verb hat in dieser Bedeutung eine entsprechende *semantische Selektionsrestriktion*.

D.h. sowohl syntaktisches als auch semantisch-relationales Wissen ist geeignet, bestimmte lexikalische Mehrdeutigkeiten im Syntagma auszufiltern.

2.1.2 Strukturelle Mehrdeutigkeiten

Es gibt eine ganze Reihe von strukturellen Mehrdeutigkeiten syntaktischer und auch rein semantischer Art.

(2) *a. Gebildete Frauen und Männer haben bessere Chancen.*
Les femmes cultivées et les hommes ont de meilleures chances.
Les femmes et les hommes cultivés ont de meilleures chances.

b. Scorsese zeigte den Film seiner Crew.
Scorsese showed the film of his crew..
Scorsese showed the film to his crew.

(2.a) ist ein Beispiel einer *Attachment*-Ambiguität, wobei es mehrere mögliche Bezugspunkte eines Wortes oder einer Struktur gibt. In dem Beispiel sind die beiden Alternativen - *gebildet* bezieht sich auf *Frauen* allein oder auf die ganze N-Koordination *Frauen und Männer* - auch mit unterschiedlichen Übersetzungen assoziiert, was hier daran liegt, dass das Französische einer anderen Wortordnung folgt als das Deutsche und andere Kongruenzregeln hat, mit der Folge, dass die Ambiguität im Deutschen bei der Übersetzung *disambiguiert* werden muss.

Ähnliches ist der Fall in Beispiel (2.b), das eine *funktionale* Ambiguität, die auch *Label*- oder *Etiketten*-Ambiguität genannt wird, beinhaltet: *seiner Crew* im Deutschen ist ambig zwischen Dativ- und Genitivlesart und den entsprechenden semantischen Rollen. Im Englischen muss die Mehrdeutigkeit in diesem Fall aber aufgelöst werden.

2.1.3 Referentielle Mehrdeutigkeiten

Referentielle Bezüge gehen häufig über die Satzgrenze hinaus. Ihre Auflösung ist oft wichtig für die Übersetzung:

(3) *Die Katze spielt mit der Maus. Sie mag das nicht.*
Le chat joue avec la souris. Il / Elle n'aime pas cela.

In (3) gibt es Gründe, das Pronomen *sie* auf die *Katze* zu beziehen (Parallelität der Konstruktion), als auch solche, die nahelegen, es auf die *Maus* zu beziehen (Weltwissen). In manchen Kontexten wird die eine, in manchen die andere Lösung favorisiert sein, in jedem Fall muss die Beziehung bei der Übersetzung ins Französische wegen der Genus-Unterschiede zwischen *chat* und *souris* aufgelöst werden.

2.2 *Translation mismatches*

Nach einem Vorschlag aus Kameyama et al. (1991) sind *translation mismatches* Übersetzungsschwierigkeiten, die aus systemischen Unterschieden der ineinander zu übersetzenden Sprachen resultieren: Dann, wenn eine Sprache keine Übersetzungsäquivalent der gleichen Form und mit demgleichen Bedeutungsumfang für ein Wort, eine Phrase oder einen Satz vorsieht, ist es notwendig, zusätzliches Wissen aus dem Kontext zur Disambiguierung abzuleiten und eine entsprechend spezifischere Form für die Formulierung in der Zielsprache zu wählen, oder, falls das nicht möglich ist, auch eine allgemeinere Form zu wählen. Es ist eben nicht immer möglich, wie aus den Übersetzungswissenschaften hinlänglich bekannt ist, für Wörter, Phrasen, Sätze und auch Texte in jedem Fall eine Übersetzung mit genau gleichem Informationsgehalt zu finden. Dies kann in der Maschinellen Übersetzung nicht anders sein.

Nicht alle formal-strukturellen Unterschiede zwischen den Sprachen sind auch gleichzeitig Übersetzungsschwierigkeiten. Die folgenden sind oft genannte Unterschiede.

2.2.1 Lexikalische Divergenz

Sie liegt vor bei unterschiedlicher Strukturierung der Wortfelder. Bekannte Beispiele sind das Fehlen von Substantiven für *Rappe* und *Schimmel* im Französischen oder die in Durrell (2000) beschriebenen Felder zu *Boden/Erde* etc. im Deutschen und *soil/earth* etc. Im Englischen mit ähnlichen Bedeutungen, aber unterschiedlichen Zusammenordnungen.

Stilistisch stellt eine Wortlücke natürlich ein Problem dar, aber inhaltlich nicht notwendigerweise. Französisch *cheval blanc* für *Schimmel* oder das deutsche Kompositum *Jungbulle* für Spanisch *novillo* etc. sind inhaltlich durchaus akzeptable Übersetzungen. Die richtigen Übersetzungen für Wörter wie *Boden* zu finden ist jedoch viel schwieriger, weil es zwar Übersetzungen als Substantiv im Englischen gibt, diese aber das Wortfeld anders strukturieren und es deshalb auf die genaue Bedeutung ankommt und diese erst aus dem Kontext abgeleitet werden muss.

2.2.2 Thematische Divergenz und Scrambling

Thematische Divergenz liegt vor, wenn die Kasusrahmen von Wörtern nicht gleichförmig übersetzt werden (vgl. Dorr (1994); Hutchins and Somers (1992)), wie in (4):

(4) *Mir gefällt die Aufführung.*
I like the performance.

Diese Divergenz stellt kein Übersetzungsproblem dar, wenn bekannt ist, welcher Kasusrahmen vorliegt und das Lexikon vorgibt, welche Kasus (oder Funktionen oder Rollen) in welche übergehen (hier indirektes Objekt in Subjekt und Subjekt in direktes Objekt).

Es kann aber natürlich bei Verwendungsmehrdeutigkeit ein Problem sein, zu bestimmen welche Kasus oder Rollen wie besetzt sind (vgl. (2.b) oben mit der formalen Ununterscheidbarkeit von Dativ und Genitiv). Außerdem ist eine Voraussetzung für die korrekte Übersetzung (wenigstens in einem linguistisch konzipierten Übersetzungssystem), dass das Lexikon detailliert die Abbildung der Kasus, Funktionen oder Rollen beschreibt; dieses ist in jedem Fall ein Problem der Quantität.

Scrambling, d.h. die zulässige unterschiedliche Anordnung von Konstituenten an der Satz-Oberfläche stellt häufig ein schwieriges Problem dar in der Übersetzung, weil Sprachen unterschiedlichen Anordnungsprinzipien folgen und die zu wählende Anordnung im Zielsatz oft von Wissen über die pragmatische Informationsstruktur des Satzes abhängig ist (z.B. vom Wissen *welche Information neu und welche es nicht ist*):

(5) *Pierre remet le bouquet à la femme.*
a. Pierre überreicht der Frau den Strauß.
b. Pierre überreicht den Strauß der Frau.

2.2.3 Hinzufügen, Tilgen, Umkehren von Teilstrukturen

In der Regel stellen Strukturveränderungen, wie sie die folgenden Beispiele illustrieren, zwar Anforderungen an die Expressivität des bilingualen Lexikons, aber keine besonderen an die inhaltliche Auswertung des umgebenden Textes.

(6) *a. Pierre traverse la rivière en nageant.*
Pierre durchschwimmt den Fluß.
b. Pierre raucht gerne.
Pierre likes to smoke.

(6.a) ist ein Beispiel für *Inkorporation* (des Partizipialausdrucks in das Verb im Deutschen) und (6.b) für das sogenannte *head switching* (bei dem die Übersetzung des Kopfs der Ausgangsstruktur, *smoke/rauchen*, in der Zielstruktur abhängig wird von der Übersetzung eines Komplements der Ausgangsstruktur, *like to/gerne* (vgl. u.a. Sadler and Thompson (1991); Kaplan et al. (1989)).

Gerade Inkorporation und vor allem Head switching machen deutlich, dass, neben der adäquaten Disambiguierung übersetzungsrelevanter Mehrdeutigkeiten, eine Voraussetzung für die qualitativ gute Maschinelle Übersetzung ist, solche Strukturveränderungen adäquat repräsentieren zu können. Dabei spielt eine Rolle, auf welcher Ebene die zu übersetzenden Texte und Sätze überhaupt repräsentiert werden.

2.3 Repräsentationen

Im Rahmen von RBMT sind verschiedene Vorschläge für geeignete Repräsentationen für Texte und Sätze und die Ebene der Übersetzung gemacht worden (Zu einem Überblick vgl. Hutchins and Somers (1992); Trujillo (1992)). Sehr häufig werden die Sätze des Inputs syntaktisch analysiert und den Analysestrukturen syntaktische Strukturen der Zielsprache zugewiesen, aus denen dann Sätze der Zielsprache generiert werden ‚die die Strukturanforderungen erfüllen. Es gibt aber auch Ansätze und Systeme, bei denen der Input auf einer 'höheren' semantischen oder konzeptuellen Ebene repräsentiert und dann übersetzt wird. Dabei entstehen die Repräsenationen typischerweise entsprechend der Montague'schen Vorgehensweise aus weniger abstrakten syntaktischen Strukturen. Die Möglichkeiten, die es dabei prinizipiell gibt und die auch fast alle ihren Niederschlag in konkreten Systemen fanden, werden häufig in einem Schaubild in der Form eines Dreiecks oder einer Pyramide dargestellt. Solche Zusammenstellungen gehen auf einen Vorschlag von Vauquois zurück:

2.3.1 Architekturschema nach Vauquois

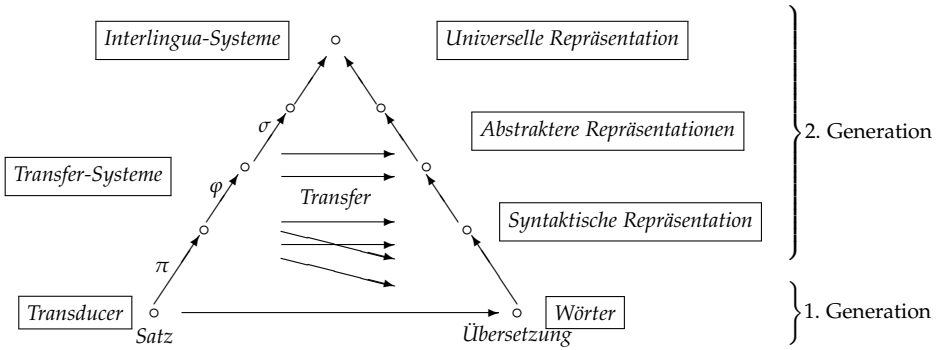


Abbildung 1: Regel-basierte Architekturen (vgl. Vauquois (1975))

An der Basis der Struktur finden sich die *Transducer*-Übersetzungsmodelle, bei denen keine oder nur eine marginale Analyse stattfindet. Das sind insbesondere die schon

genannten Wort-zu-Wort-Übersetzungsansätze der 1. MÜ-Generation. Bekannt geworden aus dieser Zeit ist vor allem der *Georgetown*-Demonstrator des *Georgetown Automatic Translation*-Projekts (GAT), auf den bzw. das das kommerzielle Übersetzungssystem SYSTRAN zurückgeht. Ein anderes kommerzielles System, das auf einen Prototypen aus der 1. Generation zurückgeht, ist LOGOS (cf. Stoll (1986); Trabulsi (1989); Drouin (1989)). Bei den sog. *Transfer-Systemen* wird der Input wie beschrieben einer syntaktischen oder weitergehenden Analyse unterzogen. Die Ergebnisse werden in Strukturen der Zielsprache *transferiert* und aus diesen werden, je nach Abstraktionsgrad der Struktur, mit mehr oder weniger Aufwand, die Zielsätze generiert. Die meisten kommerziellen Systeme heute sind im Wesentlichen solche Systeme der 2. Generation (auch die aktuellen Versionen der SYSTRAN-Sprachpaare). In diesem Rahmen sind sie zumeist Vertreter des eher Syntax- statt Semantik-orientierten Transfers. Die in die Vauquois-Struktur eingefügten nicht-horizontalen Pfeile deuten an, dass nicht alle Systeme einer völlig symmetrischen Architektur folgen. Manche vermeiden es, die Analysestrukturen zuerst in zielsprachliche Strukturen derselben Abstraktionsebene zu überführen, indem sie aus den Analysestrukturen direkt (in der Regel durch Anwendung eines Systems von Ersetzungs-Regeln) als Ergebnis des Transfers den Zielsatz oder eine oberflächennahe Repräsentation des Zielsatzes erzeugen, ohne zuvor entsprechende abstraktere Zielstrukturen erzeugt zu haben. Die weiter unten besprochene Architektur des *Logic based Machine Translation* Projekts (LMT) der IBM verfolgt beispielsweise einen solchen Transfer-Ansatz.

