Überlegungen zu einer engeren Verzahnung von Terminologiedatenbanken, Translation Memories und Textkorpora

Uwe Reinke
Universität des Saarlandes

1 Interaktion der Komponenten von Translation Memory-Systemen

TM-Systeme besitzen im wesentlichen zwei 'Wissensquellen': Terminologiedatenbank und Referenzmaterial, d.h. eine TM-Datenbank oder eine maschinenlesbare Sammlung von Texten und ihren Übersetzungen. Bislang ist das Maß an Interaktion zwischen diesen beiden 'Wissensquellen' allerdings eher gering. So werden Terminologiedatenbanken lediglich benutzt, um in den zu übersetzenden Texten Termini zu identifizieren und die entsprechenden zielsprachlichen Benennungen zur Verfügung zu stellen. Keines der verfügbaren kommerziellen Systeme setzt vorhandene Terminologie ein, um die Retrieval-Leistung seiner TM-Komponente zu verbessern.

Ähnlich stellt sich die Situation bei der Verwendung vorhandener maschinenlesbarer Texte und ihrer Übersetzungen dar. Diese werden in erster Linie als 'Ansammlung von Übersetzungseinheiten' behandelt. Die in ihnen 'eingebettete' Terminologie bleibt ungenutzt. Zwar ist die rechnergestützte Terminologiegewinnung noch ein sehr junges Forschungsfeld, dennoch finden sich in der Literatur einige interessante Ansätze, die bislang allerdings nicht in kommerzielle TM-Systeme Eingang gefunden haben und lediglich in einigen meist nicht-kommerziellen Konkordanzwerkzeugen und Programmen zur Terminologieextraktion implementiert sind.

Im folgenden möchte ich einige Möglichkeiten aufzeigen, wie die Retrieval-Leistung von TM-Systemen durch Nutzung der vorhandenen terminologischen Informationen und Parallelkorpora[1] verbessert werden könnte. Dabei werde ich zwischen 'expliziten' und 'impliziten' terminologischen Informationen unterscheiden. 'Explizite terminologische Informationen' sind die für andere Komponenten des TM-Sy-

stems direkt zugänglichen Termini der Terminologiedatenbank. 'Implizite terminologische Informationen' sind die für die Komponenten des TM-Systems nicht zugänglichen, in die Texte des Parallelkorpus eingebetteten Termini.

2 Optimierung von TM-Systemen durch bessere Nutzung expliziter terminologischer Informationen

Derzeit finden sich in der Literatur die beiden folgenden Vorschläge:

- 'Generalisierung' der im TM gespeicherten Übersetzungseinheiten mit Hilfe der in der Terminologiedatenbank vorhandenen Benennungen
- Verwendung von Termini zur Alignierung von Einheiten unterhalb der Satzebene.

2.1 Explizite terminologische Informationen zur Generalisierung der Übersetzungseinheiten im Translation Memory

In einem Aufsatz, der in einer Sonderausgabe der Zeitschrift *Machine Translation* zur rechnergestützten Humanübersetzung erschienen ist, schlagen Langé et al. [LGD97] vor, die Übersetzungseinheiten eines TM mit Hilfe der in der Terminologiedatenbank des TM-Systems verfügbaren Termini zu 'abstrahieren' und bekannte Termini durch Variablen zu ersetzen. Das Ziel dieses Ansatzes besteht darin, den Recall von TM-Komponenten – d.h. die 'Fähigkeit', relevante Übersetzungseinheiten zu finden - zu verbessern. Das folgende Beispiel aus dem erwähnten Aufsatz verdeutlicht den Grundgedanken:

(1) Proceed with *installation checking*.

(2) Proceed with *customization*.

(3) Proceed with *X*.

Da die kursiv markierten Termini mit ihren zielsprachlichen (ZS) Entsprechungen in der Terminologiedatenbank vorhanden sind, reicht es nach Langé et al. aus, wenn das TM das abstrahierte Segment (3) enthält. Langé et al. gehen davon aus, daß:

> "a sentence that has been skeletonized to include variable parts is more general, and should therefore be found more frequently in the translation memory than fully instantiated sentences" [LGD97:46].

Allerdings haben Tests von TM-Systemen gezeigt, daß einfache paradigmatische Modifikationen – d.h. Änderungen, die (nahezu) keinen Einfluß auf Syntax und Länge einer Übersetzungseinheit besitzen[2] – i.d.R. nicht zu Retrieval-Problemen führen (vgl. z.B. [Rei94], [Rös/War97]). Der einzige Vorteil einer 'Generalisierung' besteht möglicherweise darin, daß sich auf diese Weise der Umfang von TM-Datenbanken reduzieren läßt. Dies setzt jedoch voraus, daß die zu bearbeitenden Texte einen hohen Anteil an syntaktisch identischen ausgangssprachlichen (AS) Sätzen enthalten. Andererseits nennen Langé et al. selbst bereits eine Reihe von Schwierigkeiten, die bei einer 'Generalisierung' von TM-Einheiten gelöst werden müßten:

- Benennungsüberschneidung: Auf eine Wortfolge der Übersetzungseinheit 'passen' mehrere Termini (z.B. '*Install the receiving antenna support.* → *receiving antenna* vs. *antenna support*')
- Terminologische Varianten: Identifikation verschiedener Instanzen einer Benennung (z.B. morphosyntaktische Varianten)
- Auswahl von ZS-Termini bei Synonymie
- Kongruenzprobleme: Angleichung von Kasus und Numerus beim Ersetzen der Platzhalter.

