

# Translation-Memory-Module automatischer Übersetzungssysteme

*Uta Seewald-Heeg (Hochschule Anhalt)*

*Rita Nübel (IAI Saarbrücken)*

## 1 Systemtypen

Neben vollautomatisch arbeitenden Übersetzungswerkzeugen, bei deren Einsatz der Humanübersetzer (oder allgemein Benutzer) während des Übersetzungsprozesses nicht in den maschinellen Ablauf eingreifen kann,<sup>1</sup> und satzspeicherbasierten Systemen, d.h. Translation Memories (TM),<sup>2</sup> die selbst keinerlei Übersetzung vornehmen, sondern den Übersetzer bei der Recherche nach Textsegmenten unterstützen, die bereits im Zuge anderer Übersetzungen oder im Rahmen der Übersetzung des gerade in Arbeit befindlichen Dokuments übersetzt wurden, ist inzwischen ein Systemtyp entwickelt worden, der die Vorteile eines satzspeicherbasierten Systems mit denen eines vollautomatischen Übersetzungssystems verbindet. Systeme dieses Typs, die also Systemkomponenten auf unterschiedlichen Verarbeitungsstufen integrieren, bezeichnen wir nachfolgend als Integrierte Systeme.<sup>3</sup> Konfigurationen integrierter Systeme lassen sich durch die Kombination eines Translation-Memory-Systems mit einem vollautomatischen Übersetzungssystem erzeugen, wie dies beispielsweise bei der Kombination von Transit, dem TM der Firma Star, mit dem Übersetzungssystem Logos geschehen ist.<sup>4</sup>

Integrierte Systeme sind aber nicht nur das Ergebnis der Zusammenführung zweier sonst auch selbständig einsetzbarer ‚Großsysteme‘, sondern existieren mittlerweile in der Gestalt vollautomatischer Übersetzungssysteme, die um eine Translation-Memory-Komponente erweitert wurden.

Systemkonzeptionen dieser Art sind eine unmittelbare Konsequenz der eingeschränkten Leistungsfähigkeit vollautomatisch arbeitender maschineller Übersetzungssysteme. Wie die von [Nübel/Seewald-Heeg98] zusammengetragenen Ergebnisse der Evaluation der linguistischen Performanz vollautomatischer maschineller Übersetzungssysteme gezeigt hat, liegt der Prozentsatz der von den Systemen vollständig korrekt übersetzten Sätze ohne vorherige Systemoptimierung durch Hinzuladen benutzerdefinierter Wörterbucheinträge oder Zusatzwörterbücher bei den leistungsfähigsten Systemen durchschnittlich nur knapp über 30%. Für qualitativ hochwertige Übersetzungen ist beim Einsatz solcher Systeme

daher ein relativ hoher Posteditationsaufwand erforderlich. Die Erweiterung der Systeme durch ein Satzarchiv kann zu einer Effizienzsteigerung führen, da mit einem Satzarchiv neben der automatisch erstellten Übersetzung nun auch die Gesamtheit der bereits posteditierten oder manuell übersetzten Texte zur Verfügung gestellt wird.

## **2 Übersetzungssysteme mit Übersetzungsspeicher-Modul**

Unter den kommerziellen Übersetzungssystemen verfügen gegenwärtig die Systeme T1 von Langenscheidt und Personal Translator plus (PT plus) von Linguattec über ein Übersetzungsspeicher-Modul.

Das Übersetzungsspeicher-Modul von T1 ist wie jenes von PT plus eine Datenbank, die in diesem Zusammenhang meist als Übersetzungsarchiv oder Satzarchiv bezeichnet wird. Das Übersetzungsarchiv enthält Satzpaare, die jeweils aus einem Ausgangssprachlichen Satz und dessen Übersetzung in die jeweilige Zielsprache bestehen. Die Übersetzung eines Ausgangssatzes ist entweder das Ergebnis der automatischen Übersetzung oder aber eine vom Übersetzer erstellte bzw. von ihm posteditierte Übersetzung. Der Benutzer kann sich jede im Satzarchiv enthaltene Übersetzung eines Satzes anzeigen lassen, wobei die Suche sowohl vom Ausgangssprachlichen Text als auch von dessen Übersetzung aus möglich ist.

