Volume 24 – Number 3 – 2009 – ISSN 0175-1336

JLCL

Journal for Language Technology
and Computational Linguistics

Maschinelle Übersetzung – von der Theorie zur Anwendung

Machine Translation – Theory and Applications

Herausgegeben von / *Edited by*
Uta Seewald-Heeg und Daniel Stein

GSCL Gesellschaft für Sprachtechnologie und Computerlinguistik www.gscl.info

Maschinelle Übersetzung – von der Theorie zur Anwendung

Machine Translation – Theory and Applications

JLCL Impressum

JLCL	Journal for Language Technolgy and Computational Linguistics
ISSN 0175-1336	Offizielles Organ der Gesellschaft für Sprachtechnologie und Computerlinguistik / *German Society for Language Technology and Computational Linguistics*
Herausgeber	Gesellschaft für Linguistische Datenverarbeitung e. V. (GLDV) Prof. Dr. Alexander Mehler, Universität Bielefeld, *alexander.mehler@uni-bielefeld.de* Prof. Dr. Christian Wolff, Universität Regensburg *christian.wolff@sprachlit.uni-regensburg.de*
Band 24 – 2009 – Heft 3	Maschinelle Übersetzung - von der Theorie zur Anwendung
Herausgeber	Prof. Dr. Uta Seewald-Heeg, Hochschule Anhalt, Köthen Daniel Stein, Ludwig-Maximilians-Universität München
Anschrift der Redaktion	Prof. Dr. Christian Wolff, Universität Regensburg Institut für Information und Medien, Sprache und Kultur D-93040 Regensburg
Wissenschaftlicher Beirat	Vorstand, Beirat und Arbeitskreisleiter der GSCL *http://www.gscl.info/*
Erscheinungsweise	2 Hefte im Jahr, halbjährlich zum 31. Mai und 31. Oktober. Preprints und redaktionelle Planungen sind über die Website der GLDV einsehbar (*http://www.gldv.org*).
Einreichung von Beiträgen	Unaufgefordert eingesandte Fachbeiträge werden vor Veröffentlichung von mindestens zwei ReferentInnen begutachtet. Manuskripte sollten deshalb möglichst frühzeitig eingereicht werden. Die namentlich gezeichneten Beiträge geben ausschließlich die Meinung der AutorInnen wieder. Einreichungen sind an die Herausgeber zu übermitteln.
Bezugsbedingungen	Für Mitglieder der GLDV ist der Bezugspreis des LDV-Forums im Jahresbeitrag mit eingeschlossen. Jahresabonnements können zum Preis von 25,- € (inkl. Versand), Einzelexemplare zum Preis von 15,- € (zzgl. Versandkosten) bei der Redaktion bestellt werden.
Satz und Druck	Uta Seewald-Heeg, Daniel Stein und Christian Wolff, mit *LaTeX* (*pdfeTeX / MiKTeX*) und *Adobe InDesign CS3 V 5.0.2*, Druck: Druck TEAM KG, Regensburg

Inhaltsverzeichnis JLCL

JLCL – Volume 24 – Number 3 – 2009
Maschinelle Übersetzung – von der Theorie zur Anwendung

Uta Seewald-Heeg

Vorwort

Maschinelle Übersetzung (MÜ) ist in den vergangenen Jahren in der Sprachenindustrie wieder in den Mittelpunkt des Interesses gerückt. Nicht neue bahnbrechende technologische Ansätze, die eine bessere Qualität der maschinell übersetzten Texte liefern, sondern die Verfügbarkeit großer Datenmengen als Trainingskorpora für statistische Systeme sowie die Integration von maschineller Übersetzung in vorhandene Arbeitsabläufe versprechen sowohl verbesserte Kommunikationsabläufe in global agierenden Unternehmen als auch zeitnahere und kostengünstigere Übersetzungen selbst an Stellen, an denen hochwertige professionelle Übersetzungsqualität gefordert ist.

Bill Gates äußerte 2005 die Ansicht, maschinelle Übersetzung von Texten aus dem Computerbereich erreiche bereits die Qualität humanübersetzter Texte dieses Fachgebiets (vgl. [Gates(2005)]), eine These, die allzu leicht widerlegt werden kann. Denn wer Anfang 2009 Hilfe aus der Microsoft Wissensdatenbank über den *Release Candidate 1* des *Internet Explorer 8* in deutscher Sprache benötigte, erhielt einen maschinell übersetzten Text, dessen Qualität dem Ratsuchenden kaum weiterhalf (vgl. [Ries(2009)]). Dennoch sollten derartige Beispiele aber nicht dazu herangezogen werden, um über die grundsätzliche Eignung maschineller Übersetzung in bestimmten Bereichen zu urteilen, wie zahlreiche Einsatzgebiete in Unternehmen zeigen. Auch die Lokalisierungsindustrie setzt seit einiger Zeit auf Maschinelle Übersetzung. Denn durch die Integration in bestehende Übersetzungsabläufe, in denen Translation-Memory-Technologie eingesetzt wird, birgt Maschinelle Übersetzung Einsparpotentiale. Voraussetzung hierfür ist allerdings, dass das MÜ-System auf die Unternehmens- bzw. Fachgebietsterminologie zugreift und Texte verarbeitet werden, die hinsichtlich ihrer syntaktischen Struktur keinen zu hohen Komplexitätsgrad aufweisen. Entsprechend dieser Erwartung veröffentlichen Anbieter von Translation-Memory-Technologie ebenso wie Entwickler von Übersetzungsprogrammen zunehmend Schnittstellen zu verschiedenen Anwendungen: Anbieter von MÜ-Software bieten vermehrt die Integration in Office-Produkte und E-Mail-Dienste an, während Translation-Memory-Hersteller mit Entwicklern von MÜ-Systemen kooperieren oder ihre Schnittstellen zur Integration entsprechender Systeme bereitstellen.

Doch nicht nur die Integration von MÜ in vorhandene Übersetzungsabläufe, auch die Integration verschiedener Ansätze im Bereich der MÜ versprechen Qualitätszugewinne, so dass sich Maschinelle Übersetzung gerade auch im akademischen Umfeld wieder zu einem viel beachteten Gegenstand der computerlinguistischen Forschung entwickelt hat. Hier stehen die verschiedenen Ansätze, regelbasierte MÜ (*Rule Based Machine Translation*, RBMT), statistische MÜ (*Statistical Machine Translation*, SMT) und beispielbasierte MÜ (*Example Based Machine Translation*, EBMT) einander gegenüber. Vor

allem die Verfügbarkeit großer paralleler Korpora, die statistischen MÜ-Systemen als Trainingsdatenbasis dienen, haben dazu beigetragen, dass statistische MÜ gegenüber den in ihrem Entwicklungsaufwand sehr kostspieligen regelbasierten Systemen für Entwickler und Anwender interessant geworden ist. So basiert denn auch das von Google veröffentlichte Übersetzungsangebot Google Translate im Internet auf einem statistischen Ansatz. Als Trainingsdaten dienten hier große Dokumentenmengen der Vereinten Nationen (UN) in den offiziellen Sprachen der UN und Dokumente der Europäischen Union. Für Sprachen, in denen keine großen Datenmengen paralleler Texte vorliegen, verspricht dieses Verfahren allerdings weit weniger schnell zufriedenstellende Ergebnisse, so dass bereits Ansätze einer Integration der verschiedenen MÜ-Methoden erprobt werden.

Maschinelle Übersetzung ist heute in global agierenden Unternehmen wie Volkswagen (vgl. [Porsiel(2008a)] und [Porsiel(2008b)]) für die standortübergreifende weltweite Kommunikation zu einem integrativen Bestandteil der Unternehmenskommunikation geworden. Aus Gründen der Datensicherheit vertrauen solche Unternehmen aber vielfach nicht auf kostenfreie Internetangebote, sondern setzen auf hauseigene Implementierungen, die dem Markt ebenfalls neue Impulse verleihen.

Die im vorliegenden Band unter dem Titel „Maschinelle Übersetzung – Von der Theorie zur Anwendung“ versammelten Beiträge basieren auf Vorträgen, die im Rahmen des gleichnamigen Workshops des Arbeitskreises „Maschinelle Übersetzung“ der GSCL im Juni 2008 an der Hochschule Anhalt in Köthen gehalten wurden. Einleitend beschreibt Daniel Stein in seinem Beitrag „Maschinelle Übersetzung – ein Überblick“ die historische Entwicklung der verschiedenen Ansätze der MÜ. Einen Einblick in die jüngsten Entwicklungen der Integration von computerunterstützter Übersetzung mithilfe von Translation-Memory-Technologie und Maschineller Übersetzung gewährt Dino Azzano in seinem Beitrag „CAT und MÜ – getrennte Welten?“. Kurt Eberle illustriert in seinem Beitrag „Integration von regel- und statistikbasierten Methoden in der Maschinellen Übersetzung“, wie regelbasierte mit statistischen Verfahren kombiniert und zur Auflösung linguistischer Mehrdeutigkeiten erfolgversprechend eingesetzt werden können. Unter dem Titel „METIS-II: Low-Resource MT for German to English“ erläutert Michael Carl am Beispiel der Implementierung der Übersetzungsrichtung Deutsch-Englisch die Prinzipien der im METIS-II-Projekt implementierten Methoden der maschinellen Übersetzung. Wie bereits METIS-I zielte das Projekt METIS-II darauf ab, maschinelle Übersetzungen auf der Grundlage einsprachiger Textkorpora mit getaggten und lemmatisierten Texten der Zielsprache und zweisprachiger Lexika in den jeweiligen Sprachen der für die MÜ gewünschten Übersetzungsrichtungen zu ermöglichen. Dieser auf Pattern-Matching-Methoden beruhende beispielbasierte Ansatz ist vor allem für Sprachpaare interessant, für die keine großen parallelen Korpora verfügbar sind. Der Beitrag „The Automatic Translation of Film Subtitles. A Machine Translation Success Story“ von Martin Volk zeigt, dass sich eine Textsorte wie Filmuntertitel für die Über-

setzung mit einem MÜ-System eignet und für das Sprachpaar Schwedisch-Dänisch eindrucksvolle Ergebnisse liefert. Um die Eignung einer anderen Textsorte, allerdings für die computergestützte Übersetzung, geht es im letzten Beitrag dieses Bandes. Heribert Härtinger präsentiert die Ergebnisse einer Untersuchung von Patentschriften und deren Eignung für die Übersetzung mit Translation-Memory-Systemen auf der Grundlage eines Korpus deutscher und spanischer Patentschriften.

Bei der Gegenüberstellung von Übersetzungsergebnissen verschiedener MÜ-Verfahren mit von Humanübersetzern übersetzten Texten – wie im Beitrag von Michael Carl und Martin Volk erwähnt – rücken auch Methoden der Evaluierung, die automatisiert erfolgen und in ihren Ergebnissen möglicherweise über bisherige Verfahren wie das von IBM entwickelte Verfahren BLEU ([Papineni(2002)]) oder NIST hinausgehen, in das Interesse der MÜ-Forschung und dürften hier auch künftig von Bedeutung sein.

Literatur

[Gates(2005)] Gates, Bill. "Remarks by Bill Gates, Chairman and Chief Software Architect.", 2005, Accessed 25.03.09. `http://www.microsoft.com/presspass/exec/billg/speeches/2005/10-14Princeton.aspx`.

[Papineni(2002)] Papineni, Kishore. "BLEU: a method for automatic evaluation of machine translation." In *ACL-2002: 40th Annual meeting of the Association for Computational Linguistics*. 2002, 311–318.

[Porsiel(2008a)] Porsiel, Jörg. "Machine translation at Volkswagen: a case study." In *Multilingual*. 2008a, 58–61.

[Porsiel(2008b)] ———. "Maschinelle Übersetzung bei Volkswagen. Sprache als betriebswirtschaftlicher Faktor." *MDÜ* 5: (2008b) 44–47.

[Ries(2009)] Ries, Uli. "Übelsetzt: Microsoft verstört IE8-Nutzer mit wirrem Hilfstext.", 2009. `http://www.computerzeitung.de/articles/uebelsetzt_microsoft_verstoert_ie8-nutzer_mit_wirrem_hilfstext:/2009006/31815447_ha_CZ.html?thes=`.

Daniel Stein

Maschinelle Übersetzung – ein Überblick

Die Idee der formalen Manipulation von Sprachen geht auf die philosophischen Traditionen von Geheim- und Universalsprachen, wie sie Ramon Llull oder Gottfried Wilhelm Leibniz begründet haben, zurück. Bis heute ist die Maschinelle Übersetzung (MÜ) Königsdisziplin der Sprachverarbeitung geblieben: Die Fortschritte seit den ersten praktischen Versuchen sind auf den ersten Blick nur bescheiden. Dabei haben sich im Verlauf der Jahrzehnte zahlreiche unterschiedliche Ansätze zur MÜ gebildet. Nach einer von linguistischen Theorien dominierten Phase stehen seit Beginn der 1990er Jahre wiederentdeckte mathematische Methoden im Vordergrund. Im vorliegenden Beitrag werden die wichtigsten Ansätze eingebettet in ihren historischen Kontext vorgestellt. Besonderes Augenmerk gilt dabei dem regelbasierten und dem statistischen Ansatz.

1 Geschichtlicher Hintergrund

Die ersten Systeme zur maschinellen Übersetzung entstanden kurz nach dem Zweiten Weltkrieg und stellen damit eine der ältesten Anwendungen für Computer überhaupt dar. Um die aktuellen Entwicklungen in der MÜ angemessen beurteilen zu können, ist es wichtig, über Hintergrundwissen zu deren geschichtlicher Entwicklung zu verfügen.

1.1 Geheim- und Universalsprachen als Vorgänger der MÜ

Die Geschichte der MÜ beginnt mit den ersten Gedanken zur formalen Manipulation von Sprachen. Ein wichtiger Vordenker auf diesem Feld war der Katalane Ramon Llull, der schon im 13. Jahrhundert eine Art logischer Maschine sowie eine formale Sprache erdacht hatte. Der berühmteste Vertreter im deutschsprachigen Raum wurde Gottfried Wilhelm Leibniz, der mit seiner Monadentheorie (1696) die Sprache in kleinste Teile zu zerlegen versuchte, um sie aus diesen neu und umfassend aufzubauen (vgl. hier und im Folgenden Gardt (1999)).

Die formalen Arbeiten an der Sprache spalteten sich schnell in zwei unterschiedliche Schulen auf: Universalsprachen und Geheimsprachen. Die Wissenschaft der Universalsprachen hing dem Versuch an, eine Sprache zu entwickeln, die entweder alle denkbaren Gedanken rechnerisch erschließbar machte oder die zumindest für alle Dinge auf der Welt eine ontologisch exakte Bezeichnung habe. Ziel dieser Bemühung war zum einen

eine religiös motivierte Aufhebung der babylonischen Sprachverwirrung. Zum anderen aber erhoffte man sich durch das Beenden der Verständigungsprobleme auf der Welt die Einkehr von Frieden. Ein besonders für die MÜ interessanter Denker war Johann Joachim Becher. Der Universalgelehrte veröffentlichte 1661 eine Publikation mit dem Titel „Allgemeine Verschlüsselung der Sprachen“ und eröffnete seinen Zeitgenossen „Eine geheimschriftliche Erfindung, bisher unerhört, womit jeder beim Lesen in seiner eigenen Sprache verschiedene, ja sogar alle Sprachen, durch eintägiges Einarbeiten erklären und verstehen kann“ (vgl. Becher (1962)). Trotz der offensichtlichen Nähe des von Becher vorgestellten Systems zu den ersten tatsächlichen maschinellen Übersetzungssystemen ist der Einfluss der Universalsprachtheorien auf die Theorien der MÜ bislang eher gering; Wesentlicher war von Beginn an die Wissenschaft der Geheimsprachen, die Kryptologie.

Im Zweiten Weltkrieg spielte die Dechiffrierung feindlicher Funksprüche eine wichtige Rolle. Für das Knacken des Codes der deutschen ENIGMA war in erster Linie das britische Team um Alan Turing in Bletchley Park verantwortlich. Mittels statistischer Methoden, ausgewertet von auf Relais basierenden Rechenmaschinen, legten die Wissenschaftler hier, ohne es zu wissen, den Grundstein für die praktische MÜ. Auf den in Bletchley Park gewonnenen Erfahrungen aufbauend führten Warren Weaver und Andrew Booth einen Briefwechsel, der als Geburtsstunde der MÜ gilt. Dort schrieb Weaver etwa „[...] it is very tempting to say that a book written in Chinese is simply a book written in English which was coded into the ‚Chinese Code‘. If we have useful methods for solving almost any cryptographic problem, may it not be that with proper interpretation we already have useful methods for translation?“ (vgl. den Nachdruck des Memorandums in Weaver (1955)).

1.2 Die Evolution der MÜ-Systeme

Jedoch erwiesen sich die aus der Kryptologie entliehenen mathematischen Ansätze als nicht adäquat für die weitaus komplexere Aufgabe der Übersetzung. Als Folge daraus wurden die ersten Systeme entwickelt, die sich anhand von Wörterbüchern und sparsam eingesetzten syntaktischen Operationen mit der MÜ beschäftigten. Diese wiesen nun erstaunliche Parallelen zu den 1661 vorgestellten Entwürfen von J.J. Becher auf und sind aus heutiger Sicht von bemerkenswerter Naivität gekennzeichnet. Nicht ohne Selbstironie wurde Bechers Schrift 1962 auch mit dem Untertitel „Ein Programmierversuch aus dem Jahre 1661“ (Becher (1962)) neu aufgelegt. Die Bedrohungsszenarien des Kalten Krieges lösten jedoch in Regierungs- und Militärkreisen eine regelrechte Euphorie über die zu erhoffenden Möglichkeiten der MÜ aus und so wurden bis 1966 Unsummen in die Entwicklung von Übersetzungssystemen mit der Sprachrichtung Russisch=>Englisch investiert. Dann jedoch folgte mit einem Paukenschlag das weitreichende Ende dieser Phase: Der 1964 von der US-Regierung, dem CIA und der National

Science Foundation in Auftrag gegebene Automatic Language Processing Advisory Committee (ALPAC)-Report sah die MÜ als zu kostspielig, von den Ergebnissen her unnütz und auch langfristig ohne Hoffnung an (vgl. Hutchings (1996)). Bis auf wenige praktisch orientierte Forschungsgruppen in den USA und Europa kam die Forschung zur MÜ nahezu vollständig zum Erliegen.

Als Reaktion auf die Ausdünnung der Forschungslandschaft konzentrierte man sich vermehrt auf eine Verwissenschaftlichung des Diskurses und die Einbeziehung linguistischen Fachwissens, vor allem auf semantische Analysen. Die hiermit erzielten Erfolge sorgten Mitte der 1970er Jahre wieder für einen Aufschwung, der, getragen von der rasanten Entwicklung der Technologie und der Einführung und zunehmenden Verbreitung der Heimcomputer zu Beginn der 1980er Jahre in einen kontinuierlichen Aufwärtstrend mündete.

Ende der 1980er Jahre veröffentlichte eine Forschergruppe der IBM um Peter F. Brown einen Aufsatz, der erneut statistische Methoden als Grundlage für ein MÜ-System vorstellte. Die verbesserte Rechenleistung und die zunehmende Verfügbarkeit großer, maschinenlesbarer zweisprachiger Korpora hatten eine signifikante Änderung der Ausgangssituation nach sich gezogen. Binnen kürzester Zeit konzentrierte sich die Mehrheit der Forschungen auf die statistischen Ansätze, mit denen man Erfolge erzielen konnte, die mit denen der etablierten, regelbasierten Systeme vergleichbar waren – nur dass man zu deren Erstellung keine 10 Jahre Entwicklungszeit und kein Fachwissen von Linguisten benötigte. Ein paar Tage Zeit und große bilinguale Korpora (Bitexte) genügten für einen Prototypen.

Seit den Jahren ihres Entstehens hat auch die statistisch basierte MÜ einige Entwicklungsphasen durchlaufen und stößt mittlerweile an ihre systembedingten Grenzen. Daher beschäftigt sich die gegenwärtige Entwicklung vor allem mit einer Integration von statistischen und regelbasierten Verfahren, so genannten hybriden Systemen.

2 Typologie

Im Laufe der Jahre haben sich verschiedene Ansätze zur MÜ herausgebildet. Die wichtigsten Vertreter sind heute die regelbasierte und die statistische Übersetzung. Von einigen werden sie immer noch als Konkurrenten begriffen, üblicher ist heute jedoch die Sicht, dass sämtliche Ansätze gewisse Werkzeuge zur Verfügung stellen, die undogmatisch miteinander kombiniert werden können. Im Folgenden werden neben den beiden Hauptvertretern auch die geläufigsten alternativen Ansätze vorgestellt.

2.1 Regelbasierte MÜ

Der regelbasierte Ansatz (RBMT = Rule-Based Machine Translation) ist heute der klassische Ansatz zur MÜ und findet sich in den meisten kommerziellen Systemen

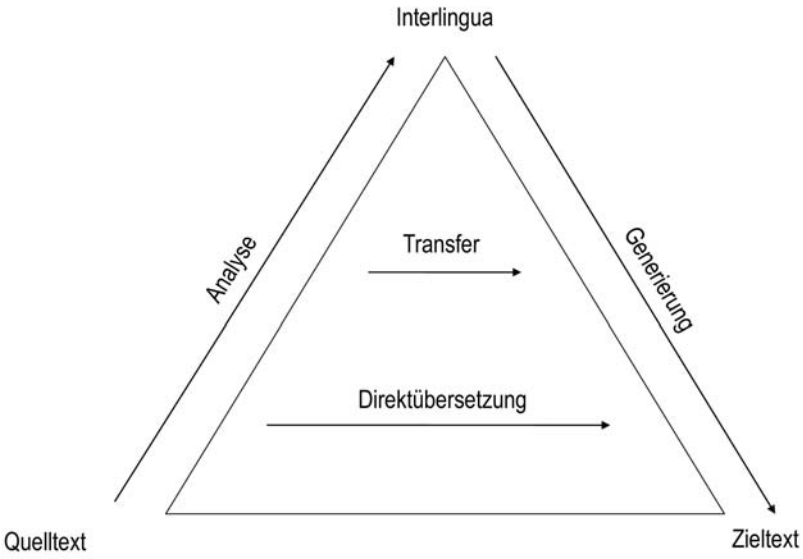
wieder. Die von regelbasierten Systemen produzierten Ergebnisse reichen von kurios bis nützlich, ganz in Abhängigkeit davon, um welches Sprachpaar es sich handelt und ob eine Fachsprache unterstützt wird und entsprechend Fachterminologie eingepflegt wurde, oder ob es sich um ein allgemeinsprachliches System handelt. Ein RBMT-System erarbeitet eine Übersetzung in drei aufeinanderfolgenden Stufen: Analyse, Transfer und Synthese (bzw. Generierung). Man unterscheidet drei (lose) Grade an Komplexität dieser drei Stufen, die Auswirkung auf die Übersetzungsqualität ist jeweils deutlich.

Direkte Übersetzung Bei der direkten Übersetzung handelt es sich um ein System für simple Wort-zu-Wort-Übersetzungen. Diese werden meist über eine syntaktische Komponente oberflächlich an die Satzstellung der Zielsprache angepasst. Die meisten Ergebnisse sind nur in eingeschränkten Anwendungsszenarien zu gebrauchen, was auch daran liegt, dass es für einen großen Teil der Wörter mehr als eine mögliche Übersetzung gibt. Des Weiteren handelt es sich bei vielen von Leerzeichen getrennten Wörtern um Elemente von Mehrwortlexemen, die zumeist nicht wörtlich zu übersetzen sind, wie z.B. ‚ins Gras beißen'.

Transferübersetzung Bei der Transferübersetzung werden zusätzlich morphologische und semantische Informationen in die Übersetzung mit einbezogen, außerdem ist auch die syntaktische Komponente elaborierter. Für alle drei Quellen an zusätzlichen Informationen gilt, dass die Grenze nach oben offen zu sein scheint und sich zehntausende von Regeln und Kombinationen definieren lassen. Allerdings zeigt die Praxis, dass es einen Punkt gibt, ab dem höhere Komplexität nicht mehr dazu beiträgt, die Qualität der Übersetzungen zu verbessern. Stattdessen beginnen interne Konflikte und sich widersprechende Regeln neue Fehler zu produzieren.

Interlingua Übersetzung Der dritte Grad an Komplexität ist die so genannte Interlingua-Übersetzung, ein bis heute utopisches Ideal, das auf der Annahme beruht, es gäbe eine universelle und völlig sprachunabhängige Art der Kodierung von sprachlichen Informationen. Diese abstrakte universalsprachliche Repräsentation würde dann das Ziel und die Quelle sämtlicher Übersetzungssysteme sein. So wäre es möglich, die Informationen aus einem Text vollständig von der Ausgangssprache zu lösen und einen neuen, vom Ausgangstext völlig unabhängigen aber gleichwertigen Text in der Zielsprache zu generieren. Unglücklicherweise ist so eine universelle Sprache bis heute nicht entdeckt worden, auch wenn bereits Llull und Leibniz, wie beschrieben, daran forschten.

Die folgende Grafik stellt den jeweils zu leistenden Aufwand in den drei Phasen der MÜ für die unterschiedlichen Komplexitätsphasen dar.



2.2 Statistikbasierte MÜ

1988 stellt der IBM-Wissenschaftler Peter Brown dem überraschten Publikum auf der Second TMI Conference der Carnegie Mellon University einen rein statistischen Ansatz zur MÜ vor (SMÜ, bzw. SMT = Statistical Machine Translation) (vgl. Brown et al. (1988)). SMÜ basiert auf dem Gedanken, dass Übersetzungsentscheidungen anhand von bedingten Wahrscheinlichkeiten getroffen werden können. Anstelle aufwändiger Regelwerke werden große parallele Korpora benötigt.

2.2.1 Funktionsweise

Die Funktionsweise eines SMÜ-Systems basiert auf der folgenden Überlegung: Wir versuchen den beliebigen englischen Satz **e** ins Französische zu übersetzen. Alle möglichen und unmöglichen französischen Sätze **f** sind potentielle Übersetzungen des einen engli-

schen Satzes e.[1] Einige davon sind jedoch wahrscheinlicher als andere. **p(f | e)** beschreibt die Wahrscheinlichkeit, dass f eine Übersetzung von e ist.

Des Weiteren gehen wir davon aus, dass der Sprecher von e zwar Muttersprachler ist, sich e aber im Geiste erst als f gedacht und diese Vorlage dann übersetzt hat. Diese etwas umständliche Voraussetzung dient dazu, die tatsächliche Aufgabe eines SMÜ-Systems zu definieren: Das Ziel lautet, das ursprünglich gedachte f zu finden, die so genannte *wahrscheinlichste Übersetzung*.

Dieser gedachten Situation muss man die Unmöglichkeit, alle beliebigen Sätze einer Sprache verfügbar zu haben, entgegenstellen. Daher wird in der SMÜ mit Näherungen gearbeitet, mit Modellen. Ein zweisprachiges aliniertes Korpus bildet das Übersetzungsmodell, das alle möglichen Übersetzungen zwischen beiden Sprachen repräsentiert. Alle vorhandenen Sätze stellen jeweils potentielle Übersetzungen voneinander dar, die einander zugewiesenen haben jedoch die höchste Wahrscheinlichkeit. Ein einsprachiges Korpus in der Zielsprache stellt das Sprachmodell dar und repräsentiert hier alle gültigen Sätze einer Sprache. Da die Zahl aller möglichen Sätze auch hier noch zu groß ist, wird auch das Sprachmodell weiter abstrahiert und man arbeitet auf der Wortebene oder mit Wortsequenzen. Auch das Übersetzungsmodell muss weiter abstrahiert werden, dazu wird es in ein Lexikonmodell und ein Alinierungsmodell aufgeteilt. Ersteres beschreibt die Richtigkeit von Wort(sequenzen)übersetzungen – je wahrscheinlicher ein Wort eine Übersetzung eines anderen ist, desto höher sein Wert. Das zweitgenannte beschreibt die Richtigkeit von Satzstellungen. Je wahrscheinlicher eine Satzstellung eine Übersetzung einer anderen ist, desto höher ihr Wert. Ein Suchalgorithmus ermittelt nun den Satz, dessen Produkt der Werte von Satzgültigkeit (Sprachmodell), Wortübersetzung (Lexikonmodell) und Satzstellung (Alinierungsmodell) am höchsten ist. Das Ergebnis ist die wahrscheinlichste Übersetzung.

Die Wahrscheinlichkeiten, mit denen gerechnet wird, sind nicht „einfach da“, sondern müssen vom Computer geschätzt werden. Dazu wird in der Regel der Satz von Bayes angewendet:

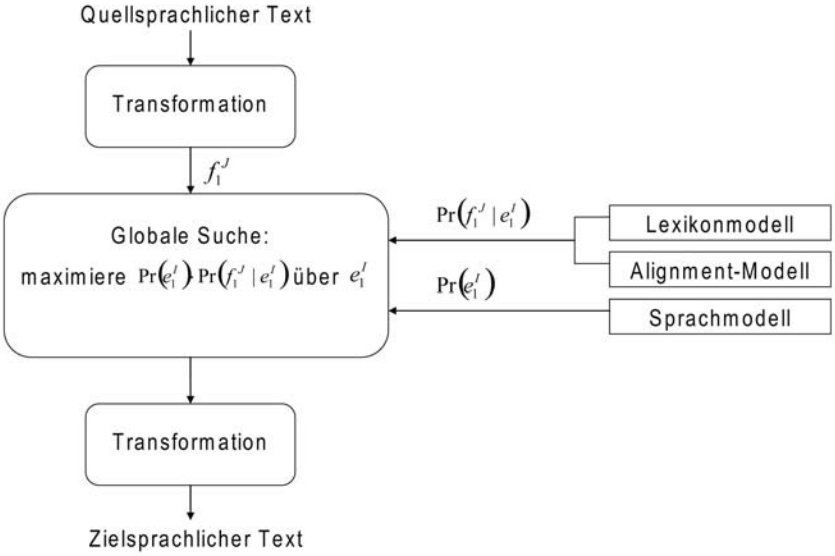
$$Pr(e|f) = \frac{Pr(e) * Pr(f|e)}{Pr(f)} \tag{1}$$

Der Satz kann reduziert werden auf die Suche nach dem Maximalwert der beiden Terme Pr(e) und Pr(f | e), wobei der erste bedeutet „Wahrscheinlichkeit, dass jemand e so gesagt hat“ und der zweite „Wahrscheinlichkeit, dass jemand e so nach f übersetzt hätte“:

$$\hat{e} = argmaxPr(e) * Pr(f|e) \tag{2}$$

[1] Die Beispielsprachen Englisch und Französisch beziehen sich auf das von Brown verwendete englisch-französische *Hansard-Korpus*, welches Protokolle des kanadischen Parlaments enthält.