In gewisser Weise markiert die direkte Übersetzung das eine Extrem der Transfer-Systeme, mit minimaler Analyse (die zumeist immerhin die Abbildung in Grundformen mit morphologischer Kennzeichnung vorsieht), während die *Interlingua*-Übersetzung an der Spitze des Vauquois-Dreiecks das andere Extrem markiert. *Interlingua* ist dabei eine inhaltliche Analyse, die von jeder sprachspezifischen Beschreibung abstrahiert und als solche nicht nur Ergebnis der Analyse, sondern gleichzeitig, ohne weitere Transfernotwendigkeit, Grundlage der zielsprachlichen Generierung ist. Ein Vertreter dieser Interlingua-Architektur ist das *UNITRAN*-System (vgl. Dorr (1993, 1994)). Es ordnet den Texten und Sätzen sog. *lexical conceptual structures* (LCS) zu (vgl. Jackendoff (1983, 1990)) und generiert daraus die Zielsätze und -texte. (Es gibt andere Vorstellungen von Interlingua, die eher an der ESPERANTO-Philosophie orientiet sind, die aber im Zusammenhang mit dem Umgang mit Mehrdeutigkeiten keinen neuen Aspekt einbringen und deshalb hier weggelassen werden).

Die Bezeichnungen π, φ und σ am Analyse-Schenkel des Dreiecks sollen an die entsprechend benannten Projektionen zwischen den LFG-Analyse-Ebenen erinnern (vgl. Kaplan and Bresnan (1982)) und damit andeuten, dass neben (und zwischen) Konstituentenstruktur- und semantischer Ebene eine Reihe von Abstraktionsebenen denkbar sind, wie die funktionale Ebene der LFG oder Entsprechendes, aber auch unterschiedliche Grade der semantischen Auswertung, bis hin zu einer konzeptuellen

Interlingua.[1]

2.3.2 Mehrdeutigkeit, Mismatches und Repräsentation

Wie ist der Zusammenhang zwischen Mehrdeutigkeit, Mismatches und Repräsentation? Je abstrakter die Repräsentation der Sätze ist, desto geringer ist offensichtlich der strukturelle Unterschied zwischen Quell- und Zielrepräsentation. Das veranschaulicht die Verjüngung des Vauquois-Dreiecks nach oben; zwei Beispiele:

• Tempus- und Aspektinformation

Auf der Ebene der syntaktischen Repräsentation sind analytische und synthetische Tempus- und Aspektinformationen in der Regel als solche noch erhalten und damit strukturell voneinander verschieden.

Auf der Ebene der funktionalen Repräsentation (der LFG beispielsweise) und darüber, sind die Unterschiede nur noch als unterschiedliche Feature-Werte repräsentiert oder (bei angenommener Bedeutungsgleichheit auf der semantischen Ebene) nicht mehr vorhanden, vgl. (7) und dessen funktionale Repräsentationen (8).

(7) *Pierre würde den Wein nicht mögen.*
Pierre n'aimerait pas le vin.

(8)

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

• Rollen-Information

Semantische Repräsentationen behalten in der Regel Perspektiven, wie sie für den Zusammenhang von Individuen in Subkategoriesierungsrahmen etc. eingenommen werden, in der Form bei (vgl. beispielsweise die vorgeschlagenen Repräsentationen der *Diskursrepräsentationstheorie* (DRT) in Kamp and Reyle (1993) oder der *Situationstheorie* in Barwise and Perry (1983)). Deshalb bleiben unterschiedliche Perspektiven wie beispielsweise in der Head switching-Übersetzung in (6.b) auf dieser Ebene bzw. diesen Ebenen, erhalten. Zielt die semantische Repräsentation aber auf die den Sätzen zugrundeliegende Konzeptualsierung, wie in UNITRAN, kann der strukturelle Unterschied durch die Abbildung in nicht-sprachnahe semantische Operatoren und Basiskonstrukte vermieden werden, wie in der folgenden LCS-orientierten Repräsentation (9) von (6.b):

[1] Im ursprünglichen Vorschlag von Vauquois finden sich solche Projektionen natürlich nicht, sondern nur Pfeile entlang der Schenkel des Dreiecks, die zeigen, wie lange die Wege für Analyse und Generierung werden können.

(9) *gerne(pierre,λ x.(rauchen(x)))*
like(pierre,λ x.(smoke(x)))

Die Distanzverringerung zwischen Transfer-In- und Output, die man erzielt durch eine Analyse der Sätze, die sich auf immer abstraktere Repräsentationsebenen bezieht, wird in der Regel allerdings erkauft durch einen immer größeren Disambiguierungsaufwand. (Um von der spezifischen Form abstrahieren zu können, muss, wenigstens dann, wenn dieser Form mehrere Inhalte der jeweiligen Ebene zugeordnet werden können, entschieden werden, welcher der möglichen Inhalte gemeint ist). Nicht umsonst wird das wohl bekannteste *Interlingua*-System, *Kant* (und das spätere *Mikrokosmos*), als Repräsentant von *Knowledge based Machine Translation* etikettiert (KBMT, vgl. Carbonell et al. (1992); Onyshkevych and Nirenburg (1995); Nirenburg et al. (1996)).

Manche Sprachen sind strukturell eng benachbart und verwenden dieselben Mehrdeutigkeiten. Deshalb brauchen Mehrdeutigkeiten einer ganzen Reihe von Arten oft gar nicht aufgelöst zu werden, um eine korrekte Übersetzung zu wählen: So sind Wörter wie *Drucker* und *printer* zwar mehrdeutig, umfassen aber im wesentlichen die selben Bedeutungen, können also ineinander übersetzt werden. Neben solchen lexikalischen Mehrdeutigkeiten gibt es auch viele strukturelle Mehrdeutigkeiten, die bei vielen Übersetzungsrichtungen nicht aufgelöst werden müssen.Ein prominentes Beispiel sind die für die Semantik ansonsten so wichtigen Skopusambiguitäten:

Unabhängig davon, ob (10) die Lesart (10.a) oder (10.b) im Kontext erhält, wird die Übersetzung ins Englische in der Regel die aus (10.c) sein.

(10) *Viele Hunde jagen eine Katze.*
a. viel(x,hund,ein(y,katze,jagen(x,y)))
b. ein(y,katze,viel(x,hund,jagen(x,y)))
c. Many dogs chase a cat.

Aus dieser Einsicht heraus ist von Kay und anderen auch das Konzept der *variablen Analysetiefe* für die Maschinelle Übersetzung vorgeschlagen worden, mit der Perspektive, die Übersetzungsmaschine als *negociator* zu sehen, die in Abhängigkeit der Übersetzungsaufgaben regelt, wie tief analysiert werden soll (vgl. Kay et al. (1994)).

Mit dieser Konzeption stellt sich die Frage, wie mit Ambiguitäten umgegangen werden soll, die nicht aufgelöst werden brauchen. Wie werden sie repräsentiert? Es gibt unterschiedliche Vorgehensweisen, auch abhängig von den verschiedenen Repräsentationsebenen.

Syntaktische Mehrdeutigkeiten werden in den allermeisten System-Typen aufgelöst, auch wenn sie dies nicht müssten. Aufgelöst werden sie meistens nach einer Präferenzheuristik auf der Basis von semantisch-sortalem Wissen und einem Grundbestand an Weltwissen.

Semantische Mehrdeutigkeit von Wörtern findet sich in vielen Transfersystemen nicht direkt, sondern als Menge verschiedener Übersetzungsmöglichkeiten (wie *lock* und

castle zu *Schloss*), eventuell versehen mit Gewichten oder kontextuellen Übersetzungsbedingungen oder mit beidem. Strukturell-semantische Mehrdeutigkeit, die nicht Folge syntaktischer Mehrdeutigkeit ist, wird in den meisten kommerziellen, aber auch in vielen klassischen Forschungssystemen nicht behandelt.

Seit den frühen 90er Jahren sind vermehrt, vor allem im Spektrum der DRT, Vorschläge entstanden, Mehrdeutigkeiten *unterspezifiziert*, also kompakt und unaufgelöst, zu repräsentieren. Forschungsseitig ist das früh und mit viel Wahrnehmung in der Literatur vor allem in den RBMT-Prototypen des VERBMOBIL-Projekts realisert worden (vgl. Wahlster (2000), speziell Emele et al. (2000)). Für den kommerziellen Bereich ist aufgrund der unterschiedlichen Veröffentlichungslage schwer abzuschätzen, in welchen Systemen es entsprechende Repräsentationen gibt.

Bei der Skizzierung von Integrationsmöglichkeiten im übernächsten Abschnitt beziehen wir uns auf das System *translate*, für das es solche Repräsentationen und entsprechende Veröffentlichungen gibt.

Bevor das geschieht, ist aber zu beleuchten, welcher Philosophie die Vorschläge der dritten MÜ-Generation folgen und welches Potenzial sich daraus für hybride Entwicklungen ableiten lässt.

3 Daten-getriebene Maschinelle Übersetzung

Seit Ende der 80er Jahre sind Übersetzungsarchitekturen vorgestellt worden, die bewusst auf linguistisches A-priori-Wissen verzichten und versuchen, Übersetzungssysteme (allein) aus Sprach- und Übersetzungsdaten abzuleiten. Solche Ansätze haben natürlich eine sehr hohe Attraktivität, weil sie versprechen, Systeme weitaus ökonomischer herstellen zu können.

3.1 Das statistische Übersetzungsmodell

Der statistische Ansatz ist aus den Erfahrungen mit statistischer Spracherkennung entstanden und ist, zumindest was das 'klassische' *Source-Channel*-oder *Noisy-Channel*-Modell anbelangt eine mehr oder weniger direkte Übertragung auf das Übersetzungsproblem (vgl. Brown et al. (1990, 1992)).