Nachfolgend werden zunächst einige Merkmale der Benutzerschnittstelle des Satzarchivs von T1, anschließend jene von PT plus dargestellt und einer Evaluation unterzogen.

### **2.1 Das Übersetzungsspeicher-Modul von T1**

Das Satzarchiv von T1 lässt sich in verschiedene Module untergliedern, so dass bestimmte Satzarchivmodule beispielsweise in Abhängigkeit von der Fachrichtung und Textsorte eines zu übersetzenden Textes oder abhängig vom Auftraggeber ausgewählt werden können, während andere für einen bestimmten Übersetzungsauftrag unberücksichtigt bleiben, das System als Referenzmaterial bei der Suche nach im Archiv enthaltenen Sätzen also nur in den ausgewählten Modulen nachschlägt.

Das Übersetzungsarchiv von T1 lässt sich sowohl in Kombination mit der automatischen Übersetzung als auch ohne diese einsetzen. Beim kombinierten Einsatz des Satzarchivs mit der maschinellen Übersetzung kann der Benutzer spezifizieren, ob Sätze, die in der im Text vorliegenden Form oder einer ähnlichen Form im Übersetzungsarchiv gespeichert sind, nicht mehr maschinell übersetzt und deren Übersetzung automatisch in den Zieltext übernommen werden sollen. Soll das Übersetzungsarchiv dagegen separat eingesetzt werden, muss vom Benutzer die Option „T1 Übersetzung – Nie“ aktiviert werden (vgl. Abb. 1).

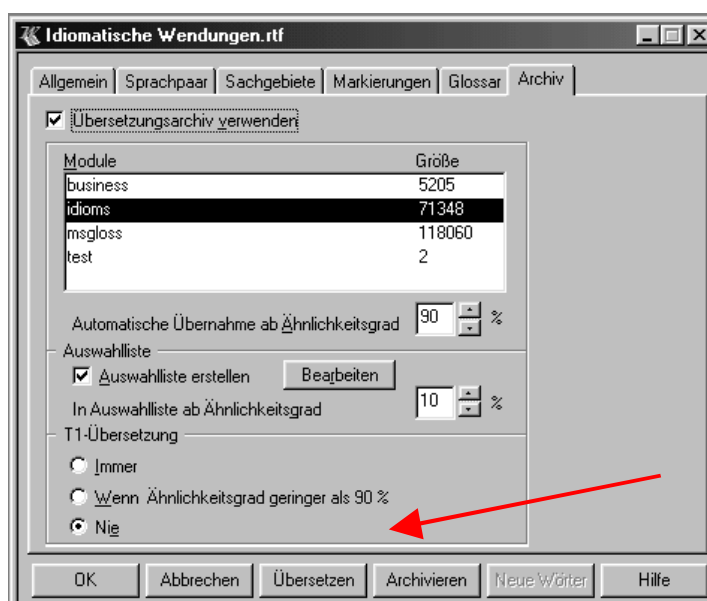


Abb. 1: Benutzerschnittstelle des Satzarchivmoduls von T1.

In einem als „Auswahlliste“ bezeichneten Bildschirmfenster des Übersetzungsarchivs werden bei der Bearbeitung der Suchergebnisse neben dem zu übersetzenden Satz bis zu drei im Satzarchiv als gleich oder ähnlich identifizierte Sätze bzw. deren Übersetzung angezeigt. Der Inhalt der einzelnen Teilfenster ist editierbar, und per Mausklick lässt sich die im Archiv gefundene und gegebenenfalls vom Benutzer bearbeitete Übersetzung in den Zieltext übernehmen.