Die folgende Darstellung (vgl. Stein et al. (2006)) illustriert den Aufbau eines SMÜ-Systems anhand der verwendeten Modelle:



2.2.2 SMÜ-Typen

Die Analyse von ganzen Sätzen ist in der SMÜ wenig sinnvoll: Wie oft findet sich schon der zu übersetzende Satz vollständig in den zugrundeliegenden Korpora wieder? Solange ein SMÜ-System nicht tatsächlich über ein Korpus verfügt, das alle (oder wenigstens annähernd alle) möglichen Sätze einer Sprache enthält, ist es sinnvoll, die zu betrachtende Einheit zu verkleinern. SMÜ-Typen lassen sich nach der Ebene unterscheiden, auf der sie Texte analysieren. Man unterscheidet allgemein zwischen wortbasierter und phrasenbasierter SMÜ.

Wortbasierte SMÜ Die ursprüngliche Variante der SMÜ analysiert die Trainings- und Testdaten auf der Ebene der Wörter. Das bedeutet, dass ein Wort in der Quellsprache einem Wort in der Zielsprache entsprechen muss. Gelgentlich kommt es auch vor, dass ein Wort in der Quellsprache sich nur durch mehrere Wörter in der Zielsprache übersetzen lässt, wie Englisch „slap“ => Spanisch „dar una botifada“. Dies ist mit der wortbasierten SMÜ zwar möglich. Die Umkehrrichtung jedoch, also dass mehrere Wörter in der Quellsprache zusammen nur ein Wort in der Zielsprache ergeben (dar una botifada => slap) ist durch die Wortbasiertheit unmöglich. Jedem Wort in der

Quellsprache *muss* also mindestens ein Wort in der Zielsprache entsprechen. Ein weiteres, verwandtes Problem ist, dass zusammengehörende Wörter nicht zusammen übersetzt werden können. Besonders störend wirkt sich das unter anderem bei Klammerverben aus, da diese, unabhängig voneinander betrachtet, stark abweichende Bedeutungen haben können (vgl. das alleinstehende ‚ab' in: „Ich *reiste* schon nach vierzehn Tagen wieder *ab*"). Dieses Problem wirkt sich auch auf Sprachen aus, die eine stark voneinander abweichende Syntax verwenden, beispielsweise was die Position des finiten Verbs angeht.

Phrasenbasierte SMÜ Um der genannten Probleme Herr zu werden, entwickelten sich unterschiedliche neue Ansätze der SMÜ heraus. Heute gängige Systeme arbeiten in der Regel auf der Ebene von Phrasen. Diese sind jedoch nicht – wie der Name nahe legt – linguistisch motiviert. Im Gegenteil werden die Trainings- und die Testdatensätze maschinell in Gruppen einer bestimmten Größe geteilt und müssten daher eigentlich einfach Wortsequenzen genannt werden. Durch die Betrachtung dieser Art von Phrasen ist es innerhalb der phrasenbasierten SMÜ somit möglich, mehrere Wörter mit einem zu übersetzen und umgekehrt. Ein weiterer Vorteil der Betrachtung von Wortsequenzen ist es, dass der erweiterte Kontext die Möglichkeit eröffnet, bestimmte Disambiguierungsentscheidungen zu treffen. So zum Beispiel wäre die wortbasierte SMÜ nicht in der Lage, zu entscheiden, welche Übersetzung von „pretty" in den Fällen „pretty much" und „pretty girl" die richtige wäre. Es gibt verschiedene Möglichkeiten, die Ebene von Phrasen zu behandeln, je nach System und Größe der Sequenzen ist es auch möglich, die erwähnten Unterschiede zwischen Quell- und Zielsyntax zu überbrücken.

2.2.3 Vorzüge und Nachteile der SMÜ

Es ist als ein großer Vorteil der SMÜ zu werten, dass ein funktionierendes System in weitgehender Unkenntnis der zu verwendenden Sprachen und ihrer Eigenheiten erarbeitet werden kann. Durch den Verzicht auf linguistisches Fachwissen und dessen aufwändige Modellierung (die sich über Jahrzehnte erstrecken kann) ist es möglich geworden, verhältnismäßig robuste Systeme in kurzer Zeit und für wenig Geld zu erstellen. Diese können dann auch für Sprachen verfügbar gemacht werden, die bisher nicht über die für ein regelbasiertes System notwendigen Ressourcen verfügen. Die einzige Bedingung ist, dass genügend alinierte mehrsprachige Korpora vorhanden sind. Dies ist zum Beispiel bei den meisten Sprachen der Europäischen Union der Fall, da sie über das Korpus der Protokolle des Europäischen Parlaments, EuroParl, verfügen. Auf dieser Grundlage kann man mittels der SMÜ in kürzester Zeit Systeme zur Verfügung stellen, deren Qualität mit jener der etablierten regelbasierten Systeme vergleichbar ist. Im Gegensatz zu diesen ist die SMÜ sogar im Vorteil, wenn es um die Lösung lexikalischer Ambiguitäten oder arbiträrer Redewendungen geht, allerdings nur, wenn

diese auch in genügender Zahl im Trainingsmaterial repräsentiert werden. Daher ist die schlichte Regel der SMÜ die folgende: „Größere Korpora bringen bessere Ergebnisse."

Die Nachteile der SMÜ ergeben sich beinahe vollständig aus ihren Vorteilen: Da sämtliche Übersetzungen aus nicht mehr nachvollziehbaren Berechnungen auf der Grundlage des unüberschaubaren Trainingsmaterials basieren, ist es so gut wie unmöglich, einzelne Fehlerquellen auszumachen. Eine Korrektur bestimmer systematisch falscher Ergebnisse ist im Gegensatz zu regelbasierten Systemen nur schwer möglich. Des Weiteren ist trotz der weitgehenden Sprachunabhängigkeit von SMÜ-Systemen anzumerken, dass bei bestimmten Kombinationen von Quell- und Zielsprache schwerwiegende Probleme auftauchen können, etwa wenn es sich um Sprachen mit stark unterschiedlicher Struktur (Flexion, Satzbau, Prodrop etc.) handelt. Gerade zusammengehörige Sprachbestandteile, die mehrere Wörter voneinander entfernt sind – beispielsweise die deutschen Verbklammern – werden von den SMÜ-Systemen schlichtweg ignoriert. Dies führt häufig zu Übersetzungen, in denen ausgerechnet das entscheidende Verb fehlt. Auch die Notwendigkeit großer Korpora ist ein Problem nicht nur für die genannten kleineren Sprachen. Denn die meisten aktuell verfügbaren zweisprachigen Korpora entstammen Fachsprachen wie der Gesetzesgebung und deren Fachtermini sind in den Korpora weit überrepräsentiert. So ist es auch kein Wunder, dass die SMÜ in für spezielle Fachsprachen entwickelten Systemen die besten Ergebnisse erbringt. Darauf aufbauend ist auch das nächste Problem offensichtlich. Die Regel „Größere Korpora bringen bessere Ergebnisse" deutet schon den ungeheuren Datenhunger der SMÜ an: Ein Korpus kann einfach nicht groß genug sein.

2.3 Beispielbasiert

Neben dem statistikbasierten Ansatz ist der beispielbasierte (EBMT = Example Based Machine Translation) einer der gegenwärtig meist diskutierten. Die Grundlage der EBMT ist der der SMÜ gleich. Gearbeitet wird nämlich auf einem Korpus von parallelen Texten. Die Herangehensweise an dieses Korpus ist jedoch eine grundverschiedene: Anstelle ein möglichst großes Korpus zu analysieren um die, auf Grundlage der vorhandenen Daten, wahrscheinlichste Übersetzung zu erlangen, vergleicht das EBMT-System Teile des zu übersetzenden Textes mit einem verhältnismäßig viel kleineren Korpus nach dem Analogieprinzip. Das EBMT-System identifiziert verwertbare Teile und rekombiniert diese für die Übersetzung. Abschließend wird versucht, die auf Beispielen basierenden Übersetzungsbruchstücke in zusammenhängende Sätze zu transformieren. Aufgrund dieser Verfahrensweise wird die EBMT häufig mit so genannten Translation Memory (TM)-Systemen in Zusammenhang gebracht. Dies ist jedoch nur bedingt zutreffend, da es sich bei TM-Systemen um interaktive Unterstützung für menschliche Übersetzer handelt, während ein EBMT-System vollkommen autonom arbeitet (vgl. Somers (2003)).

2.4 Kontextbasiert

Der Ansatz der kontextbasierten MÜ (CBMT = Context Based Machine Translation) ist verhältnismäßig neu und arbeitet wie SMÜ und EBMT auf der Grundlage von Korpora. Im Unterschied zu den genannten Ansätzen benötigt die CBMT jedoch ausschließlich möglichst große einsprachige Korpora der Zielsprache. Grundlage des Übersetzungsprozesses ist hier ein umfangreiches zweisprachiges Vollformenlexikon. Dieses ermittelt für jedes Wort alle möglichen Übersetzungsvarianten und lässt diese in alternativen Übersetzungen intern weiterführen. Um nun die korrekten von den falschen Übersetzungen zu unterscheiden, werden diese auf Basis von N-Grammen mit dem Zielkorpus abgeglichen. Die Variante, die mehr oder längere Treffer im Korpus hat, wird weitergeführt. Unmögliche und unwahrscheinliche Übersetzungen werden so zuverlässig gefunden und ausgeschlossen. Des Weiteren wird auf dieser Ebene auch im Rahmen des gegebenen Kontextes, also der N-Gramm-Größe, disambiguiert (vgl. Carbonell et al. (2006)).

2.5 Wissensbasiert

Ein oft diskutiertes Problem der MÜ ist, dass zum Übersetzen ein gewisses Maß an Weltwissen unabdingbar scheint. Zum Beispiel ist es schwer, einen der alternativen Sätze „Das Schloss liegt auf dem Berg/Tisch.“ korrekt zu übersetzen, wenn man nicht weiß, woran man erkennen kann, um welche Form von Schloss es sich handelt. Der wissensbasierte Ansatz (KBMT = Knowledge Based Machine Translation) versucht, Wissen dieser Form in einer Datenbank zu organisieren. Dies ist jedoch bislang nur für Spezialgebiete möglich. Aufgrund der metasprachlichen Organisation von Wissen gilt die wissensbasierte Übersetzung als Spezialfall der regelbasierten Interlingua.

2.6 Hybride Ansätze

Unter hybriden Ansätzen versteht man MÜ-Systeme, die versuchen, die Vorteile verschiedener Ansätze in einem System zu vereinen. Dies betrifft vor allem die SMÜ. Es gibt zahllose Entwürfe, SMÜ durch vorgeschaltete syntaktische Analysen oder semantische Operationen zu verbessern. Dies bietet sich vor allem bei für die SMÜ ungünstigen Sprachkombinationen an. Ungünstig, etwa weil die Sprachen unterschiedlich stark flektieren, einen deutlich voneinander abweichenden Satzbau haben oder weil zum Beispiel eine der beteiligten Sprachen nur über sehr kleine Korpora verfügt.

2.6.1 Ein hybrides System als Beispiel

Ein hybrides System stellt de Gisbert in seinem Papier „Improving Statistical Machine Translation by Classifying and Generalizing Inflected Verb Forms“ (de Gispert Ra-

mis et al. (2005)) vor. Wie beschrieben kann sich unterschiedlich starke Flexion von Quell- und Zielsprache als ungünstig für ein SMÜ-System erweisen. Spanisch ist eine stark flektierende Sprache, für das Englische say/said können im Spanischen decir/digo/dices/dice/dicen usw. vorkommen, ganz zu schweigen von den Varianten mit Hilfsverbgefüge. Dies verkleinert die statistische Basis für Wortübersetzungen erheblich. Der entstehende negative Effekt zeigt sich sowohl bei der Übersetzungsqualität als auch beim Trainingsprozess: Die grammatischen Informationen, die das System aus den Trainingsdaten ziehen kann, sind äußerst gering.

Dabei ist es jedoch möglich, die beschriebenen Probleme anhand von morphologischen Methoden zu umgehen. Verwendet man ein phrasenbasiertes SMÜ-System, müssen dazu in einem ersten Schritt die aus den Trainingsdaten erstellten Phrasen analysiert werden. Anschließend wird eine Auswahl der Phrasen – solche, die die Hauptverben innerhalb des Satzes in sich bergen – entsprechend den Ergebnissen der morphologischen Analyse linguistisch klassifiziert. In einem weiteren Schritt werden diese Phrasen einander in Tupeln zugewiesen. Das heißt, ein Tupel beinhaltet die jeweilige Phrase in beiden Sprachen und zusätzlich linguistische Informationen zu diesen, beispielsweise über Numerus, Genus und Kasus.

Die aus diesem Vorgang gewonnenen klassifizierten parallelen Tupel werden nun dazu verwendet, unbekannte Verbformen über Generalisierung zu erschließen. Dies geschieht am Beispiel des englischen Satzes „we would have payed it“. Das Korpus beinhaltet für die englische Verbklasse V[pay] beispielsweise die folgenden drei dementsprechenden Tupel:

T1=(V[pay],V[pagar])
T2=T(V[pay],V[hacer] el pago)
T3=T(V[pay] it, lo V[pagar])

Trotz der drei verschiedenen Treffer ist die spezielle Form „we would have payed it“ nicht vertreten. In diesem Fall listet das System alle Fälle und deren Frequenz auf, in denen die Klasse pay übersetzt wurde und die dazu dienen können, „we would have payed it“ zu übersetzen (vgl. Tabelle 1).

Die Klassifizierung der Phrasen nach linguistischen Merkmalen bestimmt unter anderem das Genus der darin vorhandenen Verben. Also erkennt das System, dass es sich bei „we would have payed it“ um die 1. Person Plural handelt. Für jede der in der obigen Tabelle angegebenen Varianten generiert das System daraufhin ein neues Verb mit dem angegebenen Geschlecht. Diese werden in einer Tabelle, gewichtet nach den Wahrscheinlichkeiten der Wörter, von denen sie stammen, angegeben (vgl. Tabelle 2).

In uneindeutigen Fällen, wie beispielsweise der Übersetzung von ‚you‘ entweder in der 2. Person Singular oder die 2. Person Plural, ist das System so programmiert, alle möglichen Varianten zu ermitteln und dem mit monolingualen Korpora zusätzlich

Tabelle 1:

T1 = (V[pay] , V[pagar])		
I would have payed	habría pagado	3
you would have payed	habrías pagado	1
you would have payed	pagarías	1
T2 = (V[pay] , V[hacer] el pago)		
* would have payed it	–	0
T3 = (V[pay] it , lo V[pagar])		
I would have payed it	lo habría pagado	1

Tabelle 2:

T1	we would have payed	habríamos pagado	4/6
T2	we would have payed	pagaríamos	1/6
T3	we would have payed it	lo habríamos pagado	1/6

trainierten Sprachmodell die Entscheidung zu überlassen. Eine alternative Form wäre die so genannte erweiterte Generalisierung (*Extended Generalization*). Sie behandelt speziell das Problem, das auftritt, wenn genau eine exakte Realisation (*perfect match*) einer Verbform in den Trainingsdaten vorkommt, diese jedoch als Übersetzung sehr unwahrscheinlich erscheint. Normalerweise wird diese vom System dennoch als richtige Übersetzung erkannt und andere, wahrscheinlichere Tupel, die jedoch erst gebildet werden müssten, werden vom System nicht mehr berücksichtigt. Hier besteht die Verbesserung einfach darin, bei entsprechenden Fällen dennoch in allen Tupeln des Test-Sets nach anderen Übersetzungsmöglichkeiten zu suchen.

Zur Evaluation wurden Übersetzungen vom Englischen ins Spanische in vier verschiedenen Modi angefertigt. Die erste Übersetzung wurde hergestellt, ohne eine der beschriebenen Implementierungen hinzuzuschalten (*Baseline*). Bei der zweiten wurden die Verben zwar klassifiziert, nicht aber generalisiert (*Verb class*). Der dritte Versuch schließt eine Generalisierung ein (*Verb class+gen*), der letzte verwendet die erweiterte Generalisierung (*Verb class+genEX*). Die Ergebnisse werden nach den gängigen Maßen Word Error Rate (WER) und BLEU-Score (Bilingual Evaluation Understudy) evaluiert (vgl. Tabelle 3).

Die Ergebnisse geben ein recht eindeutiges Bild wieder: Die reine Klassifizierung der Phrasen und die Zuweisung derselben untereinander in Tupeln haben in allen Bereichen bereits deutliche Verbesserungen gegenüber dem ursprünglichen Systemaufbau ermöglicht. Jedoch hat die Weiterverwendung der linguistisch verwertbaren Daten dieser Klassifizierung anhand von Generalisierung und erweiterter Generalisierung nur

Tabelle 3:

	Dev set		Test set	
	WER	BLEU	WER	BLEU
baseline	21,32	0,698	23,16	0,671
Verb class	19,37	0,728	22,22	0,686
Verb class+gen	19,27	0,727	21,65	0,692
Verb class+gen ex	19,25	0,729	21,62	0,689

noch geringe Steigerungen des BLEU-Score beziehungsweise Senkungen der WER nach sich gezogen. Dies liegt mitunter sicherlich daran, dass die Verbesserung durch die Klassifizierung der Phrasen die gesamte Übersetzung betrifft, während hingegen die (erweiterte) Generalisierung von unbekannten Verben entsprechend nur die Übersetzung derjenigen Sätze verbessern kann, die auch unbekannte Fälle enthalten. Dieser Ansatz belegt zweierlei: Erstens, dass sich schon mit wenig Aufwand und einem Minimum an linguistischer Information bedeutende Verbesserungen an einem SMÜ-System vollziehen lassen. Und zweitens, dass man häufig auch durch komplexere Ansätze nur minimale Fortschritte erzielen kann und sich bestimmte Ansätze auch gegenseitig im Weg stehen können. Es ist in jedem Fall noch viel Entwicklungspotential für eine linguistisch aufgewertete SMÜ vorhanden.

3 Perspektiven

Die MÜ-Forschung hat in den vergangenen Jahrzehnten schon einige Hochs und Tiefs mitgemacht. Die Aussicht auf vollautomatische Qualitätsübersetzungen versetzte (von Johann Joachim Becher ausgehend bis heute) Forscher, Geldgeber und Laien regelmäßig in Euphorie, die sich, nachdem man mit den neuen Methoden ebenfalls nicht zum Ziel kam, schnell wieder verflüchtigte und einer regelrechten Depression wich. Der gegenwärtige Aufwärtstrend begann mit der Veröffentlichung des statistischen Ansatzes von Brown und erreichte seinen vorläufigen Höhepunkt, als sich die beiden Softwareriesen Google und Microsoft in den letzten Jahren mit ihren MÜ-Systemen auf den globalen Markt begaben. Bei Google arbeitet ein reines SMÜ-System, Microsoft setzt auf eine Zwischenlösung: computerbezogene Texte werden vom hauseigenen SMÜ-System übersetzt, alles andere durch Ergebnisse des regelbasierten Systransystems ergänzt. Die Ergebnisse der beiden Systeme unterscheiden sich im Endeffekt nicht von den bisherigen: Zuweilen unterhaltsam, meist zumindest nützlich. Auch die EU investiert – nachdem die EG in den 1980er Jahren viel Geld mit einem ungenügenden System (Eurotra) in den Sand gesetzt hatte – erstmals wieder in ein größeres MÜ-System. Das Projekt EuroMatrix soll ein hybrides System entwickeln, das zwischen den Sprachen aller Mitgliedsstaaten der EU übersetzt. Ob dieses ehrgeizige Ziel erreicht werden kann, ist noch nicht absehbar.

Weder die regelbasierten noch die rein empirischen Modelle versprechen noch nennenswerte Verbesserungen für die Zukunft, doch sie bieten reichhaltige Werkzeuge für neue Verfahren, um vielleicht endlich den ersehnten Qualitätssprung in der MÜ zu erreichen.

Literatur

Becher, J. J. (1962). *Zur mechanischen Sprachübersetzung. Ein Programmierversuch aus dem Jahre 1661. Allgemeine Verschlüsselung der Sprachen.* Kohlhammer.

Brown, P. F., Cocke, J., Pietra, S. A. D., Pietra, V. J. D., Jelinek, F., Mercer, R. L., and Roossin, P. S. (1988). A statistical approach to french/english translation.

Carbonell, J., Klein, S., Miller, D., Steinbaum, M., Grassiany, T., and Frey, J. (2006). Context-based machine translation.

de Gispert Ramis, A., Mariño, J. B., and Crego, J. M. (2005). Improving statistical machine translation by classifying and generalizing inflected verb forms.

Gardt, A. (1999). *Geschichte der Sprachwissenschaft in Deutschland. Vom Mittelalter bis ins 20. Jahrhundert*. de Gruyter.

Hutchings, J. (1996). ALPAC: The (in)famous report. In *MT News International. Newsletter of the International Association for Machine Translation*, volume 14. International Association for Machine Translation.

Somers, H. (2003). An Overview of EBMT. In Carl, M. and Way, A., editors, *Recent advances in Example-Based Machine Translation*, pages 3–57. Kluwer, Dordrecht.

Stein, D., Bungeroth, J., and Ney, H. (2006). Morpho-syntax based statistical methods for automatic sign language translation.

Weaver, W. (1955). Translation. In Locke, W. N. and Booth, D. A., editors, *Machine Translation of Languages. Fourteen Essays*. Technology Press of MIT, New York.

Dino Azzano

CAT und MÜ – Getrennte Welten?

Im vorliegenden Artikel werden die Zusammenhänge zwischen *computerunterstützter Übersetzung* (*Computer Assisted Translation*, CAT) und *maschineller Übersetzung* (MÜ) untersucht. Im Mittelpunkt stehen die Systeme zur computergestützten Übersetzung sowie ihre Integrierbarkeit mit maschinellen Übersetzungssystemen. Eingangs werden einige terminologische Unterscheidungen getroffen, um die wichtigsten Begrifflichkeiten zu klären. Darüber hinaues werden die Hauptunterschiede zwischen CAT und MÜ erwähnt. Ein Überblick über die wichtigsten Komponenten eines CAT-Systems sowie über die gängigsten Produkte auf dem Markt dient als Grundlage für die Beschreibung der Integrationsmöglichkeiten. Vier Beispielsprozesse veranschaulichen die konkrete Arbeitsweise. Abschließend werden Vorteile und Nachteile einer Integration von CAT und MÜ besprochen

1 Terminologie

Einige terminologische Vorbemerkungen dienen zur Unterscheidung der wichtigsten Begrifflichkeiten.

1.1 CAT und MÜ

Computerunterstützte Übersetzung oder maschinengestützte Übersetzung, auch als Computer Assisted Translation, oder Computer Aided Translation (CAT) bekannt, definiert eine Übersetzungsmethode, bei der die Übersetzung in der Regel von einem Menschen mit Unterstützung eines Programms gefertigt wird. Über Ausnahmen wird im Kapitel 5 berichtet.

Maschinelle Übersetzung (MÜ) oder automatische Übersetzung, auch als Machine Translation (MT), Automated oder Automatic Translation bekannt, definiert dagegen eine Übersetzung, die von einer Übersetzungssoftware erstellt wird, gegebenenfalls ohne menschliches Eingreifen.

Insbesondere die englische Bezeichnung MT wird allerdings manchmal in einem sehr allgemeinen Sinn verwendet und schließt ein (nach Reinke (2003)):

MAHT: Machine-Aided Human Translation, auch als MAT (Machine Aided Translation) bekannt.

HAMT: Human-Aided Machine Translation.

FAMT: Fully Automatic Machine Translation, auch als FAHQMT (Fully Automatic High Quality Machine Translation) bekannt.

Angesichts dieser Gliederung, die den Anteil des menschlichen Übersetzers am Übersetzungsablauf berücksichtigt, wäre die CAT in der MAHT anzusiedeln. Jedoch wird MT meistens nur im engeren Sinne von HAMT oder FAMT verwendet, vgl. zum Beispiel Hutchins (2003) und Trujillo (1999). Deswegen ist die CAT eher als eigenständige Übersetzungsmethode zu betrachten und nicht als Unterkategorie der MÜ.

1.2 MÜ-Methoden

Aufgrund der unterschiedlichen Methoden, die der MÜ zu Grunde liegen, wird sie wie folgt gegliedert (nach Trujillo (1999), Carl and Way (2003) und Eberle (2006)).[1]

RBMT: Rule-Based Machine Translation, regelbasierte maschinelle Übersetzung: Der Ausgangstext wird analysiert, und diese Analyse wird mittels eines Satzes von linguistischen Regeln und eines Wörterbuchs in Strukturen der Zielsprache übersetzt, aus denen der Zieltext generiert wird.

EBMT: Example-Based Machine Translation, beispielbasierte maschinelle Übersetzung: Die Übersetzung eines Ausgangstexts wird mittels Regeln aus einem alignierten zweisprachigen Korpus rekonstruiert.

SMT: Statistical Machine Translation, statistische maschinelle Übersetzung: Die Übersetzung wird mittels statistischer Wahrscheinlichkeitsmodelle aus einem zweisprachigen Korpus erzeugt, wobei linguistisches Wissen nicht unbedingt einbezogen wird.

1.3 CAT, TM und Match

Die Software, um genau zu sein das Software-Paket, welches für die CAT verwendet wird, hat unterschiedliche, zum Teil aus dem Englischen übernommene Bezeichnungen: CAT-Tool, CAT-System, integriertes System sowie Translator's Workstation. Diese Bezeichnungen sind zueinander synonym.

Ein weiterer häufiger Begriff ist Translation Memory System (TM-System). In Kapitel 3 wird das Translation Memory – eine Komponente des CAT-Systems, die Satzpaare

[1] Auf diese Gliederung wird im vorliegenden Artikel nur oberflächlich und vereinfacht eingegangen. Für eine umfassende Einführung und Unterscheidung, siehe TRUJILLO 1999. Für eine kürzere Einführung, siehe EBERLE 2006.

bestehend aus ausgangssprachlichen Sätzen und deren Übersetzung speichert, – näher beschrieben. Die Benennung TM-System nimmt also eine einzelne Komponente für die Bezeichnung des ganzen Software-Pakets her (pars pro toto).

Auf der Ebene des Translation Memorys sind drei weitere Begriffe einzuführen. Das erneute Vorkommen desselben Ausgangssegments, für das im Translation Memory eine Übersetzung vorhanden ist, wird 100% Match oder *Perfect Match* genannt.[2] Das Vorkommen eines ähnlichen Ausgangssegments, für das im Translation Memory eine Übersetzung vorhanden ist, wird *Fuzzy-Match* genannt.

Das Vorkommen eines Ausgangssegments, für das im Translation Memory überhaupt keine Übersetzung vorhanden ist, wird *No Match* genannt. Dabei kann der Ausgangstext vollständig neu sein oder nicht ähnlich genug, um als Fuzzy-Match erkannt zu werden.

2 Hauptunterschiede

Der wichtigste Punkt bei einer Gegenüberstellung von CAT und MÜ ist, dass ein CAT-System Übersetzungen wiederverwendet und nicht neu erstellt. Aus diesem Grund sind CAT-Systeme eher Information-Retrieval-Systemen ähnlich (siehe Reinke (2003)). Hinzu kommt, dass in einem CAT-System die Übersetzung in der Regel von einem Menschen erstellt wird. Das System beschränkt sich darauf, beim erneuten Vorkommen desselben oder eines ähnlichen Ausgangssegments die bereits erstellte Übersetzung vorzuschlagen.

Zwar sind modernere CAT-Systeme in der Lage, gewisse Textanpassungen vorzunehmen, zum Beispiel bei Zahlen, Tags oder Satzzeichen. Diese Ersetzungen können jedoch nicht als Neutexterzeugung definiert werden. Nur das Teilsatz-Matching, das vorerst lediglich von einigen Systemen angeboten wird, kann als Brückenschlag zur MÜ gesehen werden, wobei diese Aussage angefochten werden kann, weil das Kennzeichen dieser Funktionalitäten die kleinere Match-Einheit ist (eine Phrase statt eines ganzen Segments), und nicht die Synthese einer neuen Übersetzung samt grammatischen Anpassungen.

Die Ähnlichkeit von CAT und EBMT ist zwar unübersehbar. Allerdings werden in der CAT – in Gegensatz zu EBMT – keine automatischen Anpassungen von der Software vorgenommen, welche linguistisches Wissen voraussetzen. Sie bleiben eine Aufgabe des Humanübersetzers (siehe Trujillo (1999)).