Zu Recht weisen Langé et al. darauf hin, daß diese Schwierigkeiten für jede Art der Termextraktion und -erkennung typisch sind und daher eigentlich sowieso in den entsprechenden Komponenten rechnergestützter Übersetzungshilfen gelöst werden müßten [LGD97:49]. M.E. liegt das entscheidende Problem der von Langé et al. vorgeschlagenen 'Generalisierung' jedoch vielmehr in den vielfältigen Unterschieden der Oberflächenrepräsentationen in AS und ZS. Zur Verdeutlichung mögen die folgenden Sätze dienen, die mögliche deutsche Übersetzungen von (1) und (2) darstellen:

(1a) *Überprüfung der Installation* fortsetzen.
⇒ *X* fortsetzen.

(1b) Setzen Sie die *Überprüfung der Installation* fort.
⇒ Setzen Sie *X* fort.

(1c) *Überprüfen* Sie als nächstes die *Installation*.
⇒ ??

(1d) *Überprüfung des Leitungssystems fortsetzen.*
⇒ *X* fortsetzen.

(2a) Restliche *benutzerdefinierte Einstellungen festlegen*.
⇒ Restliche *X* (??)

(2b) *Legen* Sie die restlichen *benutzerdefinierten Einstellungen fest*.
⇒ ??

(2c) *Legen* Sie als nächstes die *benutzerdefinierten Einstellungen fest*.
⇒ ??

Erstens können AS/ZS-Einheiten natürlich nur dann in der von Langé et al. vorgeschlagenen Form abstrahiert werden, wenn sich Termini auf 'AS- und ZS-Seite' gleichermaßen durch Variablen ersetzen lassen. Andererseits können „[s]prachliche Repräsentationen eines Begriffes […] von Sprache zu Sprache zwischen Fachwendung, Mehrwortbenennung und einfacher Benennung variieren" [Schm96:200]. Wird also z.B. ein in der AS durch ein einfaches Substantiv oder durch eine Mehrwortbenennung repräsentierter Begriff in der ZS durch eine Verbalphrase wiedergegeben (siehe (1c), (2a-c)), scheidet eine Generalisierung der Übersetzungseinheit spätestens dann aus, wenn diskontinuierliche Strukturen verwendet werden.

Zweitens müssen Strategien für den Umgang mit Polysemien und Homonymien gefunden werden. Dies gilt sowohl für Termini (vgl. die Wiedergabe von *installation checking* in (1a-c) vs. (1d)) als auch für 'Nicht-Termini' (vgl. die Wiedergabe von *proceed* (1a,b,d) vs. (1c) und (2a,b) vs. (2c)).

2.2 Explizite terminologische Informationen zur Erkennung von Satzfragmenten

Der Aufsatz von Langé et al. enthält einen weiteren Vorschlag zur Nutzung expliziter terminologischer Informationen, der sich einer für die (computerlinguistische) Weiterentwicklung von TM-Systemen m.E. weitaus wichtigeren Frage zuwendet, nämlich der Aufgabe des Retrievals von Übersetzungseinheiten unterhalb der Satzebene. Hierzu müssen zumindest zwei Probleme gelöst werden:

- die Identifikation von ‘Satzfragmenten’ in AS und ZS mit Hilfe eines geeigneten Segmentierungsverfahrens
- die Alignierung der identifizierten Fragmente, wobei es sich bei diesen Fragmenten um Einheiten verschiedenster Strukturebenen (Teilsätze, Phrasen unterschiedlicher Komplexität) handeln kann.

Für die Alignierung auf Satzebene werden häufig sog. ‘Anker’ verwendet, d.h. Zeichenketten, die auf der ‘AS-Seite’ und der ‘ZS-Seite’ eines Satzpaares identisch oder sehr ähnlich sind. Hierzu zählen u.a. Datumsangaben, Zahlen, Eigennamen sowie die sogenannten ‘cognates’, d.h. AS- und ZS-Wörter, „[that] share ‘obvious’ phonological or orthographic and semantic properties, with the result that they are likely to be used as mutual translations“ [SFI92:71]. In TM-Systemen scheint es naheliegend, zusätzlich auch die in den Terminologiedatenbanken der Systeme verfügbaren Termini als ‘Anker’ einzusetzen. Eben dies schlagen Langé et al. für die Alignierung von Satzfragmenten vor, wobei sie für die vorausgehende Segmentierung der Sätze jedoch ein sehr einfaches Verfahren vorsehen, das sich auf einige wenige Heuristiken beschränkt:

> “we envisage that it *[d.h. ein aus Phrasen/Satzfragmenten bestehendes TM; U.R.]* will be triggered only in simple cases, for example when the splitting of a sentence into two bricks is made easier by the presence of an unambiguous marker such as a conjunction, or punctuation marks” [LGD97:43].

Daß diese auf einfachen Oberflächenmarkierungen beruhende Segmentierung nicht ausreicht, um aus Parallelkorpora Einheiten unterhalb der Satzebene zu extrahieren, mag das folgende Beispiel verdeutlichen, das der deutschsprachigen Leistungsbeschreibung eines Mobilfunksystems und seiner englischen Übersetzung entnommen wurde:

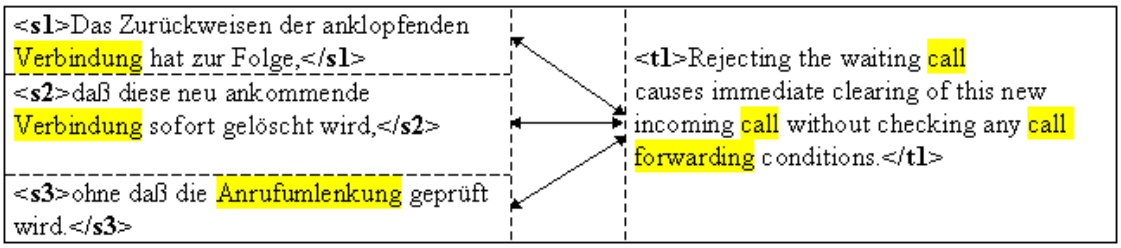
AS: Das Zurückweisen der anklopfenden Verbindung hat zur Folge, daß diese neu ankommende Verbindung sofort gelöscht wird, ohne daß die Anrufumlenkung geprüft wird.