## 2.2 Das Übersetzungsspeicher-Modul von PT plus

Ähnlich wie T1 bietet PT plus die Möglichkeit, neben der automatischen Übersetzung auch ein vom Benutzer definiertes Satzarchiv mit dem darin enthaltenen zweisprachigen Referenzmaterial zu verwenden. Der Benutzer kann sein Satzarchiv entweder mit den Ergebnissen der automatischen Übersetzungskomponente von PT plus füllen oder mit einer manuell nacheditierten Version dieser Übersetzungen. Die Größe der jeweils definierten Übersetzungseinheiten ist sowohl bei T1 als auch bei PT plus nicht an Phrasen- oder Satzgrenzen gebunden, sondern kann bei der Bearbeitung des Satzarchivs individuell bestimmt werden.<sup>5</sup> Die möglichen Parameter zur Optimierung der Übersetzung bzw. zur Einschränkung des Suchraums bei der Satzarchivfunktion (Auswahl der Satzarchivmodule, Benutzer usw. [vgl. Abb. 2]) sind mit jenen bei T1 vergleichbar.

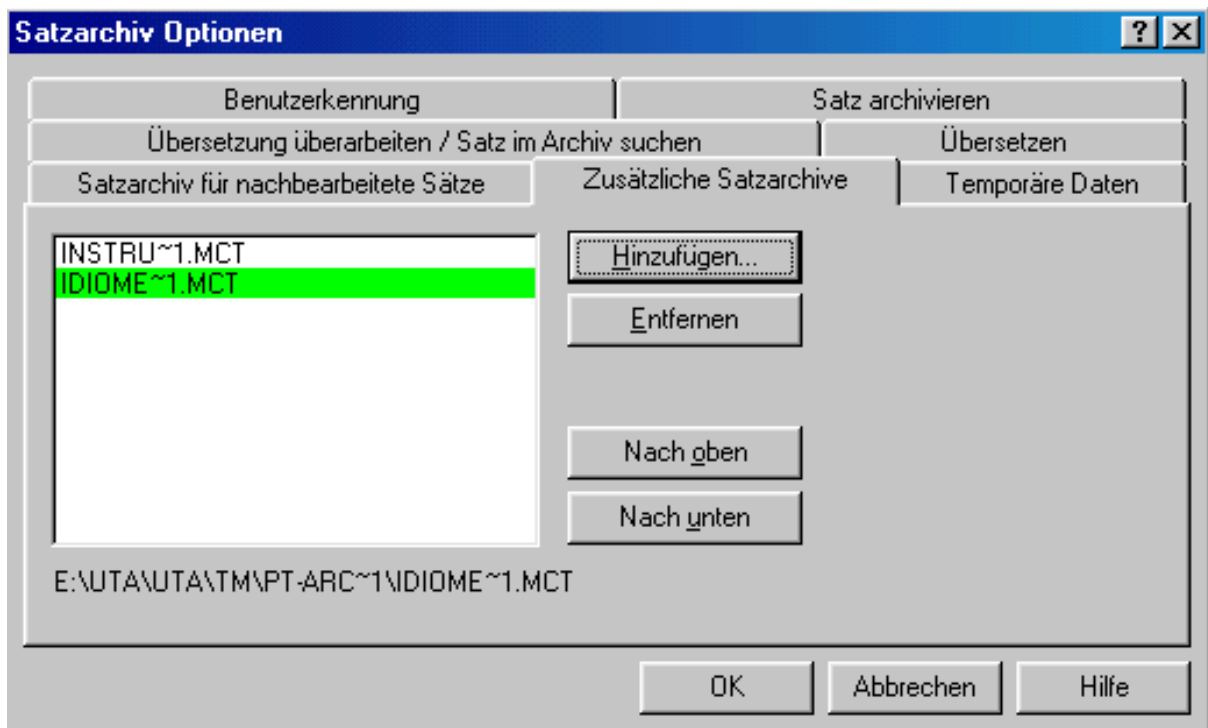


Abb. 2: Benutzerschnittstelle des Satzarchivmoduls von PT plus 2000.