Der zweite wichtige Punkt bei der Unterscheidung von MÜ-Systemen und CAT-Systemen ist der Entscheidungsträger. Bei den letzteren kann der Übersetzer immer eingreifen und bestimmen, ob eine angebotene Übersetzung bzw. vorgenommene Anpassung übernommen werden soll oder nicht (siehe Somers (2003a)). Es sind zwar Funktionen vorhanden, welche die ungeprüfte Übernahme von 100% Matches ermöglichen,

[2] Der Begriff *Perfect Match* kann auch andere Bedeutungen haben, zum Beispiel in SDL Trados. Im Allgemeinen sind weitere produktspezifische Begrifflichkeiten möglich.

aber selbst in diesem Fall liegt die Entscheidung beim Übersetzer, diese Funktionalität zu verwenden.

3 Komponenten eines CAT-Systems

Die Benennung CAT-System weist auf die Tatsache hin, dass dieser Systemtyp mehrere Komponenten verbindet, welche unterschiedliche und zum Teil unabhängige Funktionen ausführen.[3]

Die einzelnen Komponenten sind vom Produkt abhängig, es gibt jedoch einige, die bei allen Produkten zu finden sind und das Kernstück solcher Systeme bilden.

Der Rahmen dieses Artikels ermöglicht lediglich eine teilweise Aufzählung. Für eine vollständige und detaillierte Beschreibung, siehe Massion (2005).[4]

Übersetzungsspeicher: Er wird auch als Translation Memory (TM) bezeichnet und ist die wichtigste Komponente solcher Systeme. Bei den meisten Systemen handelt es sich um eine Datenbank, in einigen Fällen kommt aber auch eine Sammlung von Referenzdateien zum Einsatz.[5] Der Inhalt der Datenbank besteht aus Segmentpaaren, wobei ein ausgangssprachliches und mindestens ein zielsprachliches Segment vorhanden sind. Dazu kommen teilweise konfigurierbare Informationseinheiten, zum Beispiel zum Autor, Erzeugungsdatum, Fachgebiet und andere.

Editor: Anwendung, in der die Übersetzung angefertigt wird. Je nach Produkt kann der Editor integriert sein oder auf eine externe Anwendung zur Textverarbeitung zurückgreifen (in der Regel MS Word).

Formatfilter: Diese Filter ermöglichen die Bearbeitung unterschiedlicher Formate im Editor. Ihre Aktualisierung seitens der Hersteller, die saubere Trennung von Inhalt und Format sowie deren korrekte Zusammenführung nach der Übersetzung und schließlich ihre Konfigurierbarkeit seitens der Anwender können entscheidende Kaufkriterien sein.

Alignment-Komponente: Sie dient zum Aufbau eines Übersetzungsspeichers auf der Basis von Paralleltexten, das heißt Paare von ausgangssprachigen und zielsprachigen Segmenten, wenn ihre Übersetzung ohne CAT-System angefertigt wurde.

[3] Ein Beispiel liefert das Produkt SDL Trados: Das Paket beinhaltet unter anderem die Anwendung zur Terminologieverwaltung SDL Trados MultiTerm. Diese Anwendung kann jedoch auch eigenständig verwendet werden, ohne jegliche Anbindung an das Translation Memory.

[4] Für diesen Artikel, wenn nicht anders angegeben, wurden Across Personal Edition (4.00), Déjà Vu X Professional (7.5), Heartsome Translation Studio (7.0), MemoQ (3.2), MultiTrans (4.3), SDL Trados Freelance (8.3.0), STAR Transit NXT (Informationen entnommen aus dem Benutzerhandbuch) sowie Wordfast (5.5) berücksichtigt.

[5] Zum Beispiel bei STAR Transit. Auch die so genannten TextBases von MultiTrans, eines Produktes des kanadischen Herstellers MultiCorpora, weisen gewisse Ähnlichkeiten mit diesem Ansatz auf.

KOMPONENTE	FUNKTION
Translator's Workbench	Übersetzungsspeicher
TagEditor	Integrierter Editor
Filter Settings u.a.	Formatfilter
WinAlign	Alignment Komponente
Synergy	Projektmanagement-Anwendung
MultiTerm	Terminologie-Datenbank

Tabelle 1: Komponenten eines CAT-Systems

SYSTEM	MÜ-INTEGRATION
Across	Ja
Heartsome	Ja (über API)
Déjà Vu	Ja (über API)
SDL Trados	Ja
Transit	Ja
Wordfast	Ja

Tabelle 2: Technische Integration von CAT und MÜ

Projektmanagement-Anwendung: Viele Produkte bieten auch eine Software, welche die Verwaltung (im weiteren Sinne) der Übersetzungsprojekte vereinfacht. Bei Unternehmenslösungen sind die Funktionalitäten dieser Komponenten erwartungsgemäß besonders ausgebaut.

Terminologie-Datenbank: Analog zu dem Übersetzungsspeicher beinhaltet sie die ein- oder mehrsprachigen Terminologieeinträge. Je nach Produkt kann sie eine einfache Liste sein oder mit zusätzlichen Feldern und Informationen (einschließlich Bilder) versehen werden.

Ein Beispiel anhand eines marktüblichen Produkts (SDL Trados) liefert Tabelle 1.

Abschließend folgt eine alphabetische Auflistung der gängigsten – keineswegs aber aller erhältlichen – Systeme (vgl. Lagoudaki (2006)): Across, Déjà Vu, Heartsome, MemoQ, MultiTrans, SDL Trados, Transit, Wordfast.

4 Integration

Die Integration zwischen CAT-Systemen und MÜ-Systemen hat in den letzten Jahren an Bedeutung gewonnen und wird mittlerweile von mehreren Systemen angeboten, wie Tabelle 2 zeigt (aus Massion (2008)):

CAT-System	MÜ-System
Across	Language Weaver
SDL Trados	SDL, Systran, Logos
Transit	Logos, Reverso, Systran
Wordfast	Systran, Power Translator ...

Tabelle 3: Zusammenarbeit zwischen Herstellern von CAT und MÜ-Systemen

Allerdings ist nur bei einigen Systemen eine direkte Integration vorhanden. In den übrigen Fällen muss die Interaktion über eine API-Implementierung (Application Programming Interface) erfolgen, also über eine Programmierschnittstelle, die von einer Software zur Verfügung gestellt wird, und mit deren Hilfe andere Programme an die Software angebunden werden können.

Die Integration über API birgt die Gefahr, dass bei Software-Updates die Anbindung nicht mehr funktioniert. Damit solche Kompatibilitätsprobleme gelöst werden können, ist eine kontinuierliche Pflege der API notwendig.

Die Integration der Systeme hat auch die Zusammenarbeit zwischen den Herstellern von CAT-Systemen und MÜ-Systemen gefördert. Tabelle 3 im Kapitel 4.1 bietet einen Überblick. Es muss jedoch vorausgeschickt werden, dass bei großen Unternehmenslösungen weitere, im vorliegenden Artikel nicht aufgelistete Integrationen möglich sind.

4.1 CAT- und MÜ-Systeme

Im vorliegenden Artikel wird die Integration ausschließlich aus dem Blickwinkel der CAT-Systeme betrachtet. Es wird nicht darauf eingegangen, welche MÜ-Systeme über Schnittstellen zu CAT-Systemen verfügen.

Seit 2007 besteht eine Partnerschaft zwischen Across Systems GmbH und Language Weaver Inc. Language Weaver ist ein statistisches maschinelles Übersetzungssystem. Die Integrationsmöglichkeit besteht vorerst nur für den Across Language Server, also nicht für die Across Einzelplatzversion. Die Anbindung erfolgt über eine Dynamic Link Library (DLL).

SDL Trados verfügt über eine Exportfunktion, die ein mit Systran bzw. Logos kompatibles Format bietet. Für nähere Informationen siehe 5.1.2. Die Firma SDL verfügt außerdem über eine eigene Lösung, den SDL Enterprise Translation Server. Seit der Version 8.3 bietet SDL Trados auch für Einzelplatzversionen den Zugang über das Internet zu diesem maschinellen Übersetzungssystem.

Abbildung 1: Transit Ressourcenleiste

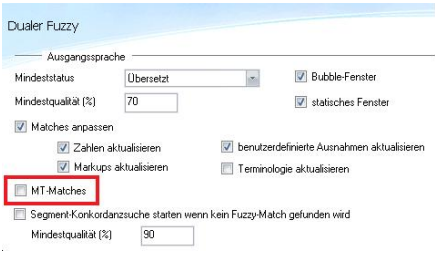


Abbildung 2: Abzug für MÜ-Matches in Transit NXT

Mit der neuen Version NXT[6] bietet Transit eine erweiterte Integration mit MÜ-Systemen (vgl. StarAG (2008)), die kundenspezifisch erfolgen wird. Durch die Schaltfläche `Masch. Übers.` auf der Ressourcenleiste kann eine Verbindung zum gewünschten MÜ-System hergestellt werden (siehe Abbildung 1).

Standardmäßig behandelt Transit NXT die Übersetzungen aus einem MÜ-System gesondert, indem sie nicht automatisch übernommen werden. Dies lässt sich in den Grundeinstellungen der Funktion `Dualer Fuzzy` durch die Option `MT Matches` ändern, siehe Abbildung 2.

Wird die Option aktiviert, werden 100% Matches, die auf maschinellen Übersetzungen gründen, automatisch übernommen.

Unter den CAT-Systemen stellt Wordfast in gewisser Hinsicht einen Spezialfall dar. Dieses Produkt ist keine eigenständige Anwendung, sondern ein Add In für MS Word.[7] Aus diesem Grund kann Wordfast prinzipiell mit allen MÜ-Systemen kommunizieren, die ebenfalls zumindest auch als Add-In in MS Word arbeiten können.

Die Marktentwicklung belegt das wachsende Interesse zum Thema Integration. Weitere Hersteller von CAT-Systemen werden solche Funktionalitäten künftig auch anbieten, zum Beispiel MultiCorpora. In der Version 4.4 von MultiTrans wird eine Integration mit @promt und Systran möglich sein.

Die Integration zwischen Heartsome und Asia Online war zum Redaktionsschluss noch nicht fertig. Sie wird über eine API erfolgen.

[6] Transit NXT ist Ende 2008 zu Redaktionsschluss auf den Markt gekommen. Aus diesem Grund waren für den vorliegenden Artikel nur allgemeine Informationen verfügbar und das Produkt wird nicht näher beschrieben.
[7] Mit der Version 6.0 bietet Wordfast eine eigenständige Anwendung.

4.2 EBMT-Methoden in CAT-Systemen

Etliche CAT-Systeme integrieren EBMT-Methoden, um ihre Retrieval-Leistung zu verbessern.

Déjà Vu bietet eine eigene Funktion namens EBMT[8], welche unter bestimmten Bedingungen eingreift:

Das Ausgangssegment im Translation Memory und das im Text unterscheiden sich lediglich durch einen Terminus. Beispiel: „Hauptschalter ausschalten“ und „Näherungsschalter ausschalten“.

Beide Termini sind in der Terminologiedatenbank und verfügen über eine Übersetzung in die Zielsprache: „main switch“ und „proximity switch“.

Im Translation Memory ist schon ein Fuzzy-Match für das Ausgangssegment vorhanden: „Switch off the main switch“.

Unter diesen Bedingungen ersetzt Déjà Vu automatisch die Übersetzung des alten Terminus im Fuzzy-Match durch die Übersetzung des neuen Terminus: „Switch off the proximity switch“. Dadurch wird das Fuzzy-Match zu einem Perfect Match. Dabei wird jedoch keine grammatische Anpassung vorgenommen, welche unter Umständen notwendig sein könnte.

Genau über die gleiche Funktion wie EBMT von Déjà Vu verfügen auch Transit – sie wird aber *Terminologie aktualisieren* benannt – sowie Heartsome – mit dem Namen *Quick Translation*.

In Déjà Vu ist EBMT ein Teil der Funktion *Assemble*, die Subsegmente zusammenführen kann, um die Übersetzung eines Segmentes anzubieten.

Über eine mit Assemble vergleichbare Funktion verfügt MemoQ: Die Übersetzung kann ebenfalls aus Fragmenten zusammengeführt werden, wenn für das ganze Segment keine Entsprechung gefunden werden konnte.

Dies zeigt, dass die Zusammenführung unterschiedlicher Technologien in eine Anwendung ebenfalls möglich ist.

5 Prozesse

Die Prozessintegration von CAT-Systemen und MÜ-Systemen kann sich auf verschiedene Weise realisieren. Zunächst muss definiert werden, welche Software im Mittelpunkt des Prozesses steht: Im vorliegenden Artikel werden die CAT-Systeme betrachtet, möglich sind aber auch die MÜ-Systeme. Für eine nähere Beschreibung siehe Geldbach and Seewald-Heeg (2006).

[8] Der Begriff EBMT wird im Benutzerhandbuch von Déjà Vu verwendet. Auf die Frage, ob diese Benennung für diese Funktion optimal ist, wird nicht eingegangen.

Selbst wenn das CAT-System im Mittelpunkt steht, sind unterschiedliche Prozesse denkbar. In diesem Artikel werden zwei Varianten beschrieben. Die Qualität der gelieferten maschinellen Übersetzung wird nicht thematisiert.

5.1 Sukzessive Bearbeitung

In diesem Prozess bearbeiten das CAT- und das MÜ-System die Dateien sukzessive. Eine gleichzeitige Bearbeitung durch die Interaktion von beiden Systemen ist nicht möglich.

Im Allgemeinen lässt sich der Prozess folgendermaßen skizzieren: Die Ausgangsdatei wird durch das CAT-System vorbereitet. Nach der Übersetzung durch das MÜ-System erfolgen die Korrektur des maschinell übersetzten Textes sowie die etwaige Vervollständigung seitens des Humanübersetzers wieder im CAT-System, aus dem die Zieldatei erzeugt wird.

5.1.1 Across

Dank der Partnerschaft mit Language Weaver Inc. bietet Across die Integration mit einem MÜ-System. Es ist zwar kein Export aus der Anwendung notwendig (vgl. 5.1.2 SDL Trados), trotzdem erfolgt die Bearbeitung durch das MÜ-System und das CAT-System in zwei getrennten Schritten.[9]

Die Bearbeitung von Übersetzungen erfolgt in Across projektbasiert. Ein Schritt der Projektvorbereitung ist die Vorübersetzung. Der Text wird mit dem Übersetzungsspeicher abgeglichen und die Übersetzungen oberhalb einer definierten Ähnlichkeitsgrenze – unter `Tools > Profileinstellungen > crossTank > Erweiterte Einstellungen > VorÜbersetzung (ab)` – werden ins Dokument eingefügt. Wenn der Übersetzer mit seiner Tätigkeit beginnt, liegt das Dokument zum Teil schon in der Zielsprache vor, es sei denn, aus dem Übersetzungsspeicher kam kein einziges Match. Die Vorübersetzungsfunktion ist auch in den anderen CAT-Systemen vorhanden.

Bei der Vorübersetzung in Across kann Language Weaver für diejenigen Matches verwendet werden, die keine 100% Matches sind. Das heißt, nur die perfekten Treffer werden aus dem Übersetzungsspeicher genommen. Der Rest wird hingegen mittels Language Weaver übersetzt. Diese Übersetzungen von Language Weaver werden nicht unmittelbar in den Übersetzungsspeicher aufgenommen. Sie sollen vom Übersetzer geprüft und bestätigt werden. Ein spezielles Symbol dient zur Unterscheidung dieser Übersetzungen von jenen aus dem Übersetzungsspeicher. In der Analyse (Report) für das Dokument (oder die Dokumente) werden diese automatisch übersetzten Segmente gesondert ausgewiesen.

[9] Damit die maschinelle Übersetzung mit Language Weaver eingesetzt werden kann, müssen gewisse Vorarbeiten geleistet werden, insbesondere rund um das Trainingskorpus für das System. Eine Beschreibung dieser Vorarbeiten ist außerhalb des Rahmens dieses Artikels. Alle Informationen über Across sind von Keller entnommen.

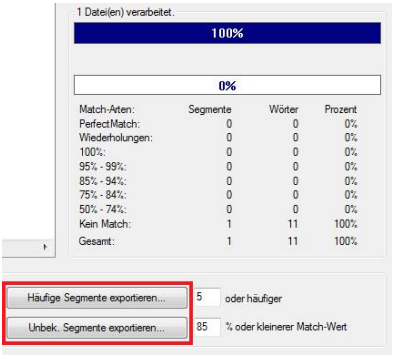


Abbildung 3: Analyse mit Export-Funktionen in SDL Trados

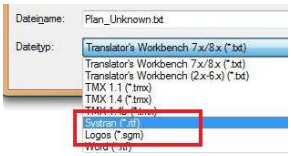


Abbildung 4: Verfügbare Exportformate in SDL Trados

5.1.2 SDL Trados

Das folgende Beispiel basiert auf SDL Trados Freelance, das heißt die Einzelplatzversion von SDL Trados für selbstständige Übersetzer. Bevor die Übersetzung einer Ausgangsdatei begonnen wird, ist die Analyse der Ausgangsdatei ein zwar nicht notwendiger aber üblicher Schritt. Die Analyse ist ein Abgleich zwischen der Ausgangsdatei und dem Übersetzungsspeicher und beziffert, wie viele Segmente des zu übersetzenden Textes im Übersetzungsspeicher schon enthalten sind.[10]

Nach Abschluss der Analyse wird in SDL Trados ein Fenster mit den Analyseergebnissen (Abbildung 3) angezeigt:

Mit den Funktionen `Häufige Segmente exportieren` und `Unbek. Segmente exportieren` lassen sich die gewünschten Segmente aus der Datei extrahieren und exportieren. Dieser Schritt löst den Text als Einheit auf. Als Dateitypen für den Export stehen unter anderem zwei MÜ relevante Formate zur Verfügung (Abbildung 4):

Die gewünschten Segmente werden in ein Format exportiert, das durch Systran bzw. Logos bearbeitet werden kann. Die Bearbeitung im jeweiligen MÜ-System kann in

[10] Auf eine genauere Beschreibung einer Analyse wird verzichtet. Sie kann, je nach Konfiguration und CAT-System, auch weitere Informationen liefern wie z.B. interne Wiederholungen.

```
<TrU>
<CrD>15122005, 09:34:52
<CrU>MT!
<Att L=Kunde>
<Txt L=Auftragsnummer>
<Seg L=DE-CH>Stabiles und sicheres
<Seg L=IT-CH>Strumento di lavoro st
</TrU>
```

Abbildung 5: Attribut-Wert maschinell übersetzter Segmente in Translator's Workbench

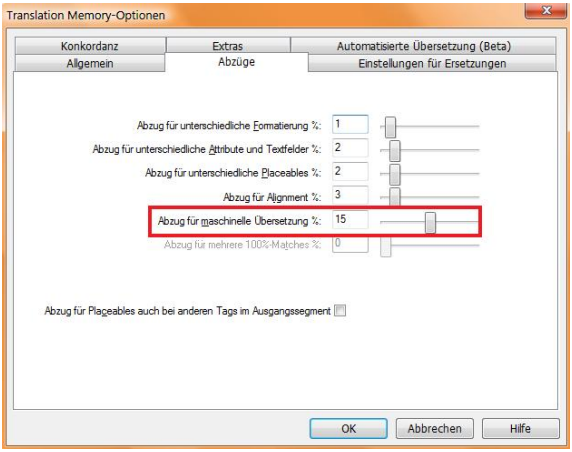


Abbildung 6: Abzüge für maschinell übersetzte Segmente in Translator's Workbench

diesem Artikel nicht behandelt werden und kann im Detail in der Dokumentation des jeweiligen Programms nachgeschlagen werden. Für einen Überblick über Systran siehe Geldbach and Seewald-Heeg (2006).

Nach der Übersetzung im MÜ-System werden die Segmente in den Übersetzungsspeicher importiert und dabei automatisch mit einem speziellen Attributwert versehen. Der Import erfolgt über `Datei > Import` im Hauptfenster von Translator's Workbench. Das Attributfeld heißt `CrU` (Creation User) und bekommt den Wert `MT!` (Machine Translation), siehe Abbildung 5.

Die Übersetzungseinheiten aus dem MÜ-System können mit einem speziellen Abzug versehen werden, um zum Beispiel die ungeprüfte Übernahme zu vermeiden. Dafür ist es erforderlich, `Optionen > Translation Memory-Optionen > Abzüge` in Translator's Workbench auszuwählen (Abbildung 6).

Alle Abzüge, einschließlich `Abzug für maschinelle Übersetzung`, sind durch Schieber frei konfigurierbar. Wenn Translator's Workbench im Attributfeld `CrU` den Wert

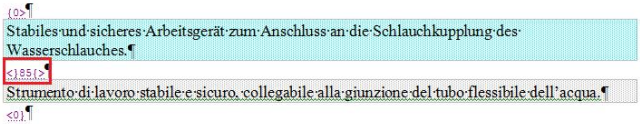


Abbildung 7: Match aus MÜ-System

`MT!` findet, wird der Abzug angewendet. Dies hat wichtige Folgen für die Übersetzung. Verschiedene Szenarien sind vorstellbar:

1. Der zu übersetzende Text ist gleich dem im Übersetzungsspeicher vorhandenen Text.
2. Der zu übersetzende Text ist dem im Übersetzungsspeicher vorhandenen Text ähnlich.
3. Der zu übersetzende Text ist im Übersetzungsspeicher nicht vorhanden.

Das letzte Szenario wird nicht näher beschrieben und ist uninteressant, da der Abzug irrelevant ist. Kein Ähnlichkeitswert kann dabei vermindert werden.

Das erste Szenario würde im Normalfall ein 100% Match zurückgeben, weil kein Unterschied im Text den Ähnlichkeitswert vermindert. Durch den Attributwert `MT!` und die Abzugseinstellung von 15%, wird hingegen nur ein Fuzzy-Match (85%) angeboten (Abbildung 7):

Der graue Hintergrund ist eine Besonderheit derjenigen Fuzzy-Matches, für die ein Abzug wegen ihres Ursprungs aus einem MÜ-System angewendet wurde. Sonst wäre der Hintergrund hellgelb.[11]

Das zweite Szenario unterscheidet sich nur geringfügig vom ersten. Der Ähnlichkeitswert des Fuzzy-Match, zum Beispiel 93%, wird durch den Abzug zusätzlich vermindert und wird 78%. Sollte der Ähnlichkeitswert durch den MÜ-Abzug unter die Mindestähnlichkeitsgrenze fallen, die vom Übersetzer unter `Optionen > Translation Memory-Optionen > Allgemein > Minimaler Match-Wert` in Translator's Workbench eingestellt worden ist, wird das Segment als No Match angeboten.

Nachdem das Segment vom Übersetzer überprüft und gegebenenfalls überarbeitet worden ist, wird es bestätigt und im Übersetzungsspeicher gesichert. Ein zusätzliches Attributfeld wird hinzugefügt, `ChU` (Change User) und bekommt als Wert die Identifikationskennung (User ID) des Übersetzers (Abbildung 8):

Wenn der Ausgangstext wieder in einem zu übersetzenden Text vorkommt, wird er als 100% Match angeboten. Der Abzug greift nicht mehr. Auf diese Weise werden

[11] Standardmäßige Farbeinstellungen, welche angepasst werden können.

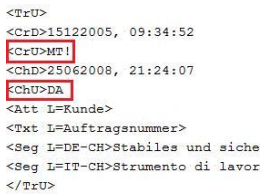


Abbildung 8: Attributfeld Change User

Übersetzungen aus dem MÜ-System nach Überprüfung wie alle anderen Übersetzungen behandelt.

5.2 Gleichzeitige Bearbeitung

Die Dateien werden vom MÜ-System und vom CAT-System gleichzeitig bearbeitet. Eine Interaktion von beiden Systemen ist möglich.

5.2.1 SDL Trados

Ab der Version 8.3 bietet SDL Trados Freelance eine Beta-MÜ-Funktion, die über das Internet arbeitet.[12] Diese Funktion kann über `Optionen > Translation Memory-Optionen > Automatisierte Übersetzung (Beta)` aktiviert werden, siehe Abbildung 9.

Translator's Workbench kontaktiert den SDL Automated Translation Server, der eine maschinelle Übersetzung zurückliefert. Sie wird ins Zielsegment zur weiteren Bearbeitung hinzugefügt. Im Gegensatz zu Wordfast (5.2.2) erfolgt die maschinelle Übersetzung in jedem Fall über das Internet. Auf Vor- und Nachteile dieser Möglichkeit, auch im Hinblick auf Datensicherheit, wird hier nicht eingegangen.

Die maschinelle Übersetzung kann schon bei der Vorübersetzung zum Einsatz kommen, vorausgesetzt, dass der Match-Wert für zu übersetzende Segmente (unter `Extras > Übersetzung` in Translator's Workbench) unterhalb 100% eingestellt ist. In diesem Fall wird für diejenigen Segmente, die kein 100% Match oder Fuzzy-Match haben, eine automatische Übersetzung hinzugefügt. Diese Segmente werden im Vorübersetzungsbericht nicht gesondert ausgezeichnet und gelten als unübersetzt. In der zweisprachigen Datei erscheinen sie aber als 1% Match, siehe Abbildung 10.

Die maschinelle Übersetzung kann außerdem während der interaktiven Übersetzung zum Einsatz kommen. Sie greift nur dann ein, wenn für ein Ausgangssegment weder ein 100% Match noch ein Fuzzy-Match gefunden werden konnte.

[12] Diese Funktion steht nicht für alle Sprachpaare zur Verfügung.

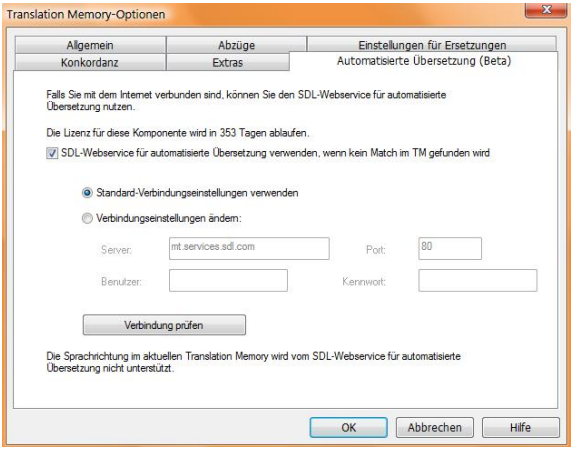


Abbildung 9: SDL Trados Automated Translation

TU <cf bold="on">**Näherungsschalter**</cf> ·auch·Näherungsinitiator·oder·Annäherungsschalter· genannt,·verwenden·Sensoren·die·auf·Annäherung,·das·heißt·ohne·direkten·Kontakt· berührungsfrei·reagieren. 1 Approximation·switch,·also·approximation·initiator·or·approach· switch·named,·use·-freely·react·sensors·that·to·approach,·that·is·without·direct·contact·contacts. TU·TU Näherungsschalter·werden·bei·technischen·Prozessen·zur·Positionserkennung·von·

Abbildung 10: Vorübersetzter Text

Abbildung 11: SDL Trados MÜ-Match

Abbildung 12: Wordfast Symbolleiste in MS Word

Die maschinelle Übersetzung im Zielsegment wird im Editor durch einen grauen Hintergrund gekennzeichnet. Darüber hinaus werden in Translator's Workbench folgende Informationen angezeigt, siehe Abbildung 11: Der Status `Automatisch übersetzt` in der Statusleiste; Der Fuzzy-Wert von `1%`.

5.2.2 Wordfast

Das CAT-System Wordfast ist in MS Word integriert.[13] Wordfast besteht aus einer Word-Vorlage (`wordfast.dot`), welche eine Sammlung von Makros beinhaltet und als Add-In verwendet wird. Ein Add-In ist ein „Programm zum Hinzufügen von [...] Befehlen oder Features" (WORD 2003) in ein Hauptprogramm.

Über solch ein Add-In für MS Word verfügen auch diverse MÜ-Systeme, welche sonst auch als unabhängige Anwendung arbeiten, zum Beispiel translate pro, Systran, Personal Translator, @promt, T1. Dies ist die Voraussetzung für die gleichzeitige Bearbeitung mit Wordfast.

Wordfast verwendet MS Word als Editor, wobei Formate aus anderen Programmen ebenfalls bearbeitet werden können.[14] Wenn Wordfast korrekt installiert worden ist, erscheint in MS Word eine zusätzliche Symbolleiste (Abbildung 12).[15]

Bevor die Übersetzung begonnen werden kann, müssen die Aufrufparameter für das jeweilige MÜ-System in `Wordfast > Setup > MT` definiert werden (Abbildung 13):

[13] In der Version 6.0 von Wordfast ist ein proprietärer Editor verfügbar. Da zum Redaktionsschluss nur eine Prerelease Version verfügbar war, und da das Add-In-Konzept parallel beibehalten wird, wird im vorliegenden Artikel die Version 5.5 beschrieben.

[14] Weitere Formate, die Wordfast 5.5 direkt bearbeiten kann, sind MS Excel, MS PowerPoint, MS Access und getaggte Formate (zum Beispiel aus SDL Trados S-Tagger). HTML, SGML, XML können nur mit +Tools bearbeitet werden. Mit der Version 6 wird die Formatsunterstützung ausgeweitet.

[15] Diese Beschreibung bezieht sich auf Wordfast 5.5 in Verbindung mit MS Word 2003.

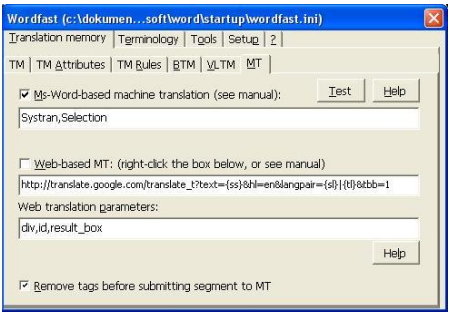


Abbildung 13: MÜ-System in Wordfast

SYSTEM	ABZUG
Across	Nein, aber spezielles Symbol
SDL Trados	Ja
Transit	Ja
Wordfast	Ja

Tabelle 4: CAT-Systeme und Abzug

In Abbildung 13 sind beispielsweise die Aufrufparameter von Systran angegeben. außerdem können auch im Internet verfügbare automatische Übersetzungsdienste verwendet werden (Option `Web-based MT`). Die Suche nach der Übersetzung erfolgt in folgender Reihenfolge: Zunächst wird der Wordfast-Übersetzungsspeicher abgefragt. Wird kein Treffer gefunden, wird eine Übersetzung vom MÜ-System angefordert und in das Dokument eingefügt.

