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

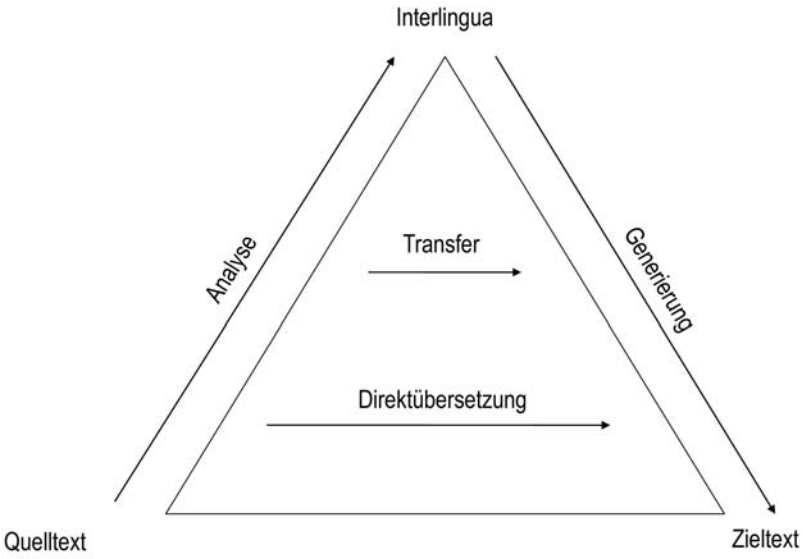
wieder. Die von regelbasierten Systemen produzierten Ergebnisse reichen von kurios bis nützlich, ganz in Abhängigkeit davon, um welches Sprachpaar es sich handelt und ob eine Fachsprache unterstützt wird und entsprechend Fachterminologie eingepflegt wurde, oder ob es sich um ein allgemeinsprachliches 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:

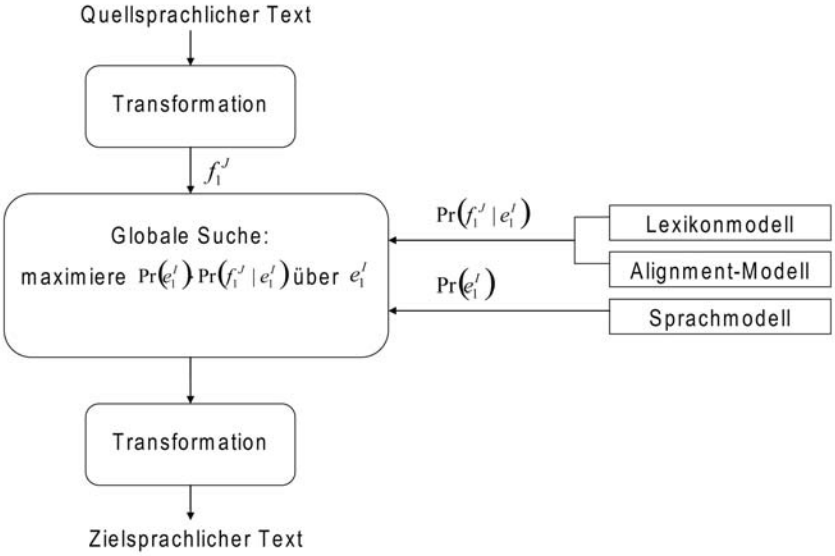
$$Pr(e|f) = \frac{Pr(e) * Pr(f|e)}{Pr(f)} \tag{1}$$

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.