ZS: Rejecting the waiting call causes immediate clearing of this new incoming call without checking any call forwarding conditions.

In der Terminologiedatenbank seien die folgenden Termini enthalten, die für den Alignierungsprozeß als ‘Anker’ zur Verfügung stehen:

Anklopfen	⇔	call waiting
Anrufumlenkung	⇔	call forwarding
Verbindung	⇔	call; connection

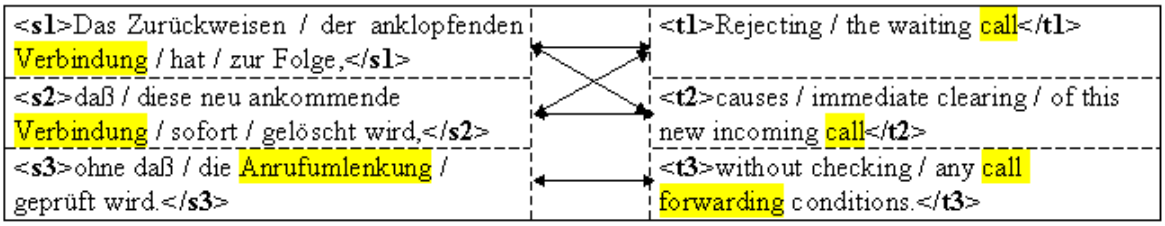
Bei Sprachen mit einem vergleichsweise geringen Anteil an eindeutigen Oberflächenmarkierungen führt das von Langé et al. vorgeschlagene Verfahren wohl eher selten zu einer für den Übersetzer brauchbaren Alignierung von Satzfragmenten. In dem angeführten Beispiel erhält man folgendes Resultat:



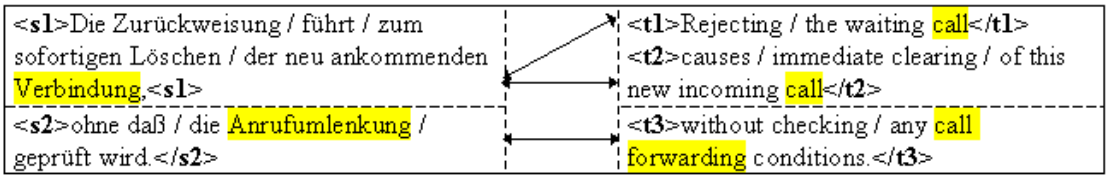
Für die Satzsegmentierung in TM-Systemen bedarf es also eines Verfahrens, das in der Lage ist, auch solche Teilsatzgrenzen zu erkennen, die nicht explizit an der Satzoberfläche markiert sind. Ein Beispiel für einen entsprechenden Formalismus ist das bereits in den 80er Jahren als Segmentierungsmodul für das MÜ-System SUSY entwickelte zweistufige regelbasierte Parser-Konzept PHRASEG [Schm86]. Aufbauend auf den Ergebnissen einer vorangegangenen morphosyntaktischen Analyse sowie einer Wortartendisambiguierung werden auf der ersten Stufe jene Wortklassen zu Einheiten verbunden, zwischen denen keine Teilsatzgrenzen auftreten können. Auf der zweiten Stufe werden diese sogenannten 'Phrasings' dann zu teilsatzwertigen Einheiten zusammengefaßt. Das Verfahren wurde für die Sprachen Deutsch, Englisch und Französisch implementiert und kann ohne weiteres auf andere Sprachen ausgeweitet werden. Eine detaillierte Beschreibung enthält [Schm86]. Dort wird auch darauf hingewiesen, daß „das Verfahren zusammen mit einer morphologischen Analyse und der Wortklassenvereindeutigung als eigenständige syntaktische Analyse innerhalb eines Systems zur computergestützten Übersetzung verwendet werden [kann]" [Schm86:156].[3]

Die folgende Tabelle zeigt, wie das Satzpaar des Beispiels unter Anwendung der PHRASEG-Regeln segmentiert würde. Die Segmentierung des deutschen Satzes in drei Teilsätze bleibt dabei aufgrund der vorhandenen Oberflächenmarkierungen unverändert. Der englische Satz wird jedoch ebenfalls in drei Teilsätze zerlegt. Das Beispiel macht aber auch deutlich, daß die verfügbaren Benen-

nungen aus der Terminologiedatenbank für eine korrekte Zuordnung der Satzfragmente nicht unbedingt ausreichend sind.



Das Beispiel verdeutlicht ferner, daß sich AS- und ZS-Fragmente nicht immer im Verhältnis 1:1 entsprechen. So können z.B. einander zuzuordnende AS- und ZS-Fragmente der Stufe 1 (‘Phrasings’; durch ‘/ ’ getrennt) Teil unterschiedlicher Fragmente der Stufe 2 (‘Teilsätze’) sein (z.B. die Verbalphrasen *hat zur Folge* und *causes* in <s1> bzw. <t2>). Die einander zugeordneten Teilsätze $<s_i>$ und $<t_j>$ dürfen also nicht ohne weiteres als Übersetzungseinheiten verstanden werden. Darüber hinaus dürften häufig auch stufenübergreifende Zuordnungen zwischen ‘Phrasing’-Ebene und Teilsatzebene vorkommen. So entspricht in der folgenden Abwandlung des bisherigen Beispiels das AS-Fragment *die Zurückweisung* dem ZS-Teilsatz <t1>. Hier scheint eine 1:2-Alignierung von <s1> mit <t1> und <t2> sinnvoll.