### 3 Nachschlagen im Satzarchiv

Die wesentliche Leistung eines Satzarchivmoduls beruht auf effizienten Suchalgorithmen, die einen Satz eines Quelltextes mit den entsprechenden quellsprachlichen Einträgen der Datenbank vergleichen. Satzspeichersysteme zeichnen sich dadurch aus, dass die Systeme auf eine Suchanfrage nicht nur vollständige Übereinstimmungen, sogenannte 100%-Matches, liefern, sondern aufgrund der jeweils unterschiedlichen Realisierung von Fuzzy-Match-Algorithmen auch Sätze finden, die Abweichungen gegenüber dem zu übersetzenden Text aufweisen können. Dabei ist der maximale Grad der Abweichung zwischen zu übersetzendem Satz und im Satzspeicher enthaltenen Satz vom Benutzer jeweils in Form eines Prozentwertes anzugeben, so dass bei der Suche des Satzes 1.1 aus Beispiel (1) im Satzarchiv auch ein Satz wie 1.2 gefunden werden kann.

(1) 1.1: *This chapter gives an introduction to XII operations.*

1.2: *This chapter is an introduction to XII operations.*

Da sich satzspeicherbasierte Systeme in der Bewertung der Übereinstimmung zwischen zu übersetzendem Satz und im Satzarchiv enthaltenen Referenzsätzen zum Teil deutlich voneinander unterscheiden, sollen die Suchergebnisse der Systeme T1 und PT plus nachfolgend anhand eines kleinen Testkorpus dokumentiert und einander gegenübergestellt werden.

#### 3.1 Das Testmaterial

Um die grundlegende Funktionsweise der beiden Übersetzungssysteme beim Zugriff auf ihr Satzspeichermodul bzw. bei der Suche (Retrieval) in dem jeweiligen Satzarchiv zu überprüfen, wird zunächst ein Satzarchiv angelegt, das nur zwei Referenzsätze enthält, eine Kapitelüberschrift und der erste Satz des dazugehörigen Kapitels (vgl. Bsp. (2)). Durch die Beschränkung des Satzarchivs auf nur zwei Sätze ist es möglich, die Suchergebnisse der Systeme bei Abweichungen von zu übersetzenden Sätzen gegenüber dem Referenzmaterial im Satzarchiv bereits auf der Basis eines kleinen Textkorpus zu systematisieren. Das als Testmaterial verwendete Textkorpus umfasst 34 Testsätze, von denen sieben den ersten Referenzsatz (vgl. Tab. 1) und die übrigen 27 Sätze den zweiten Referenzsatz des Satzarchivs (vgl. Tab. 2) modifizieren.

(2) *Chapter 2 Getting Started*

*This chapter is an introduction to X11 operations.*

Das Verhalten der Systeme bei der Suche in einem besonders großen Satzarchiv wurde im vorliegenden Test nicht untersucht, da die von den Systemen vorgenommene Bewertung der Ähnlichkeit zwischen Testsatz und Referenzsatz im Vordergrund stand.