5.3 Auszeichnung im Fokus

Wie in den Kapiteln 4.1, 5.1.1 sowie 5.1.2 bereits erwähnt, verwenden diverse CAT-Systeme Abzüge für die Treffer aus MÜ-Systemen bzw. heben sie gesondert hervor. Damit wird der Übersetzer auf sie hingewiesen und kann sie entsprechend prüfen. Eine Übersicht der Kennzeichnungsmöglichkeiten liefert Tabelle 4.

Across, das sonst ebenfalls einen Ähnlichkeitswert für Fuzzy-Matches angibt, verzichtet auf einen Abzug und zeichnet die Treffer mit einem speziellen Symbol aus.

6 Vorteile und Nachteile

Die Integration von CAT- und MÜ-Systemen bietet wichtige Vorteile der maschinellen Übersetzung. In erster Linie können damit größere Übersetzungsvolumina im Vergleich zur Humanübersetzung bearbeitet werden.

Beachtliche Kosteneinsparungen können ebenfalls erzielt werden. Selbst wenn maschinell übersetzte Texte von Humanübersetzern geprüft werden sollen, werden sie in der Regel als Fuzzy-Match (SDL Trados) oder als Sondertreffer (Across) angeboten. Dafür sind Preisstaffelungen möglich, die günstiger sind als eine Neuübersetzung.

Diesen Vorteilen stehen jedoch auch Nachteile gegenüber. Wenn der Korrekturaufwand für die Übersetzer hoch ist, könnte die beabsichtigte Prozessbeschleunigung nicht erzielt werden. Dabei zeigen Vergleichstests, dass die Korrektur einer schlechten und insbesondere einer mittelmäßigen (maschinellen) Übersetzung mehr Zeit in Anspruch nimmt als eine Neuübersetzung. Für eine detaillierte Behandlung dieses Themas siehe Krings (1998). Der erhöhte Zeitaufwand kann in höheren Kosten resultieren.

Das entscheidende Kriterium für den erfolgreichen Einsatz der kombinierten Lösungen (MÜ-System und CAT-System) ist also die Qualität der maschinellen Übersetzung. Dafür ist ein entsprechender nicht unerheblicher Aufwand einzubringen, wie zum Beispiel in Trojanus (2002) und Geldbach and Seewald-Heeg (2006) beschrieben. Eine unbedachte Integration kann hingegen die Erwartungen enttäuschen.

Literatur

Atril (2003). Madrid.

Carl, M. and Way, A. (2003). Introduction. In *Recent Advances in Example-Based Machine Translation*, pages XVII–XXXI. Kluwer Academic Publisher.

Doug, A. (2003). Why translation is difficult for computers. pages 119–142. John Benjamins Publishing Company, Amsterdam, Philadelphia.

Eberle, K. (2006). Maschinelle Übersetzung - Hopp oder top? In der Dolmetscher und Übersetzer e.V., B., editor, *MDÜ - Fachzeitschrift für Übersetzer*, volume 4 of *9-15*. BDÜ.

Geldbach, S. and Seewald-Heeg, U. (2006). MÜ ist so gut wie ihre Funktionalität, Prä- und Postedition. In *MDÜ - Fachzeitschrift für Übersetzer*, pages 9–15. BDÜ.

Hutchins, J. (2003). Commercial Systems: The state of the Art. In Somers, H., editor, *Computers and Transaltion: a translator's guide*, pages 161–174. John Benjamins Publishing Company, Amsterdam, Philadelphia.

Keller, N. Die Anbindung des MÜ-Systems Language Weaver an Across. Unveröffentlichter Vortrag beim Workshop "Maschinelle Übersetzung – Von der Theorie zur Anwendung" der GLDV, AK Maschinelle Übersetzung. Köthen, Hochschule Anhalt, 4.07.2008.

Krings, H. (1998). *Texte reparieren*. Gunter Narr Verlag.

Lagoudaki, E. (2006). Translation Memories Survey 2006: Enlightning users' perspective.

Lange, C. A. and Bennett, W. S. (2000). Combining Machine Translation with Translation Memory at Baan. In Sprung, R. C., editor, *Translating into Success*, pages 203–218. John Benjamins Publishing Company, Amsterdam, Philadelphia.

Massion, F. (2005). *Translation Memory Systeme im Vergleich*. Doculine Verlag, Reutlingen.

Massion, F. (2008). Integration durch Standards. In *Produkt Global*, pages 22–25. Hüthig, Heidelberg.

Microsoft (2003). Online-Hilfe von Microsoft Word.

Reinke, U. (2003). *Translation Memories: Systeme – Konzepte – Linguistische Optimierung*. Peter Lang Verlag.

Seewald-Heeg, U. (2007). Vielfalt auf dem Markt. 4:12–25.

Somers, H. (2003a). An overview of EBMT. In *Recent Advances in Example-Based Machine Translation*, pages 3–57. Kluwer Academic Publisher.

Somers, H. (2003b). Translation memory systems. pages 31–47. John Benjamins Publishing Company, Amsterdam, Philadelphia.

StarAG (2008). *Transit NXT - Benutzerhandbuch (Vorabversion)*. Ramsen.

Trojanus, K.-H. (2002). Anspruch und Wirklichkeit. In *MDÜ - Fachzeitschrift für Übersetzer*, pages 19–24. BDÜ.

Trujillo, A. (1999). *Translation Engines: Techniques for Machine Translation*. Springer Verlag, London.

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

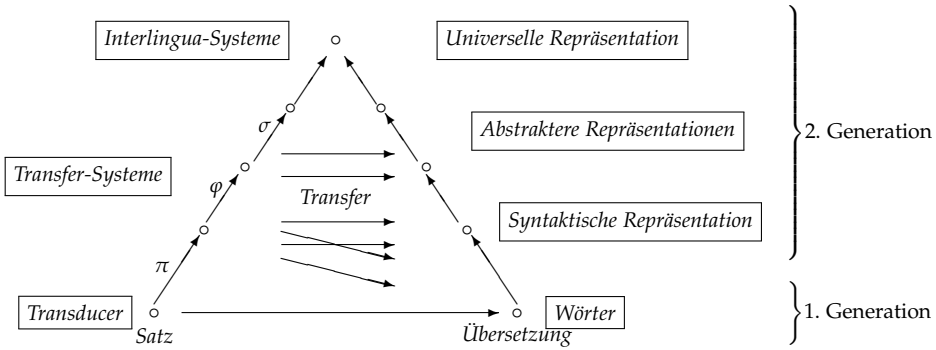


Abbildung 1: Regel-basierte Architekturen (vgl. Vauquois (1975))

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:

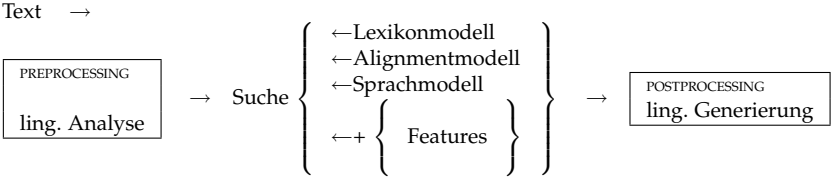
$$\hat{e}_1^I = \underset{e_1^I}{\operatorname{argmax}} \{ \sum_{m=1}^{M} \lambda_m h_m(e_1^I, f_1^J) \}$$

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

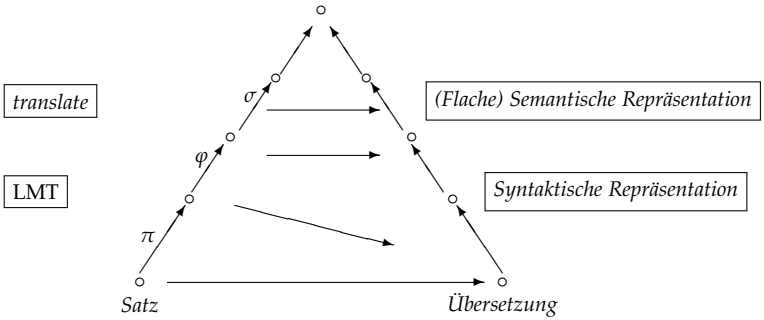


Abbildung 2: Transfer in *translate*

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:

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

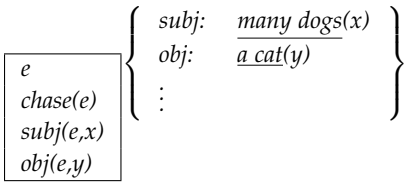
Danach wird eine Struktur, bei der eine Basisrepräsentation (z.B. des Verbs) modifiziert wird, durch eine Reihe von Modifikatoren (z.B. die Repräsentationen der Verbargumente und Adjunkte) in der Weise übersetzt, dass die Übersetzungen der Modifikatoren die Übersetzung der Basisrepräsentation modifizieren, 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) 

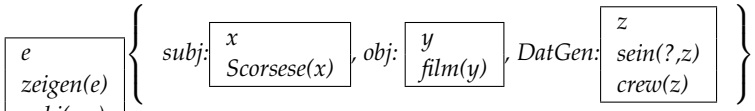
Wenn AC, das hier leer ist, bei der Übersetzung nicht weiter spezifiziert wird, ist die Zielrepräsentation bezüglich der Anwendungsreihenfolge, d.h. hier bzgl. der Skopuslesart, so neutral wie die Ausgangsrepräsentation. D.h. der Default-Transferalgorithmus unterstützt die ambiguitätserhaltende Übersetzung.

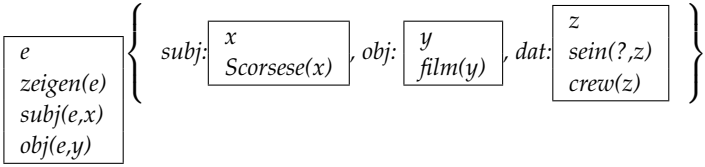
5.4 Partielle Disambiguierung

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*.

Michael Carl

METIS-II: Low-Resource MT for German to English

1 Abstract

METIS-II was a EU-FET MT project running from October 2004 to September 2007, which aimed at translating free text input without resorting to parallel corpora. The idea was to use 'basic' linguistic tools and representations and to link them with patterns and statistics from the monolingual target-language corpus. The METIS-II project has four partners, translating from their 'home' languages Greek, Dutch, German, and Spanish into English.

The paper outlines the basic ideas of the project, their implementation, the resources used, and the results obtained. It emphazises on the German implementation.

2 Introduction

Starting in October 2004, METIS-II was the continuation of METIS-I (IST-2001-32775) Dologlou et al. (2003). Like METIS-I, METIS-II aims at translating free text input by taking advantage of a combination of statistical, pattern-matching and rule-based methods. The METIS-II project has four partners, each translating from their 'home' languages Greek, Dutch, German, and Spanish into English.

The following goals and premises were defined for the project:

1. use 'basic' NLP tools and resources,
2. use bilingual hand-made dictionaries,
3. use a monolingual target-language corpus,
4. use translation units within the sentence boundary,
5. allow different tag sets for SL and TL possible,

Crucially, parallel corpora are not required, and their usage was excluded within METIS-II. The rationale behind this was to develop prototypes of MT systems which would be suitable to translate 'small languages', i.e. language pairs for which parallel texts are difficult to come by. A basic set of NLP tools is nonetheless required for these languages, albeit very basic. The availability of the monolingual target language corpus,

from which statistical language models are computed, makes METIS-II a data-driven MT system. These facts set METIS-II apart from mainstream SMT/EBMT systems.

With these goals and requirements, a number of implementations are possible. The METIS-II partners decided therefore to test and compare various implementations of the ideas, which will be outlined in this paper.

Hence, METIS-II consists of a number of modules which can be investigated horizontally, from source language to target language, or vertically, dividing the task into source-language analysis, lexical transfer, target language word-order generation and word-token generation. While the development of the four horizontal translation directions are to a large extent free-standing and independent efforts of the respective METIS-II partners, the consortium has also developed an exchange and interface format to communicate intermediate (i.e. vertical) processing results between the different parallel modules METIS-II (2006, 2007).

In this paper we aim at presenting METIS-II from a vertical and from a horizontal perspective. We discuss each of the parallel processing steps for all language modules involved, thereby showing their common and diverging characteristics.

The project has used a broad set of tools for source text analysis that were available or else easily obtainable by the partners. The Spanish analysis module experiments on using as few linguistic resources as possible - essentially only a lemmatizer and PoS tagger. The Dutch module adds a shallow parser to detect phrases and clauses while the German module includes also "topological" information. The Greek module seeks a more complete syntactic analysis of input.

The Spanish module uses only a bilingual dictionary that had been extracted from a printed Spanish-English dictionary. The Dutch-English dictionary was also compiled from external sources and the Greek-English dictionary was compiled from preexisting machine-readable dictionaries and augmented manually by the most frequent entries from the Hellenic National Corpus. The German-English dictionary is the largest of all the reported sizes and has been collected from unnamed sources over a long period of time. It covers words and both continuous and discontinuous phrases. Unlike other dictionaries, the German dictionary is preprocessed before use essentially through morphological analysis and generation of variants.

Section 6 describes the main resources used for generation and section 7 explains the way(s) how translations are generated in METIS-II. METIS-II follows a "generation-heavy" approach Habash (2004), where most of the hard translation issues are addressed during the generation phase.

The basic resource for generation are target language models, which are extracted from a huge target language corpus (the BNC) and which assist in selecting — and in some cases also in generating the word order of — the best translations. In this respect, the METIS-II core approach resembles Whitelock's (1991; 1992) 'shake-and-bake' method where the "target texts are constructed from a bag of TL basic expressions,

whose elements are derived from the analysis of the source text and a set of equivalences of basic expressions"(Whitelock, 1991, p:1). However, while Whitelock uses logical and semantic constraints for 'baking' a target text from the basic expressions, METIS-II relies on statistical and pattern-based language models extracted from the target corpus to consolidate and verify target sentences.

Section 6 shows how the target-language corpus was preprocessed and how language models were conceptualized and extracted from the corpus. These models are built in idiosyncratic ways, with significant differences across language pairs. The Spanish module uses sequences of lemma/tag to validate insertions, deletions and permutations of words, the Greek and Dutch modules consolidate TL word order based on patterns and templates and the German module uses statistical *n*-grams.

Section 7 deals with the actual translation, the "decoding" of the source language. The overall translation method in METIS-II is creating a set of possible translation solutions and then using statistical methods to find the most probable translations. The language models play a crucial role in the selection process. Section 9 provides a detailed comparison of the differences and similarities across these modules.

Section 8 presents an evaluation of the translation systems using two test sets, the test suites used during development and a EUROPARL fragment, using BLEU, NIST and TER. Results for each language pair, using a well consolidated system such as Systran, are used as topline reference measure to gauge METIS-II results.

3 Background of METIS-II Implementations

In this section we briefly describe the basic ideas behind the implementations of the four translation directions. A linguistically minimal approach is favoured by the Spanish module, while the other modules employ a shallow parser to detect phrases and clauses. The Dutch and Greek modules assume some kind of structural isomorphism of phrases and clauses between the source and the target language, while the German module employs flat re-ordering rules.

3.1 Spanish to English

The approach followed by the Spanish-to-English METIS-II system strives to use as little linguistic resources as possible. The motivation in this case is not the lack of resources for processing Spanish but the desire to experiment in the leanest possible conditions, so that our findings can be applied to other, possibly smaller languages with fewer resources available. Consistently with this purpose, the preprocessing of the Spanish input requires only a tool able to lemmatize and assign morphological tags to each word of the sentence. The Spanish sentence is thus tokenized, tagged and lemmatized, but it is not chunked or analyzed in terms of constituency.

3.2 Dutch to English

For the Dutch-to-English translation pair was chosen an approach that requires a number of tools in order to perform a shallow source language analysis: a tagger, a lemmatizer, and a shallow parser (including a clause detector). We required the target-language corpus to be preprocessed with the same means, so equivalent tools for the target language are needed off line(Vandeghinste (2008)).

3.3 Greek to English

What is crucial within the Greek-to-English METIS-II approach is the notion of pattern, that is, phrasal segments that serve as the basis for modelling both the source (SL) and the target (TL) languages. The patterns roughly correspond to phrasal constituents of a varying size and type, ranging from clauses to sub-clausal level patterns (chunks and contained tokens). This approach, because it reflects the recursive character of natural language is expected to assist more effectively the translation process. Besides, even within the Statistical Machine Translation paradigm that strictly aimed to avoid using phrasal segments, the potential beneficial role of phrase-based models has now been recognized (Carpuat and Wu (2007)).

3.4 German to English

The German METIS-II architecture uses rule-based techniques to generate a graph of partial translation hypotheses and employs statistical techniques to rank the best translation(s) in their context. Word tokens are generated for the *n*-best translations.

The core idea is similar to Brown and Frederking (1995) who use a statistical English Language Model to combine partial translations produced by three symbolic MT systems. In contrast to their approach, we build the search graph with flat re-ordering rules.

The re-ordering rules generate an acyclic AND/OR graph which allows for compact representation of many different translations. A beam search algorithm tries to find most likely paths in the AND/OR graph. A similar idea for generation was suggested by Langkilde and Knight (1998) who use 2-gram language models to find the best path in a word lattice. Unlike a usual statistical decoder (Germann et al. (2001); Koehn (2004)), our ranker, hence, does not modify the graph and it does not generate additional paths which are not already contained in the graph.

4 Morphological processing

Each of the source languages modules in METIS-II has their individual preprocessing and SL analysis tools which are described in this section. In line with the requirements and philosophy of the project, all language modules use a lemmatizer and PoS tagger to

process the source language input. In addition Dutch and Greek use a shallow parser to detect phrases and clauses and German recognizes topological fields. Besides the source language analysis, we have also implemented a reversible lemmatizer for the target language (English) which was used throughout for generation in METIS-II.

lemma	#	PoS	chunks	clauses
{lu=das,	wnrr=1,	c=w,sc=art,	phr=np;subjF,	cls=hs;vf}
,{lu=haus,	wnrr=2,	c=noun,	phr=np;subj,	cls=hs;vf}
,{lu=werden,	wnra=3,	c=verb,vt=fiv,	phr=vg_fiv,	cls=hs;lk}
,{lu=von,	wnrr=4,	c=w,sc=p,	phr=np;nosubjF,	cls=hs;mf}
,{lu=Hans,	wnrr=5,	c=noun,	phr=np;nosubj,	cls=hs;mf}
,{lu=kaufen,	wnra=6,	c=verb,vt=ptc2,	phr=vg_ptc,	cls=hs;rk}

Table 1: Analysis for the German sentence "Das Haus wurde von Hans gekauft" (The house was purchased from Hans).

The German source-language analysis produces a flat sequence of feature bundles which contain chunking and topological information of the sentence Müller (2004). An example of the German analysis is given in table 1.

Among other things, the analysis comprises of a unique word number, the lemma and part-of-speech of the word, as well as morphological and syntactic information. It also contains chunking and topological information. The parser produces a linguistically motivated, flat macro structure of German sentences, as coded by the `cls` feature.

Within the METIS-II project, we have implemented a reversible lemmatizer for English (Carl et al. (2005)) which reads CLAWS5-tagged words and generates a lemma together with two additional features indicating the orthographic properties (O) and the index of the inflection rule (IR). The IR-index serves to memorize the inflection rule which was applied to generate the lemma. Lemmatization rules are used to strip off or modify regular inflection suffixes from word tokens. Table 2 plots two lemmatization examples. A lemmatization lexicon is used for the irregular cases.

TAG	token	⇔	lemma	TAG_O_IR	IR	suffix mapping		
VVG	sniffing	⇔	sniff	VVG_l_1	1	`ffing`	↔	`ff`
VVG	DRESSING	⇔	dress	VVG_c_3	3	`ssing`	↔	`ss`

Table 2: Left: input and output of lemmatization and token-generation, Right: corresponding bi-directional inflection rule which can be used for lemmatization and for token generation.

The lemmatizer uses a single table of 128 lemmatization rules (two of which are shown on the right side in table 2). Each rule specifies the removal or replacement of an ending, conditionally on the TAG of the word and its suffix. Lemmatization and token generation is 100% reversible: a token set {token,TAG} is equivalent to a lemma

set {lemma,TAG,O,IR} and both sets can be transformed into each other without loss of information, by reversing the lemmatization rule.

However, during token generation, we usually want to produce word forms from incomplete lemma sets {lemma,TAG}, where the inflection rule *IR* is not known. To generate an educated guess which IR would produce the desired word form, we have counted for each lemma suffix the inflection rules which generated the lemma. A word form would then be generated from a lemma by looking at the ending of the lemma and by applying the most likely reversed inflection rule. With slightly more than 20,000 lemma suffixes the reversible lemmatizer achieves a precision of more that 99.5%. In order to achieve this precision we had to add a few additional tags to the original CLAWS5 tagset, and then re-tagged the BNC[1] with the enhanced tagset. Table 2 plots two lemmatization examples.

5 Bilingual Dictionary

Apart from the resources required for the monolingual source language analysis, there are two other types of resources that were used in METIS-II: a bilingual transfer dictionary and the monolingual target-language corpus. For Spanish, Dutch and Greek the dictionary was compiled from external resources and adapted to the needs of METIS-II.

The German-English METIS-II dictionary contains more than 629,000 entries collected over the past 20 years. In its editable form, dictionary entries are represented as full forms and both language sides are independent. That is, a single word can translate into a single word, a phrase or a discontinuous phrase as in table 3. The German verb *einsperren* for instance, translates into a discontinuous English verb *lock* ⟨*so.*⟩ *away*. Entries are coded as flat trees: while the word(s) of the entries represent the leaves of the tree, the features *DE* and *EN* in table 3 are their 'mother nodes', which provide information about the type of the entry.

German	*DE*	English	*EN*
einsperren	verb	lock ⟨so.⟩ away	verb
Anweisung ausführen	verb	execute statement	verb
von ⟨etw.⟩ Kenntnis nehmen	verb	take note of	verb

Table 3: Examples from the German-to-English dictionary

The dictionary undergoes a number of preprocessing steps before the entries can be mapped on a German lemmatized and analysed sentence. The source and the target language sides of the dictionary pass through a multi-layered fully automatic compilation step. For the SL side this involves:

[1]Section 6 gives more information on this corpus.

5.1 Morphological analysis and lemmatization of the 'leaves'

With the lacking context of words in a dictionary, the morphological analyser MPROMaas (1996) provides the following ambiguous readings for the word *ausführen*.

lemma	PoS	agreement	morph. structure
ausführen	noun	sg, acc;dat;nom, neut	aus_$führen
ausführen	verb, fin	plu, 1;3, pres	aus_$führen
ausführen	verb, inf		aus_$führen
ausfahren	verb, fin	plu, 1;3, past, subj	aus_$fahren

The symbol '_$' marks the detachable prefix *aus*, and thus illustrates the structure of the word. These readings are then disambiguated and filtered based on the type of the entry.

5.2 Checking internal consistency of the entries

By means of a set of patterns we control whether the analyses of the words (i.e. the leaves of the entry, as in the table above) are consistent with its type. A dictionary entry is consistent if at least one of its readings can be consolidated by a pattern associated to its type; otherwise the entry will be marked obsolete. This process also disambiguates readings and filter those readings that are intended by its type (e.g keeping only the *verb,inf* reading of *ausführen*). The process makes sure that the representations of the entries are consistent with the analysed words of an input text.

5.3 Variant generation

Variants are generated to extend the coverage of the dictionary for nominal and verbal expressions. A variant is an additional translation relation that covers a different realization of a dictionary entry. The verb *ausführen*, for instance, matches a main-clause verb in a non-compositional tense while the variation *führen . . .aus* matches in a subordinate clause. For nominal expressions morpho-syntactic variation for compounding, as e.g.: *Abfertigung des Gepäcks → Gepäckabfertigung*, but also coordination, and synonyms are generated (Carl and Rascu (2006)).

6 Target Language Modelling

We have experimented with various ways to use the implicit knowledge encoded in the monolingual target language corpus, and generated different language models. All language models are based on the BNC[2]. The BNC is a tagged collection of texts making use of the CLAWS5 tagset which comprises roughly 70 different tags. As pointed out in

[2]The British National Corpus (BNC) consists of more than 100 million words in more than 6 million sentences `http://www.natcorp.ox.ac.uk/`

section 4, to ensure reversibility of the lemmatized forms we had to add a few tags to the tagset and re-tag the BNC accordingly. The re-tagged BNC was then lemmatized before building the language models. For target language modelling there were, thus, three types of information available: (i) the original word form, (ii) the lemma and (iii) the PoS tag of the words.

In the German-to-English module, we have generated statistical *n*-gram language models. The language models (LMs) were generated using the CMU language modelling toolkit[3] or SRILM toolkit. The functions provided with these toolkits were adapted and integrated into a beam search algorithm as described in section 7. We have experimented with the following parameters:

- number of sentences arbitrarily extracted from the BNC:
  - 100K, 1M, 2M and 5M
- different kinds of statistical language models:
  - token-based LM: using the surface word forms
  - lemma-based LM: using the lemmatized word forms
  - tag-based LM: using the CLAWS5 tags
  - lemma-tag co-occurrence statistics
- 3 and 4-gram for token and lemma LMs and 4 to 7-gram CLAWS5-tag LMs

7 Translating with METIS-II

In line with the different philosophies and the variety of resources, decoding works differently for each of the language pairs. This section illustrates how translations are actually produced for German ↔ English.

In the German-to-English approach, rule-based devices generate an acyclic AND/OR graph, which allows for compact representation of many different translations. A statistical beam-search tries to find the best translation in that graph. Starting from a SL sentence, the graph is constructed in three rule-based steps. The graph is then traversed and translations are ranked. Finally word tokens are generated for the *n*-best translations. The architecture consists of the following five steps:

7.1 German SL Analysis

The *Analyser* lemmatizes and morphologically analyses the SL sentence. It produces a (flat) grammatical analysis of the sentence, detecting phrases and clauses and potential subject candidates as described in section 4, table 1.

[3]This toolkit can be downloaded from `http://www.speech.cs.cmu.edu/SLM_info.html`

7.2 Dictionary Lookup

The analysed SL sentence is then matched on the transfer dictionary. The procedure retrieves ambiguous and/or overlapping entries and stores them in the graph. Matching proceeds on morphemes and lemmatized forms and suited to retrieve discontinuous entries, cf. section 5.

Due to the complexity of discontinuous matches, we only allow discontinuous matches for verbal and nominal entries. In Carl and Rascu (2006) we have described various strategies to reject matched entries if they do not obey a predefined set of criteria.

For verbal entries, various permutations of the words are possible, according to whether the entry occurs in a subordinate clause or in a main clause. We use the field and chunk annotation in the German analysis to validate and filter or reject the matched entries. These criteria are further developed in Anastasiou and Culo (2007) making use of the German topological fields.

To account for a maximum number of different contexts, the dictionary generates all translation hypotheses which are then filtered and graded by the *Ranker* in the context of the generated sentence.

7.3 Word-Order Generation

This step inserts, deletes, moves, and permutes items or chunks in the AND/OR graph according to the TL syntax by means of a rule-based device. The rules take into account phrase and clause segmentation of the SL language sentence as well as word grouping resulting from the dictionary lookup. The modifications in the graph are such that each path containes exactly once the translation(s) of all the words of the source language sentence.

As in the so-called "generation-heavy" translation (Habash (2004)), the rules produce numerous partial translation hypotheses. For our German-to-English module we have currently ca. 50 rules, which are described in more detail in Carl (2007). This "symbolic overgeneration" is then constrained by a statistical ranker making use of several statistical feature functions.

7.4 Ranking and Translation Selection

In this step, the AND/OR graph is traversed to find the most likely translations as a path through the graph. Ranking is a beam search algorithm which estimates each node in the path with a set of feature functions (Och and Ney (2002)) and keeps those target sentence $\hat{e}$ with the highest probability according to equation (1).

$$\hat{e} = \texttt{argmax} \sum_n \sum_m w_m h_m(\cdot) \tag{1}$$

In equation (1), h_m is a feature function and w_m is a weighing coefficient, while n is the number of non-overlapping translation units matching the SL sentence (including those inserted or deleted in the generation module). Given the rich annotation of our data, there are numerous possibilities for the selection of feature functions, some of which are described in section 6. In the METIS-II evaluations reported in sections 8 we compare different ways to compute translation units and their mapping into the target language.

7.4.1 Token Generation

This step (cf. section 4) generates surface word-forms from the lemmas and PoS tags.

8 Evaluation of METIS-II

The evaluation of METIS-II was performed on two test sets, one consisting of data that had been used throughout the project for development purposes and one consisting of unseen data gathered from a previously existing bilingual corpus (Vandeghinste et al. (2008)). To measure results we used BLEU (Papineni et al. (2002)), NIST (Doddington (2002)) and TER (Snover et al. (2006)). The first two metrics measure edit distance using n-grams, while TER (Translation Error Rate) measures the amount of editing that a human would have to perform to get the translation right.

Each language group constructed a development set consisting of 200 sentences, with material evenly distributed among four different categories: 56 sentences illustrating grammatical phenomena (defined by each site), 48 sentences from newspapers; 48 sentences from encyclopedia articles, or similar sources of non-specialized texts, which provides a homogeneous evaluation framework. We compared Metis translations of this set with Systran translations. Systran is a syntactic transfer, rule-based MT system that has been under development since 1968, with a huge amount of funding from companies and institutions and large development teams. It uses large repositories of rule sets, large dictionaries, full parsers, elaborated algorithmic principles, etc. METIS-II, on the other hand, has been built in 3 years within 4 university groups, as an exploratory effort to build a hybrid MT system with no parallel corpus. Its architecture and components have been subject to much experimentation during the process. It is therefore reassuring that its results, though clearly worse than those obtained with Systran, stand up to the comparison.