Das Modell ist eine Kombination aus drei Basismodellen: dem *Alignment-Modell* (das die Wahrscheinlichkeit für Wörter angibt, in bestimmten Positionen zu erscheinen), dem *Sprachmodell* (das die Wahrscheinlichkeit angibt, mit der die Wörter einer Sprache als Nachfolger anderer erscheinen) und dem *Übersetzungsmodell* (das die Wahrscheinlichkeit angibt, mit der Wörter in solche der Zielsprache in spezifischen Kontexten übersetzt werden). Die Kontexte sind dabei Folgen von n Wörtern, sog. n-Gramme.

Die folgende Formel beschreibt die auszuwählende Zielwortfolge (den Zielsatz) als diejenige Folge $\hat{e}_1^I$ (bestehend aus den Wörtern $e_1, \ldots, e_I$), die den höchsten Wahr-

scheinlichkeitswert hat, gegeben den Quellsatz f_1^J (bestehend aus den Wörtern f_1,..., f_J), wobei die Wahrscheinlichkeit unter Zuhilfenahme der Bayes'schen Formel aus den einzelnen Wahrscheinlichkeiten nach den drei Basismodellen errechnet wird (wobei in der gegebenen einfachen Version Alignment- und Sprachmodell integriert sind):

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

Das Noisy-Channel-Modell war sehr erfolgreich und ist Basis vieler in der Folge entstandener Verfeinerungen, unter anderem der für VERBMOBIL entwickelten statistischen Prototypen (vgl. Vogel et al. (2000)). [2]

3.2 Die beispielbasierte Übersetzung

Die beispielbasierte Übersetzung (*Example Based Machine Translation*: EBMT) ist aus der *Translation Memory*-Technologie entstanden. Translation Memories speichern Sätze und ihre Übersetzung zur (automatischen) Verwendung in späteren Übersetzungen (vgl. Schäler (1996)). Diese Methode verfeinert EBMT, indem nicht (nur) Sätze gespeichert werden, sondern (häufig in Sätzen vorkommende) Sequenzen von Wörtern, mit den jeweiligen im untersuchten Datenmaterial verwendeten Übersetzungen, die wieder Sequenzen von Wörtern sind. Bei der EBMT-Übersetzung wird dann für einen (neuen) Satz eine möglichst 'beste' Überdeckung aus solchen gespeicherten *Beispielen* berechnet und aus deren Zielteilen der Zielsatz (vgl. Sumita et al. (1990); Maruyama and Watanabe (1992)).

4 Auf der Suche nach hybriden Systemen

Hybride Systeme, also solche, die sich aus verschiedenen Systemen bedienen, können in unterschiedlicher Weise konstruiert werden, *schwach integrierend* und *stark integrierend* vgl. Eisele et al. (2008).

Ein *schwach integrierender* Ansatz sieht ein *Multi-System* vor, das im Wesentlichen aus einer Reihe von konkurrierenden MÜ-Systemen und einer Kontrollkomponente besteht, wobei die MÜ-Systeme parallel den Input übersetzen und die Resultate von der Kontrollkomponente zu einem Übersetzungsvorschlag aufbereitet werden, der dann ausgegeben wird. Das Aufbereiten der Ausgabe kann einfach aus dem Vergleich der Ergebnisse und Auswahl nach bestimmten Präferenzkriterien bestehen, wenn die Ergebnisse analytisch strukturiert sind. Die Ausgabe kann aber auch aus Teilen verschiedener Ergebnisse zusammengesetzt werden, ähnlich dem Vorgehen bei der EBMT. Ein frühes,

[2] Die Verwendung von *e* und *f* in diesem und späteren Modellen bezieht sich darauf, dass statistische Übersetzungsforschung zu Beginn vor allem unter Verwendung des englisch-französischen-Hansard-Korpus, der elektronisch verfügbaren kanadischen Parlamentstexte durchgeführt wurde.

wenn nicht das erste System dieser Art ist der (erste) Verbmobil-Demonstrator, bei dem mehrere SMT- und RBMT-Systeme verwendet wurden (vgl.Wahlster (2000)).

Stark integrierende Ansätze versuchen SMT- und RBMT-Komponenten bzw. Methoden unterhalb der Eingabe-/Ausgabe-Ebene zu kombinieren, also beispielsweise die morphologische Analyse des RBMT-Systems im SMT-System zu nutzen oder Konkurrenz auf Teil-Analyse-Ebene zu installieren und dergleichen.

Wir skizzieren im Folgenden einige, in den letzten Jahren entstandene, stark integrierende Ansätze. Gekennzeichnet sind diese zumeist dadurch, dass sie von einem Architekturtyp als Basis ausgehen und diesen durch Verfahren oder Information aus anderen Architekturen ergänzen.

4.1 Maximum-Entropie-Modell und linguistische Features

Eines der Hauptprobleme (rein) datengetriebener statistischer Ansätze zum Lernen von Sprachen und Übersetzungen ist das sog. *Sparse-Data-Problem*, weil die elektronisch verfügbaren Daten nicht ausgewogen genug sind, um das Sprach- bzw. Übersetzungsverhalten als solches ausgewogen in Wahrscheinlichkeiten abzubilden. Dieses Problem wird noch gravierender, wenn sich die erzeugten Modelle auf einzelne Wörter und Wortformen beziehen wie beim Source-Channel-Modell in seiner Grundform. D.h. Phänomene wie die Zusammenschau mehrerer Wörter (bei Funktionsverbgefügen und Mehrwortausdrücken aller Art) oder die Abstraktion auf Klassen von Wörtern (desselben Lemmas, desselben semantischen Typs) spielen bei der Berechnung der Wahrscheinlichkeiten und beim Suchalgorithmus zur Bestimmung einer besten Übersetzung keine Rolle. Die Verwendung von Grundformen widerspricht der behavioristischen 'A posteriori'-Philosophie, die die Konzeption des Source-Channel-Modells, wenn nicht geleitet, so doch beeinflusst hat (das erste IBM-SMT-System heißt bezeichnenderweise *CANDIDE*). Schließlich ist an dem Ansatz auch (oder, je nach Standpunkt, vor allem) attraktiv, detailliertes und damit kostenintensiv herzustellendes Sprachwissen nicht als Vorarbeit in das Übersetzungssystem investieren zu müssen, sondern es über Training und Anwendung des Systems als Ableitung umsonst zu erhalten. Bei der Übersetzung von einzelnen Wörtern abstrahieren zu können, und bei Bedarf die Übersetzungsrelation für (zusammenhängende) Wortgruppen definieren zu können, widerspricht der Philosophie nicht. (Wortgruppen sind in einer Zeichenkette konkret vorhanden und keine abstrakten Ableitungen). Deshalb ist die sog. *Fertilität* (*fertility*), die die Übersetzung durch mehrere Wörter thematisiert, schon in den ersten Papieren zur SMT als Möglichkeit miteinbezogen worden. Das Problem des Source-Channel-Ansatzes ist es, dass es nur schwer möglich ist, darüberhinaus weitere Informationen in den Modell-Entwurf mitaufzunehmen, selbst wenn dies gewollt wird.

In einem Aufsatz von 2002, der viel Aufmerksamkeit gefunden und viele Modelle in der Folge beeinflusst hat, schlagen Och und Ney vor, den Source-Channel-Ansatz, der

letztlich nur zwei statistische Informationstypen (mit einigen parametrischen Verschiebungen) zulässt, durch ein Maximum-Entropie-Modell zu ersetzen, das erlaubt, beliebig viele statistische Parameter in die Berechnung der wahrscheinlichsten Übersetzung miteinzubeziehen (vgl. Och and Ney (2002)). Der entscheidene Punkt an der wie folgt vorgeschlagenen Auswahlfunktion ist insofern die zahlenmäßig nicht begrenzte Verwendbarkeit sog. *Feature-Funktionen*, h_m:

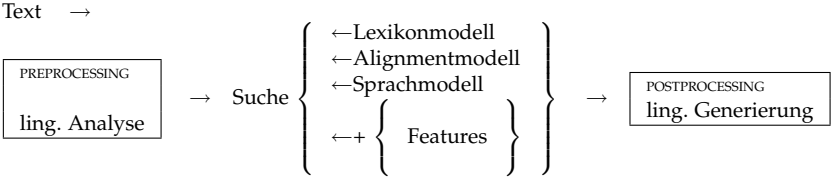
$$\hat{e}_1^I = \underset{e_1^I}{\operatorname{argmax}} \{ \sum_{m=1}^{M} \lambda_m h_m(e_1^I, f_1^J) \}$$

Diese Feature-Funktionen können durchaus auch linguistisches Wissen beschreiben, wobei es keine Rolle spielt, ob (für einzelne h_m) dieses Wissen datengetrieben aus (auch einsprachigen) Korpora abgeleitet wurde oder konventionell regelbasiert zugeordnet wird. Dieser Ansatz gestattet es also in stark-integrierender Weise, regelbasiertes Wissen in ein grundsätzlich statistisches System aufzunehmen. Relationale Features können beispielsweise kategoriale Gleichheit zwischen Quell- und Zielausdruck bevorzugen oder semantische Ähnlichkeit oder auch einzelsprachliche Erwartungen zum syntaktischen und semantischen Zusammenhang von Syntagmen.

4.2 Regelbasierte Vor- und Nachbereitung: SMT auf abgeleiteten Repräsentationen

Einen anderen Weg der 'Hybridisierung' verfolgen Vorschläge wie die *Dependency treelet translation* (vgl. Quirk et al. (2006)). Dabei werden die Quell- und Zielsätze des Korpus, aus dem das SMT-System gebildet wird, mit einzelsprachlichen Grammatiken analysiert, und das SMT-System bezogen auf die Ergebnisse der Analyse (bei der dependency treelet translation sind das Dependenzbäume) trainiert, d.h. es werden dort Analysen bzw. die Elemente, aus denen diese bestehen, aufeinander bezogen:

Dependency treelet translation



Der Vorteil aller Ansätze dieser Art liegt darin, dass der Übergang zu Abstraktionen bedeutet, dass das Modell, um genügend signifikant zu sein, mit kleineren Korpora auskommt.[3] Der Nachteil dieser Ansätze liegt darin, dass sie Vorwissen verlangen

[3] Das Ausgewogenheitsproblem reduziert sich dadurch allerdings nur bei Phänomenen, die durch die Abstraktionen thematisiert und damit abgepuffert werden, also beispielsweise seltene Wortformen durch morphologische

und dass die Analyse der Sätze fehlerhaft sein kann. Das Mehrdeutigkeitsproblem der Sprache wirkt sich hier, je nach Tiefe der Analyse, gravierend aus.

4.3 Klassen von Beispielen: Beispieltypen

Vorschläge, die in eine ähnliche Richtung weisen, wie die im letzten Abschnitt, aber aus einer anderen Perspektive heraus, sind solche wie das HIERO-Modell von Chiang (vgl.Chiang (2006), HIERO für *hierarchical phrase based translation*).