3 Optimierung von TM-Systemen durch bessere Nutzung impliziter terminologischer Informationen

TM-Systeme könnten die in den vorhandenen maschinenlesbaren Texten und ihren Übersetzungen verfügbaren impliziten terminologischen Informationen auf mindestens zweierlei Weise nutzen. Beispielsweise ließen sich die Texte als Basis zur automatisierten Gewinnung zweisprachiger Terminologie verwenden, um den Bestand der Terminologiedatenbank zu erweitern, oder sie könnten der Extraktion zweisprachiger ‘Anker’ für die Alignierung von TM-Einheiten unterhalb der Satzebene dienen und so die Retrieval-Leistung der TM-Komponente verbessern.

3.1 Implizite terminologische Informationen zur Erweiterung der Terminologiedatenbank

Die übersetzungsvorbereitende Terminologiearbeit, d.h. das 'Füllen' einer Terminologiedatenbank mit der für eine qualitativ hochwertige Übersetzung notwendigen Terminologie, stellt eine der zeitaufwendigsten Arbeitsschritte des Übersetzungsprozesses dar. Angesichts knapper Zeitvorgaben und fehlender linguistischer Datenverarbeitungsverfahren bzw. unzureichendem Wissen über den sinnvollen Einsatz bereits verfügbarer Verfahren erhalten Übersetzer oftmals bestenfalls Wortlisten mit fragmentarischen Informationen, die keinesfalls dazu dienen können, die terminologische Konsistenz des ZS-Textes zu sichern.

Andererseits ist gerade in den letzten Jahren ein wachsendes Maß an angewandten Forschungsarbeiten zu verzeichnen, die sich mit der rechnergestützten Extraktion von Terminologie oder der automatisierten Erstellung von Wörterbüchern aus maschinenlesbaren Textkorpora beschäftigen und verschiedene Werkzeuge und Methoden zur Vereinfachung dieser arbeitsintensiven Prozesse hervorgebracht haben.

Verfahren zur Terminologieextraktion lassen sich grob nach statistischen, linguistischen und hybriden Ansätzen unterteilen [Drou97]. Nach der Anzahl der beteiligten Sprachen können monolinguale und bilinguale Ansätze unterschieden werden.[4]

Bilinguale Extraktionsverfahren basieren auf Parallelkorpora und verwenden entweder ein probabilistisches 'Übersetzungsmodell' (rein statistische Verfahren; siehe hierzu auch [BPPM93]) oder berechnen Assoziationen zwischen potentiellen AS-Termini und ihren Entsprechungen 'auf der ZS-Seite' des Korpus, wobei die AS-Termkandidaten zunächst anhand linguistischer Muster identifiziert werden (hybride Verfahren). Dabei werden zuvor i.d.R. die beiden Seiten des Parallelkorpus auf Satzebene aligniert, um dann anhand von Kookkurrenzwerten in den alignierten Einheiten die wahrscheinlichsten Übersetzungskandidaten zu ermitteln (vgl. z.B. [Dai94]).[5]

Betrachtet man die Leistungsfähigkeit der Extraktionsverfahren, so scheinen die Ergebnisse einiger statistischer Verfahren im Vergleich zu linguistischen Ansätzen einen sehr viel höheren Anteil an Noise zu enthalten (vgl. z.B. [HJKH96:148]).[6] Ferner können Werkzeuge, die ausschließlich statistische Methoden verwenden, i.d.R. keine Mehrwortbenennungen extrahieren (vgl. z.B. [Rapp96] und [Bro97]). Andererseits sind rein linguistische Verfahren nicht nur

per Definition sprachabhängig, sondern auch nicht ohne weiteres in der Lage, einfache, ausschließlich aus Stammwörtern bestehende Benennungen zu identifizieren [LBBL96]. Darüber hinaus enthalten auch die Resultate linguistischer Extraktionsverfahren z.T. einen relativ hohen Noise-Anteil [Pea98], was vermutlich auf eine zu starke Verallgemeinerung der Wortbildungsmuster fachsprachlicher Benennungen zurückgeführt werden kann. So verifiziert Pearsons Arbeit z.B. die Hypothese, daß fachsprachliche Wortbildungsmuster innerhalb einer Sprache von Fachgebiet zu Fachgebiet sowie zwischen verschiedenen Kommunikationsebenen variieren können.

Als zentrale Anwendungsfelder für Software zur Termextraktion nennen [LBBL96] die Bereiche Übersetzen, Terminologiearbeit und Dokumentenmanagement und weisen darauf hin, daß diese drei Bereiche sehr unterschiedliche Benutzeranforderungen aufweisen. Starke Unterschiede in den Anforderungen an eine Termextraktionssoftware bestehen jedoch bereits innerhalb des für diesen Aufsatz relevanten Bereichs des Übersetzens. Die Anwendungsmöglichkeiten reichen hier von der automatischen Generierung zweisprachiger Wörterbücher für beispielbasierte MÜ-Systeme bis zur rechnergestützten Erstellung von Terminologiesammlungen und projektspezifischen Glossaren für die computergestützte Humanübersetzung. Während die für die Humanübersetzung nötige Genauigkeit der Glossare i.d.R. eine umfangreiche manuelle Aufbereitung der maschinell gewonnenen Ergebnisse erfordert, sind für die Generierung von Systemwörterbüchern zur Unterstützung des Alignierungsprozesses in beispielbasierten MÜ-Systemen sehr viel niedrigere Precision-Werte akzeptabel (vgl. [Bro97]).