In table 4 we plot the results of the German-to-English METIS-II system in two different experimental settings.

In the first experiment (METIS-II$_1$), we used a basic set of generation rules (cf. section 7). In the second experiment (METIS-II$_2$), we further developed and refined some generation rules for handling adverbs and negation particles, such as ‘never’, ‘usually’,

	Development set			Europarl test set	
	METIS-II$_1$	METIS-II$_2$	Systran	METIS-II$_2$	Systran
BLEU	0.186	0.223	0.313	0.282	0.396
NIST	5.48	5.32	6.36	6.68	8.05
TER	—	—	—	55.97	42.93

Table 4: DE-EN results for METIS-II and Systran on the Development and the Europarl test set.

extraposition of prenominal adjectives (e.g., "der vom Baum gefallene Apfel" would become "The apple fallen from the tree"), and "um ... zu" constructions. In the ranker we used lemma language models with 3 and 4-grams and tag language models with 4, 5, 6, and 7-grams. We varied weights between 0.01 and 10 for each of the feature functions and kept the combination which provided the best results. This setting was also used to evaluate the Europarl test set. The public version of Systran (Babelfish), however, performs even better than our best setting.

	Europarl	development	difference
NL-EN	0.1925	0.2369	0.0444
DE-EN	0.2816	0.2231	-0.0585
EL-EN	0.1861	0.3661	0.1800
ES-EN	0.2784	0.2941	0.0157

Table 5: Cross-language results on the development and Europarl test set (BLEU).

Table 5 shows that ES-EN is the system that has the most stable performance across test sets, while EL-EN shows the greatest variation. The most surprising result is DE-EN's, which performs better on the Europarl corpus than on the development set. A partial explanation may be that DE-EN has used Europarl type of text to tune lexical weights. Also, the DE-EN development set was chosen to contain hard translation problems so that also Systran performs more poorly on it than on the Europarl test set.

9 Comparison of decoders

This section resumes and compares the characteristics of the METIS-II decoders by looking at how hypotheses about TL word order are generated and how the most likely translation is selected.

9.1 Greedy vs. exhaustive translation modelling

Spanish, Dutch and Greek follow an incremental, non-monotonic approach to 'shake-and-bake', where the target sentence is piece by piece constructed from portions of the 'bag of TL expressions' (Whitelock (1991)) and each portion is in itself locally validated through the target language model. In contrast, the German decoder first produces all possible translation hypotheses in a compact graph representation and then uses language models and a beam searcher to select the best translation as a path through the graph.

9.2 Algorithmic vs. rule-heuristic word re-ordering

The Dutch and German modules employ rules to generate hypotheses of possible TL word-order — particularly for long distance movements. Spanish and Greek chose an algorithmic way to permute TL expressions. The latter approach has potentials of making the systems more language independent, while it is hard to correctly produce discontinuous translation in an algorithmic manner, which seems to be particularly important for Dutch and German.

9.3 Isomorphism vs. local changes

Dutch and Greek assume structure-isomorphism of phrases and clauses in the source and target language, while Spanish and German rely on local re-arrangements of the TL expressions. The former method requires a synchronization of the source- and target language resources, while for the latter, in principle, SL and TL resources may be processed and prepared independently.

9.4 SL vs. TL information for word order hypotheses

Permutation and re-arrangement of TL expressions for the German module is based exclusively on SL information from which these expressions were derived, while for Spanish TL word order hypotheses are based only on the TL information of the expressions. Due to the isomorphism assumption, the Dutch and Greek modules hypothesize TL word order based to some extend on the correlation of SL and TL information.

9.5 Top-down vs. bottom-up vs. flat re-ordering

The Greek module generates translations top-down by applying first larger, more abstract clause pattern models and then establishing the correct word order within each chunk. The Dutch module proceeds bottom-up incrementally consolidating word order from lower level phrases to higher level phrases. Spanish and German use flat re-ordering rules.

10 Conclusions

The paper reports on the underlying ideas, implementation and results of the EU-FET MT project METIS-II running from October 2004 to September 2007. METIS-II aimed at translating free text input using basic linguistic resources and a monolingual target language corpus.

With only a limited amount of work (about 12 man years) we have developed four language pairs, Dutch, German, Greek and Spanish into English. While results of METIS-II are not as good as a well-established MT system such as Systran, which we have chosen as topline reference, they can be considered of an acceptable quality. The paper shows that METIS-II provides a solid framework that can be easily adapted to new language pairs, that can be tuned to particular domains, and that can be upgraded with additional resources as they become available.

The paper describes the language processing tools and bilingual dictionaries of METIS-II which rely on shallow linguistic representations. Within METIS-II we have developed and explored various innovative language models and the paper points out how the models are exploited during translation. While we also give a comparative evaluation of the modules, we feel it is too early to draw ultimate conclusions on the best parameter settings.

We view METIS-II in the bigger context of self-learning systems that learn to translate from textual resources. Instead of learning relations between surface word forms, we maintain that the learned parameters must include linguistic properties of words and sentences for the system to tackle the hard problems of machine translation. Appropriate adaptive and dynamic representation of these parameters together with suitable reasoning mechanisms will ultimately help overcome the shortcomings of today's SMT systems. METIS-II has explored some of the possible avenues, and pointed to further directions that can be followed.

References

Anastasiou, D. and Culo, O. (2007). Using Topological Information for detecting idiomatic verb phrases in German. In *Proceedings of the Conference on Practical Applications in Language and Computers (PALC)*, pages 49–58, Lodz, Poland.

Brown, R. and Frederking, R. (1995). Applying statistical English language modelling to symbolic machine translation. In *Proceedings of the 6th International Conference on Theoretical and Methodological Issues in Machine Translation (TMI)*, pages 221–239, Leuven, Belgium.

Carl, M. (2007). METIS-II: The German to English MT System. In *Proceedings of the 11th Machine Translation Summit*, Copenhagen, Denmark.

Carl, M. and Rascu, E. (2006). A dictionary lookup strategy for translating discontinuous phrases. In *Proceedings of the European Association for Machine Translation*, pages 49–58, Oslo, Norway.

Carl, M., Schmidt, P., and Schütz, J. (2005). Reversibl1e Template-based Shake & Bake Generation. In *Proceedings of the Example-Based Machine Translation Workshop held in conjunction with the 10th Machine Translation Summit*, pages 17–26, Phuket, Thailand.

Carpuat, M. and Wu, D. (2007). How Phrase Sense Disambiguation outperforms Word Sense Disambiguation for Statistical Machine Translation. In *Proceedings of the 11th International Conference on Theoretical and Methodological Issues in Machine Translation (TMI-07)*, pages 43–52, Skövde, Sweden.

Doddington, G. (2002). Automatic Evaluation of Machine Translation Quality using N-gram Co-occurrence Statistics. In *Proceedings of the second Human Language Technologies Conference (HLT-02)*, pages 128–132, San Diego.

Dologlou, I., Markantonatou, S., Tambouratzis, G., Yannoutsou, O., Fourla, A., and Ioannou, N. (2003). Using monolingual corpora for statistical machine translation. In *Proceedings of EAMT/CLAW 2003*, pages 61–68, Dublin, Ireland.

Germann, U., Jahr, M., Knight, K., Marcu, D., and Yamada, K. (2001). Fast Decoding and Optimal Decoding for Machine Translation. In *Proceedings of the 39th ACL and 10th Conference of the European Chapter*, pages 228–235, Toulouse, France.

Habash, N. (2004). The use of a structural n-gram language model in generation-heavy hybrid machine translation. In *Proceeding 3rd International Conference on Natural Language Generation (INLG '04), volume 3123 of LNAI, Springer*, pages 61–69.

Koehn, P. (2004). Pharaoh: a Beam Search Decoder for Phrase-Based Statistical Machine Translation Models. In *Proceedings of AMTA, the Association for Machine Translation in the Americas*, pages 115–124, Washington, DC, USA.

Langkilde, I. and Knight, K. (1998). The Practical Value of n-grams in generation. In *In Proceedings of the 9th International Natural Language Workshop (INLG '98)*, Niagara-on-the-Lake, Ontario.

Maas, H.-D. (1996). MPRO - Ein System zur Analyse und Synthese deutscher Wörter. In Hausser, R., editor, *Linguistische Verifikation, Sprache und Information*. Max Niemeyer Verlag, Tübingen.

METIS-II (2006). Validation/Evaluation framework. Public Report, D5.1, European Commission, FP6-IST-003768, Brussels. `http://www.ilsp.gr/metis2/files/Metis2_D5.1.pdf` [25.Aug.2008].

METIS-II (2007). Validation & Fine-Tuning Results for the first Prototype. Public Report, D5.2, European Commission, FP6-IST-003768, Brussels. `http://www.ilsp.gr/metis2/files/Metis2_D5.2.pdf`[25.Aug.2008].

Müller, F. H. (2004). *Stylebook for the Tübingen Partially Parsed Corpus of Written German (TÜPP-D/Z)*. `http://www.sfb441.uni-tuebingen.de/a1/pub.html`[25.Aug.2008].

Och, F. J. and Ney, H. (2002). Discriminative Training and Maximum Entropy Models for Statistical Machine Translation. In *Proceedings of the 40th annual ACL Conference*, pages 295–302, Philadelphia, PA.

Papineni, K., Roukos, S., Ward, T., and Zhu, W. (2002). BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th ACL*, pages 311–318.

Snover, M., Dorr, B., Schwartz, R., Micciulla, L., and Makhoul, J. (2006). A study of translation edit rate with targeted human annotation. In *Proceedings of Association for Machine Translation in the Americas (AMTA 2006)*, pages 223–231.

Vandeghinste, V. (2008). A Hybrid Modular Machine Translation System. Phd thesis, Netherlands Graduate School of Linguistics.

Vandeghinste, V., Dirix, P., Schuurman, I., Markantonatou, S., Sofianopoulos, S., Vassiliou, M., Yannoutsou, O., Badia, T., Melero, M., Boleda, G., Carl, M., and Schmidt, P. (2008). Evaluation of a Machine Translation System for Low Resource Languages: METIS-II. In *Proceedings of the Sixth International Language Resources and Evaluation (LREC)*, page 96, Marrakech, Morocco.

Whitelock, P. (1991). *Shake-and-Bake Translation*. Unpublished Draft.

Whitelock, P. (1992). Shake-and-Bake Translation. In *Proceedings of the COLING92*.

Heribert Härtinger

Textsortenbezogene linguistische Untersuchungen zum Einsatz von Translation-Memory-Systenen an einem Korpus deutscher und spanischer Patentschriften

Patentschriften stellen eine häufig übersetzte Textsorte dar, zählen aber trotz des hohen Grades ihrer sprachlichen Standardisierung bislang nicht zu den typischen Einsatzgebieten von CAT-Tools. Die hier vorgestellte Studie untersuchte an einem Korpus deutscher und spanischer Patentschriften den Zusammenhang zwischen linguistischen Textsortenmerkmalen und dem Einsatznutzen integrierter Übersetzungssysteme. Im Mittelpunkt der Untersuchung standen die Analyse textsortentypischer Rekurrenzmuster mit Blick auf die erwartbaren Konsequenzen für die Retrieval-Effektivität kommerzieller Translation-Memory-Systeme sowie die Frage nach Textsortencharakteristika, die sich auf die Verwertbarkeit der Suchergebnisse auswirken können. Das zweisprachige, nach den Erfordernissen der Fragestellung ausgewählte Korpus bestand aus 60 vollständigen Textexemplaren und diente sowohl der Registrierung textinterner und textexterner Rekurrenzen als auch der Bewertung ihrer Retrieval-Relevanz anhand exemplarischer Satzinhaltsvergleiche. Die Analyse erfolgte aus der Perspektive einer integrierten Übersetzungsumgebung mit der Möglichkeit der Konkordanzsuche und eingebundener terminologisch-phraseographischer bzw. textographischer Datenbank, so dass auch textsortentypische Rekurrenzen unterhalb der Satzgrenze im Ergebnis berücksichtigt werden konnten. Als Testsoftware diente die *Translator's workbench* der Firma SDL/Trados.

1 Einleitung

Die Effizienz von Translation-Memory-Werkzeugen hängt von personen- und systembezogenen Parametern (Arbeitsstil des Übersetzers, linguistische Leistungsfähigkeit des Systems) und insbesondere von textbezogenen Faktoren ab (Reinke (2004)). Textbezogene Faktoren mit entscheidendem Einfluss auf die Effizienz der Systeme sind die terminologische und stilistische Konsistenz der Ausgangstexte (Glover and Hirst (1996), Merkel (1996)) und insbesondere die Häufigkeit, mit der sich Sätze, Teilsätze und längere Syntagmen innerhalb des zu übersetzenden Textes wiederholen (textinterne Rekurrenzen) oder bereits im Referenzmaterial des Übersetzungsspeichers vorhanden sind (textexterne Rekurrenzen).

Bestimmend für die Retrieval-Effektivität von TM-Systemen ist neben dem quantitativen Parameter des Rekurrenzgrades aber auch die inhaltliche Frage, in welchem Maße ein im Speicher aufgefundenes Segment dem Informationsbedürfnis des Übersetzers Rechnung trägt. Den Grad der Übereinstimmung eines nachgewiesenen AS-/ZS-Segmentpaares mit den übersetzerischen Informationsbedürfnissen bezeichne ich mit Reinke (1999) in Analogie zum informationswissenschaftlichen Relevanzbegriff als *Retrieval-Relevanz*. Das qualitative Kriterium der Relevanz muss zwar keineswegs mit dem Grad der formalen Übereinstimmung korrelieren, bestimmt aber den Formulierungs- bzw. Rekontextualisierungsaufwand des Übersetzers. Reinke (2004) weist daher zu Recht auf die geringe Aussagekraft einer bloßen quantitativen Evaluierung der Retrieval-Effektivität von TM-Systemen unter Anwendung gängiger informationswissenschaftlicher Kenngrößen (v. a. *Recall* und *Precision*) hin und schlägt ein System von Ähnlichkeitskriterien für eine qualitative Bewertung der Retrieval-Ergebnisse vor.

Sowohl die Rekurrenzquoten als auch die Relevanz der Suchergebnisse sind in hohem Maße textsortenabhängig. Dennoch liegen bislang nur vereinzelte textsortenspezifische Untersuchungen (z. B. Brungs (1996) vor, die der Frage nach der Effizienz von Translation-Memory-Systemen durch die Analyse textsortentypischer linguistischer Parameter nachgehen.

2 Zielsetzung

Die hier vorgestellte Studie zielte darauf ab, durch linguistische Untersuchungen an einem zweisprachigen Textkorpus zu verallgemeinerbaren Aussagen in Bezug auf den Nutzen von Translation-Memory-Systemen bei der Übersetzung von Patentschriften zu gelangen. Die Fragestellung bewegt sich somit im Schnittfeld von intra- und interlingual orientierter korpusbasierter Fachsprachenlinguistik einerseits und übersetzungstechnologischen Fragestellungen andererseits und nimmt dabei eine häufig übersetzte Textsorte in den Blick, die bislang offensichtlich nicht zu den typischen Einsatzgebieten von TM-Werkzeugen gehört.[1]

[1] Meines Wissens gibt es keine repräsentative Umfrage, die eine statistische Aussage zur Verwendungshäufigkeit von Translation Memories bei der Übersetzung von Patentschriften ermöglichen würde. Eine Umfrage des Verfassers bei einer Reihe größerer und mittlerer Übersetzungsdienstleister sowie bei spezialisierten Einzelübersetzern weist aber darauf hin, dass in der Praxis der Patentübersetzung derzeit nur Terminologieverwaltungssysteme in nennenswertem Umfang eingesetzt werden. Dies könnte u. a. auf den Umstand zurückzuführen sein, dass zum einen bei dieser Textsorte die Übersetzung von Folgeversionen keine Rolle spielt (vgl. 4.2) und zum anderen Patentschriften als geistiges Eigentum verbriefende Urkunden bis vor wenigen Jahren i. d. R. als Papierausdrucke oder als gescannte PDF-Bilddateien an den Übersetzungsdienstleister übermittelt wurden. Bezeichnenderweise war bis zum Jahr 2008 auch beim europäischen Marktführer für Patentübersetzungen der Workflow durch den Umgang mit nicht maschinenlesbaren Texten bei gleichzeitigem Verzicht auf ein systematisches Terminologiemanagement und auf Übersetzungsspeicherprogramme charakterisiert (http://www.lifepr.de/pressemeldungen/sdl-stuttgart/boxid-49513.html [14. März 2009]). Diese Befunde liegen auch auf einer Linie mit dem Ergebnis einer von Höcker (2003) durchgeführten Studie

Konkret sollte an einem größeren Korpus deutsch- und spanischsprachiger Patentschriften untersucht werden, inwieweit sich textsortenimmanente linguistische Strukturen mit Einfluss auf die Effizienz von TM-Programmen nachweisen lassen. Zu fragen war dabei nicht nur nach dem Wiederholungsfaktor der Texte, sondern auch nach textsortenbezogenen Parametern, die sich auf die Relevanz der Retrieval-Ergebnisse auswirken können. In zweiter Linie war die Frage zu klären, ob und wie es möglich ist, durch eine textsortenabhängige Konfiguration (z. B. von Segmentierungsparametern) die Einsatzbedingungen kommerziell vertriebener Programme zu optimieren und auf diese Weise zusätzliche Produktivitätssteigerungen und/oder Qualitätsverbesserungen zu erzielen. Schließlich sollten am Beispiel der verwendeten Software auch mögliche Defizite marktgängiger Programme aufgezeigt und Desiderate an die Software-Entwicklung abgeleitet werden. Die skizzierte Fragestellung fällt in ein noch weitgehend unbetretenes Forschungsfeld, da die bisher vorliegenden linguistischen Untersuchungen zur Textsorte Patentschrift entweder einzelsprachlich ausgerichtet (Dederding (1982b) und Dederding (1982a), Schamlu (1985a) und Schamlu (1985b), Liu (1992)) oder aber auf andere Sprachenpaare bzw. Sprachgruppen spezialisiert waren (z. B. Raible (1972), Barb (1982), Göpferich (1995a), Scheel (1997a) und Scheel (1997b), Gläser (562), Göpferich (2006)). Gänzlich neu ist auch die sprachliche Analyse der Textsorte mit Blick auf die Parameter des Nutzens einer integrierten Übersetzungsumgebung.

3 Methodik

Methodische Grundlage des Forschungsprojektes war die korpusbasierte Erfassung, Kategorisierung und Auswertung textsortentypischer linguistischer Merkmale, die im Übersetzungsprozess für die Retrieval-Leistung und die Effizienz von Translation-Memory-Systemen bestimmend sind. Das zweisprachige, nach den Erfordernissen der Fragestellung ausgewählte Korpus bestand dabei aus 60 vollständigen Exemplaren der Textsorte Patentschrift und diente sowohl der Registrierung textinterner und textexterner Rekurrenzen als auch der Bewertung ihrer Retrieval-Relevanz anhand exemplarischer Satzinhaltsanalysen, wobei hier auch übersetzungsmethodische Fragen mit Blick auf die besonderen Bedingungen der Textsorte (Raible (1987), Engberg (1999), Göpferich (2006)) zu berücksichtigen waren.

Die Analyse erfolgte aus der Sicht einer integrierten Übersetzungsumgebung in Form eines Translation-Memory-Systems mit der Möglichkeit der Konkordanzsuche unterhalb der Satzgrenze und eingebundener terminologisch-phraseographischer bzw. textographischer Datenbank, da auch textsortentypische Rekurrenzen unterhalb der Satzgrenze im Ergebnis berücksichtigt werden sollten. Für die untersuchte Textsorte war dies besonders wichtig, da sich der hohe sprachliche Konventionalisierungs- und

zur Häufigkeit des TM-Einsatzes bei deutschen Übersetzern, der zufolge die Nicht-User bei der Angabe ihrer Spezialisierung am häufigsten die Textsortenklasse der juristischen Texte (78%) nannten.

Normierungsgrad von Patentschriften auch auf syntagmatischer Ebene niederschlägt. Hier galt es das von Kühtz (2007) vorgelegte phraseologische Klassifizierungsmodell für die intralinguale Analyse fruchtbar zu machen und zugleich auf interlinguale Fragestellungen anzuwenden. Praktische Lösungsansätze bieten hier auch die Ergebnisse der angewandten fachsprachenbezogenen Phraseologieforschung (Budin and Galinski (1992), Hohnhold (1992), Schmitz (1996)).

Die Korpusanalyse umfasste drei Ebenen:

1. Linguistische Analyse nach dem Aspekt textsortentypischer textinterner Rekurrenzen auf Satz- und Teilsatzebene sowie auf der Ebene komplexer Syntagmen

2. Linguistische Analyse nach dem Aspekt textsortentypischer intertextueller Rekurrenzen auf Satz- und Teilsatzebene

3. Linguistische Analyse nach dem Aspekt textinterner und textexterner Rekurrenzen in Form textsortentypischer Formulierungsmuster und textsortentypischer fachsprachlicher Phraseologismen unterhalb der Satzebene

Für die Beurteilung der Retrieval-Relevanz ist die Beschreibung des Ähnlichkeitsverhältnisses zwischen dem zu übersetzenden AS-Segment und im Speicher abgelegten AS-Segmenten erforderlich. Wie Reinke (999a) ausführt, sind hierbei nicht nur formale, sondern auch inhaltliche Kriterien zu berücksichtigen.[2] So können z. B. TM-Einheiten, die Paraphrasen des zu übersetzenden AS-Segments darstellen oder sich nur durch einen abweichenden Explizitheitsgrad von ihm unterscheiden, selbst dann von hoher Relevanz sein, wenn die formale Übereinstimmung relativ gering ist und das TM-System einen niedrigen Match-Wert ermittelt bzw. bei entsprechend niedrigem Schwellenwert die vorhandenen TM-Einheiten gar nicht erst anbietet. Für den Zweck der vorliegenden Untersuchung wurden Rekurrenzen daher nicht im engen Sinne der älteren Textlinguistik (wie z. B. noch bei de Beaugrande and Dressler (1981) nur als ausdrucksseitige Erscheinung im Sinne einer referenzidentischen Wiederholung lexikalischer Einheiten (Rekurrenz als Kohäsionsphänomen) verstanden, sondern im erweiterten textlinguistischen Sinne als eine Wiederaufnahme von Inhaltsseitigem und/oder Ausdrucksseitigem (Linke and Nussbaumer (2000)). Entsprechend der Zielsetzung dieser Studie wurden

[2]Reinke hierzu: „Formal ließen sich die Unterschiede zwischen ‚Suchanfragen' und ‚Suchergebnissen' einfach in Form von mehr oder weniger umfangreichen Ersetzungen, Hinzufügungen, Auslassungen und Umstellungen (Verschiebungen) von Zeichenketten beschreiben. Ein ‚Treffer' wäre demzufolge umso relevanter, je geringer das Ausmaß dieser Veränderungen ist. Dies entspricht jedoch nicht unbedingt dem ‚Informationsbedürfnis' des Übersetzers, das in erster Linie darin besteht, aus der Menge der in einem TM vorhandenen AS/ZS-Segmentpaare jene herauszufinden, die im Vergleich zum aktuell zu übersetzenden AS-Segment identische oder zumindest möglichst ähnliche Inhalte aufweisen, so daß die ‚ZS-Seite' der gefundenen TM-Einheit wahrscheinlich mit möglichst geringem Aufwand in die aktuelle Übersetzung eingebettet werden kann." (Reinke (999a): 104)

dabei nur solche Fälle berücksichtigt, in denen sich die inhaltsseitige Rekurrenz zugleich in einer Identität bzw. Ähnlichkeit der Zeichenkette niederschlägt. Der Fall der pragmatischen Bedeutungsgleichheit oder -ähnlichkeit bei vollständiger ausdrucksseitiger Substitution war damit aus der Untersuchung ausgeschlossen.

Auf die statistische Analyse der im Korpus nachzuweisenden textinternen Rekurrenzen wurde aus mehreren Gründen verzichtet: Zum einen erfasst die Analysekomponente des verwendeten Translation-Memory-Programms nur vollständige textinterne Wiederholungen auf Satzebene, nicht dagegen textinterne Ähnlichkeiten (*fuzzy matches*) auf Satzebene oder Rekurrenzen unterhalb der Satzgrenze, so dass die Aussagekraft der quantitativen Daten sehr beschränkt bliebe. Zum anderen hätten selbst zuverlässige statistische Angaben noch immer einen geringen Aussagewert im Hinblick auf den ebenfalls effizienzbestimmenden Aspekt der Retrieval-Relevanz.

Auch gegen die statistische Gesamtauswertung textexterner Rekurrenzen innerhalb der Sprachkorpora und der ermittelten Ähnlichkeitswerte (Match-Werte) sprachen mehrere Gründe: So hätte sie schon deshalb keine für die Praxis repräsentativen Werte liefern können, weil der Anteil textexterner Rekurrenzen in hohem Maße vom Umfang des Referenzmaterials sowie von textthemen- und autorenbezogenen Faktoren (auf die Berufspraxis des Übersetzers übertragen: von der Größe des Übersetzungsspeichers und vom Grad der Spezialisierung auf bestimmte Fachgebiete und Auftraggeber) abhängt. Unabhängig davon hätte auch bei den textexternen Rekurrenzen eine statistische Bezifferung nur wenig Aussagekraft in Bezug auf die Retrieval-Relevanz, da die mit der Analysefunktion von Translation-Memory-Programmen ermittelten Ähnlichkeitswerte das Ergebnis eines einfachen und ausschließlich an der Textoberfläche orientierten Algorithmus sind und nur sehr bedingt den Ähnlichkeitsurteilen von Humanübersetzern entsprechen (Seewald-Heeg and Nübel (1999)).

Aus den genannten Gründen konzentrierte sich die Untersuchung auf die qualitative Beschreibung effizienzbestimmender linguistischer Parameter unter Berücksichtigung textsortenbezogener Übersetzungsstrategien und mit Blick auf die Frage, welche Arten der Ähnlichkeit und der Mehrdeutigkeit textsortentypisch bzw. textsortenuntypisch sind. Da generalisierbare Ergebnisse erzielt werden sollten, wurden im Rahmen der Analyse ausschließlich textsortenbezogene, nicht aber textthematisch bedingte Rekurrenzen erfasst.

Das Gesamtkorpus bestand aus 60 ungekürzten (jeweils 30 original spanischsprachigen und 30 original deutschsprachigen) Patentschriften aus den Jahren 2000 bis 2008 mit insgesamt 5.250 Sätzen und 220.000 Wörtern und umfasste inhaltlich ein breites Spektrum von Fachgebieten (Fahrzeugtechnik, Elektrotechnik, Metalltechnik, Kunststofftechnik, Medizintechnik, Medizin und Chemie), so dass die Möglichkeit einer Themenabhängigkeit der Ergebnisse auch formal weitgehend ausgeschlossen war. Jeweils fünf Patentschriften stammten von derselben Anwaltskanzlei, weil v. a. mit Blick auf die Analyse intertextueller Rekurrenzen das für den Berufsübersetzer relevante Phänomen

möglicher Formulierungspräferenzen wiederkehrender Auftraggeber (Patentanwälte bzw. Anwaltsbüros) im Korpus abgebildet werden sollte. Das für die empirischen Untersuchungen herangezogene Translation-Memory-Werkzeug war die *Translator's workbench* der Firma SDL/Trados (Version 7.0.0).[3]

4 Ergebnisse und Diskussion

Die Analyseergebnisse sollen im Folgenden anhand ausgewählter Beispiele skizziert werden. Im Falle der textinternen Rekurrenzen (Abschnitt 4.1) wird dabei - mit Ausnahme von Beispiel 4 - exemplarisch von der Übersetzungsrichtung Deutsch-Spanisch ausgegangen.

4.1 Textinterne Rekurrenzmuster

Der hohe Normierungsgrad der Textsorte (vgl. Schamlu (1985a), Gläser (562), Göpferich (2006)) kommt auch innerhalb des Einzeltextes zum Tragen. Wie die folgenden Beispiele belegen, weisen Patentschriften intratextuell eine Vielzahl textsortenimmanenter Wiederholungen und Ähnlichkeiten auf. Zu diskutieren sind dabei insbesondere die Art der Ähnlichkeit und die daraus resultierende Verwertbarkeit der Suchergebnisse. Die zu diesem Zweck durchgeführten Satzinhaltsvergleiche stützen sich insbesondere auf die von Reinke (2004) vorgeschlagene Typologie von Ähnlichkeitskriterien.

In den Satzbeispielen sind die übereinstimmenden Passagen jeweils durch Fettdruck hervorgehoben. Die durchgeführten empirischen Tests basieren auf der Annahme, dass die Übersetzung der Chronologie des Textablaufs folgt. Die angegebenen Match-Werte beziehen sich also auf den Fall, dass der jeweils zuerst genannte Beispielsatz (1a, 2a usw.) den bereits im Speicher enthaltenen Referenzsatz darstellt und der jeweils zweite Beispielsatz (1b, 2b usw.) der zu übersetzende Testsatz ist.

4.1.1 Rekurrenzen auf Satz- und Teilsatzebene

Beispiel 1: Rekurrenzen zwischen dem ersten Satz der Beschreibung (Gattungsangabe) und dem Oberbegriff des Hauptanspruchs Textsortentypisch für deutsche Patentschriften ist die wörtliche oder weitgehend wörtliche Wiederaufnahme des ersten Satzes der Beschreibung im Teiltext Ansprüche, wo er, gekürzt um die Einleitungsphrase und ergänzt um Bezugsnummern, in Form einer komplexen Nominalphrase als Oberbegriff des Hauptanspruchs fungiert[4]:

[3] Eine ausführliche Beschreibung des Systems im Vergleich mit anderen marktgängigen Systemen findet man in Seewald-Heeg (2005).