Die Idee ist, EBMT flexibler zu gestalten, indem Beispiele nicht einfach Teilstrings von Sätzen sind, die aus bilingualen Korpora (nach bestimmten Häufigkeitskriterien) extrahiert und aufeinander bezogen werden, sondern (linguistisch) strukturiert sein sollen oder können. In HIERO sehen solche Beispiele - *Phrasen* - Variablen für Konstituenten vor, die bei der Satzübersetzung durch andere Beispiele instantiiert werden können. D.h. ein Satz wird in eine hierarchische Struktur von Phrasen analysiert, deren beispielbasierte Übersetzungen entsprechend der Bezugsinformtion zum Zielsatz zusammengebaut werden.

Die folgende Regel ist typisch für diese Art von rekursiver Übersetzungsinformation. Sie thematisiert die Übersetzung der englischen Possessivkonstruktion mit Genitiv-s durch eine Konstruktion mit de-PP im Französischen.

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

Anders als bei Ansätzen wie der dependency treelet translation wird ein Satz bei solchen Vorschlägen nicht (notwendig) in alle seine Teile analysiert, sondern bestimmte Abschnitte bleiben unanalysiert; es findet, wenn man so will, eine syntaktische Analyse auf weniger fein granulierter Ebene statt. Die Ebene der Granulation gibt dabei, ebenfalls anders als bei der dependeny treelet translation, nicht die linguistische Analysekompetenz vor, sondern die durch die Korpus-Daten bestimmte Unterscheidungsnotwendigkeit. Die Vorteile sind die entsprechend geringeren Kosten, die Nachteile sind zu erwartende Fehler dort, wo feiner granulierte Analysen als solche oder Konsequenzen daraus benötigt werden.

Sinnvoll scheinen Modelle, die flexibel tiefere Analysen durchführen können, wo das nötig erscheint, und dies vermeiden, wo es nicht nötig erscheint, und damit die Fehlinterpretationen, wie sie aus tieferen Analysen resultieren können minimieren.

Abstraktion, semantische Selektionsbeschränkungen durch semantische Klassifizierung etc.; aber nicht was die Übersetzung selten vorkommender Sätze eines bestimmten Typs betrifft, z.B. spezielle Frageformen etc.; dazu wäre notwendig, verschiedene Konstruktionen in einer Klasse zusammenfassen zu können.

5 Integration statistisch gewonnener Information in RBMT am Beispiel *translate*

Ein RBMT-System besitzt Komponenten zur morphologischen Analyse des Inputs, d.h. es kann einen Input *taggen* und den Wortformen ihre morphologische Klasse und Grundform zuweisen. Es besitzt Komponenten für die syntaktische Analyse oder für tiefere Analysen und bietet damit die Voraussetzung für Verfahren wie die dependency treelet translation. Darin liegt die Chance, Übersetzungen von Teilstrukturen, strukturierten Phrasen, aus Korpora zu lernen. Das statistische Modell ist, gegeben ein Korpus einer bestimmten Größe, um so besser, je weniger idiosynkratisch die Ausdrücke (im Sinne von Repräsentationen) sind, die potenziell aufeinander bezogen werden. Am besten geeignet sind offensichtlich Systeme, die erlauben, Sätze, abhängig von Zwecken, Repräsentationen unterschiedlicher Ebenen und Abstraktionsgrade zuzuweisen. Das Übersetzungssystem *translate* erlaubt solche Repräsentationen. Wir zeigen im folgenden, welche Integrationen statistisch aus Korpora gewonnener Information geeignet erscheinen bzw. in diesem System implementiert sind.

5.1

translate *translate* ist ein kommerzielles Übersetzungssystem (vgl. `http://lingenio.de/Deutsch/Produkte/Uebersetzungssysteme.htm`). Es geht zurück auf das *Logic based Machine Translation* LMT-Projekt der IBM, das Ende der 80er Jahre aufgelegt wurde zu dem Zweck, ein modulares, linguistisch prinzipienbasiertes Übersetzungssystem mit möglichst breiter grammatischer und lexikalischer Abdeckung für viele Sprachen zu erstellen (vgl. McCord (1989)). Das Deutsch-Englisch-System des LMT-Projekts wurde erstmals als Produkt 1996 veröffentlicht, unter dem Namen *Personal Translator*. *translate* ist eine Weiterentwicklung. LMT sieht Transfer auf der Ebene von Analysen der sog. *slot grammar* vor, einer unifikationsbasierten Dependenzgrammatik (vgl. McCord (1991)). Es erlaubt, lexikalische Einträge semantisch zu klassifizieren und semantische Selektionsbeschränkungen zu formulieren, sodass im Zusammenspiel dieser Informationen bestimmte strukturelle und lexikalische Lesarten ausgeschlossen bzw. präferiert werden können. Pronomen können zur Übersetzung satzübergreifend aufgelöst werden (vgl. Lappin and McCord (1990)).

Um weitergehende semantische Auswertung und strukturell einfachere Transferrelationen zu ermöglichen, sieht die Weiterentwicklung zu *translate* eine Abbildung von slot-grammar-Dependenzanalysen zu unterspezifizierten semantischen Repräsentationen vor, entsprechend der folgenden Graphik in Abb. 2.

Wie Abb. 2 zeigt, generiert LMT Zielsätze relativ direkt aus den syntaktischen Dependenzanalysen (die Ergebnis sind einer Analyse im Sinne der LMT-typischen Projektion π). In *translate* können den syntaktischen Dependenz-Analysen flache semantische Repräsentationen zugewiesen und in entsprechende Repäsentationen der Zielsprache

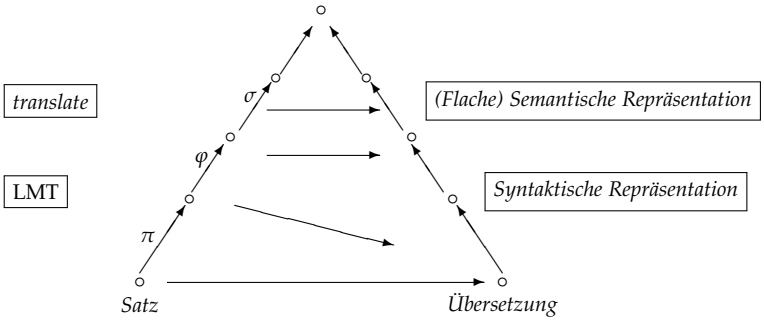


Abbildung 2: Transfer in *translate*

übersetzt werden, aus denen dann die Zielsätze generiert werden. Flache semantische Analysen können bei Bedarf weiter spezifiziert werden, sodass Transfer auf verschiedenen Ebenen der semantischen Auswertung stattfinden kann. (Die Verwendung der LMT-typischen Variablen φ und σ deuten an, dass die erreichten Ebenen im Spezifikationsgrad der funktionalen Ebene der LFG, aber auch Auswertungen bis hin zu einer vollspezifizierten semantischen Analyse im Sinne der σ-Projektion entsprechen können.)

5.2 Flache Semantik

translate verwendet für die semantische Repräsentation eine Kodierung der sog. *flachen unterspezifizierten Diskursrepräsentationsstrukturen* (FUDRSen) der *Flat underspecified discourse representation theory* (FUDRT). FUDRT ist eine Erweiterung der *Underspecified discourse representation theory* (UDRT) (Reyle (1993)) mit dem Zweck, neben Skopusambiguitäten auch andere Arten von strukturellen Mehrdeutigkeiten sowie lexikalische Mehrdeutigkeiten kompakt notieren und bei Bedarf graduell disambiguieren zu können (vgl. Eberle (1997)).

5.2.1 Lexikalische Repräsentationen

In FUDRT sind die semantischen Repräsentationen von Lexemen Funktionen, die schrittweise durch den Kontext ausgewertet werden, wie im folgenden für die Repräsentation des mehrdeutigen Substantivs *Drucker*:

$\underline{\text{drucker}}(x) := l_{x@PROF}$: [x | druck_arbeiter(x)] $\vdash_D$ x@¬ARTEFACT

$\underline{\text{drucker}}(x) := l_{x@EGERAET}$: [x | druck_geraet(x)] $\vdash_D$ x@¬HUMAN)

Danach ist die semantische Repräsentation von *Drucker* eine funktionale Charakterisierung $\underline{\text{drucker}}(x)$ (wobei der funktionale Charakter eines Prädikats PREDICATE durch Unterstreichen, $\underline{\text{PREDICATE}}$, gekennzeichnet wird), die ausgewertet werden kann im Sinne von *druck_arbeiter(x)*, falls (aus dem Kontext) ableitbar ist (per Default), dass das charakterisierte Objekt *x* kein künstliches Objekt (d.h. x@¬ARTEFACT) ist (denn dann muss es ein Mensch sein, der den Beruf *Drucker* hat). Wenn im Gegensatz dazu abgeleitet werden kann, dass *x* kein Mensch sein kann (d.h. x@¬HUMAN), muss es sich, bei Zutreffen der Kennzeichnung also um ein *druck_geraet* handeln.

Die Auswertung entsprechend der Definitionen der funktionalen Charakterisierung findet als *lazy evaluation* statt, sobald die als auslösend gekennzeichnete Information vorliegt (d.h. x@¬ARTEFACT und x@¬ARTEFACT wirken wie eine *freeze*-Bedingung, vgl. Narain (1990)). Auswertungen können auch ohne echtes Erreichen eines solchen auslösenden Wissenszustands in eine Repräsentation aufgenommen werden, und zwar dann, wenn im Rahmen einer (von der umgebenden Kontrollkomponente) erzwungenen disjunktiven Ausdifferenzierung der Repräsentation die einschlägigen Annahmen zum jeweils betrachteten Fall hinzugenommen werden, soweit das jeweils widerspruchsfrei möglich ist, und die Konsequenzen dieser Spezifizierung berechnet und ebenfalls hinzugefügt werden, so wie dies bei *constraint propagation* üblich ist. Im Falle der funktionalen Charakterisierungen sind das dann die Auswertungen, die durch Hinzunahme der *freeze*-Bedingungen begründet werden.

5.2.2 Satzrepräsentationen

Sätze werden in FUDRT als Menge partieller Repräsentationen repräsentiert. Im Unterschied zur UDRT sind partielle Repräsentationen aber nicht notwendigerweise DRSen oder Mengen von DRSen, sondern können auch *DRS-Modifikatoren* sein (also Funktionen, die sich auf DRSen oder DRS-Modifikatoren beziehen), wobei die *Art der Applikation* in Grenzen unterspezifiziert sein kann. Damit ist es möglich, neben Skopusambiguitäten auch *Attachment-* und funktionale Ambiguitäten zu repräsentieren (und eine Reihe weiterer Ambiguitäten, vgl. Eberle (2004)).