<b>Referenzsatz</b>			
<i>Chapter 2 Getting Started</i>			
<b>Nr.</b>	<b>Testsatz</b>	<b>Match-Wert<sup>*)</sup> in %</b>	
		<b>T1</b>	<b>PT</b>
1	Chapter 2 Getting started	100	100
2	Chapter 2 Getting started	75	100
3	Chapter 2 Getting Started.	75	100
4	Chapter 1 Getting Started	75	95
5	Chapter 2 Ø	50	0
6	Ø Getting Started	50	0
7	Ø Getting started	25	0

*Tab. 1: Retrievalergebnisse von T1 und PT plus. Die mit <sup>\*)</sup> markierte Spalte enthält die vom jeweiligen System ermittelte Übereinstimmung zwischen Test- und Referenzsatz.*

<b>Referenzsatz</b>			
<i>This chapter is an introduction to X11 operations.</i>			
<b>Nr.</b>	<b>Testsatz</b>	<b>Match-Wert in %</b>	
8	This chapter is an introduction to X11 operations.	100	100
9	This chapter <b>gives</b> an introduction to X11 operations.	87	91
10	This <b>manual</b> is an introduction to X11 operations.	87	86
11	This chapter is an introduction to X11 operations <b>Ø</b>	87	100
12	This chapter is an introduction to X11 <b>functions</b> .	87	82
13	This chapter is an introduction to X11 <b>Ø</b> .	75	87
14	This chapter is an introduction to X11 <b>Ø</b>	87	87
15	This chapter is an <b>XXX</b> to X11 operations.	87	82
16	This chapter is an <b>overview over</b> X11 operations.	75	71
17	This chapter is a <b>list of</b> X11 operations.	62	75
18	This chapter is a <b>brief</b> introduction to X11 operations.	77	89
19	This chapter is a <b>brief</b> introduction to X11 <b>Ø</b> .	62	80
20	<b>The following</b> chapter is an introduction to X11 operations.	77	78
21	<b>Chapter 2</b> is an introduction to X11 operations.	75	84
22	<b>Chapter 2</b> is a <b>brief</b> introduction to X11 operations.	55	76
23	<b>Chapter 2 of this manual</b> is an introduction to X11 operations.	54	86
24	This chapter is an introduction to X11 <b>operations and X11 functions</b> .	72	77
25	This chapter is an introduction to <b>Ø</b> operations <b>of X11</b> .	66	93
26	This <b>Ø</b> is an introduction to X11 operations.	87	91
27	<b>See the</b> introduction to X11 operations.	50	75
28	X11 operations are introduced in this chapter.	12	53
29	X11 is <b>introduced in</b> this chapter.	25	0
30	chapter This an is to introduction operations X11.	75	91

31	is an introduction This chapter to operations. X11	87	74
32	X11 introduction This chapter is an to operations.	<b>100</b>	87
33	X11 introduction This <b>an</b> chapter is <b>an</b> to operations.	<b>100</b>	97
34	X11 This <b>an</b> chapter is <b>an</b> to operations.	<b>100</b>	86

Tab. 2: Retrievalergebnisse von T1 und PT plus.  $\emptyset$  markiert die Auslassungen in den Testsätzen relativ zum Referenzsatz.

## 4 Ergebnisse des Retrievals im Übersetzungsspeicher-Modul von T1

Wie aus den in Tabelle 1 und in Tabelle 2 dargestellten Retrievalergebnissen hervorgeht, steht die Bewertung der Ähnlichkeit eines zu übersetzenden Satzes relativ zu einem im Archiv enthaltenen Referenzsatz bei T1 im Zusammenhang mit der Anzahl der zwischen Test- und Referenzsatz übereinstimmenden Wörter.

Im Unterschied zur automatischen Übersetzung, wo in der ersten Analysephase, der Phase der Segmentierung eines Satzes in seine Bestandteile, die Interpunktionszeichen von den ihnen vorangehenden Zeichenketten separiert werden, leistet der Retrieval-Algorithmus von T1 diese Segmentierung nicht mehr, so dass Sätze wie Satz (3) aus Tab. 1 oder Satz (11) aus Tab. 2 relativ zum Referenzsatz nach der prozentualen Bewertung des Systems bereits deutlich voneinander abweichen. Obschon die hier vorliegenden Abweichungen in Satz (3) lediglich durch den zusätzlichen und in Satz (11) alleine durch den fehlenden Punkt am Satzende verursacht werden, gibt T1 bei Satz (3) als Match-Wert 75% an, während das System die Ähnlichkeit von Satz (11) zum Referenzsatz mit 87% bemisst. Der Grund hierfür ist die Tatsache, dass der Match-Algorithmus des T1-Satzarchivs jeden Bestandteil eines Satzes zu gleichen Teilen in die Bewertung für die Übereinstimmung zweier Sätze einfließen lässt.