Insgesamt ist offensichtlich, daß die eindeutige Identifikation von Termini in maschinenlesbaren Korpora und das Erstellen begriffsorientierter Terminologiedatenbanken Aufgaben sind, ”that must, in all cases, be carried out by humans during the last stages” [LBBL96:294]. Hierfür nimmt man sich in der Übersetzungsbranche aber leider nur selten Zeit. Immerhin könnte die in Parallelkorpora eingebettete Terminologie jedoch zumindest zur Unterstützung des Retrievals in TM-Systemen genutzt werden. Im folgenden soll daher untersucht werden, ob sich ein einfaches statistisches Verfahren zur Extraktion von Übersetzungskandidaten [Rapp96] zur Gewinnung von ‘Ankern’ und somit zur Unterstützung der Alignierung von Satzfragmenten eignet.[7]

3.2 Implizite terminologische Informationen zur Erkennung von Satzfragmenten

Rapps Verfahren zur Extraktion von Übersetzungskandidaten setzt ein auf Satzebene aligniertes und lemmatisiertes Parallelkorpus voraus und geht von den beiden folgenden Annahmen aus:

(4) Extrahiert man aus dem Korpus alle Satzpaare, in denen eine AS-Wortform *s* vorkommt, und bestimmt die Häufigkeit aller ZS-Wortformen in den extrahierten Sätzen, so weisen gebräuchliche Übersetzungen (nach häufigen Funktionswörtern) die größte Häufigkeit auf.

(5) Die Häufigkeiten von AS-Wörtern im AS-Teilkorpus und ihren ZS-Entsprechungen im ZS-Teilkorpus sind idealerweise annähernd identisch. M.a.W.: Der Quotient aus beiden Häufigkeiten beträgt idealerweise eins.

Diese Annahmen lassen sich in der folgenden Formel ausdrücken [Rapp96:106]:

$$a_t = \begin{cases} f_{st} \cdot f_s / f_t & \text{für } f_s \leq f_t \\ f_{st} \cdot f_t / f_s & \text{für } f_s > f_t \end{cases}$$

Die 'Wahrscheinlichkeit' a_t, mit der ein Wort *t* eine Übersetzung des Wortes *s* ist – oder in Rapps Worten die 'Aktivität' von *t* – hängt ab von der Häufigkeit f_{st}, mit der *t* in den alignierten Satzpaaren gemeinsam mit *s* vorkommt, sowie von der Relation zwischen den Korpushäufigkeiten f_s und f_t der Wörter *s* bzw. *t*. Dabei führt die Berücksichtigung von Hypothese b) dazu, daß die ersten Positionen der Rangliste der Übersetzungskandidaten nicht von häufigen Funktionswörtern besetzt werden.

In einem Experiment habe ich das von Rapp beschriebene 'Aktivitätsmaß' auf ein kleines deutsch-englisches Parallelkorpus angewandt, das aus Texten der technischen Leistungsbeschreibung eines Mobilfunksystems besteht. Jede 'Seite' des Korpus umfaßt ca. 10.000 Wörter. Dieser Wert mag unter korpuslinguistischen Gesichtspunkten äußerst niedrig erscheinen, andererseits ist dieser Umfang für kleinere Übersetzungsprojekte durchaus realistisch. Das Korpus wurde in einem vorbereitenden teilautomatischen Arbeitsschritt mit einem kommerziellen Alignment-Werkzeug aligniert. Für die Lemmatisierung wurde MPRO (s.o.)

eingesetzt. Die Häufigkeitswerte wurden mit Hilfe von WORD BASIC-Makros ermittelt.

Die beiden Beispiele in Tabelle 1 und 2 verdeutlichen, daß dieses einfache statistische Verfahren keine Mehrwortbenennungen extrahieren kann. Die Identifikation von ZS-Mehrwortbenennungen ist insbesondere dann schwierig, wenn die einzelnen Komponenten vergleichsweise häufig auch separat oder in anderen Mehrwortbenennungen auftreten (Tabelle 1).

Deutsche Lemmata (s) mit Korpushäufigkeit (f_s)	Rang	Englischer Übersetzungs-kandidat (t)	Häufigkeit von t in den alignierten Sätzen (f_{st})	Korpushäufigkeit von t (f_t)	'Aktivität' von t (a_t)
Kennungs-anforderung 6	1	failure	3	5	2,50
	2	cause	3	8	2,25
	3	correlation	1	6	1,00
	4	previous	2	12	1,00
	5	identity	5	34	0,88
	6	recovery	1	7	0,86
	7	request	5	36	0,83
	8	include	1	8	0,75
	9	interworking	1	4	0,67
	10	identify	1	12	0,50

Tab. 1: Kennungsanforderung ⇔ identity request

Das Beispiel in Tabelle 2 zeigt jedoch auch, daß sich bei ZS-Mehrwortbenennungen zumindest dann brauchbare 'Anker' für die Alignierung von Satzfragmenten ergeben können, wenn die Bestandteile der Mehrwortbenennung nur selten separat oder in anderen Mehrwortbenennungen vorkommen. In dem in Tabelle 2 dargestellten Beispiel könnte die deutsche Benennung *Korrelationstabelle* zusammen mit dem an erster Stelle rangierenden Bestandteil der englischen Benennung *correlation table* als 'Anker' verwendet werden.