(12) veranschaulicht wie die Attachment-Ambiguität in (11) repräsentiert wird:

(11) *Bilder der Kanzlerin beim Außenminister.*

(12) $\underline{bilder}(X)\left\{ \; ngen: \underline{kanzlerin}(y), \; xprep(bei): \underline{außenminister}(z) \; \right\}$

ngen und *xprep* sind (an den zugrundeliegenden syntaktischen Constraints orientierte) unterspezifizierte Beschreibungen der semantischen Rolle, die die entsprechenden DRS-Modifikatoren spielen. *ngen* umfasst die Rollen, die mit Genitiv ausgedrückt werden, sodass die Kanzlerin, *y*, Ursache der Bilder sein kann (*Subjekt/Agens*), oder Inhalt (*Objekt*) etc.; die bei-PP kann sich auf *y* beziehen oder auf *X*, wobei die Rollenbezeichnung, (das *x* in *xprep*), deutlich macht, dass nicht nur die Art der Beziehung (welche Rolle die PP spielt) unterspezifiziert ist, sondern auch der Bezugspunkt als solcher (das kann die Repräsentation des Head-Nomens selber sein oder eine rechts stehende nominale Modifikation in der Repräsention der Nomenprojektion, wobei in (12) dafür nur noch die Repräsentation der Genitiv-Rolle in Betracht kommt). [4]

In *translate* sind bislang nicht alle Repräsentationsmöglichkeiten von FUDRSen kodiert. Insofern sind die Verfahren in den folgenden Abschnitten Spezifikationen von Integrationsmöglichkeiten, geben aber nicht in jedem Fall den implementierten Zustand wieder. [5]

Verfügbar sind aktuell die folgenden Informationstypen bzw. Informationsberechnungsverfahren:

- Semantische Dependenzstruktur
  Abstraktion der syntaktischen Dependenzstruktur entsprechend einer nicht weiter spezifizierten FUDRS (die rekursiv die Prädikat-Argument-Struktur beschreibt).

- Informationsstruktur
  bestehend aus Relationen zwischen den partiellen Repräsentationen zur Fokus-Hintergrundstrukturiering im Zusammenhang mit Fokus-Adverbien.

- Akzessibilitätsstruktur
  Die (partielle) Hierarchie der partiellen Repräsentationen definiert (partielle) Zugänglichkeitsrelationen, die bei der Pronomenauflösung benutzt werden (vgl. Eberle (2003)).

- Verfeinerte Informationsstrukturierung bei Bedarf.

- Skopusauflösung bei Bedarf.

[4] Zu Details der Terminologie und den Repräsentations- und Interpretationsmöglichkeiten vgl. Eberle (2004), zur Repräsentation der Attachment-Ambigutiät Eberle et al. (2008).

[5] Den Zusammenhang zwischen FUDRSen und den verwendeten Kodierungen beschreibt Eberle (2002).

5.3 Transfer

Neben dem LMT-typischen Transfer auf syntaktischen Dependenzstrukturen besitzt *translate* auch eine Komponente für die Übersetzung auf Ebene der verwendeten FUDRS-Kodierungen.

Der dabei benutzte Default-Algorithmus hat folgende Gestalt:

$$\tau(\text{BasicRep}^{\left\{\begin{array}{l}\text{rel}_1\text{: Functor}_1,\\ \vdots\\ \text{rel}_n\text{: Functor}_n\end{array}\right\}}_{\text{AC}}) := \tau_n(\text{BasicRep})^{\left\{\begin{array}{l}\tau_r(\text{rel}_1)\text{:}\tau(\text{Functor}_1),\\ \vdots\\ \tau_r(\text{rel}_n)\text{: }\tau(\text{Functor}_n)\end{array}\right\}}_{\tau_r(\text{AC})}$$

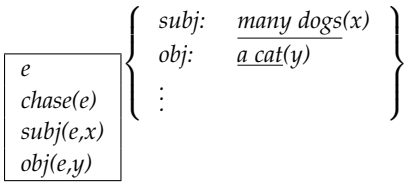
Danach wird eine Struktur, bei der eine Basisrepräsentation (z.B. des Verbs) modifiziert wird, durch eine Reihe von Modifikatoren (z.B. die Repräsentationen der Verbargumente und Adjunkte) in der Weise übersetzt, dass die Übersetzungen der Modifikatoren die Übersetzung der Basisrepräsentation modifizieren, wobei die Art der Modifikation die Übersetzung der Art der ursprünglichen Modifikation ist. Rekursive Transferstrategien dieser Gestalt sind mehrfach vorgeschlagen worden (z.B. Zajac (1989, 1990); Dorna et al. (1994)), zumeist im Zusammenhang mit getypten Featurestrukturen für syntaktisch-funktionale Beschreibungen. *AC* steht für *application constraints* (zur Art und Reihenfolge der Applikationen). Typischerweise werden diese bei der Übersetzung isomorph (modulo Umbenennungen) übernommen, wie im folgenden Beispiel das die Skopusambiguität aus (10) wieder aufnimmt:

(13) *Viele Hunde jagen eine Katze.*

Gegeben die Repräsentation des Satzes wie in (14) erhält man unter Anwendung des Algorithmus entsprechend der Gleichung in (14) die Struktur der Übersetzung:

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

Unter Anwendung der Default-Werte für τ_r und der Default-Spezifikationen im bilingualen Lexikon, ergibt sich daraus die Repräsentation (15):

(15) 

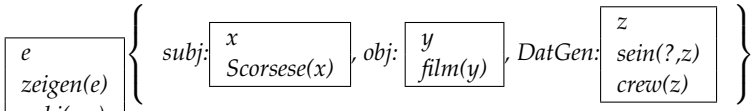
Wenn AC, das hier leer ist, bei der Übersetzung nicht weiter spezifiziert wird, ist die Zielrepräsentation bezüglich der Anwendungsreihenfolge, d.h. hier bzgl. der Skopuslesart, so neutral wie die Ausgangsrepräsentation. D.h. der Default-Transferalgorithmus unterstützt die ambiguitätserhaltende Übersetzung.

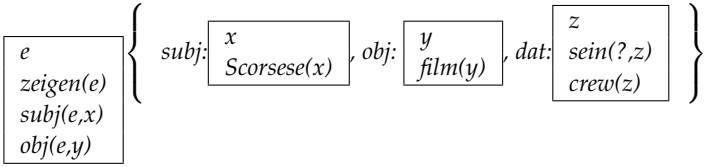
5.4 Partielle Disambiguierung

Beispiele wie (1.b), (2) machen deutlich, dass nicht immer ambigutätserhaltend übersetzt werden kann. (16) wiederholt die *funktionale Ambiguität* der Genitiv-Modifikation des Beispiels (2.b), die bei der Übersetzung ins Englische aufgelöst werden muss (mit Übersetzung als *of*- oder *to*-PP).:

(16) *Scorsese zeigte den Film seiner Crew.*
a. Scorsese showed the film of his crew..
b. Scorsese showed the film to his crew.

Die Übersetzungen (16.a) und (16.b) gründen auf Spezifikationen der Repräsentation (17) der Art (17.a) und (17.b)

(17) 

a. 

b. $\left[\begin{array}{l} e \\ zeigen(e) \\ subj(e,x) \\ obj(e,y) \end{array}\right] \left\{ (1)\ subj: \left[\begin{array}{l} x \\ Scorsese(x) \end{array}\right], (2)\ obj: \left[\begin{array}{l} y \\ film(y) \end{array}\right], (3)\ gen: \left[\begin{array}{l} z \\ sein(?,z) \\ crew(z) \end{array}\right] \right\}$ $\{first(2,3)\}$

Nach der Interpretation (17.a) spielt *seiner Crew* die Rolle eines (freien) Dativs und wird in der Konsequenz mit *to his crew* übersetzt. Um Unterschied dazu spielt *seiner Crew* in (17.b) die Rolle eines Genitivs, der sich auf die Repräsentation von *den Film* bezieht. In der FUDRT-Terminologie wird diese Spezifikation durch den zusätzlichen Applikationsconstraint *first(2,3)* festgehalten, der für die Repräsentation verlangt, dass der Modifikator 3 (also die *Crew*) eine Typerhöhung erfährt und vor Anwendung des Funktors 2 (der *Film*) auf die Verbrepräsentation auf die Repräsentation von 2 anzuwenden ist. Dies ist gleichbedeutend damit, die Repräsentation 3 in die Funktoren des Modifikators 2 aufzunehmen, wie in der folgenden Repräsentation (17.b'), die aus (17.b) folgt und bedeutungsgleich zur konventionellen DRS (17.b'') vereinfacht werden kann:

(17) b'. $\left[\begin{array}{l} e \\ zeigen(e) \\ subj(e,x) \\ obj(e,y) \end{array}\right] \left\{ (1)\ subj: \left[\begin{array}{l} x \\ Scorsese(x) \end{array}\right], (2)\ obj: \left[\begin{array}{l} y \\ film(y) \end{array}\right] \left\{ gen: \left[\begin{array}{l} z \\ sein(?,z) \\ crew(z) \end{array}\right] \right\} \right\}$

b''.

e x y z	
Scorsese(x)	zeigen(e)
film(y)	subj(e,x)
sein(?,z)	obj(e,y)
crew(z)	gen(y,z)

Korrespondierend erhält man für (17a) die DRS (17.a'):

(17) a'.

e x y z	
Scorsese(x)	zeigen(e)
film(y)	subj(e,x)
sein(?,z)	obj(e,y)
crew(z)	dat(e,z)

Wodurch werden solche Spezifikationen ausgelöst? Einerseits durch die Notwendigkeit aus Übersetzungsmöglichkeiten auswählen zu müssen, die bezogen auf den jeweiligen Ausdruck nicht bedeutungserhaltend, sondern bedeutungseinschränkend sind. Die Übersetzung des zwischen Genitiv und Dativ unterspezifizierten Kasusmorphems bzw.

der entsprechenden unterspezifizierten Rolle *DatGen* erfolgt mit *to* oder mit *of*, abhängig davon, ob *DatGen* als *dat* interpretiert wird oder als *gen*. Dieser Zusammenhang wird im Übersetzungssystem repräsentiert wie die lexikalischen Auswertungsregeln oben, mit Annotation von Konditionen.

Allerdings sind diese Regeln in diesem Fall nicht Teil eines Lexikoneintrags, sondern Teil der Definition von τ_r in der Datenbasis des Übersetzungsmoduls.

$$\begin{array}{lcll} \tau_r(DatGen) : \tau(Val) & := & pobj(to) : \tau(Val) & \text{if C} \vdash_D \text{DatGen=dat} \\ \tau_r(DatGen) : \tau(Val) & := & pobj(of) : \tau(Val) & \text{if C} \vdash_D \text{DatGen=gen} \end{array}$$

Annotierte Konditionen wirken in zwei Weisen. Einerseits als auslösende Faktoren innerhalb der lazy evaluation, d.h. wird die entsprechende Bedingung aus dem Kontext abgeleitet, findet die assoziierte Repräsentationsverfeinerung automatisch statt – und eine eventuell sich ergebende spezifische(re) Übersetzung ist die Folge. Ist es umgekehrt notwendig, bei der Übersetzung eine spezifischere Charakterisierung zu wählen (*to* oder *of* beispielsweise), ist es zur Erhaltung der Konsistenz notwendig, die inhaltlichen Konsequenzen dieser Spezifizierung zu notieren. Wie beim abduktiven Schließen wird dabei eine der möglichen inhaltlichen Situationen, aus denen eine entsprechende Spezifikation folgt, als Begründung der Spezifikation herangezogen, sprich eine entsprechende annotierte Kondition als faktisch angenommen.[6] Der folgende Beispieltext illustriert das Ineinandergreifen dieser Ableitungen im Sinne der Propagierung von Constraints:

(18) *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.