[4] In der aktuellen Fassung der deutschen Patentverordnung und im aktuellen Merkblatt für Patentanmelder des Deutschen Patent- und Markenamtes wird diese Formulierungskonvention nur implizit nahe gelegt. Dass in fast allen deutschen Korpustexten die wörtliche Wiederholung streng eingehalten wurde, könnte auf dem

(1a)	(Kontext: Gattungsangabe im ersten Satz der Beschreibung) Die vorliegende Erfindung betrifft eine **Vorrichtung zur stufenlosen Regulierung des Aufstellens einer Antriebstandemachse eines Fahrzeuges oder einer selbstfahrenden Arbeitsmaschine, die einen Fahrzeugrahmen und mindestens eine Antriebsachse aufweist, wobei jeweils an den Enden der Antriebsachse eine Antriebstandemachse mit einem beweglichen Tandemachsengehäuse angeordnet ist und über die Antriebstandemachse das von der Antriebsachse eingeleitete Moment mechanisch auf die in dem Tandemachsengehäuse angeordneten Räder verteilt wird.**
(1b)	(Kontext: Hauptanspruch) **Vorrichtung zur stufenlosen Regulierung des Aufstellens einer Antriebstandemachse eines Fahrzeuges oder einer selbstfahrenden Arbeitsmaschine, die einen Fahrzeugrahmen und mindestens eine Antriebsachse (12) aufweist, wobei jeweils an den Enden der Antriebsachse (12) eine Antriebstandemachse (16) mit einem beweglichen Tandemachsengehäuse (18) angeordnet ist und über die Antriebstandemachse (16) das von der Antriebsachse (12) eingeleitete Moment mechanisch auf die in dem Tandemachsengehäuse (18) angeordneten Räder verteilt wird,** dadurch gekennzeichnet, dass die Antriebsachse (12) mit der Antriebstandemachse (16) über eine Kugelrampenvorrichtung (20) koaxial verbunden ist, wobei die Kugelrampenvorrichtung (20) aus einer ersten Kugelrampenscheibe (22), die mit einem Ende (26) der Antriebstandemachse (16), welches einem Ende (28) der Antriebsachse (12) gegenüberliegt, verbunden ist, und einer zweiten Kugelrampenscheibe (24) die am Ende (28) der Antriebstandemachse (12) axial verschiebbar angeordnet ist, besteht und durch ein dem Antriebsdrehmoment entgegenwirkendes Drehmoment der Antriebstandemachse (16) der Abstand L zwischen der zweiten Kugelrampenscheibe (24) und der ersten Kugelrampenscheibe (22) vergrößert wird, wobei die zweite Kugelrampenscheibe (24) mit einem Kolben (30) in Wirkverbindung steht und durch die Axialbewegung der zweiten Kugelrampenscheibe (24) ein Druck mit einem Wert P1 in einem Volumen (32) erzeugt wird, wobei P1 an ein Regelventil (34) geleitet wird und das Regelventil (34) die Höhe eines Systemdrucks P3 oder P4 in Abhängigkeit von P1 regelt und ein resultierender Druck mit dem Wert P2 zur Steuerung einer Ausgleichsvorrichtung (36) zur Erzielung eines dem Aufstelleffekt entgegenwirkenden Ausgleichsmoments dient.

(Quelle: EP 1 712 381 A1)

Die komplexe Nominalphrase des Oberbegriffs in (1b) ist als elliptischer Teilsatz zu verstehen im Sinne von: *Geschützt werden soll eine Vorrichtung ...* (Schamlu (1985a)). Von der formelhaften Wendung dadurch gekennzeichnet, dass sie in maschinenlesbaren Ausgangstexten meist durch eine Absatzmarke getrennt ist, die der Segmentierungsalgorithmus der *Translator's workbench* standardmäßig als Segmentende einstuft. Die komplexe Nominalphrase wird in diesem Fall zum eigenständigen Retrieval-Segment und - je nach dem Grad der Ähnlichkeit mit dem ersten Satz der Beschreibung und je nach dem Verhältnis zwischen dem Umfang von Bezugsziffern und dem Umfang des Gesamtsegmentes - in der Regel als gut verwertbarer *fuzzy match* erkannt. Im vorliegenden Beispiel läge bei Trennung durch Absatzmarke der Match-Wert bei 77 Prozent, ohne Trennung durch Absatzmarke dagegen unterhalb des kleinsten einstellbaren Match-Wertes von 30 Prozent, so dass kein Treffer mehr möglich wäre. Es kann daher sinnvoll

Umstand beruhen, dass in älteren Versionen des Merkblattes diese Formulierungsweise explizit empfohlen wurde (vgl. Schamlu (1985a)).

sein, bei fehlender Absatzmarke eine solche einzufügen oder aber mit benutzerdefinierten Segmentierungsregeln zu arbeiten. Alternativ lässt sich die Rekurrenz auch mit Hilfe der Konkordanzsuchfunktion auffinden.

Die vergleichende Satzinhaltsanalyse ergibt eine Abweichung auf zwei Ebenen: So erfolgt in Satz (1b) zum einen eine elliptische Bezugnahme auf den in Satz (1a) mit Initiator angekündigten Erfindungsgegenstand und zum anderen eine Informationsverlagerung (hier: Expansion) mit Erhöhung des Explizitheitsgrades gegenüber (1a) durch die Hinzufügung der Bezugsnummern und die Angabe der kennzeichnenden Merkmale. Die Retrieval-Relevanz ist in jedem Falle hoch, da trotz der Nichtidentität der Satzinhalte das gesamte durch Fettdruck hervorgehobene Syntagma in (1a) ohne syntaktische Umstellungen in die Zieltextversion übernommen werden kann. Einzufügen sind nur noch als Placeables die Bezugsnummern.

Auch die Ausführungsverordnungen zum spanischen Patentgesetz und die für spanische Patentanmelder herausgegebene Informationsbroschüre[5] weisen auf den engen inhaltlichen und formalen Zusammenhang zwischen der Einleitung der Beschreibung und dem Hauptanspruch hin. Dementsprechend waren gut verwertbare Rekurrenzen zwischen dem Anfangsteil der Beschreibung und dem Oberbegriff des Hauptanspruchs auch in mehr als der Hälfte der spanischen Korpustexte nachweisbar.

Beispiel 2: Rekurrenzen zwischen weiteren Gliederungspunkten der Beschreibung und kennzeichnenden Teilen der Ansprüche Als textsortentypisch erwies sich auch die Übernahme von Merkmalsbeschreibungen der Abschnitte ‚Lösung der Aufgabe'bzw. ‚Beschreibung bevorzugter Ausführungsbeispiele'in die kennzeichnenden Teile der Ansprüche:

Das Beispiel gibt einen Fall wieder, in dem die häufige Erscheinung einer wörtlichen Wiederaufnahme nicht vorliegt. Die *Translator's workbench* gibt für den Hauptanspruch allerdings noch immer einen Match-Wert von 49 Prozent an.

Die linguistische Analyse zeigt, dass in (2b) der durch Fettdruck hervorgehobene Teil des Satzinhalts von (2a) lediglich durch die Eingliederung in einen komitativen modalen Nebensatz (*wobei das* ?) um eine syntaktische Ebene nach unten gerückt wird. Nimmt man nur die verglichenen Teilsätze in den Blick, so liegen hier aus inhaltlicher Sicht Paraphrasen mit Inhaltsverlagerung insbesondere in Form von Hinzufügungen vor (Angabe der Bezugsnummern und Hinzufügung textsortentypischer Redundanzen zur Steigerung des Explizitheitsgrades). Der Umfang der semantischen Informationsverlagerung ist aus Sicht der übersetzerischen Verwertbarkeit relativ gering. Trotz der Verschiebung auf satzsyntaktischer Ebene sind bei einer Übersetzung in die Sprachrichtung Deutsch-Spanisch nur geringfügige strukturelle Veränderungen erforderlich. Während nämlich im deutschen Ausgangstext die Stellung der Verben ([*verschlossen*]

[5]Manual informativo para los solicitantes de patentes (www.oepm.es)

(2a)	(Kontext: Die erfindungsgemäße Lösung greift auf die bekannten zusammendrückbaren Flaschen mit Flickflüssigkeit zurück.) Diese werden jedoch erfindungsgemäß nicht von Hand zusammengedrückt, sondern die [sic] Flickflüssigkeit enthaltende **flüssigkeitsspeichernde Volumen** ist **mit mindestens einer ersten zu öffnenden Verschlußstelle dicht verschlossen** und **befindet sich in einem Druckbehälter, der mit von einer Druckgasquelle geliefertem Druckgas in einem sich unter der Wirkung des Druckgases vergrößernden Kompressionsraum beaufschlagt wird und die erste Verschlußstelle geöffnet wird,** wobei **das Volumen über die geöffnete erste Verschlußstelle und einen Zuführschlauch die in ihm befindliche Flickflüssigkeit so lange an den Reifen abgibt, bis keine Flickflüssigkeit mehr gefördert wird,** wonach der Druck in einem Raum**, der über dem Druckbehälter angeordnet ist, abfällt und die Druckdifferenz zwischen einem Zuführweg für das Druckgas und dem Raum ansteigt und mindestens eine zweite zu öffnende Verschlußstelle, die zwischen dem Zuführweg und dem Raum angeordnet ist, aufgrund der angestiegenen Druckdifferenz geöffnet wird und einen Weg von der Druckgasquelle über den Zuführweg an den Reifen freigibt.**
(2b)	(Kontext: Hauptanspruch) Vorrichtung zum Beheben einer Reifenpanne mit einer in den zu reparierenden Reifen einzuführenden Flickflüssigkeit, die sich in einem komprimierbaren flüssigkeitsspeichernden Volumen (100) befindet, wobei das **flüssigkeitsspeichernde Volumen** (100) **mit mindestens einer ersten zu öffnenden Verschlußstelle** (60) **dicht verschlossen** ist **und** sich **in einem Druckbehälter (56)** befindet**, der mit von einer Druckgasquelle** (28) **geliefertem Druckgas in einem sich unter der Wirkung des Druckgases vergrößernden Kompressionsraum** (54) **beaufschlagt wird und die erste Verschlußstelle** (60) **geöffnet wird,** so daß **das Volumen** (100) **über die geöffnete erste Verschlußstelle** (60) und **einen Zuführschlauch** (16) **die in ihm befindliche Flickflüssigkeit so lange an den Reifen** (12) **abgibt, bis keine Flickflüssigkeit mehr gefördert wird,** wobei die Vorrichtung (10) weiterhin einen Raum (58) aufweist, **der über** oder in **dem Druckbehälter (56) angeordnet ist,** wobei in dem Raum (58) nach der Abgabe der Flickflüssigkeit der Druck **abfällt und die Druckdifferenz zwischen einem Zuführweg** (46) **für das Druckgas und dem Raum** (58) **ansteigt und mindestens eine zweite zu öffnende Verschlußstelle** (62)**, die zwischen dem Zuführweg** (46) **und dem Raum** (58) **angeordnet ist, aufgrund der angestiegenen Druckdifferenz geöffnet wird und einen** direkten **Weg von der Druckgasquelle** (28) **über den Zuführweg** (46) **an den** mit dem Druckgas zu füllenden **Reifen** (12) **freigibt**.

(Quelle: EP 98 948 965.3)

sein, sich befinden) in Haupt- und Nebensatz unterschiedlich ist und im Übrigen eine Reduzierung des Match-Wertes zur Folge hat, bleibt im Spanischen die Verbstellung in Haupt- und Nebensatz identisch und muss bei der Übersetzung von (2b) gegenüber der Zieltextversion von (2a) nicht abgeändert werden. Auch in diesem Fall ist also die Verwertbarkeit des Suchergebnisses de facto höher, als die statistische Analyse des Match-Wertes es vermuten lassen würde.

Beispiel 3: Rekurrenzen zwischen weiteren Gliederungspunkten der Beschreibung und kennzeichnenden Teilen der Ansprüche Umgekehrt kann es vorkommen, dass bei identischer Verbstellung im deutschen Ausgangstext (vgl. die durch Fettdruck hervorgehobenen Passagen) in der spanischen Zielsprache systemabhängige syntaktische Veränderungen zwischen den jeweiligen ZS-Versionen vorzunehmen sind (zweimalige

Nebensatzverkürzung durch Gerundialkonstruktion bei der Übersetzung von (3a) vs. 1-mal konjunktionaler Nebensatz und 1-mal Gerundialkonstruktion bei der Übersetzung von (3b)).

(3a)	(3b)
(Kontext: Beschreibung/Lösung der Aufgabe) Bei einer erfindungsgemäßen Vorrichtung zur Aufbereitung von Abfällen weist eine Zerkleinerungstrommel im Bereich ihrer Drehachse zwei sich gegenüberliegende Öffnungen auf, wobei **die erste Öffnung zum Eintrag der aufzubereitenden Abfälle und zum Austrag** von **zerkleinerten organischen Bestandteilen des Abfalls und die zweite Öffnung zum Austrag** der **abgetrennten anorganischen Bestandteile des Abfalls** dient, **wobei die erste Öffnung einen größeren Durchmesser aufweist als die zweite Öffnung und der Transport der abgesunkenen zerkleinerten anorganischen Bestandteile zur zweiten Öffnung mittels eines Schneckenaustrags erfolgt.**	(Kontext: Hauptanspruch) Vorrichtung zur Aufbereitung von Abfällen mit organischen Anteilen [?], dadurch gekennzeichnet, dass **die erste Öffnung** (34) **zum Eintrag der aufzubereitenden Abfälle und zum Austrag** der **zerkleinerten organischen Bestandteile des Abfalls und die zweite**, gegenüberliegende **Öffnung** (32) **zum Austrag** von **abgetrennten anorganischen Bestandteilen des Abfalls** ausgebildet ist, **wobei die erste Öffnung** (34) **einen größeren Durchmesser aufweist als die zweite Öffnung** (32) **und der Transport der abgesunkenen zerkleinerten anorganischen Bestandteile zur zweiten Öffnung** (34) **mittels eines Schneckenaustrags erfolgt.**

(Quelle: EP 98 108 158.1)

(Anm.: Die Verteilung der Beispielsätze auf mehrere Absätze wurde zur besseren Übersicht vorgenommen und entspricht - mit Ausnahme der Absatzmarken vor und nach der Wendung *dadurch gekennzeichnet* - nicht dem Original.)

Bei der Übersetzung von (3b) ist trotz eines relativ geringen Match-Wertes von 54 Prozent (bei eingefügter Absatzmarke vor dem Kennzeichnungsteil in (3b)) die Relevanz des Suchergebnisses (3a) als hoch einzustufen. Bei der dokumentarischen Übersetzung der Nebensätze von (3b) sind lediglich die geringfügige Erhöhung des Explizitheitsgrades sowie die Ersetzungen auf lexikalischer Ebene (Kontextsynonyme *dienen* vs. *ausgenommen sein*) und im Bereich der Textdeixis (bestimmter vs. unbestimmter Artikel) zu berücksichtigen.

Beispiel 4: Rekurrenzen zwischen weiteren Gliederungspunkten der Beschreibung und kennzeichnenden Teilen der Ansprüche in spanischen AS-Texten

Derselbe Rekurrenztyp (Wiederaufnahme lösungsbezogener Beschreibungselemente im Kennzeichnungsteil der Ansprüche) wurde auch in den spanischen Korpustexten häufig registriert:

Auch in diesem Fall erhöht die Einfügung der Absatzmarke vor dem Kennzeichnungsteil den Match-Wert deutlich (von 49 Prozent auf 75 Prozent), wobei allerdings zwischen den deutschen Zieltextversionen von (4a) und (4b) eine Anpassung der zielsprachlichen Verbstellung erforderlich wäre.

(4a)	(4b)
(Kontext: Beschreibung) La **superficie externa** del tallo **presenta unas ranuras longitudinales, que abarcan, aproximadamente, la mitad superior de la altura del tallo.**	(Kontext: Unteranspruch) Tallo femoral para prótesis total de cadera, según la reivindicación 1, caracterizado por el hecho de que su **superficie externa presenta unas ranuras** (6) **longitudinales, que abarcan, aproximadamente, la mitad superior de la altura del tallo.**

(Quelle: P 9 100 003)

Beispiel 5: Rekurrenzen zwischen der Beschreibung von Merkmalen vorteilhafter Ausführungen und der Beschreibung der Figuren

Zu den Charakteristika der Textsorte gehört auch die Beschreibung der Merkmale verschiedener Ausführungsbeispiele (5a) und die nachfolgende Bezugnahme auf diese Merkmale in der Erläuterung der zeichnerischen Darstellungen (5b):

Der Satzinhaltsvergleich zeigt eine funktionale Verschiebung bei nur geringer Informationsverlagerung zwischen (5a) und (5b) (vgl. v. a. die Bezugnahme auf die zeichnerischen Darstellungen in (5b)). Abgesehen von den vergleichsweise geringfügigen lexikalisch-semantischen Veränderungen (v. a. Bezugsnummern, Kontextsynonyme, Text- bzw. Situationsdeixis im Bereich der Artikel, Modalpartikel) handelt es sich bei den Beispielen um Paraphrasen mit Verschiebungen auf transphrastischer Ebene. Konkret liegt hier in syntaktischer Hinsicht eine Expansion (Ausweitung eines Satzgefüges zu einer Satzfolge) vor. Bei satzweiser Suchanfrage während der Übersetzung von (5b) wäre bei der Arbeit mit der *Translator's workbench* kein Match in Bezug auf (5a) möglich. Erst bei manueller Segmenterweiterung auf alle vier Sätze von (5b) wird ein Match-Wert von 40 Prozent in Bezug auf den vorangehenden Einzelsatz (5a) erreicht. Dies demonstriert deutlich die Wünschbarkeit satzübergreifender Erkennungsalgorithmen.

Zuweilen wurden im deutschen Sprachkorpus auch Fälle syntaktischer Expansionen bzw. Reduktionen dieses Typs registriert, in denen die Satzfolge nicht durch Punkt, sondern durch Semikolon getrennt war. Eine Erkennung ist in diesen Fällen dann sehr wahrscheinlich, sofern bei der Konfiguration der Segmentierungsparameter das Semikolon nicht als Segmentende definiert wird. Bei der *Translator's workbench* entspricht dies der Standardeinstellung.

4.1.2 Rekurrenzen auf der Ebene komplexer Nominalphrasen

Der Zweck von Patentschriften ist die juristisch tragfähige Absicherung von Schutzrechten. Die allgemein fachsprachentypische funktionale Eigenschaft der Ökonomie tritt hier deshalb zugunsten der Eindeutigkeit stärker in den Hintergrund als bei den meisten anderen Fachtextsorten. Es überrascht daher nicht, dass insbesondere im terminologischen Bereich die totale Rekurrenz (im engeren textlinguistischen Sinne) konventionellerweise

(5a)	(5b)
(Kontext: Beschreibung bevorzugter Ausführungsbeispiele) In einer weiteren vorteilhaften Ausgestaltung der Erfindung ist **die Ausgleichsvorrichtung** eine **Drehkolbenzylinderanordnung,** **wobei** ein **Gehäuse eines mit Druck beaufschlagbaren Drehkolbenzylinders fest mit** einem **Achsgehäuse der Antriebsachse verbunden ist** und **ein drehbeweglicher Kolben des Drehkolbenzylinders ein nachgeschaltetes Planetengetriebe mit einer Planetenachse und Planetenrädern** antreibt **,** wobei **die Planetenachse des Planetengetriebes fest mit dem** Achsgehäuse oder dem **Gehäuse des Drehkolbenzylinders verbunden ist** und **die Planetenräder auf einen Zahnkranz, der mit dem Tandemachsengehäuse verbunden ist,** einwirken.	(Kontext: Beschreibung der zeichnerischen Darstellungen) In dem dargestellten Ausführungsbeispiel besteht **die Ausgleichsvorrichtung** 36 aus einer **Drehkolbenzylinderanordnung** 38, **wobei** das **Gehäuse** 50 **eines mit Druck beaufschlagbaren Drehkolbenzylinders** 40 **fest mit** dem **Achsgehäuse** 14**der Antriebsachse** 12 **verbunden ist.** **Ein drehbeweglicher Kolben** 42 **des Drehkolbenzylinders** 40 treibt **ein nachgeschaltetes Planetengetriebe** 44 **mit einer Planetenachse** 46 **und Planetenrädern 48** an. **Die Planetenachse** 46 **des Planetengetriebes** 44 ist dabei **fest mit dem Gehäuse** 50 **des Drehkolbenzylinders** 40 **verbunden**. **Die Planetenräder** 48 wirken dagegen **auf einen Zahnkranz** 52 ein**, der mit dem Tandemachsengehäuse** 18 **verbunden ist.**

(Quelle: EP 1 712 381)

andere Kohäsionsmittel wie die Substitution durch Synonyme, Hyponyme oder Hyperonyme und v. a. Pro-Formen weitgehend verdrängt.[6] Dies gilt im Deutschen und – in geringerer Ausprägung - im Spanischen auch für den Fall sehr komplexer Nominalphrasen, deren vollständige Wiederaufnahme im jeweiligen Textzusammenhang hochgradig redundant erscheinen kann.

Ein eindrucksvolles Beispiel hierfür liefert die spanische Offenlegungsschrift mit dem Titel „Máquina de soldar por láser para soldadura de perfiles sobre componentes estructurales de gran tamaño“ (ES 2161113 A1), in der die Nominalphrase des Titels insgesamt 18-mal in voller Länge auftritt (jeweils 1-mal im Titel und im Abstract, 1-mal

[6] Auf die partielle Rekurrenz trifft dies in deutschen Patentschriften nur mit Einschränkungen zu (vgl. Dederding 1982b). Im Zusammenhang mit der Verwendung von Pro-Formen scheint es meiner Korpusanalyse zufolge zwischen deutschen und spanischen Patentschriften klare Unterschiede zu geben, die weitergehende Untersuchungen rechtfertigen würden.

im einleitenden Satz der Beschreibung[7], 1-mal im zweiten Satz der Beschreibung, 1-mal im Initiator des Hauptanspruchs, 13-mal als Initiator der Nebenansprüche) und damit mehr als 7 Prozent des gesamten Textumfangs stellt.

Da Patentschriften dokumentarisch übersetzt werden und Rekurrenzen dieser Art deshalb stets im vollen Wortlaut wiederzugeben sind, ermöglicht die Integration von Terminologieverwaltungssystemen hier mitunter erhebliche Effizienzvorteile. Aber auch die Suchalgorithmen des TM-Systems führen hier häufig zu Treffern. Deshalb ist es in Bezug auf die Übersetzung des Teiltexts ‚Ansprüche' von Vorteil, wenn im Ausgangstext durch die Einfügung einer Absatzmarke jeweils nach dem Oberbegriff von Haupt- und Nebenansprüchen ein Segmentende signalisiert wird. In dem genannten Beispiel würde die komplexe Nominalphrase so - auch unabhängig von einer Registrierung in der Terminologieverwaltungskomponente - in 13 Fällen (Oberbegriff der Nebenansprüche) auch unmittelbar von der Translation-Memory-Software als *full match* erkannt

4.2 Intertextuelle Rekurrenzen auf der Ebene satzwertiger Formulierungsstereotype

Die Übersetzung von Folgetexten, d. h. korrigierter oder aktualisierter Textversionen, spielt in der Berufspraxis des Patentübersetzers keine nennenswerte Rolle. Übersetzt wird in aller Regel nur eine Textfassung, nämlich das vom zuständigen Patentamt geprüfte und erteilte Patent, dessen Fassung nachträglich selbst im Falle offensichtlicher Fehler nicht mehr geändert werden darf (Dybdahl (2004)). Themenabhängige intertextuelle Rekurrenzen können aber dennoch z. B. in Fällen auftreten, in denen mehrere Patentanmeldungen auf denselben Stand der Technik Bezug nehmen und die betreffenden Anträge Formulierungen aus früheren einschlägigen Patentschriften entweder identisch oder paraphrasierend übernehmen. Intertextuelle Rekurrenzen dieser Art sind insbesondere dann nahe liegend, wenn mehrere Ausgangstexte von demselben Anmelder bzw. demselben Anwaltsbüro stammen. Derartige Fälle sind dem Verfasser zwar aus der eigenen Übersetzungspraxis bekannt, waren aber in den Korpustexten nicht nachweisbar.

Durchaus im Korpus nachzuweisen waren aber themenunabhängige Rekurrenzen auf Satz- und Teilsatzebene, die unmittelbar aus der makrostrukturellen und sprachlichen Stereotypie von Patentschriften resultieren und damit textsortenimmanent sind. Die Rede ist von einer Vielzahl stark konventionalisierter, funktional und makrostrukturell gebundener und in aller Regel als metakommunikative Elemente fungierender Äußerungen, die je nach dem Grad ihrer Vorgeprägtheit entweder als referentiell-propositionale Phraseologismen (Festgeprägtheit), als satzwertige Routineformeln (weitgehend stabile

[7] In spanischen Patentschriften wird der Teiltext ‚Descripción' (‚Beschreibung') konventionell mit einem elliptischen Satz in Form einer Nominalphrase eingeleitet, die den Titel der Anmeldung wiedergibt und häufig im zweiten Satz der Beschreibung wiederholt und ausformuliert wird. Zum Teil ergeben sich hier gut verwertbare fuzzy matches.

Formelhaftigkeit) oder aber als satzwertige Formulierungsmuster (variable Musterhaftigkeit) einzustufen sind. Die Übergänge zwischen diesen Kategorien sind fließend (Kjær (1991), Stein (2001), Kühtz (2007)), so dass im Folgenden keine Zuordnung vorgenommen werden soll.

Bei den registrierten Rekurrenzen handelte es sich keineswegs nur um wiederkehrende Textbausteine in Anmeldungen derselben Anwaltskanzlei, sondern häufig auch um identische oder ähnliche Formulierungsstereotype verschiedener Autoren. Der Grund für dieses Phänomen ist der hohe Normierungs- und Standardisierungsgrad der Textsorte, der wiederum darauf zurückzuführen ist, dass die inhaltliche, strukturelle und sprachliche Gestaltung von Patentanmeldungen zum einen von historisch gewachsenen Konventionen und zum anderen von gesetzlichen Regelungen beeinflusst wird.[8]

Für einige in deutschen Patentschriften besonders häufig auftretende Stereotype wurden in den spanischsprachigen Korpustexten weder semantische noch funktionale Entsprechungen nachgewiesen (vgl. 4.2.1 und 4.2.2). Dies belegt zwar, dass trotz der Harmonisierung im Bereich der Europapatente kulturabhängige Argumentations- und Formulierungsschablonen fortbestehen, spricht aber nicht gegen den Nutzen integrierter Übersetzungssysteme bei dieser Textsorte, da im Rahmen einer dokumentarischen Übersetzung sprachlich standardisierte Ausgangstextsegmente auch beim Fehlen funktionaler Äquivalente in die Zielsprache übertragen werden und somit in späteren Übersetzungssituationen als Referenzmaterial zur Verfügung stehen.

Es folgt eine exemplarische Übersicht über die in beiden Sprachkorpora nachgewiesenen Formulierungsstereotype. Die Klassifizierung erfolgt dabei nach dem übersetzungsmethodisch wichtigen - weil für die Austauschbarkeit bzw. Verwertbarkeit der Retrieval-Segmente entscheidenden - Aspekt der kommunikativen Funktion. Die Anordnung der Unterkapitel basiert auf dem Ablaufschema der Textsorte (vgl. auch Göpferich (2006)). Die Angaben zur makrostrukturellen Lokalisierung der einzelnen Formeln innerhalb der Korpustexte erschienen mir vor allem deshalb wichtig, weil diese Information bei der Erstellung von Datenbankeinträgen eine eigene Datenkategorie bilden sollte (vgl. 4.2.5).

4.2.1 Metakommunikativer Verweis auf die wesentlichen Merkmale der Erfindung

Erläuterung: Die zu schützenden Erfindungsmerkmale werden unter Bezugnahme auf den Hauptanspruch (= unabhängiger Anspruch 1) bzw. - im Falle der gleichzeitigen Patentierung von Vorrichtungen und Verfahren – unter Bezugnahme auf die beiden Hauptansprüche erwähnt.

[8] Die von den zuständigen Patentbehörden herausgegebenen, auf nationalen Gesetzen und Verordnungen beruhenden Richtlinien für die Abfassung von Patentschriften sind schon deshalb in sprachlicher und argumentativer Hinsicht normstiftend, weil ihre Nichteinhaltung die Verweigerung der Patenterteilung zur Folge haben kann. (Näheres bei Göpferich (2006))

Makrostrukturelle Einbettung: Teiltext ‚Beschreibung'; Gliederungspunkt ‚Lösung der Aufgabe'

Formulierungsstereotype deutsch:

Okkurrenzen gesamt: 17 / Zahl der nachgewiesenen Varianten: 9
Beispiele:

- *Gelöst wird diese Aufgabe durch eine Vorrichtung mit den Merkmalen des Schutzanspruches 1.* (4/2)[9]
- *Zur Lösung dieser Aufgabe dient eine Vorrichtung gemäß den Merkmalen des unabhängigen Anspruchs 1.* (2/1)

Formulierungsstereotype spanisch:

Okkurrenzen gesamt: Im spanischsprachigen Korpus waren keine inhaltlich oder funktional entsprechenden satzwertigen Formulierungsstereotype nachweisbar.