Da der Abgleich zwischen zwei Sätzen allein auf der Basis eines Zeichenkettenvergleichs erfolgt, erhalten Sätze wie Satz (28) und Satz (29) (Tab. 2) einen sehr niedrigen Match-Wert, obschon sie inhaltlich dem Referenzsatz sehr ähnlich sind. Der in Satz (28) und (29) ausgedrückte Sachverhalt wird im Unterschied zum Referenzsatz im Passiv ausgedrückt, was an der Satzoberfläche zu veränderten Wortformen und damit zu einem niedrigen Match-Wert führt.



Beim Abgleich der zu übersetzenden Sätze berücksichtigt T1 nicht die Reihenfolge der Wörter in einem Satz, sondern überprüft lediglich das Auftreten der einzelnen Wörter im Referenzmaterial. Dieses Verfahren bzw. die Tatsache, dass Sätze als ungeordnete Mengen von Wörtern begriffen werden, führt dazu, dass beliebige Permutationen von Wörtern, bis hin zu syntaktisch völlig sinnlosen Kombinationen, als mit dem Referenzmaterial vollständig identisch, d.h. mit einer Ähnlichkeit von 100% bewertet werden (vgl. Tab. 2, Sätze (32) bis (34)). In Anbetracht dieses Ergebnisses muss bezweifelt werden, ob die Systemfunktion des T1, mit der Sätze ab einem bestimmten Match-Wert, also beispielsweise bei einem Ähnlichkeitsgrad von 100%, automatisch in die Übersetzung übernommen werden können, sinnvoll eingesetzt werden kann.

Obschon die Match-Ergebnisse von T1 mit einem Wert von 100% an dieser Stelle besonders auffallen, stellt sich auch in Bezug auf andere Systeme die generelle Frage nach der Bewertung dessen, was von Systemseite als vollständig identisch, also als 100%-Match, identifiziert wird. In diesem Zusammenhang ist daher auch die Frage berechtigt, ob die Ergebnisse des Personal Translator beim Abgleich der Sätze (32) bis (34), der hier jeweils Werte von 87%, 97% und 86% liefert, oder auch jene des Translation-Memory-Systems von Trados, der Translator's Workbench, die hier Werte von 67%, 71% und 75% errechnet, aufschlussreicher sind, handelt es sich bei den fraglichen Testsätzen doch um Wortaneinanderreihungen, in denen zwar jedes Wort des betreffenden Referenzsatzes enthalten ist und deren Länge die des Referenzsatzes nicht unterschreitet, deren Reihenfolge jedoch nichts mit der Abfolge der Wörter im Referenzsatz des Satzarchivs zu tun hat.

Wie die Sätze (35) und (36) aus Tab. 3 belegen, führt beispielsweise die veränderte syntaktische Position der Negationspartikel *not* in (36) zu einer Umkehrung der in Satz (35) ausgedrückten Bedeutung, was seinen Niederschlag in der betreffenden Übersetzung finden muss. Einen vergleichbaren Effekt hat die Verdoppelung der Negationspartikel *not* in Satz (38) im Vergleich zu (37). Eine Übernahme der Übersetzung von (37) für Satz (38) hätte somit wie schon im vorhergehenden Fall einen schwerwiegenden Übersetzungsfehler zur Folge. – Selbst wenn man davon ausgeht, dass im realen Übersetzungsfall nur selten Beispielsätze auftreten, die zwar alle Elemente des Referenzsatzes, jedoch in anderer Abfolge enthalten, muss bereits an dieser Stelle unterstrichen werden, dass ein 100%-Match unter linguistischen Gesichtspunkten eine zu hinterfragende Größe ist und – das zeigen die Ergebnisse – insbesondere bei der Suche im Satzarchiv von T1 als sehr