Deutsche Lemmata (s) mit Korpushäufigkeit (f_s)	Rang	Englischer Übersetzungs-kandidat (t)	Häufigkeit von t in den alignierten Sätzen (f_{st})	Korpushäufigkeit von t (f_t)	'Aktivität' von t (a_t)
Korrelationstabelle 5	1	correlation	3	6	2,50
	2	table	3	8	1,88
	3	access	2	10	1,00
	4	contact	1	5	1,00
	5	TMSI	8	48	0,83
	6	acknowledgement	1	4	0,80

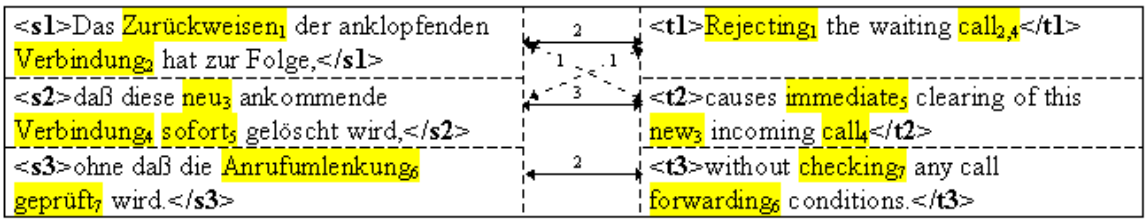
Tab. 2: Korrelationstabelle ⇔ correlation table.

Daß einfache statistische Verfahren zur Extraktion von Übersetzungskandidaten durchaus geeignet sein könnten, um 'Anker' für die Alignierung von Satzfragmenten zu gewinnen, mag ein weiterer Blick auf das bereits mehrfach angeführte deutsch-englische Mobilfunkbeispiel belegen. Tabelle 3 enthält für alle Lemmata des deutschen Satzes die mit Hilfe des zuvor beschriebenen Verfahrens extrahierten Übersetzungskandidaten.

Deutsche Lemmata (s) mit Korpushäufigkeit (f_s)		Englischer Übersetzungskandidat (t)	Häufigkeit von t in den alignierten Sätzen (f_{st})	Korpushäufigkeit von t (f_t)	'Aktivität' von t (a_t)
zurückweisen	9	reject	8	15	4,80
anklopfend	20	accept	9	18	8,10
Verbindung	186	call	225	375	111,60
Folge	4	clearing	2	4	2,00
neu	45	new	40	47	38,30
ankommend	16	incoming	15	21	11,43
sofort	5	immediate	3	8	1,88
löschen	17	cancel	9	13	6,88
Anrufumlenkung	17	forward	15	25	10,20
prüfen	9	check	8	12	6,00

Tab. 3: 'Aktivitäten' für alle 'Rang-1-Übersetzungskandidaten' des Beispiels

Die Tabelle zeigt, daß sich für unser Beispiel in der Mehrzahl der Fälle brauchbare 'Anker' ergeben (grau hinterlegt). Für die übrigen Lemmata konnten keine korrekten Übersetzungen ermittelt werden. Die zusätzlichen 'Anker' führen in unserem Beispiel zu einer eindeutigen Zuordnung der Teilsätze.



4 Einbindung in die bestehenden Arbeitsabläufe

Die vorherigen Abschnitte sollten verdeutlichen, daß das in TM-Systemen verfügbare explizite und implizite terminologische Wissen genutzt werden kann, um Einheiten unterhalb der Satzebene zu alignieren und somit die Performanz von TMs zu erhöhen. Die Nutzung expliziten terminologischen Wissens scheint zwar naheliegend, jedoch dürften die in der Terminologiedatenbank verfügbaren Be-

nennungen nicht immer ausreichen, um brauchbare (eindeutige) Alignierungen zu erhalten. Ein weiterer Schritt könnte daher darin bestehen, Verfahren zur Extraktion von Übersetzungskandidaten in TM-Systeme und Alignierungssoftware zu integrieren, um zusätzliche 'Anker' für den Alignierungsprozeß zu gewinnen. Hier sind jedoch detaillierte Untersuchungen notwendig, um einen Ansatz zu finden, der akzeptable Ergebnisse liefert, ohne allzu zeitaufwendig zu sein.

Darüber hinaus stellt sich die Frage, wie die angesprochenen Verfahren zur Alignierung von Teilsätzen in die Arbeitsabläufe von TM-Systemen eingebettet werden können. Hier sind mindestens drei verschiedene Teilprozesse zu unterscheiden:

(a) vor der eigentlichen Übersetzungsphase:

– die Aufbereitung von bereits übersetztem Textmaterial

(b) während des Übersetzens:

– Aufbereitung des zu übersetzenden Textmaterials

– Aufbereitung von neuen Übersetzungseinheiten, die dem TM hinzugefügt werden sollen.

Muß vor Beginn der eigentlichen Übersetzung zunächst ein TM aus vorhandenen Texten und deren Übersetzungen aufgebaut werden, so müssen die Algorithmen zur Zuordnung von AS- und ZS-Satzfragmenten natürlich in die entsprechenden Alignment-Werkzeuge integriert sein, wobei jedoch Sätze und Satzfragmente in separaten Arbeitsschritten aligniert werden sollten. Auf diese Weise kann das Ergebnis der automatischen Satzzuordnung bei Bedarf zunächst manuell korrigiert werden, bevor die Verarbeitung der Satzfragmente beginnt. Je nach Textumfang und Verarbeitungsmethode kann die Alignierung von Einheiten unterhalb der Satzebene einen hohen Aufwand an Rechnerzeit erfordern. Sofern jedoch das Ergebnis ohne weitere manuelle Korrekturen in das TM übernommen wird, sind längere Rechnerzeiten akzeptabel, da die Alignierung der Fragmente vor der eigentlichen Übersetzungsphase stattfindet und Übersetzer in ihrer eigentlichen Tätigkeit somit nicht behindert werden.