4.2.2 Ersterwähnung vorteilhafter Ausgestaltungen

Erläuterung: Die erste Erwähnung der vorteilhaften Ausgestaltungen der Erfindung erfolgt unter metakommunikativem Verweis auf die Unteransprüche (= abhängige Ansprüche).

Makrostrukturelle Einbettung: Teiltext ‚Beschreibung'; Gliederungspunkt ‚Lösung der Aufgabe' (Gliederungssignal als Terminator des Abschnitts)

Formulierungsstereotype deutsch:

Okkurrenzen gesamt: 20 / Zahl der nachgewiesenen Varianten: 7
Beispiele:

- *Weitere vorteilhafte Ausgestaltungen sind in den Unteransprüchen beschrieben.* (5/3)
- *Weitere Merkmale der Erfindung sind in den Unteransprüchen enthalten.* (2/1)

Formulierungsstereotype spanisch:

Okkurrenzen gesamt: Im spanischsprachigen Korpus waren keine inhaltlich oder funktional entsprechenden satzwertigen Formulierungsstereotype nachweisbar.

[9] Die erste Zahl in Klammern gibt die Gesamtzahl der Okkurrenzen für die betreffende Formulierung an; die zweite Zahl bezieht sich auf die Anzahl der verschiedenen Patentanwaltskanzleien, auf die sich die Okkurrenzen verteilen.

4.2.3 Kataphorischer Verweis auf die Beschreibung konkreter Ausführungsbeispiele unter Bezugnahme auf die zeichnerischen Darstellungen

Makrostrukturelle Einbettung: Teiltext ‚Beschreibung' ; Abschnitt ‚Beschreibung eines oder mehrerer Ausführungsbeispiele' (Gliederungssignal als Initiator des Abschnitts)

Formulierungsstereotype deutsch:

Okkurrenzen gesamt: 20 / Zahl der nachgewiesenen Varianten: 13
Beispiele:

- *Weitere Einzelheiten, Merkmale und Vorteile der Erfindung ergeben sich aus den in den folgenden Figuren dargestellten und beschriebenen Ausführungsbeispielen.* (3/3)
- *Die Erfindung wird nachfolgend anhand bevorzugter Ausführungsformen unter Bezugnahme auf die Zeichnung beschrieben.* (2/1)

Formulierungsstereotype spanisch: Okkurrenzen gesamt: 16 / Zahl der nachgewiesenen Varianten: 10
Beispiele:

- *Para complementar la descripción que se está realizando y con objeto de ayudar a una mejor comprensión de las características del invento, de acuerdo con un ejemplo preferente de realización práctica del mismo, se acompaña como parte integrante de dicha descripción un juego de dibujos en donde, con carácter ilustrativo y no limitativo, se ha representado lo siguiente:* (6/3)
- *Las características y las ventajas del dispositivo objeto de la presente invención resultarán evidentes a partir de la descripción detallada de una realización preferida del mismo que se dará, de aquí en adelante, a modo de ejemplo no limitativo, con referencia a los dibujos que se acompañan, en los cuales:* (2/1)

4.2.4 Markierung der beschriebenen Ausführungsformen als Beispiele zur Ausweitung des Schutzumfangs

Makrostrukturelle Einbettung: Teiltext ‚Beschreibung' ; Abschnitt ‚Beschreibung eines oder mehrerer Ausführungsbeispiele' (Gliederungssignal als Terminator des Teiltextes ‚Beschreibung')

Formulierungsstereotype deutsch: Okkurrenzen gesamt: 1 / Zahl der nachgewiesenen Varianten: 1
Beispiel:

- *Es versteht sich, dass die vorstehend genannten und die nachstehend noch zu erläuternden Merkmale der Erfindung nicht nur in der jeweils angegebenen Kombination, sondern auch in anderen Kombinationen oder in Alleinstellung verwendbar sind, ohne den Rahmen der Erfindung zu verlassen.* (1/1)

Formulierungsstereotype spanisch: Okkurrenzen gesamt: 15 / Zahl der nachgewiesenen Varianten: 6

Beispiele:

- *Descrita suficientemente la naturaleza de la invención, así como la manera de realizarse en la práctica, debe hacerse constar que las disposiciones anteriormente indicadas son susceptibles de modificaciones de detalle en cuanto no alteren el principio fundamental.* (4/2)

- *Se hace constar que cuantas modificaciones puedan ser introducidas en el objeto de la presente invención, sin alterar su esencialidad característica, se considerarán incluidas en él.* (3/2)

4.2.5 Schlussfolgerungen und praktische Aspekte

Wie aus dieser Übersicht hervorgeht, erfüllen die meisten der nachgewiesenen Formulierungsstereotype die kommunikative Funktion spezifischer, zuweilen nur in einem der beiden Sprachkorpora vorkommender Kohäsionsmittel, die teils zusätzlich als Gliederungssignale fungieren. Aus linguistischer Sicht handelt es sich bei den rekurrierenden Ausgangstextsegmenten teils um identische Wiederholungen der Zeichenkette und teils um Expansionen/Reduktionen von Vergleichssegmenten in Form von Paraphrasen mit oder ohne Informationsverlagerung.

Bei der Frage nach der Verwertbarkeit der Treffer ist im Falle semantischer Abweichungen je nach Fall zu prüfen, ob im Rahmen des Satzinhaltsvergleichs die kommunikative Funktion so sehr Vorrang vor der semantischen Dimension hat, dass die Austauschbarkeit dennoch gegeben ist. Nach dem Aspekt der Retrieval-Relevanz lassen sich bei den erfassten Formulierungsstereotypen dabei zwei Fälle unterscheiden: der Fall der ausdrucksseitigen und semantisch-funktionalen Identität zum einen und der Fall der ausdrucksseitigen Nichtidentität bei a) funktional oder b) semantisch und funktional identischem Satzinhalt zum anderen. Die Austauschbarkeit der entsprechenden Zieltextsegmente im Übersetzungsprozess ist nicht selten auch im zweiten Fall gegeben, da die Oberflächenstruktur dieser metakommunikativen Äußerungen den Schutzumfang häufig nicht berührt.[10]

[10] Zu Recht weist Engberg (1999) darauf hin, dass im Zuge einer differenzierten Übersetzungsstrategie auch bei juristischen Fachtexten die Ersetzung konventionalisierter Formen durch zielkulturell übliche Stereotype sinnvoll sein kann und der dokumentarischen Funktion einer Übersetzung nicht automatisch widerspricht.

Die durchgeführten empirischen Tests machten deutlich, dass es auch im Falle funktional korrespondierender (und somit theoretisch austauschbarer) Stereotype wegen starker Abweichungen auf der Ausdrucksseite häufig zu Retrieval-Problemen kommt. Typische Ursachen hierfür waren lexikalische Ersetzungen durch Kontextsynonyme (vgl. Testsätze (1a) und (1b)) und insbesondere syntaktische Umstellungen z. B. mit Fokusverschiebung und Aktiv-Passiv-Konversen (vgl. Testsatz (1c)):

- Referenzsatz (1): *Vorteilhafte Ausgestaltungen sind in den abhängigen Ansprüchen definiert.*
- Testsatz (1a): *Vorteilhafte Ausgestaltungen sind in den Unteransprüchen beschrieben.*
- Match-Wert (1a): 63%
- Testsatz (1b): *Vorteilhafte Ausgestaltungen der Erfindung sind in den Unteransprüchen dargelegt.*
- Match-Wert (1b): 46%
- Testsatz (1c): *Die Unteransprüche beinhalten vorteilhafte Ausgestaltungen der Erfindung.*
- Match-Wert (1c): kein Match[11]

Besonders deutlich wird die Retrieval-Problematik auch in den häufig nachgewiesenen Fällen, in denen bei gleicher funktionaler Wertigkeit ausgeprägte Oberflächenunterschiede mit stark abweichenden Satzstrukturen und Segmentlängen bestehen, die häufig auf die Verwendung konventionalisierter Redundanzen (vgl. Testsatz (2a)) zurückzuführen sind:

- Referenzsatz (2): *Weitere Merkmale der Erfindung ergeben sich aus der folgenden Beschreibung und den zugehörigen Zeichnungen, in denen Ausführungsbeispiele der Erfindung schematisch dargestellt sind.*
- Testsatz (2a): *Die Einzelheiten, weitere Merkmale und andere Vorteile der Erfindung ergeben sich aus der nachfolgenden Beschreibung von Ausführungsformen der Erfindung, die schematisch, d.h. unter Fortlassung aller für das Verständnis der Erfindung nicht erforderlichen Einzelheiten, in den Figuren der Zeichnungen wiedergegeben sind.*
- Match-Wert (2a): kein Match

[11] Die verwendete Version 7.0.0 der *Translator's Workbench* liefert keine Match-Werte unterhalb des kleinsten einstellbaren Schwellenwerts von 30%. Die Tests wurden mit diesem kleinstmöglichen Schwellenwert durchgeführt. Das Ergebnis „kein Match" kann also im vorliegenden Fall für jeden Vergleichswert unterhalb von 30% stehen.

Wie dieses Beispiel belegt, ist es sinnvoll, patentschriftenspezifische Standardsätze dieser Art zusätzlich in geeigneten Datenbanken zu verwalten. Schmitz (1996) empfiehlt, standardisierte Sätze und Texte wegen ihrer meist fehlenden Begrifflichkeit und aufgrund der Notwendigkeit anderer Datenkategorien nicht zusammen mit den Terminologiebeständen abzulegen, sondern sie stattdessen in speziellen Text(baustein)-Datenbanken mit der Möglichkeit der Anbindung an ein Translation-Memory-System zu verwalten. Wie hingegen Göpferich (1995b) an konkreten Beispielen demonstriert, kann es aus praktischer Sicht dennoch sinnvoll sein, zur Verwaltung textsortenspezifischer Textversatzstücke und spezifischer Zusatzinformationen eine herkömmliche Terminologiedatenbankstruktur heranzuziehen, so dass der Übersetzer bei der Arbeit nicht zwischen mehreren Datenbanken wechseln muss. Wie die Autorin aufzeigt, kann in einer solchen kombinierten Datenbank durch entsprechende Gestaltungsrichtlinien eine leichte Unterscheidung zwischen terminologischen und textographischen Datensätzen ermöglicht werden. Konsequenterweise schlägt sie vor, die begriffsbezogene Angabe zum Fachgebiet durch einen Deskriptor für die Textsorte zu ersetzen.

Unabhängig von einer separaten oder kombinierten Verwaltung erscheint es mir empfehlenswert, sowohl die makrostrukturelle Lokalisierung als auch die Funktion der jeweiligen Standardsätze in geeigneten Datenkategorien zu erfassen, da diese Informationen im Übersetzungsprozess die Rekontextualisierung erheblich erleichtern. Im Übrigen kann die Auffindbarkeit der betreffenden Versatzstücke zusätzlich verbessert werden, indem ihre kommunikative Funktion in einem separaten Dateneintrag als Stichwort eingegeben wird (Details bei Göpferich (1995b)).

4.3 Formulierungsmuster und fachsprachliche Phraseologismen unterhalb der Satzebene

4.3.1 Formulierungsmuster

Formulierungsmuster sind rekurrente und situationstypisch verwendete Form-Inhalts-Beziehungen, die in spezifischen Kommunikationssituationen zur Vermittlung wiederkehrender Inhalte bzw. zum Vollzug wiederkehrender sprachlicher Handlungen genutzt werden. Von phraseologischen Wortverbindungen und satzwertigen Routineformeln unterscheiden sie sich durch eine tendenziell stärker ausgeprägte Strukturvariabilität. (Kühtz (2007))

In beiden Sprachkorpora war eine Vielzahl textsortenspezifischer Formulierungsmuster nachweisbar, die makrostrukturell fest an einzelne Gliederungspunkte bzw. Teiltexte gebunden sind und dabei als spezifische Gliederungssignale eine klar umschriebene kommunikative Funktion erfüllen. Bei allen nachgewiesenen Formulierungsmustern gab es funktionale Entsprechungen im jeweiligen Parallelkorpus.

- Beispieltyp: *Initiatoren der Beschreibung von Merkmalen bevorzugter Ausführungsformen*

- Makrostrukturelle Einbettung: Teiltext ‚Beschreibung' ; Gliederungspunkt ‚Darstellung bevorzugter Ausführungsformen' (teils als Textbegrenzungssignal (1a/1c), teils als Wiederaufnahmesignal (1b/1d))
- Beispiele: (1a) *Eine besonders vorteilhafte Weiterbildung der Erfindung sieht vor, dass ...*
- (1b) *Eine weitere vorteilhafte Ausgestaltung des erfindungsgemäßen Verfahrens sieht vor, dass ...*
- (1c) *De acuerdo con una realización preferida de la presente invención, ...*
- (1d) *De acuerdo con otra realización preferida ...*

Insgesamt wurden zehn Typen von Formulierungsmustern erfasst, die in der folgenden Übersicht auf der Grundlage ihrer kommunikativen Funktion unterteilt sind:

Typ (kommunikative Funktion)	Makrostrukturelle Einbettung
1. Initiatoren der Nennung des Erfindungsgegenstandes (z. T. mit Spezifizierung des Fachgebiets)	Teiltext ‚Beschreibung' / Gliederungspunkt ‚Einordnung in das Fachgebiet'
2. Initiatoren der Bezugnahme auf den Stand der Technik	Teiltext ‚Beschreibung' / Gliederungspunkt ‚Beschreibung des Standes der Technik'
3. Initiatoren des intertextuellen Verweises auf früher erteilte Patente	Teiltext ‚Beschreibung' / Gliederungspunkt ‚Beschreibung des Standes der Technik'durch Verweis auf Fundstellen
4. Initiatoren der Kritik am Stand der Technik (Textbegrenzungssignal oder Wiederaufnahmesignal)	Teiltext ‚Beschreibung' / Gliederungspunkt ‚Kritik am Stand der Technik'

Auch bei identischen Formulierungsmustern und parallelen Satzstrukturen kommt es v. a. bei erheblich variierenden Segmentlängen und umfangreicheren lexikalischen Abweichungen häufig zu Retrieval-Problemen: Referenzsatz (3): Eine weitere vorteilhafte Ausgestaltung des erfindungsgemäßen Verfahrens sieht vor, dass die Trennschicht mit einem Antihafteigenschaften aufweisenden Material gebildet wird. Testsatz (3a): Eine weitere vorteilhafte Ausgestaltung des erfindungsgemäßen Verfahrens sieht vor, dass die Bogen des Bedruckstoffs und die Bogen aus dem elektrisch nicht isolierenden Material nach dem Bedrucken der Vorderseite und dem anschließenden Trocknen auf ihrer Rückseite bedruckt werden können.. Match-Wert (1a): kein Match

Auch dieses Beispiel zeigt deutlich die Wünschbarkeit einer Erkennung von Satzfragmenten. Das Auffinden dieser musterhaften Formulierungen wird bei der Testsoftware zwar durch die Verwendung der Konkordanzsuche ermöglicht; empfehlenswert ist

aber auch die Registrierung patentschriftenspezifischer Formulierungsmuster in einer terminologischen oder textographischen Datenbank. Auch hier sollten sowohl die kommunikative Funktion als auch die makrostrukturelle Lokalisierung eine eigene, die Rekontextualisierung erleichternde Datenkategorie bilden. Auch ansonsten wäre hier in praktischer Hinsicht analog zur Verwaltung von Routineformeln zu verfahren (vgl. 4.2.5).

4.3.2 Fachsprachliche Phraseologismen mit hoher Gebrauchsfrequenz

Die sprachliche Konventionalisierung der Textsorte findet ihren Niederschlag auch im fachphraseologischen Bereich, wobei sich auch hier die in Deutschland und Spanien veröffentlichten Merkblätter für Patentanmelder mit ihren Formulierungsbeispielen und empfehlungen als sprachprägend erweisen. Entsprechend der Zielsetzung der Studie wurden bei der Korpusanalyse nur textsortenspezifische und textsortentypische Fachphraseologismen erfasst. Unberücksichtigt blieben neben nicht-fachsprachlichen Phraseologismen also auch fachgebietsbezogene Kollokationen sowie die stark besetzte Klasse der fachgebietsbezogenen terminologischen Mehrwortverbindungen. Was die registrierten Phraseologismus-Typen angeht, so scheint die Klasse der strukturellen Phraseologismen (präpositionale und konjunktionale Phraseologismen sowie textkommentierende und textdeiktische Formeln) für die Textsorte nur eine vergleichsweise geringe Bedeutung zu spielen. Stark vertreten war dagegen die Klasse der referentiell-nominativen (d. h. satzgliedwertigen) Phraseologismen, die im Folgenden anhand einer kleinen Beispielauswahl charakterisiert werden sollen. 1. Substantivische Phraseologismen Dominierende Bildungsmuster: dt.: Adjektiv + substantivische Basis span.: substantivische Basis + Partizip / + präpositionales Attribut Beispiele: Deutsch Entsprechung(en) im Spanischen erfindungsgemäße Vorrichtung dispositivo propuesto/dispositivo propuesto por la invención/dispositivo de la invención gattungsgemäße Vorrichtung dispositivo del tipo indicado bevorzugte Ausführungsform realización preferida/modo de realización preferido 2. Adjektivische Phraseologismen Dominierende Bildungsmuster: dt.: Adjektiv + Partizip span. (semantische Entsprechungen): Partizip + Präpositionalattribut; Präpositionalattribut mit substantivischer Apposition; Partizip + Adverb Beispiele: Deutsch Entsprechung(en) im Spanischen einstückig ausgebildet constituido por un cuerpo monopieza / del tipo monopieza/con carácter monopieza lösbar verbunden (mit) removiblemente fijado (a/sobre) drehbeweglich verbunden (mit) unido de forma giratoria, fijado giratoriamente (a/sobre) 3 Adverbiale Phraseologismen Bildungsmuster: dt. und span.: heterogenes morphostrukturelles Erscheinungsbild; in beiden Sprachen überwiegend präpositional eingeleitet Beispiele: Deutsch Entsprechung(en) im Spanischen in schematischer Darstellung en representación esquemática in Seitenansicht en una vista lateral In Anbetracht der Vielzahl textsortentypischer Fachphraseologismen ist eine systematische Verwaltung dieser sprachlichen Einheiten in der Terminologiekompo-

nente des TM-Systems dringend zu empfehlen. Besonders geeignet für diesen Zweck ist ein Verwaltungssystem mit begriffsorientiertem Datenmodell, flexibler Eintragsstruktur und ausreichenden Feldlängen, wobei nach Ansicht von Schmitz (1996) das Prinzip der Synonymautonomie angewendet werden sollte. Praktische Hinweise zur Gestaltung der Datenkategorien bei fachphraseologischen Einträgen finden sich bei Budin/Galinski (1992).

5 Schlussfolgerungen

Obgleich die Übersetzung von Patentschriften nicht zu den routinemäßigen Einsatzgebieten von CAT-Tools zählt, erscheint die Arbeit mit einem Translation-Memory-System aus linguistischer Sicht sinnvoll. So konnten in beiden Sprachkorpora zahlreiche Typen textsortenimmanenter Wiederholungen und Ähnlichkeiten nachgewiesen werden, die dem hohen Grad der juristischen Normierung sowie der mikro- und makrostrukturellen Standardisierung von Patentschriften zu verdanken sind. Die textinterne Rekurrenz kam dabei vor allem in Form makrostrukturell bedingter Redundanzen auf Satz- und Teilsatzebene zum Ausdruck, während im Falle der textexternen Rekurrenzen ein breites Spektrum an textsortenspezifischen Routineformeln, Formulierungsmustern und satzgliedwertigen Phraseologismen zu verzeichnen war. Die hochgradige Standardisierung von Argumentationsstrukturen und Textablaufschemata und die daraus resultierende Herausbildung typischer Rekurrenzmuster berührt auch die Verwertbarkeit der Suchergebnisse im Übersetzungsprozess. Ganz besonders gilt dies für eine Vielzahl makrostrukturell gebundener Formulierungsstereotype; zum einen, weil ihre feste Lokalisierung die Rekontextualisierung erheblich erleichtert, und zum anderen, weil innerhalb des Korpus eine breite Palette funktional identischer Formeln registriert wurde, deren Austauschbarkeit selbst im Falle erheblicher lexikalischer und syntaktischer Divergenzen gegeben ist. Auch legen die bei allen Rekurrenztypen durchgeführten Satzinhaltsanalysen den Schluss nahe, dass die meisten Formen potentieller Ambiguität wegen der fachsprachlichen, strukturellen und argumentativen Merkmale der Textsorte äußerst unwahrscheinlich sind. So wurden in keinem der Sprachkorpora Fälle von syntaktischer, referentieller, elliptischer, funktionaler oder illokutiver Mehrdeutigkeit nachgewiesen, was angesichts der funktionalen Eigenschaften der Textsorte auch nicht verwundern kann. Die Retrieval-Relevanz der nachgewiesenen Rekurrenzen war deshalb insgesamt hoch und - textsortenbedingt - in vielen Fällen höher, als der Grad der formalen Übereinstimmung es vermuten ließ. Aus linguistischer Sicht sinnvoll ist die Verwendung eines integrierten Übersetzungssystems mit automatischer Terminologieerkennung und der Möglichkeit der Konkordanzsuche, weil die Einbindung terminologischer bzw. phraseologischer Datenbanken sowohl die Erkennung textsortenspezifischer Phraseologismen und Formulierungsmuster als auch die terminologische Konsistenz der Zieltexte verbessern kann und weil längere, von den Erkennungsalgorithmen marktüblicher

Translation-Memory-Systeme nicht identifizierbare Rekurrenzen unterhalb der Satzebene für die Textsorte Patentschrift besonders typisch sind, so dass auch die Verwendung der Konkordanzsuchfunktion gängiger TM-Systeme erhebliche Produktivitätsvorteile bringen kann. Unverzichtbar ist auch die systematische Verwaltung satzwertiger Formulierungsstereotype, wobei hier sowohl die Registrierung in einer separaten Datenbank als auch die Integration in das verwendete Terminologieverwaltungssystem in Frage kommen (vgl. 4.2.5). Im Hinblick auf die Retrieval-Leistung ist zwar die funktionsbedingt hohe Qualität der Ausgangstexte (geringe Häufigkeit stilistischer Variationen und terminologischer Inkonsistenzen) ein erkennbarer Vorteil; nachteilig wirkt sich allerdings auch bei dieser Textsorte der Umstand aus, dass die Erkennungsalgorithmen kommerziell vertriebener TM-Systeme derzeit in der Regel keine Identifikation von Satzfragmenten ermöglichen. Reinke (2004) schlägt hier konkrete terminologiebezogene Lösungsansätze vor. Da umfangreiche syntaktische Expansionen/Reduktionen für die Textsorte besonders typisch zu sein scheinen (vgl. 4.1.1., Bsp. 5), wäre auch die Bereitstellung satzübergreifender Erkennungsmechanismen durch Anbieter von TM-Systemen sehr zu wünschen. Ein Vergleichstest zwischen der Translator's workbench und dem System Multitrans des Herstellers Multicorpora R&D INC., das als korpus- bzw. textbasiertes Translation Memory konzipiert ist und auch die Erkennung von Segmenten unterhalb der Satzgrenze ermöglicht, erschiene vor diesem Hintergrund lohnend. Wie die Korpusanalyse gezeigt hat, resultieren Retrieval-Schwierigkeiten bei Patentschriften deutlich seltener aus morphosyntaktischen Modifikationen als aus stark variierenden Segmentlängen. Soweit das verwendete Translation-Memory-Programm es ermöglicht, kann die Retrieval-Leistung durch die textsortengerechte Konfiguration der Segmentierungsparameter allerdings in beschränktem Umfang beeinflusst werden. Wünschenswert wäre in diesem Zusammenhang auch die Möglichkeit programmseitiger Standardeinstellungen für spezifische Textsorten und Sprachenkombinationen. Darüber hinaus haben die empirischen Tests gezeigt, dass bei der Übersetzung von Patentschriften die Einstellung eines möglichst niedrigen Match-Schwellenwertes von Vorteil sein kann. Dies erscheint umso bemerkenswerter, als erfahrene TM-Nutzer in der Regel Schwellenwerte von über 70% empfehlen (Seewald-Heeg/Nübel 1999). Die Arbeit mit einer integrierten CAT-Umgebung ist bei der Übersetzung von Patentschriften auch aus praktischer Sicht empfehlenswert, da wegen der juristischen Implikationen der Textsorte (Definition des Schutzumfangs und drohende Haftungsfolgen von Übersetzungsfehlern) auf formaler und inhaltlicher Ebene mit äußerster Akribie übersetzt werden muss. Die kognitive Entlastung, die der Einsatz integrierter Übersetzungssysteme z. B. bei der Absicherung der terminologischen Konsistenz oder bei der identischen Reproduktion textintern rekurrierender Formulierungen auf Satz- und Teilsatzebene bewirken kann, wird im Falle dieser Textsorte zu einem besonders relevanten Qualitätssicherungsfaktor. Zu guter Letzt spricht für die Verwendung von CAT-Tools auch der Umstand, dass sich im Bereich der Patentschriften die Verfügbarkeit maschinenlesbarer Dateien

dank einschlägiger Online-Textdatenbanken in den letzten Jahren erheblich verbessert hat, so dass die erzielbaren Produktivitätsvorteile nicht durch das Einscannen von PDF-Bilddateien und andere Pre-Editing-Arbeiten geschmälert werden.

Literatur

Barb, W. (1982). Praktische Problematik der deutsch-englischen Patentübersetzung und rechtliche Folgen von Übersetzungsfehlern. *Mitteilungen der deutschen Patentanwälte*, 73(6):108–112.

Brungs, B. (1996). Translation Memories als Komponente integrierter Übersetzungssysteme. Eine Untersuchung anhand verschiedener Texttypen. In *Saarbrücker Studien zu Sprachdatenverarbeitung und Übersetzen*. Saarbrücken: Fachrichtung 8.6, Universität des Saarlandes; hrsg. von K.-H. Freigang und U. Reinke.

Budin, G. and Galinski, C. (1992). Übersetzungsorientierte Phraseologieverwaltung in Terminologiedatenbanken. *Terminologie et traduction*, 2(3):565–574.

de Beaugrande, R. and Dressler, W. (1981). *Einführung in die Textlinguistik*. Tübingen: Niemeyer.

Dederding, H.-M. (1982a). Verschiedene Bezeichnungen für einen technischen Gegenstand. *Mitteilungen der deutschen Patentanwälte*, 73(9):164–168.

Dederding, H.-M. (1982b). *Wortbildung, Syntax, Text. Nominalkomposita und entsprechende syntaktische Strukturen in deutschen Patent- und Auslegeschriften*. Number 34 in Erlanger Studien. Erlangen: Palm & Enke.

Dybdahl, L. (2004). *Europäisches Patentrecht. Einführung in das europäische Patentsystem*. Köln: Heymann.

Engberg, J. (1999). Übersetzen von Gerichtsurteilen: der Einfluss der Perspektive. In Sandrini, P., editor, *Übersetzen von Rechtstexten. Fachkommunikation im Spannungsfeld zwischen Rechtsordnung und Sprache*, pages 83–101. Tübingen: Narr.

Gläser, R. (556-562). Fachtextsorten der Technniksprachen: die Patentschrift. In et al., L. H., editor, *Fachsprachen. Ein internationales Handbuch zur Fachsprachenforschung und Terminologiewissenschaft*. Berlin: de Gruyter.

Glover, A. and Hirst, G. (1996). Detecting stylistic inconsistencies in collaborative writing. In Sharpes, M. and van der Geest, T., editors, *The new writing environment: Writers at work in a world of technology*, pages 147–168. London: Springer.

Göpferich, S. (1995a). *Textsorten in Naturwissenschaften und Technik. Pragmatische Typologie - Kontrastierung - Translation*. Tübingen: Narr.

Göpferich, S. (1995b). Von der Terminographie zur Textographie: computergestützte Verwaltung textsortenspezifischer Versatzstücke. *Fachsprache/Internationale Zeitschrift für Fachsprachenforschung, -didaktik und Terminologie*, 17(1-2):17–41.

Göpferich, S. (2006). Patentschriften. In et al., M. S.-H., editor, *Handbuch Translation*, pages 222–225. Tübingen: Stauffenburg.

Höcker, M. (2003). ecolore translation memory survey 2003.

Hohnhold, I. (1992). Terminologisch relevante Phraseologie in Fachtexten. *Terminologie et traduction*, 2(3):251–270.

Kjær, A. (1991). Phraseologische Wortverbindungen in der Rechtssprache? In Palm, C., editor, *Europhras*, pages 115–122. Uppsala: Almqvist & Wiksell.

Kühtz, S. (2007). *Phraseologie und Formulierungsmuster in medizinischen Texten*. Tübingen: Narr.

Linke, A. and Nussbaumer, M. (2000). Rekurrenz. In et al., K. B., editor, *Text- und Gesprächslinguistik*, Handbücher zur Sprach- und Kommunikationswissenschaft (HSK) 16.1, pages 305–315. Berlin, New York: de Gruyter.

Liu, Y. (1992). *Fachsprachliche Zeige- und Verweisungsstrukturen in Patentschriften*. München: Iudicium-Verlag.

Merkel, M. (1996). Checking Translations for Inconsistency: A Tool for the Editor. In *Expanding MT Horizons. Proceedings of the Second Conference for Machine Translation in the Americas. 2-5 October, 1996. Montreal, Canada*, pages 157–167. Washington DC: Association for Machine Translation in the Americas (AMTA).

Raible, H. (1987). Europa-Übersetzungen - ein Geschäft mit enormem Risiko. *Mitteilungen der deutschen Patentanwälte*, 78(12):225–233.

Raible, W. (1972). *Satz und Text. Untersuchungen zu vier romanischen Sprachen*. Tübingen: Niemeyer.

Reinke, U. (1999). Überlegungen zu einer engeren Verzahnung von Terminologiedatenbanken, Translation Memories und Textkorpora. *LDV-Forum*, 16(1/2):64–80.