Um das Pronomen *er* richtig übersetzen zu können, muss man wissen, ob *er* sich auf einen Menschen bezieht oder nicht. Der als Antezedent bestimmte *Drucker* kann sich auf einen Menschen beziehen oder nicht; ist das der Fall, ist anzunehmen, dass aufgrund der Selektionsbeschränkungen von *einstellen* dieses Verb als *to hire* übersetzt werden muss, sonst sicher nicht, sondern vermutlich mit *to adjust*. Im ersten Fall ist aber ein Fortgang des Textes in der Art b) nicht zulässig, weil *defekt* sich nicht auf Menschen bezieht. Bei einem Fortgang des Textes in der Art a) ist es gerade umgekehrt: Dann darf *Drucker* und *einstellen* sich gerade nicht auf einen Menschen beziehen.

[6] Gibt es mehrere unterschiedliche und sich widersprechende Konstellationen, setzt dieses Vorgehen natürlich eine *truth maintenance*-Konzeption mit *belief revision* voraus (vgl. Doyle (1979)). Dies ist in *translate* nicht implementiert und wird es auch in Zukunft aus Kostengründen nur zu einem Teil sein können.

Solche Zusammenhänge werden in *translate* typischerweise abgeleitet (so sie ableitbar sind) aus Informationen im Lexikon zum semantischen Typ und Selektionsbeschränkungen von Lesarten, notiert im Stil der oben skizzierten Auswertungsregeln, aus sortalen Zusammenhängen in der Hierarchie der semantischen Typen, und durch Regularien in der Diskurskomponente, die u.a. die Pronomenauflösung durchführt und die sortalen Konsequenzen auf die miteinander identifizierten Diskursreferenten (DRFs) propagiert.

Für (18) können wir von folgenden Angaben ausgehen:

- im Lexikon
  - Eintrag *defekt*

    • defekt(x)
    TYPE: x @ MACHINE
    τ: deficient
  - Eintrag *kündigen*

    • kündigen [subj:x @ HUMAN,obj: y]
    C: $\vdash_D$ y @ CONTRACT
    τ: terminate

    C: $\vdash_D$ empty(obj)
    τ: hand in one's notice

    ...
  - Eintrag *einstellen*

    • einstellen [subj(n),obj(n): y]
    C: $\vdash_D$ y @ HUMAN
    τ: hire

    C: $\vdash_D$ y @ ARTEFACT
    τ: adjust

    ...
- in der Diskurskomponente

  • $\vdash_D$ antecedes(DRF1,DRF2) $\Rightarrow \vdash_D$ (TYPE(DRF1) $\leftrightarrow$TYPE(DRF2))

Demnach führen die Lexikoneinträge für *defekt* und *kündigen*, unabhängig von speziellen Auswertungen, sortale Restriktionen für DRFs ein (für das Argument von *defekt* und das

Subjekt von *kündigen*). Bei den Einträgen für *einstellen* sind die Restriktionen gebunden an abzuleitende (Default)-Interpretationen und dazu passende Übersetzungen, wobei die Aufnahme der entsprechenden Spezifikationen in der oben beschriebenen Weise entweder als im Kontext fundierte Ableitung oder als widerspruchsfrei hinzunehmbare Disambiguierung mit möglicher Begründung geschieht. Ausgebeutet werden die eingeführten sortalen Restriktionen in der Diskurskomponente durch eine (schwache) Version der skizzierten Leibniz'schen Identitätsregel, sodass sortale Einschränkungen über Referenzketten propagiert werden können.

Zur Vermeidung kostenintensiver semantischer Ableitungen ist semantische Inferenz in *translate* auf solche sortalen Spezifikationen und die oben beschriebenen strukturellen Disambiguierungsmöglichkeiten beschränkt. Kontinuierliche Evaluation zeigt, dass zwischen den Alternativen 'Repräsentation ohne semantische Auswertung' und 'Repräsentation mit tiefer semantischer Auswertung' dieser Kompromiss ein sehr gutes Kosten-Nutzen-Verhältnis für Transfer-Architekturen darstellt.

Immer noch teuer ist aber bei einer solchen Transfer-Architektur mit flacher semantischer Repräsentation und Auswertung, die Voraussetzungen im Lexikon zu schaffen, d.h. genügend detaillierte semantische Klassifikationen und strukturell-semantische Übersetzungsbedingungen in möglichst breiter und gleichmäßig ausgearbeiteter Abdeckung zu formulieren. Trotzdem werden aufgrund der Beschränkung auf die Verwendung sortaler Informationen, und damit auf einen extrem kleinen Ausschnitt des semantisch-pragmatischen Weltwissens, sehr viele Übersetzungsentscheidungen letztlich inhaltlich unmotiviert oder wenig begründet bleiben müssen und damit fehleranfällig. Teuer ist dabei auch, dass solche Fehler aufgrund der notwendigen Konsistenzerhaltung bei der Textinterpretation zu Folgefehlern bei der Interpretation anderer Wörter und Strukturen führen. (Wenn in (18) bei der Kodierung von *defekt* oder *kündigen* ein Fehler gemacht wird und aufgrund dessen darauf geschlossen wird, dass das Pronomen keinen Menschen bezeichnet, folgen bei gleicher Pronomenresolution eine falsche Interpretation von *Drucker* (und eine eventuell falsche Übersetzung, z.B. ins Französische mit *imprimeur* statt *imprimante*) und *einstellen* (im Sinne von *hire/engager* statt *adjust/ajuster*).

Von ganz entscheidender strategischer Bedeutung für die Maschinelle Übersetzung ist deshalb, wie einerseits die Lexika mit semantischer Information kostengünstiger unter Verwendung automatischer Verfahren aufgebaut bzw. erweitert werden können und wie andererseits Fehlentscheidungen bei der inhaltlich nicht-begründbaren Auswahl aus Übersetzungsalternativen minimiert werden können.

5.5 Integration statistisch gewonnener Information in ein RBMT-System mit flachem semantischem Transfer

Ein System mit Transfer auf der Basis flacher unterspezifizierter semantischer Repräsentationen ist besonders geeignet für die Integration statistisch gewonnener Übersetzungsinformation:

Die Repräsentationen sind in einer Weise abstrakt, dass die Ausdifferenzierung in syntaktische Einzelfälle optimal minimiert wird und damit auch das Sparse-Data-Problem bezogen auf strukturelle Phänomene.

Die Art und Weise der Repräsentation von Wörtern in Abstraktion morphologischer Eigenschaften als möglichst flach interpretierte semantische Prädikate optimiert in ähnlicher Weise das Sparse-Data-Problem für lexikalische Phänomene.

Die Möglichkeit, bei Bedarf die Analysetiefe zu variieren und die Repräsentationen semantisch zu verfeinern und Auswertungen an Bedingungen zu knüpfen, schafft eine wohldefinierte Schnittstelle für die Integration disambiguierender Information und passt den Informationsbedarf und die Differenzierungsnotwendigkeiten des Systems optimal an die Datenlage in Korpora an. (Die Beschreibung der Mehrdeutigkeiten ist so differenziert, wie dies für die Beschreibung des Übersetzungsverhaltens im betrachteten Korpus notwendig ist).

Die wesentlichen Probleme bei Analyse und Übersetzung in solchen RBMT-Systemen sind: die Auswahl bei lexikalischen und strukturellen Mehrdeutigkeiten in der Analyse, die Bewertung von Transfer-Äquivalenten und das Lernen von relevanten Auswahlbedingungen bei der Generierung aus flachen semantischen Strukturen, vor allem die Auswahl aus Wortstellungsvarianten der Zielgrammatik.

5.5.1 Disambiguierung von Lexemen

• Statistische *word sense disambiguation*

Die Einschränkung von semantischer Auswertung aus dem letzten Abschnitt bedeutet, eine Unterscheidung zu machen zwischen Fällen, die aufgrund sortaler Eigenschaften über Selektionsbeschränkungen und Referenzketten entschieden werden können und solchen, wo dies nicht der Fall ist. Bei letzteren, die also für eine inhaltlich abgeleitete Entscheidung komplex(er)es Regel- und Hintergrundwissen voraussetzen, kann in dem beschriebenen Ansatz nur Wissen über statistisch auffällige semantische Zusammenordnungen in spezifischen Texten oder Korpora benutzt werden.

Ein sehr bekanntes Beispiel für einen solchen komplexen Zusammenhang ist Bar-Hillels *pen*-Beispiel:

(19) *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)*

Das Wort *pen* ist hier in der spezifischen Bedeutung *playpen/Laufstall*, nicht als *Schreibgerät* oder das allgemeinere *Einzäunung* zu verstehen. Dieser Zusammenhang ergibt sich für den Menschen aus Weltwissen zur Betreuung von kleinen Kindern, für die Maschine in der Regel gar nicht, weil es nicht möglich ist, für alle solchen Übersetzungsprobleme in allen Kontexten das nötige Weltwissen bereitzustellen. Das ist Bar-Hillels bekanntes Argument gegen die Möglichkeit allgemein verfügbarer Qualitätsübersetzung.

Semantische RBMT der beschriebenen Art erlaubt aber, die Wörter in Texten mithilfe seiner Analysekomponenenten recht detailliert semantisch zu klassifizieren und damit die üblichen statistischen Verfahren im Rahmen von *Word sense disambiguation* (WSD) mit sehr viel Vorwissen zu versehen, sodass entsprechende Ergebnisse optimiert werden können (vgl. Yarowsky (2000)). Umgekehrt können Texte mit den gleichen Mitteln des Systems detaillierter klassifiziert und die entsprechend abgeleiteten Klassen als Sachgebiete den dafür signifikanten Wortbedeutungen zugeordnet werden.

• Lernen von semantischen Selektionsbedingungen

Entscheidungen, die im Rahmen der Einschränkungen prinzipiell semantisch-logisch erfolgen können, setzen zuallererst detaillierte semantische Klassifizierungen der Lexeme und detaillierte semantische Selektionsbeschränkungen bei den Argumentrahmen voraus. Dafür bietet sich ein Bootstrapping-Ansatz mit den Analysekomponenten des RBMT-Systems an: Die Sätze eines Textes werden analysiert mit liberalen semantischen Vorgaben zu den Argumentrahmen. Aus den Ergebnissen und der schon vorliegenden semantischen Klassifizierung des lexikalischen Materials lassen sich statistische Selektionspräferenzen ermitteln, die dann wieder benutzt werden können, um das lexikalische Material (feiner) zu klassifizieren. Ähnliche Verfahren sind vorgeschlagen worden (u.a. im Zusammenhang mit WordNet-Information, in Schulte im Walde (2008); Schulte im Walde et al. (2008)), für eine LMT-Architektur in Bernth and McCord (2003)).

• Propagieren von semantischen Effekten entlang von Referenzketten

Es gibt mittlerweile viele Vorschläge für statistisch berechnete Pronomenauflösung (vgl. Mitkov (2002)). Das in LMT und *translate* verwendete Verfahren verwendet syntaktische Filter und aus strukturellen Phänomenen abgeleitete Präferenzen (vgl. Lappin and McCord (1990); Lappin and Leass (1994)). Es ist das anerkannte Standardverfahren Regel-basierter Pronomenauflösung. Für Versionen von *translate* wurde es um Diskursinformation im Sinne der DRT erweitert (vgl. Eberle (2003)). Es bietet sich an, solchermaßen abgeleitete Information über Unverträglichkeiten und Präferenzen in Form von Featurefunktionen in ein Maximum-Entropie-Modell der in Abschnitt 4.1 beschriebenen Art einzubauen und die Ausdifferenzierung der Auflösungspräferenzen an Korpora zu trainieren. (vgl. dazu Schiehlen (2004)).