Führt während des Übersetzens die Suche in einem TM nicht zu brauchbaren Ergebnissen (d.h. keine Treffer oder zu geringe Ähnlichkeit zwischen Treffern und Suchanfrage), so könnte der Anwender eine weitere Suche in einem aus Satzfragmenten bestehenden TM durchführen. Eine solche Suche setzt natürlich die morphologische und syntaktische Analyse des zu übersetzenden AS-Materi-

als voraus. Dies könnte entweder vor Beginn der eigentlichen Übersetzung erfolgen, so daß der gesamte AS-Text in einem Arbeitsschritt analysiert wird, oder die linguistischen Analysen werden in die Retrieval-Phase eingebunden, so daß nur jene Sätze des AS-Textes analysiert werden, für die der Anwender eine Suche nach Satzfragmenten auslöst.

Werden in der eigentlichen Übersetzungsphase ZS-Vorschläge aus dem TM modifiziert oder AS-Sätze vollständig neu übersetzt, so werden dem TM neue Übersetzungseinheiten hinzugefügt. Für diese neuen Einheiten muß unmittelbar vor dem Abspeichern natürlich zunächst die Satzsegmentierung sowie die Alignierung der erkannten Fragmente durchgeführt werden. Die hierfür benötigte Verarbeitungszeit dürfte in einem akzeptablen Bereich liegen, da jeweils nur ein AS/ZS-Segmentpaar verarbeitet wird.

Literatur

[Ahm94] Ahmad, K. (1994): Language Engineering and the Processing of Specialist Terminology. In: *Proceedings of Language Engineering Convention*, 6–7 July 1994, CNIT-La Défense. Paris. Edinburgh: European Network in Language and Speech (ELSNET).

[Ahm/Rog92] Ahmad, K./Rogers, M. (1992): Terminology Management: A Corpus-Based Approach. In: *Translating and the Computer 14. Quality Standards and the Implementation of Technology in Translation*. London: Aslib, 33–44.

[BPPM93] Brown, P./Della Pietra, St./Della Pietra, V./Mercer, R. (1993): The Mathematics of Statistical Machine Translation: Parameter Estimation. In: *Computational Linguistics*, 19(2), 263–311.

[Bro97] Brown, R. (1997): Automated Dictionary Extraction for 'Knowledge-Free' Example-Based Translation. In: *MT Yesterday, Today, and Tomorrow. Proceedings of TMI-97*, Santa Fe, July 23-25, 1997, 169–174.

[Car/Schm98] Carl, M./Schmidt-Wigger, A. (1998): Shallow Post Morphological Processing with KURD. In: *Proceedings of NeMLaP-98*. Sydney, Januar 1998.

[Dag/Chu94] Dagan, I./Church, K. (1994): Termight: Identifying and Translating Technical Terminology. In: *Proceedings of COLING 1994*. The 15th International Conference on Computational Linguistics, August 1994. Kyoto, Japan. 34–40.

[Dag/Chu97] Dagan, I./Church, K. (1997): Termight: Coordinating Humans and Machines in Bilingual Terminology Acquisition. In: *Machine Translation*, 12, 89–107.

[Dai94] Daille, B. (1994): *Approche mixte pour l'extraction de terminologie: statistique lexicale et filtres linguistiques*. Paris: Université de Paris VII, [Dissertation].

[Drou97] Drouin, P. (1997): Une méthodologie d'identification automatique des syntagmes terminologiques: l'apport de la description du non-terme. In: *META*, 42(1), 45–54.

[Gro98] Groß, B. (1998): *Vergleichende Untersuchung von Alignment-Tools*. Saarbrücken: Fachrichtung 8.6, Universität des Saarlandes (Saarbrücker Studien zu Sprachdatenverarbeitung und Übersetzen, herausgegeben von Karl-Heinz Freigang und Uwe Reinke, Band 15).

[HJKH96] Heid, U./Jauss, S./Krüger, K./Hohmann, A. (1996): Term extraction with standard tools for corpus exploration: Experience from German. In: Galinski, Ch./Schmitz, K.-D. (Hrsg.): *TKE '96: Terminology and Knowledge Engineering. Proceedings of the 4th International Congress on Terminology and Knowledge Engineering, 26–28 August 1996, Vienna*. Frankfurt/M.: INDEKS, 139–150.

[LGD97] Langé, M./Gaussier, É./Daille, B. (1997): Bricks and Skeletons: Some Ideas for the Near Future of MAHT. In: *Machine Translation*, 12, 39–51.

[LBBL96] L'Homme, M.-C./Benali, L./Bertrand, C./Lauduique, P. (1996): "Definition of an Evaluation Grid for term extraction software". In: *Terminology*, 3(2), 291-312.

[Maas96] Maas, H.-D. (1996): MPRO – Ein System zur Analyse und Synthese deutscher Wörter. In: Hausser R. (Hrsg.): *Linguistische Verifikation, Sprache und Information. Dokumentation zur Ersten Morpholympics 1994*. Tübingen: Niemeyer, 141–166.

[Mack/Han96] Macklovitch, E./Hannan, M.-L. (1996): Line 'em up: Advances in Alignment Technology and Their Impact on Translation Support Tools. In: *Expanding MT Horizons. Proceedings of the Second Conference for Machine Translation in the Americas. 2-5 October, 1996*. Montreal, Canada. Washington DC: Association for Machine Translation in the Americas (AMTA), 145–156.

[Pea98] Pearson, J. (1998): *Terms in Context*. Amsterdam, Philadelphia: Benjamins.

[Rapp96] Rapp, R. (1996): *Die Berechnung von Assoziationen: Ein korpuslinguistischer Ansatz*. Hildesheim: Olms.