Reinke, U. (1999a). Evaluierung der linguistischen Leistungsfähigkeit von Translation Memory Systemen. Ein Erfahrungsbericht. *LDV-Forum*, 16(1/2):100–117.

Reinke, U. (2004). *Translation Memories: Systeme – Konzepte – Linguistische Optimierung*. Frankfurt/M. u. a.: Peter Lang.

Schamlu, M. (1985a). *Patentschriften – Patentwesen. Eine argumentationstheoretische Analyse der Textsorte Patentschrift am Beispiel der Patentschriften zu Lehrmitteln*. München: Iudicium-Verlag.

Schamlu, M. (1985b). Zur sprachlichen Darstellung von Patentansprüchen. *Mitteilungen der deutschen Patentanwälte*, 76(3):44–47.

Scheel, H. (1997a). Sprachliche Konventionen in französischen Patentschriften. In Fleischmann, E., editor, *Translationsdidaktik. Grundfragen der Übersetzungswissenschaft*, pages 487–493. Tübingen: Narr.

Scheel, H. (1997b). Zur Makrostruktur deutscher und französischer Patentschriften. In Wotjak, G., editor, *Studien zum romanisch-deutschen und innerromanischen Sprachvergleich*, pages 143–155. Frankfurt/M.: Peter Lang.

Schmitz, K. D. (1996). Verwaltung sprachlicher Einheiten in Terminologieverwaltungssystemen. In et al., A. L., editor, *Übersetzungswissenschaft im Umbruch. Festschrift für Wolfram Wilss zum 70. Geburtstag*, pages 197–207. Tübingen: Narr.

Seewald-Heeg, U. (2005). Der Einsatz von Translation-Memory-Systemen am Übersetzerarbeitsplatz. *MDÜ (Mitteilungen für Dolmetscher und Übersetzer)*, 51(4-5):8–38.

Seewald-Heeg, U. and Nübel, R. (1999). Translation-Memory-Module automatischer Übersetzungssysteme. *LDV-Forum*, 16(1/2):16–35.

Stein, S. (2001). Formelhafte Texte. Musterhaftigkeit an der Schnittstelle zwischen Phraseologie und Textlinguistik. In Lorenz-Bourjot, M. and Lüger, H.-H., editors, *Phraseologie und Phraseodidaktik*. Wien: Edition Praesens.

Martin Volk

The Automatic Translation of Film Subtitles. A Machine Translation Success Story?

1 Introduction

Every so often one hears the complaint that 50 years of research in Machine Translation (MT) has not resulted in much progress, and that current MT systems are still unsatisfactory. A closer look reveals that web-based general-purpose MT systems are used by thousands of users every day. And, on the other hand, special-purpose MT systems have been in long-standing use and work successfully in particular domains or for specific companies.

This paper[1] investigates whether the automatic translation of film subtitles can be considered a machine translation success story. We describe various projects on MT of film subtitles and contrast them to our own project in this area. We argue that the text genre "film subtitles" is well suited for MT, in particular for Statistical MT. But before we look at the translation of film subtitles let us retrace some other MT success stories.

Hutchins (1999) lists a number of successful MT systems. Amongst them is Météo, a system for translating Canadian weather reports between English and French which is probably the most quoted MT system in practical use. References to Météo usually remind us that this is a "highly constrained sublanguage system". On the other hand there are general purpose but customer-specific MT systems like the English to Spanish MT system at the Pan American Health Organization or the PaTrans system which Hutchins (1999) calls "... possibly the best known success story for custom-built MT". PaTrans was developed for LingTech A/S to translate English patents into Danish.

Earlier (Whitelock and Kilby (1995), p.198) had called the METAL system "a success story in the development of MT". METAL is mentioned as "successfully used at a number of European companies" (by that time this meant a few dozen installations in industry, trade or banking). During the same time the European Union has been successfully using a customized version of Systran for its translation service but also later for online access by all its employees. Broad coverage systems like METAL and Systran have always resulted in a translation quality that required post-editing before publications.

[1] This is a slightly corrected and updated version of a paper first published in: Joakim Nivre, Mats Dahllöf, Beáta Megyesi (Eds.) *Resourceful Language Technology: Festschrift in Honor of Anna Sågvall Hein*, Uppsala University, 2008.

Attempts to curb the post-editing by pre-editing or constraining the source documents have gone under the name of controlled language MT. Hutchins (1999) mentions controlled language MT (e.g. at the Caterpillar company) as an example of successful employment of MT. This is an area where part of the pioneering work was done at Uppsala University by Anna Sågvall Hein and her group (Almqvist and Sågvall Hein, 1996), including the development of controlled Swedish for the automobile industry. This research subsequently led to a competitive MT system for translating from Swedish to English (Sågvall Hein et al., 2002).

The claim that web-based machine translation is a success is based on the fact that it is used by large numbers of users. Critics do not subscribe to this argument as long as the translation quality is questionable. Still, popular services including Systran (www.systran.co.uk with 14 source languages) and Google (www.google.com/translate_t with 21 language pairs) cover major Western languages like English, Spanish and French, but also Arabic and Chinese. On the other hand there are providers that have successfully occupied niche language pairs like Danish to English (Bick, 2007).

So we see that MT success stories vary considerably. We regard the following criteria as the main indicators of success:

1. A large user base (this criterion is used in web-based MT services for the general public)
2. Customer satisfaction (this criterion is used in customer-specific MT systems and usually based on improved productivity and return on investment)
3. Long-term usage of the MT system

We will check which of these criteria apply to the automatic translation of film subtitles.

2 Characteristics of Film Subtitles

When films are shown to audiences in language environments that differ from the language spoken in the film, then some form of translation is required. Larger markets like Germany and France typically use dubbing of foreign films so that it seems that the actors are speaking the local language. Smaller countries often use subtitles. Pedersen (2007) discusses the advantages and drawbacks of both methods.

Foreign films and series shown in Scandinavian TV are usually subtitled rather than dubbed. Therefore the demand for Swedish, Danish, Norwegian and Finnish subtitles is high. These subtitles are meant for the general public in contrast to subtitles that are specific for the hearing-impaired which often include descriptions of sounds, noises and music. Subtitles also differ with respect to whether they are produced online (e.g. in live

talkshows or sport reports) or offline (e.g. for pre-produced series). This paper focuses on general public subtitles that are produced offline.

In our machine translation project, we use a parallel corpus of Swedish, Danish and Norwegian subtitles. The subtitles in this corpus are limited to 37 characters per line and usually to two lines.[2] Depending on their length, they are shown on screen between 2 and 8 seconds. Subtitles typically consist of one or two short sentences with an average number of 10 tokens per subtitle in our corpus. Sometimes a sentence spans more than one subtitle. It is then ended with a hyphen and resumed with a hyphen at the beginning of the next subtitle. This occurs about 35.7 times for each 1000 subtitles in our corpus.

Example 1 shows a human-translated pair of subtitles that are close translation correspondences although the Danish translator has decided to break the two sentences of the Swedish subtitle into three sentences.[3]

(1) SV: Det är slut, vi hade förfest här. Jätten drack upp allt.
DA: Den er væk. Vi holdt en forfest. Kæmpen drak alt.
EN: *It is gone. We had a pre-party here. The giant drank it all.*

In contrast, the pair in 2 exemplifies a slightly different wording chosen by the Danish translator.

(2) SV: Där ser man vad framgång kan göra med en ung person.
DA: Der ser man, hvordan succes ødelægger et ungt menneske.
EN: *There you see, what success can do to a young person / how success destroys a young person.*

This paper can only give a rough characterization of subtitles. A more comprehensive description of the linguistic properties of subtitles can be found in (de Linde and Kay, 1999) and (Díaz-Cintas and Remael, 2007). Gottlieb (2001) and Pedersen (2007) describe the peculiarities of subtitling in Scandinavia.

3 Approaches to the Automatic Translation of Film Subtitles

In this section we describe other projects on the automatic translation of subtitles. We distinguish between rule-based, example-based, and statistical approaches.

[2] Although we are working on both Swedish to Danish and Swedish to Norwegian MT of subtitles, this paper focuses on translation from Swedish to Danish. The issues for Swedish to Norwegian are the same to a large extent.

[3] In this example and in all subsequent subtitle examples the English translations were added by the author.

3.1 Rule-based MT of Film Subtitles

Popowich et al. (2000) provide a detailed account of a MT system tailored towards the translation of English subtitles into Spanish. Their approach is based on a MT paradigm which relies heavily on lexical resources but is otherwise similar to the transfer-based approach. A unification-based parser analyzes the input sentence (including proper-name recognition), followed by the lexical transfer which provides the input for the generation process in the target language (including word selection and correct inflection).

Popowich et al. (2000) mention that the subtitle domain has certain advantages for MT. According to them it is advantageous that output subtitles can and should be grammatical even if the input sometimes is not. They argue that subtitle readers have only a limited time to perceive and understand a given subtitle and that therefore grammatical output is essential. And they follow the strategy that "it is preferable to drop elements from the output instead of translating them incorrectly" (p.331). This is debateable and opens the door for incomplete output.

Although Popowich et al. (2000) call their system "a hybrid of both statistical and symbolic approaches" (p.333), it is a symbolic system by today's standards. The statistics are only used for efficiency improvements but are not at the core of the methodology. The paper was published before automatic evaluation methods were invented. Instead Popowich et al. (2000) used the classical evaluation method where native speakers were asked to judge the grammaticality and fidelity of the system. These experiments resulted in "70% of the translations ... be ranked as correct or acceptable, with 41% being correct" which is an impressive result. Whether this project can be regarded as a MT success story depends on whether the system was actually employed in production. This information is not provided in the paper.

Melero et al. (2006) combined Translation Memory technology with Machine Translation, which looks interesting at first sight. But then it turns out that their Translation Memories for the language pairs Catalan-Spanish and Spanish-English were not filled with subtitles but rather with newspaper articles and UN texts. They don't give any motivation for this. And disappointingly they did not train their own MT system but rather worked only with free-access web-based MT systems (which we assume are rule-based systems).

They showed that a combination of Translation Memory with such web-based MT systems works better than the web-based MT systems alone. For English to Spanish translation this resulted in an improvement of around 7 points in BLEU scores (Papineni et al., 2001) but hardly any improvement at all for English to Czech.

3.2 Example-based MT of Film Subtitles

Armstrong et al. (2006) "ripped" subtitles (40,000 sentences) German and English as

training material for their Example-based MT system and compared the performance to the same amount of Europarl sentences (which have more than three times as many tokens!). Training on the subtitles gave slightly better results when evaluating against subtitles, compared to training on Europarl and evaluating against subtitles. This is not surprising, although the authors point out that this contradicts some earlier findings that have shown that heterogeneous training material works better.

They do not discuss the quality of the ripped translations nor the quality of the alignments (which we found to be a major problem when we did similar experiments with freely available English-Swedish subtitles).

The BLEU scores are on the order of 11 to 13 for German to English (and worse for the opposite direction). These are very low scores. They also conducted user evaluations with 4-point scales for intelligibility and accuracy. They asked 5 people per language pair to rate a random set of 200 sentences of system output. The judges rated English to German translations higher than the opposite direction (which contradicts the BLEU scores). Owing to the small scale of the evaluation, however, it seems premature to draw any conclusions.

3.3 Statistical MT of Film Subtitles

Descriptions of Statistical MT systems for subtitles are practically non-existent, probably due to the lack of freely available training corpora. Until recently there were no freely available subtitle collections. Both Tiedemann (2007) and Lavecchia et al. (2007) report on efforts to build such corpora with alignment on the subtitles.

Tiedemann (2007) works with a huge collection of subtitle files that are available on the internet at www.opensubtitles.org. These subtitles have been produced by volunteers in a great variety of languages. But the volunteer effort also results in subtitles of often dubious quality (they include timing, formatting, and linguistic errors). The hope is that the enormous size of the corpus will supersede the noise in practical applications. The first step then is to align the files across languages on the subtitle level. The time codes alone are not sufficient as different (amateur) subtitlers have worked with different time offsets and sometimes even different versions of the same film. Still, Tiedemann (2007) shows that an alignment approach based on time overlap combined with cognate recognition is clearly superior to pure length-based alignment. He has evaluated his approach on English, German and Dutch. His results of 82.5% correct alignments for Dutch-English and 78.1% correct alignments for Dutch-German show how difficult the alignment task is. And a rate of around 20% incorrect alignments will certainly be problematic when training a Statistical MT system on these data.

Lavecchia et al. (2007) also work with subtitles obtained from the internet. They work on French-English subtitles and use a method which they call Dynamic Time Warping for aligning the files across the languages. This method requires access to a bilingual

dictionary to compute subtitle correspondences. They compiled a small test corpus consisting of 40 subtitle files, randomly selecting around 1300 subtitles from these files for manual inspection. Their evaluation focused on precision while sacrificing recall. They report on 94% correct alignments when turning recall down to 66%. They then go on to use the aligned corpus to extract a bilingual dictionary and to integrate this dictionary in a Statistical MT system. They claim that this improves the MT system with 2 points BLEU score (though it is not clear which corpus they have used for evaluating the MT system).

This summary indicates that most work on the automatic translation of film subtitles with Statistical MT is still in its infancy. Our own efforts are larger and have resulted in a mature MT system. We will report on them in the following section.

4 The Stockholm MT System for Film Subtitles

We have built Machine Translation systems for translating film subtitles from Swedish to Danish (and Swedish to Norwegian) in a commercial setting. Some of this work has been described earlier by Volk and Harder (2007).

Most films are originally in English and receive Swedish subtitles based on the English video and audio (sometimes accompanied by an English manuscript). The creation of the Swedish subtitle is a manual process done by specially trained subtitlers following company-specific guidelines. In particular, the subtitlers set the time codes (beginning and end time) for each subtitle. They use an in-house tool which allows them to attach the subtitle to specific frames in the video.

The Danish or Norwegian translator subsequently has access to the original English video and audio but also to the Swedish subtitles and the time codes. In most cases the translator will reuse the time codes and insert the target language subtitle. She can, on occasion, change the time codes if she deems them inappropriate for the target language.

Our task is to produce Danish and Norwegian draft translations to speed up the translators' work. This project of automatically translating subtitles from Swedish to Danish and Norwegian benefits from three favorable conditions:

1. Subtitles are short textual units with little internal complexity (as described in section 2).

2. Swedish, Danish and Norwegian are closely related languages.

3. We have access to large numbers of Swedish subtitles and human-translated Danish and Norwegian subtitles. Their correspondence can easily be established via the time codes which leads to an alignment on the subtitle level.

But there are also aspects of the task that are less favorable. Subtitles are not transcriptions, but written representations of spoken language. As a result the linguistic structure of subtitles is closer to written language than the original (English) speech, and the original spoken content usually has to be condensed by the Swedish subtitler.

The task of translating subtitles also differs from most other machine translation applications in that we are dealing with creative language, and thus we are closer to literary translation than technical translation. This is obvious in cases where rhyming song-lyrics or puns are involved, but also when the subtitler applies his linguistic intuitions to achieve a natural and appropriate wording which blends into the video without disturbing. Finally, the language of subtitling covers a broad variety of domains from educational programs on any conceivable topic to exaggerated modern youth language.

We have decided to build a statistical MT (SMT) system in order to shorten the development time (compared to a rule-based system) and in order to best exploit the existing translations. We have trained our SMT system by using GIZA++ (Och and Ney, 2004)[4] for the alignment, Thot (Ortiz-Martínez et al., 2005)[5] for phrase-based SMT, and Phramer[6] as the decoder.

We will first present our setting and our approach for training the SMT system and then describe the evaluation results.

4.1 Swedish and Danish in Comparison

Swedish and Danish are closely related Germanic languages. Vocabulary and grammar are similar, however orthography differs considerably, word order differs somewhat and, of course, pragmatics avoids some constructions in one language that the other language prefers. This is especially the case in the contemporary spoken language, which accounts for the bulk of subtitles.

One of the relevant differences for our project concerns word order. In Swedish the verb takes non-nominal complements before nominal ones, where in Danish it is the other way round. The core problem can be seen in example 3 where the verb particle *ut* immediately follows the verb in Swedish but is moved to the end of the clause in Danish.

(3) SV: Du häller ut krutet.
DA: Du hælder krudtet ud.
EN: *You are pouring out the gunpowder.*

A similar word order difference occurs in positioning the negation adverb (SV: *inte*, DA: *ikke*). Furthermore, Danish distinguishes between the use of *der* (EN: *there*) and *det*

[4]GIZA++ is accessible at http://www.fjoch.com/ GIZA++.html
[5]Thot is available at http://thot.sourceforge.net/
[6]Phramer was written by Marian Olteanu and is available at http://www.olteanu.info/

(EN: *it*) but Swedish does not. Both Swedish and Danish mark definiteness with a suffix on nouns, but Danish does not have the double definiteness marking of Swedish.

4.2 Our Subtitle Corpus

Our corpus consists of TV subtitles from soap operas (like daily hospital series), detective series, animation series, comedies, documentaries, feature films etc. In total we have access to more than 14,000 subtitle files (= single TV programmes) in each language, corresponding to more than 5 million subtitles (equalling more than 50 million words).

When we compiled our corpus we included only subtitles with matching time codes. If the Swedish and Danish time codes differed more than a threshold of 15 TV-frames (0.6 seconds) in either start or end-time, we suspected that they were not good translation equivalents and excluded them from the subtitle corpus. In this way we were able to avoid complicated alignment techniques. Most of the resulting subtitle pairs are high-quality translations of one another thanks to the controlled workflow in the commercial setting.

In a first profiling step we investigated the vocabulary size of the corpus. After removing all punctuation symbols and numbers we counted all word form types. We found that the Swedish subtitles amounted to around 360,000 word form types. Interestingly, the number of Danish word form types is about 5.5% lower, although the Danish subtitles have around 1.5% more tokens. We believe that this difference may be an artifact of the translation direction from Swedish to Danish which may lead the translator to a restrictive Danish word choice.

Another interesting profiling feature is the repetitiveness of the subtitles. We found that 28% of all Swedish subtitles in our training corpus occur more than once. Half of these recurring subtitles have exactly one Danish translation. The other half have two or more different Danish translations which are due to context differences combined with the high context dependency of short utterances and the Danish translators choosing less compact representations.

From our subtitle corpus we chose a random selection of files for training the translation model and the language model. We currently use 4 million subtitles for training. From the remaining part of the corpus, we selected 24 files (approximately 10,000 subtitles) representing the diversity of the corpus from which a random selection of 1000 subtitles was taken for our test set. Before the training we tokenized the subtitles (e.g. separating punctuation symbols from words), converting all uppercase words into lower case, and normalizing punctuation symbols, numbers and hyphenated words.

4.3 Unknown Words

Although we have a large training corpus, there are still unknown words (words not seen in the training data) in the evaluation data. They comprise proper names of people or products, rare word forms, compounds, spelling deviations and foreign words. Proper names need not concern us in this context since the system will copy unseen proper names (like all other unknown words) into the Danish output, which in almost all cases is correct.

Rare word forms and compounds are more serious problems. Hardly ever do all forms of a Swedish verb occur in our training corpus (regular verbs have 7 forms). So even if 6 forms of a Swedish verb have been seen frequently with clear Danish translations, the 7th will be regarded as an unknown if it is missing in the training data.

Both Swedish and Danish are compounding languages which means that compounds are spelled as orthographic units and that new compounds are dynamically created. This results in unseen Swedish compounds when translating new subtitles, although often the parts of the compounds were present in the training data. We therefore generate a translation suggestion for an unseen Swedish compound by combining the Danish translations of its parts.

Variation in graphical formatting also poses problems. Consider spell-outs, where spaces, commas, hyphens or even full stops are used between the letters of a word, like "I will n o t do it", "Seinfeld" spelled "S, e, i, n, f, e, l , d" or "W E L C O M E T O L A S V E G A S", or spelling variations like *ä-ä-älskar* or *abso-jävla-lut* which could be rendered in English as *lo-o-ove* or *abso-damned-lutely*. Subtitlers introduce such deviations to emphasize a word or to mimic a certain pronunciation. We handle some of these phenomena in pre-processing, but, of course, we cannot catch all of them due to their great variability.

Foreign words are a problem when they are homographic with words in the source language Swedish (e.g. when the English word *semester* = "university term" interferes with the Swedish word *semester* which means "vacation"). Example 4 shows how different languages (here Swedish and English) are sometimes intertwined in subtitles.

(4) SV: Hon gick ut Boston University's School of the Performing Arts-
-och hon fick en dubbelroll som halvsystrarna in "As the World Turns".
EN: *She left Boston University's School of the Performing Arts and she got a double role as half sisters in "As the World Turns".*

4.4 Evaluating the Performance of the Stockholm MT System

We first evaluated the MT output against a left-aside set of previous human translations. We computed BLEU scores of around 57 in these experiments. In addition we computed the percentage of exactly matching subtitles against a previous human translation (How

	Exact matches	Levenshtein-5 matches	BLEU
Crime series	15.0%	35.3%	63.9
Comedy series	9.1%	30.6%	54.4
Car documentary	3.2%	22.8%	53.6
Average	9.1%	21.6%	57.3

Table 1: Evaluation Results against a Prior Human Translation

often does our system produce the exact same subtitle as the human translator?), and we computed the percentage of subtitles with a Levenshtein distance of up to 5 which means that the system output has an editing distance of at most 5 basic character operations (deletions, insertions, substitutions) from the human translation.

We decided to use a Levenshtein distance of 5 as a threshold value as we consider translations at this edit distance from the reference text still to be "good" translations. Such a small difference between the system output and the human reference translation can be due to punctuation, to inflectional suffixes (e.g. the plural -s in example 5 with MT being our Danish system output and HT the human translation) or to incorrect pronoun choices.

(5) MT: Det gør ikke noget. Jeg prøver gerne hotdog med kalkun -
HT: Det gør ikke noget. Jeg prøver gerne hotdog**s** med kalkun**,** -
EN: *That does not matter. I like to try hotdog(s) with turkey.*

Table 1 shows the results for three files (selected from different genres), for which we have prior translations (done independently of our system). We observe between 3.2% and 15% exactly matching subtitles, and between 22.8% and 35.3% subtitles with a Levenshtein distance of up to 5. Note that the percentage of Levenshtein matches includes the exact matches (which correspond to a Levenshtein distance of 0).

On manual inspection, however, many automatically produced subtitles which were more than 5 keystrokes away from the human translations still looked like good translations. Therefore we conducted another series of evaluations with translators who were asked to post-edit the system output rather than to translate from scratch. We made sure that the translators had not translated the same file before.

Table 2 shows the results for the same three files for which we have one prior translation. We gave our system output to six translators and obtained six post-edited versions. Some translators were more generous than others, and therefore we averaged their scores. When using post-editing, the evaluation figures are 13.2 percentage points

	Exact matches	Levenshtein-5 matches	BLEU
Crime series	27.7%	47.6%	69.9
Comedy series	26.0%	45.7%	67.7
Car documentary	13.2%	35.9%	59.8
Average	22.3%	43.1%	65.8

Table 2: Evaluation Results averaged over 6 Post-editors

higher for exact matches and 19.5 percentage points higher for Levenshtein-5 matches. It becomes also clear that the translation quality varies considerably across film genres. The crime series file scored consistently higher than the comedy file which in turn was clearly better than the car documentary.

There are only few other projects on Swedish to Danish Machine Translation (and we have not found a single one on Swedish to Norwegian). Koehn (2005) trained his system on a parallel corpus of more than 20 million words from the European parliament. In fact he trained on all combinations of the 11 languages in the Europarl corpus. Koehn (2005) reports a BLEU score of 30.3 for Swedish to Danish translation which ranks somewhere in the middle when compared to other language pairs from the Europarl corpus. The worst score was for Dutch to Finnish (10.3) and the best for Spanish to French translations (40.2). The fact that our BLEU scores are much higher even when we evaluate against prior translations (cf. the average of 57.3 in table 1) is probably due to the fact that subtitles are shorter than Europarl sentences and perhaps also due to our larger training corpus.

5 Conclusions

We have sketched the text genre characteristics of film subtitles and shown that Statistical MT of subtitles leads to good quality when the input is a large high-quality parallel corpus. We are working on Machine Translation systems for translating Swedish film subtitles to Danish and Norwegian with very good results (in fact the results for Swedish to Norwegian are slightly better than for Swedish to Danish).

We have shown that evaluating the system against independent translations does not give a true picture of the translation quality and thus of the usefulness of the system. Evaluation BLEU scores were about 8.5 points higher when we compared our system output against post-edited translations averaged over six translators. Exact matches and Levenshtein 5 scores were also clearly higher.

We are dealing with customer-specific MT systems covering a broad set of textual domains. The customer is satisfied and has employed our MT systems in large scale subtitle production since early 2008. The MT systems have resulted in considerable time savings in the translation process. It is by now safe to call this a Machine Translation success story.

6 Acknowledgements

We would like to thank Jörgen Aasa, Søren Harder and Christian Hardmeier for sharing their expertise, providing evaluation figures and commenting on an earlier version of the paper.

References

Almqvist, I. and Sågvall Hein, A. (1996). Defining ScaniaSwedish - a controlled language for truck maintenance. In *Proceedings of the First International Workshop on Controlled Language Applications*, Katholieke Universiteit Leuven.

Armstrong, S., Way, A., Caffrey, C., Flanagan, M., Kenny, D., and O'Hagan, M. (2006). Improving the quality of automated DVD subtitles via example-based machine translation. In *Proc. of Translating and the Computer 28*, London. Aslib.

Bick, E. (2007). Dan2eng: Wide-coverage Danish-English machine translation. In *Proc. of Machine Translation Summit XI*, Copenhagen.

de Linde, Z. and Kay, N. (1999). *The Semiotics of Subtitling*. St. Jerome Publishing, Manchester.

Díaz-Cintas, J. and Remael, A. (2007). *Audiovisual Translation: Subtitling*, volume 11 of *Translation Practices Explained*. St. Jerome Publishing, Manchester.

Gottlieb, H. (2001). Texts, translation and subtitling - in theory, and in Denmark. In Holmboe, H. and Isager, S., editors, *Translators and Translations*, pages 149–192. Aarhus University Press. The Danish Institute at Athens.

Hutchins, J. (1999). The development and use of machine translation systems and computer-based translation tools. In *Proc. of International Symposium on Machine Translation and Computer Language Information Processing*, Beijing.

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

Lavecchia, C., Smaili, K., and Langlois, D. (2007). Machine translation of movie subtitles. In *Proc. of Translating and the Computer 29*, London. Aslib.

Melero, M., Oliver, A., and Badia, T. (2006). Automatic multilingual subtitling in the eTITLE project. In *Proc. of Translating and the Computer 28*, London. Aslib.

Och, F. J. and Ney, H. (2004). The alignment template approach to statistical machine translation. *Computational Linguistics*, 30(4):417–449.

Ortiz-Martínez, D., García-Varea, I., and Casacuberta, F. (2005). Thot: A toolkit to train phrase-based statistical translation models. In *Tenth Machine Translation Summit*, Phuket. AAMT.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2001). Bleu: a method for automatic evaluation of machine translation. Technical Report RC22176 (W0109-022), IBM Research Division, Thomas J. Watson Research Center, Almaden.

Pedersen, J. (2007). *Scandinavian Subtitles. A Comparative Study of Subtitling Norms in Sweden and Denmark with a Focus on Extralinguistic Cultural References*. PhD thesis, Stockholm University. Department of English.

Popowich, F., McFetridge, P., Turcato, D., and Toole, J. (2000). Machine translation of closed captions. *Machine Translation*, 15:311–341.

Sågvall Hein, A., Forsbom, E., Tiedemann, J., Weijnitz, P., Almqvist, I., Olsson, L.-J., and Thaning, S. (2002). Scaling up an MT prototype for industrial use - databases and data flow. In *Proceedings of LREC 2002. Third International Conference on Language Resources and Evaluation*, pages 1759 – 1766, Las Palmas.

Tiedemann, J. (2007). Improved sentence alignment for movie subtitles. In *Proceedings of RANLP*, Borovets, Bulgaria.

Volk, M. and Harder, S. (2007). Evaluating MT with translations or translators. What is the difference? In *Machine Translation Summit XI Proceedings*, Copenhagen.

Whitelock, P. and Kilby, K. (1995). *Linguistic and Computational Techniques in Machine Translation System Design*. Studies in Computational Linguistics. UCL Press, London, 2 edition.

Autorenverzeichnis JLCL

Dino Azzano
Centrum für Informations- und
Sprachverarbeitung (CIS)
Ludwig-Maximilians-Universität
München
Oettingenstr. 67
80538 München
dino.azzano@nectarine.it

Michael Carl
Institut für Angewandte
Informationsforschung (IAI)
Universität des Saarlandes
Martin-Luther Str. 14
66121 Saarbrücken
carl@iai.uni-sb.de

Kurt Eberle
Lingenio GmbH
Karlsruher Str. 10
69126 Heidelberg
k.eberle@lingenio.de

Heribert Härtinger
Institut für Translation und
Mehrsprachige Kommunikation
Fachhochschule Köln
Mainzer Str. 5
50678 Köln
heribert.haertinger@fh-koeln.de

Uta Seewald-Heeg
Computerlinguistik und Fachübersetzen
Fachbereich Informatik
Hochschule Anhalt
Lohmannstraße 23
06366 Köthen
uta.seewald-heeg@inf.hs-anhalt.de

Daniel Stein
Centrum für Informations- und
Sprachverarbeitung
Ludwig-Maximilians-Universität
Postfach 200
Geschwister-Scholl-Platz 1
80539 München
ds@daniel-stein.com

Martin Volk
Institut für Computerlinguistik
Universität Zürich
Binzmühlestr. 14
CH-8050 Zürich
Schweiz
volk@cl.uzh.ch