5.5.2 Disambiguierung von Strukturen

In Hindle and Rooth (1993) ist früh vorgeschlagen worden, wie die Disambiguierung spezifischer struktureller Mehrdeutigkeiten, in dem Fall die Entscheidungen bei PP-Attachment-Ambiguität, trainiert werden.

Wie beschrieben ist Analyse auf der Ebene von FUDRSen besonders geeignet, solche Methoden auch auf andere strukturelle Mehrdeutigkeiten anzuwenden, weil der Abstraktionsgrad von vorneherein hoch ist und erlaubt, nicht interessierende formale Details auszublenden und weil es möglich ist, die Kassifizierungs- und Detaillierungsmöglichkeiten auszunützen, um signifikante Zusammenhänge festzustellen und auf der angemessenen Ebene zu repräsentieren (z.B. auf der Ebene der allgemeinen oder der detaillierten semantischen Klassifizierung oder der Wortebene im Zusammenhang mit Forderungen an Elemente von Konstruktionen und Kollokationen). Auch hierbei bietet sich ein Bootstrapping-Ansatz an, der das analytische Vorwissen des Systems, einschließlich des lexikalisch-semantischen Wissens und der vordefinierten Präferenzen, für das Training nutzt, um es durch die statistische Auswertung zu verbessern. Letztlich geht es dabei um Verfahren, den deklarativen Kern einer Grammatik für unterspezifizierte Analysen um statistische Entscheidungsregeln zu vervollständigen (vgl. Eberle and Rapp (2008)).

5.5.3 Lernen von Übersetzungsbeziehungen

Als Folge der propagierten Einschränkung bei der Verfügbarkeit semantischer Informationen gibt es auch beim Problem der Auswahl aus Übersetzungsmöglichkeiten Fälle, die sinnvoll auf der Basis strukturell-semantischen Wissens zum Satzkontext entschieden werden können und solchen, wo dies nicht der Fall ist. Letztere können, da sie an korrespondierende Analyseentscheidungen gebunden sind, wie dort im Rahmen einer verfeinerten Sachgebietserkennung abgehandelt werden. Interessanter ist an dieser Stelle der andere Fall. Wenn es gelingt, präzise operationalisierbare Bedingungen für spezifische Übersetzungen automatisch aus dem Satzkontext abzuleiten, kann damit nicht nur die Maschinelle Übersetzung signifikant verbessert werden, auch den menschlichem Nutzer entsprechender Lexika sind damit konkrete Handlungsanweisungen für die Übersetzung von Wörtern im Kontext an die Hand gegeben.

Der Vorschlag für FUDRS-Übersetzung sieht das folgende Verfahren vor, das wieder Bootstrapping benutzt, des bilingualen Lexikons in diesem Fall: Bilinguale Korpora werden mit den Mitteln des RBMT-Systems aligniert, flache Analysen von Quell- und Zielsätzen werden berechnet und Quell- und Zielstrukturen nach den Maßgaben des vorliegenden Lexikons möglichst gut aufeinander bezogen. Exemplifiziert ein Satzpaar nach dieser Aufbereitung eine neue Übersetzungsmöglichkeit für ein Wort (oder einen Mehrwortausdruck), dann wird aus dem Quellsatz ein Kontext in Begriffen der verwendeten Repräsentationssprache abgeleitet, der als signifikant vermutet wird für die

Auswahl der im Zielsatz gefundenen Übersetzung. Diese Verwendungshypothese wird anschließend gegen das Korpus und die zuvor schon verfügbaren Übersetzungsmöglichkeiten getestet. Dabei wird schrittweise die Spezifizität der getesteten Bedingungen zurückgenommen, um Bedingungen mit maximaler Abdeckung von Fällen bei gleichbleibender Verlässlichkeit der Auswahl zu bestimmen.

Das skizzierte Verfahren ist für eine Anwendung in *translate* in der Testphase (vgl. Eberle and Rapp (2008)). (20) zeigt ein für *einstellen* gefundenes Satzpaar aus dem Europarl-Korpus (vgl. Koehn (2005)):

(20) *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)*

Die erkannte Übersetzung *establish* für *einstellen* ist neu. [7]

Einbeziehen der prädikativischen Beschreibungen aller Argumente des Verbs und der Adjunkte ergibt eine erste Hypothese, (21):

(21)

- 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* [∅,obj(n):$\tau(l_4)$]
& $\tau(d\text{-}l_1)=\tau(l_0)\text{-}d(obj)\text{-}d(nadj)$
& $\tau(d\text{-}l_3)=\tau(l_0)\text{-}d(obj)\text{-}d(prep(in))$
& $\tau(d\text{-}l_2)=\tau(l_0)\text{-}d(obj)\text{-}d(prep(in))\text{-}d(prep(between))$

Entsprechend der Analyse geht der Vorschlag davon aus, dass *einstellen* mit *establish* übersetzt wird, falls die Subjektsrolle die unterspezifizierte Beschreibung fraktion erfüllt, die Objektsrolle feindseligkeit und es weitere Einschränkungen durch eine adverbiale Kennzeichnung vorübergehend und eine (vermutlich instrumental zu lesende) PP mit Argument der Art waffenstillstand gibt. Falls diese Bedingungen in einem Satz greifen, wird mit *establish* übersetzt, wobei eine Restrukturierung der Argumente entsprechend der Pfadangaben stattfindet (die hier im Stile der üblichen LFG-Transfer-Gleichungen angegeben sind).

Verallgemeinerungen, die in der Folge zu testen sind, entstehen durch Weglassen von Rollen und Kennzeichnungen aus Adjunkten bzw. durch Verallgemeinerungen entlang der systemimmanenten Hierarchie der semantischen Typen.

[7] Eine Erweiterung des Verfahrens liegt auf der Hand: es kann benutzt werden, um händisch notierte Übersetzungsbedingungen am Korpus auf ihre Signifikanz zu überprüfen.

Eine mögliche Verallgemeinerung ist etwa die folgende (für *jmd stellt einen ZUSTAND durch ein EREIGNIS ein*):

(22)

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

c: *d(subj):*l_2*:jmd*
& d(obj):l_3*:s @ STATE & d(prep(durch)):*l_4*:e @ EVENT*

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

5.5.4 Statistisch gewonnene Wortstellungsregeln

Je freier die Wortstellung der Zielsprache ist, umso schwieriger ist es in der Regel, kontextuell passende Wortstellungen zu generieren. (Es gibt auch andere Probleme bei der Generierung aus flachen semantischen Strukturen, aber das Wortstellungsproblem ist vermutlich dasjenige, für das Integration statistischen Wissens am meisten Erfolg verspricht). Bei der Übersetzung ins Deutsche von Sätzen wie in (23) hängt es neben den Referentialisierungseigenschaften der Argumente und ihrem 'Gewicht' (d.h. ihrer Länge und Informationsdichte) auch von der pragmatischen Informationsstruktur des Satzes und seines Kontexts ab, welche Anordnung die natürlichere ist.

(23) *Poirot remet la lettre à la femme.*
a. Poirot übergibt den Brief der Frau.
b. Poirot übergibt der Frau den Brief.

Wie bei einigen Aufgaben der Abschnitte zuvor kann das Wortstellungsproblem in solchen Fällen im Rahmen des vorgeschlagenen Ansatzes aus prinzipiellen Gründen nicht zureichend behandelt werden, weil wesentliche Information zur pragmatischen Informationsstrukturierung nicht zur Verfügung stehen kann.

Es gibt ermutigende Untersuchungen, formale und semantisch-klassenbezogene Kriterien für die Wortstellung aus Korpora zu lernen, die recht weit tragen (vgl. Cahill et al. (2007)). Auch hier ist anzunehmen, dass die Ergebnisse umso verlässlicher sind, je größer das linguistisch-klassifikatorische Vorwissen ist, das in die statistische Untersuchung eingeht.

6 Ausblick

Aufgrund der herausragenden Rolle, die der Mehrdeutigkeit in natürlichen Sprachen zukommt, ist die richtige Auswahl aus Interpretationsalternativen und Übersetzungsmöglichkeiten das entscheidende Problem der Maschinellen Übersetzung, neben dem

Problem der schieren Zahl von Wörtern und Übersetzungsrelationen. Regelbasierte Analyse- und Übersetzungssysteme versprechen sinnvolle Abstraktionen, um die Datenflut aus großen Korpora zu kanalisieren und in wesentliche Fälle zusammenzufassen. Tiefe Analyse mit solchen Systemen ist in vielerlei Hinsicht teuer, sehr flache Analyse dagegen wenig ergiebig auf dem Weg zu genügend abstrakten Repräsentationen. Flache unterspezifizierte semantische Repräsentationen in der Art von FUDRSen scheinen, auch in vielerlei Hinsicht, ein guter, wenn nicht bester Kompromiss in diesem Zusammenhang. Systeme mit entsprechender Analyse und Übersetzung können im Vergleich kostengünstig erstellt werden, erlauben genügend gute Abstraktion von Korpus-Daten und geben in natürlicher Weise Schnittstellen vor, über die mit kombiniert analytisch-statistischen Methoden gewonnene Information aus Korpora aufbereitet und integriert werden kann. Als Beispiele sind genannt worden: Beiträge zur Lösung der Entscheidungsprobleme im lexikalischen und strukturellen Bereich der Analyse, bei der Äquivalentwahl und bei der Generierung von Wortstellungsvarianten und Beiträge zum semi-automatischen Auf- und Ausbau der bilingualen Lexika. Durch die Zunahme der elektronischen Verfügbarkeit ein- und mehrsprachiger Korpora und den spürbar steigenden Bedarf an Übersetzungen in der globalisierten Welt nimmt die Bedeutung solcher integrierender Verfahren in der Zukunft ganz zweifellos weiter zu. Auch weil die Unausgewogenheit von Korpora und mangelnde Verfügbarkeit für viele Sprachpaare in der Zukunft ebenfalls, so ist zu vermuten, ein notorisches Problem sein wird, trotz der generellen Zunahme von Übersetzungsdaten, werden Systeme, die in umgekehrtem Zugang auf dem statistischen Modell beruhen und versuchen, dessen Verhalten durch linguistische Features zu optimieren, auf mittlere Sicht, unserer Einschätzung nach, nicht die Oberhand behalten. Allerdings wird der momentan noch mit großem Interesse verfolgte Gegensatz zwischen RBMT, SMT, EBMT und all den anderen Architekturen sich innerhalb der nächsten Jahre verwischen,so ist weiter zu vermuten, und einer unprätentiösen und vorurteilsfreien Suche nach der kostengünstigsten Architektur Platz machen, die sich analytischer und statistischer Methoden, Korpusdaten und Grammatiken bedient und solche zusammenstellt, ohne darauf zu achten, was als definierende Basis und Etikettierung des Ansatzes betrachtet wird.