[Rei94] Reinke, U. (1994): Zur Leistungsfähigkeit integrierter Übersetzungssysteme. In: *Lebende Sprachen*, 3/94, 97-104.

[Rös/War97] Rösener, Ch./Wargenau, J. (1997): *Terminologie- und Satzerkennung für Englisch und Russisch am Beispiel der Translator's Workbench von Trados*. Saarbrücken: Fachrichtung 8.6, Universität des Saarlandes (Saarbrücker Studien zu Sprachdatenverarbeitung und Übersetzen, herausgegeben von Karl-Heinz Freigang und Uwe Reinke, Bad 8)

[Schm86] Schmitz, K.-D. (1986): *Automatische Segmentierung natürlichsprachiger Sätze*. Hildesheim: Olms.

[Schm96] Schmitz, K.-D. (1996): Verwaltung sprachlicher Einheiten in Terminologieverwaltungssystemen. In: Lauer, A./Gerzymisch-Arbogast, H./Haller, J./Steiner, E.: *Übersetzungswissenschaft im Umbruch. Festschrift für Wolfram Wilss zum 70. Geburtstag*. Tübingen: Narr, 197–207.

[SFI92] Simard, M./Foster, G./Isabelle, P. (1992): Using Cognates to Align Sentences in Bilingual Corpora. In: *Proceedings of TMI-92*, Centre d' innovation en technologies de l' information, Montreal, Kanada, 67–81.

ANMERKUNGEN

1 Unter einem Parallelkorpus soll im folgenden ein *n*-sprachiges Korpus verstanden werden, das entsprechend der Anzahl der beteiligten Sprachen aus *n* Teilkorpora besteht. Dabei stellen die Texte eines der Teilkorpora Ausgangstexte für die Übersetzung in die übrigen *n-1* Sprachen dar (Teilkorpora 2, 3, …, *n*). Zur Problematik des Begriffs 'Parallelkorpus' siehe auch [Pea98:47f.].

2 Vergleichsweise geringe Längenveränderungen treten bei lexikalischen Ersetzungen nur dann auf, wenn das ersetzende Lexem im Vergleich zum ersetzten Lexem aus einer höheren oder geringeren Anzahl von Wörtern besteht.

3 Leider ist PHRASEG jedoch nicht mehr implementiert. Andererseits war PHRASEG eines der wenigen sprachdatenverarbeitenden Programme seiner Zeit, das auf einer möglichst strikten Trennung von Algorithmus und linguistischen Regeln beruhte, so daß es mir sinnvoll scheint, das in [Schm86] dokumentierte Regelwerk wiederzuver-

wenden. Dabei nutze ich für die vorausgehenden morphologischen Analysen das am Institut für Angewandte Informationswissenschaft (IAI) in Saarbrücken entwickelte Werkzeug MPRO [Maas96]. MPRO läßt sich weitgehend auf das morphologische Modul des MÜ-Systems SUSY zurückführen und umfaßt Komponenten für die Sprachen Deutsch, Englisch und Französisch. Für die Wortartendisambiguierung sowie für die eigentliche Implementierung der PHRASEG-Regeln wird der ebenfalls am IAI entwickelte Formalismus KURD verwendet, der morphosyntaktisch analysierte Sätze mit Hilfe 'flacher Operationen' (*shallow processing*) weiterverarbeitet [Car/Schm98]. Mögliche Einsatzbereiche von KURD sind z.B. Wortformendisambiguierung, Syntaxprüfung, Stilprüfung oder partielles Parsing. Deutsche und englische Disambiguierungsregeln wurden mir z.T. vom IAI zur Verfügung gestellt, so daß sich meine eigenen Bemühungen im wesentlichen auf die Umsetzung der PHRASEG-Regeln konzentrieren.

4 Einen knappen Überblick über monolinguale und bilinguale Verfahren zur Extraktion von Termkandidaten enthält [Dag/Chu97]. Detaillierte Darstellungen verschiedener statistischer Maße, die in Verfahren zur Terminologieextraktion eingesetzt werden, finden sich in [Dai94].

5 Es wird allgemein davon ausgegangen, daß

> "[t]he problems of sentence-alignment, if not entirely resolved, are fairly well understood" [Mack/Han96:147].

Tests von Alignment-Werkzeugen zeigen jedoch, daß die derzeit eingesetzten Verfahren zur Alignierung auf Satzebene bei der Erkennung komplexerer Zuordnungen häufig nicht fehlerfrei arbeiten [Gro98]. Solche komplexeren Zuordnungen sind z.B. Kontraktionen (n:1-Entsprechungen), Expansionen (1:n-Entsprechungen), Auslassungen (1:0-Entsprechungen) oder Hinzufügungen (0:1-Entsprechungen). Eine Extraktionsmethode, die auf den Ergebnissen eines wortbasierten Alignierungsverfahrens beruht, wird in [Dag/Chu94] und [Dag/Chu97] beschrieben.

6 Heid hat die Ergebnisse seines linguistischen Ansatzes den Resultaten des in [Ahm/Rog1992] und [Ahm94] beschriebenen statistischen Verfahrens gegenübergestellt, das im wesentlichen auf der Hypothese beruht, daß fachsprachliche Benennungen in Fachtextkorpora häufiger auftreten als in einem 'repräsentativen' gemeinsprachlichen Korpus.

7 Das in [Brow97] beschriebene Verfahren zur Wörterbuchgenerierung dient im übrigen ähnlichen Zwecken. Das aus einem Parallelkorpus gewonnene Wörterbuch unterstützt die Alignierung von Satzfragmenten in einem beispielbasierten MÜ-System und kann daher nicht mit den Systemwörterbüchern 'traditioneller' (regelbasierter) MÜ-Systeme verglichen werden.