Literatur

Barwise, J. and Perry, J. (1983). *Situations and Attitudes*. MIT Press, Cambridge, Massachusetts.

Bernth, A. and McCord, M. (2003). A hybrid approach to deriving selectionalpreferences. In *Proceedings of MT Summit IX*, New Orleans, USA.

Brown, P. F., Cocke, J., Pietra, S. D., Pietra, V. D., F. Jelinek, R. M., and Roossin, P. (1990). A statistical approach to machine translation. *Computational Linguistics*, 16(2).

Brown, P. F., Pietra, S. A. D., Pietra, V. J. D., Lafferty, J. D., and Mercer, R. L. (1992). Analysis, statistical transfer, and synthesis in machine translation. In *4th International Conference on Theoretical and Methodological Issues in Machine Translation*, Montreal.

Cahill, A., Forst, M., and Rohrer, C. (2007). Stochastic realisation ranking for a free word order language. In Busemann, S., editor, *Proceedings of the European Workshop on Natural Language Generation (ENLG-07)*, Dagstuhl, Germany.

Carbonell, J., Mitamura, T., and Nyberg, E. (1992). The kant perspective: A critique of pure transfer (and pure interlingua, pure statistics, ... In *Proceeedings of the International Conference on Theoretical and Methodological Issues in Machine Translation (TMI'92)*, pages 225–235, Montréal, Canada.

Chiang, D. (2006). A hierarchical phrase-based model for statistical machine translation. In *Proceedings HLT-NAACL-2006*, New York.

Dorna, M., Eberle, K., Emele, M., and Rupp, C. (1994). Semantik-orientierter rekursiver Transfer in HPSG am Beispiel des Referenzdialogs. Verbmobil-Report 39, IMS, Universität Stuttgart.

Dorr, B. (1993). *Machine Translation: A View from the Lexicon*. MIT Press, Cambridge, Massachussetts.

Dorr, B. (1994). Machine translation divergences: A formal description and proposed solution. *Computational Linguistics Journal*, 20(4):597–633.

Doyle, J. (1979). A truth maintenance system. 12:231–272.

Drouin, N. (1989). Le système logos. In A, A. A., editor, *Traduction assistée par ordinateur: perspectives technologiques, industrielles et économiques envisageables à l'horizon 1990*. editions Daicadif, Paris.

Durrell, M. (2000). *Using German Synonyms*. Cambridge University Press, Cambridge.

Eberle, K. (1997). Flat underspecified representation and its meaning for a fragment of German. Arbeitspapiere des Sonderforschungsbereichs 340 *Sprachtheoretische Grundlagen für die Computerlinguistik* 120, Universität Stuttgart, Stuttgart.

Eberle, K. (2002). Tense and aspect information in a FUDR-based German French Machine Translation System. In Kamp, H. and Reyle, U., editors, *How we say WHEN it happens. Contributions to the theory of temporal reference in natural language*, pages 97–148. Niemeyer, Tübingen. Ling. Arbeiten, Band 455.

Eberle, K. (2003). Anaphernresolution in flach analysierten Texten für Recherche und Übersetzung. In Seewald-Heeg, U., editor, *GLDV-Jahrestagung 2003*. Gardez!, Köthen.

Eberle, K. (2004). Flat underspecified representation and its meaning for a fragment of German. Habilitationsschrift, Universität Stuttgart, Stuttgart.

Eberle, K., Heid, U., Kountz, M., and Eckart, K. (2008). A tool for corpus analysis using partial disambiguation and bootstrapping of the lexicon. In Storrer, A., Geyken, A., Siebert, A., and Würzner, K.-M., editors, *Text Resources and Lexical Knowledge: Selected Papers from the 9th Conference on Natural Language Processing KONVENS 2008*. De Gruyter, Berlin.

Eberle, K. and Rapp, R. (2008). Rapid construction of explicative dictionaries using hybrid machine translation. In Storrer, A., Geyken, A., Siebert, A., and Würzner, K.-M., editors, *Text Resources and Lexical Knowledge: Selected Papers from the 9th Conference on Natural Language Processing KONVENS 2008*. De Gruyter, Berlin.

Eisele, A., Federmann, C., Uszkoreit, H., Saint-Amand, H., Kay, M., Jellinghaus, M., Hunsicker, S., Herrmann, T., and Chen, Y. (2008). Hybrid machine translation architectures within and beyond the euromatrix project. In Hutchins, J. and v.Hahn, W., editors, *12th annual conference of the European Association for Machine Translation (EAMT)*, pages 27–34, Hamburg, Germany.

Emele, M. C., Dorna, M., Lüdeling, A., Zinsmeister, H., and Rohrer, C. (2000). Semantic-based transfer. In Wahlster, W., editor, *Verbmobil: Foundations of Speech-to-Speech Translation*, pages 359–376. Springer, Berlin, Heidelberg, New York.

Hindle, D. and Rooth, M. (1993). Structural ambiguity and lexical relations. *Computational Linguistics*, 1(19):103–120.

Hutchins, W. J. (1995). Machine translation: A brief history. In Koerner, E. and Asher, R., editors, *Concise history of the language sciences: from the Sumerians to the cognitivists*, pages 431–445. Pergamon Press, Oxford.

Hutchins, W. J. and Somers, H., editors (1992). *An Introduction to Machine Translation*. Academic Press, London.

Jackendoff, R. (1983). *Semantics and Cognition*. MIT Press, Cambridge, Massachusetts.

Jackendoff, R. (1990). *Semantic Structures*. MIT Press, Cambridge, Massachusetts.

Kameyama, M., Ochitani, R., and Peters, S. (1991). Resolving translation mismatches with information flow. In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, Berkeley.

Kamp, H. and Reyle, U. (1993). *From Discourse to Logic*. Kluwer Academic Publishers, Dordrecht.

Kaplan, R. and Bresnan, J. (1982). Lexical functional grammar: A formal system for grammatical representation. In Bresnan, J., editor, *The Mental Representation of Grammatical Relations*. MIT Press.

Kaplan, R., Netter, K., Wedekind, J., and Zaenen, A. (1989). Translation by structural correspondences. In *Proceedings of E-ACL*, Manchester.

Kay, M., Gawron, J. M., and Norwig, P. (1994). *VERBMOBIL: A Translation System for Face-to-Face Dialog*. CSLI, Stanford.

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. In *Proceedings of MT Summit X.*, Phuket, Thailand.

Lappin, S. and Leass, H. (1994). An algorithm for pronominal anaphora resolution. *Computational Linguistics*, 20(4):535–561.

Lappin, S. and McCord, M. (1990). Anaphora resolution in slot grammar. *Computational Linguistics*, 16.

Maruyama, H. and Watanabe, H. (1992). Tree cover search algorithm for example-based translation. In *Proceedings of the International Conference on Theoretical and Methodological Issues in Machine Translation (TMI'92)*, pages 173–184, Montréal, Canada.

McCord, M. (1989). Design of LMT. *Computational Linguistics*, 15.

McCord, M. (1991). The slot grammar system. In Wedekind, J. and Rohrer, C., editors, *Unification in Grammar*. MIT-Press.

Mitkov, R. (2002). Automatic anaphora resolution: Limits, impediments, and ways forward. In *PorTAL*, pages 3–4.

Narain, S. (1990). Lazy evaluation in logic programming. In *Proceedings of the International Conference on Computer Languages*, pages 218–227.

Nirenburg, S., Beale, S., Mahesh, K., Onyshkevych, B., Raskin, V., Viegas, E., Wilks, Y., and Zajac, R. (1996). Lexicons in the mikrokosmos project. In *Proceedings of the Society for Artificial Intelligence and Simulated Behavior Workshop on Multilinguality in the Lexicon*, Brighton, U.K.

Och, F. J. and Ney, H. (2002). Discriminative training and maximum entropy models for statistical machine translation. In *Proceedings of the Annual Meeting of the ACL*, pages 295–302, Philadelphia, PA.

Onyshkevych, B. and Nirenburg, S. (1995). A lexicon for knowledge-based MT. *Machine Translation*, 10(1-2).

Quirk, C., Menezes, A., and Cherry, C. (2006). Dependency treelet translation; syntactically informed phrasal smt. In *Proceedings HLT-NAACL-2006*, New York.

Reyle, U. (1993). Dealing with ambiguities by underspecification: Construction, representation, and deduction. *Journal of Semantics*, 10(2):123–179.

Sadler, L. and Thompson, H. S. (1991). Structural non-correspondence in translation. In *Proceedings of E-ACL*, Berlin.

Schäler, R. (1996). Machine translation, translation memories and the phrasal lexicon: the localisation perspective. In *Proceedings of EAMT*, Vienna, Austria.

Schiehlen, M. (2004). Optimizing algorithms for pronoun resolution. In *Proceedings of the 20th International Conference on Computational Linguistics*. University of Geneva.

Schulte im Walde, S. (2008). The induction of verb frames and verb classes from corpora. In Lüdeling, A. and Kytö, M., editors, *Corpus Linguistics. An International Handbook*. Mouton de Gruyter,.

Schulte im Walde, S., Hying, C., Scheible, C., and Schmid, H. (2008). Combining em training and the mdl principle for an automatic verb classification incorporating selectional preferences. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics*, Columbus.

Stoll, C. (1986). The systran system. In IAI, editor, *Proceedings First International Conference on State of the Art in Machine Translation*, Saarbrücken.

Sumita, E., Iida, H., and Kohyama, H. (1990). Translating with examples: A new approach to machine translation. In *Proceedings of the Third International Conference on Theoretical and Methodological Issues in Machine Translation (TMI'90)*, pages 203–212, Austin, Texas.

Trabulsi, S. (1989). Le système systran. In A, A. A., editor, *Traduction assistée par ordinateur: perspectives technologiques, industrielles et économiques envisageables à l'horizon 1990*. editions Daicadif, Paris.

Trujillo, A. (1992). *Translation Engines: Techniques for Machine Translation*. Springer, London.

Vauquois, B. (1975). *La Traduction Automatique á Grenoble*. Dunod, Paris.

Vogel, S., Och, F. J., Tillmann, C., Nießen, S., Sawaf, H., and Ney, H. (2000). Statistical methods for machine translation. In Wahlster, W., editor, *Verbmobil: Foundations of Speech-to-Speech Translation*, pages 377–393. Springer, Berlin, Heidelberg, New York.

Wahlster, W., editor (2000). *Verbmobil: Foundations of Speech-to-Speech Translation*. Springer, Berlin, Heidelberg, New York.

Weaver, W. (2003). Translation. In Nirenburg, S., Somers, H., and Wilks, Y., editors, *Readings in Machine Translation*, pages 363–394. MIT Press, Cambridge Massachussetts. Reprint.

Yarowsky, D. (2000). Word sense disambiguation. In Dale, R., Moisl, H., and Somers, H., editors, *Handbook of Natural Language Processing*, pages 629–654. Marcel Dekker, New York.

Zajac, R. (1989). A transfer model using a typed feature structure rewriting system with inheritance. In *Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics*, pages 1–6, Vancouver.

Zajac, R. (1990). A relational approach to translation. In *3td International Conference on Theoretical and Methodological Issues in Machine Translation*.