<b>Referenzsatz</b>		
<i>I do not drink wine, but I drink beer.</i>		
<b>Nr.</b>	<b>Testsatz</b>	<b>Match-Wert</b>
35	I <b>do not</b> drink wine, but I drink beer.	<b>100</b>
36	I drink wine, but I <b>do not</b> drink beer.	<b>100</b>
<b>Referenzsatz:</b>		
<i>It's true that I do not drink wine.</i>		
<b>Nr.</b>	<b>Testsatz</b>	<b>Match-Wert</b>
37	It's true that I do <b>not</b> drink wine.	<b>100</b>
38	It's <b>not</b> true that I do <b>not</b> drink wine.	<b>100</b>

Tab. 3: 100%-Matches von T1.

problematisch einzustufen ist. Sätze mit dieser Bewertung dürfen unter keinen Umständen ungeprüft in eine Übersetzung übernommen werden.

#### 4.1 Der Suchalgorithmus im T1-Satzarchiv

Der für das Satzarchiv von T1 realisierte Suchalgorithmus lässt sich aus den Beispielen (Tab. 1 und 2) mit *Reverse-Engineering*-Techniken rekonstruieren. Er ist verhältnismäßig einfach. Aus seiner Struktur lässt sich seine Unzulänglichkeit unmittelbar ableiten. Neben der Anzahl der zwischen Testsatz und Referenzsatz übereinstimmenden Wörter hängt der jeweils vom System gelieferte Match-Wert von der Länge der miteinander verglichenen Sätze ab. Wie unter anderem aus Satz (24) ersichtlich ist, spielt aber nicht allein die Länge des Referenzsatzes, der als Vergleichsbasis herangezogen wird, eine Rolle. Vielmehr wird für die Ermittlung des Match-Wertes die Länge des jeweils längeren der beiden miteinander verglichenen Sätze herangezogen. Dieser Wert wird als Quotient angewendet auf die Anzahl der Wörter, die Referenzsatz und zu übersetzender Satz gemeinsam haben (siehe Abb. 1).<sup>6</sup>

## 4.2 Das Satzarchiv als „Wörterbuch-Erweiterung“ für Mehrwortausdrücke

Wenn nun – wie in Abschnitt 4.1 dargestellt – der Match-Algorithmus des Satzarchivs in T1 als zu trivial erscheint, als dass das Satzarchiv-Modul im professionellen Einsatz Funktionen eines integrierten Translation Memory übernehmen könnte, so stellt sich die Frage, ob es gleichsam als Erweiterung des Wörterbuchs fungieren kann, und zwar für Mehrwortausdrücke, deren Aufnahme im Wörterbuch nicht möglich ist. Zwar ist in T1 vorgesehen, dass Substantive, die aus mehreren Wörtern zusammengesetzt sind (Mehrwortausdrücke), in das Lexikon aufgenommen werden können, doch ist die Aufnahme von Mehrwortausdrücken auf diese Wortart beschränkt. Die Eingabe eines verbalen Idioms wie z.B. eng. *rain cats and dogs* (dt. *in Strömen gießen*) bleibt dem Benutzer mit dem Hinweis „Keine gültige Eingabe für diese Kategorie. Bitte ändern Sie die Eingabe oder die Kategorie.“ verwehrt.

Der Aufnahme von Mehrwortausdrücken in das Satzarchiv des T1 stehen dagegen keinerlei Einschränkungen entgegen. Sie ist in der gleichen Form möglich, wie dies auch mit beliebigen Sätzen oder anderen Satzfragmenten geschehen kann. Im Lieferumfang von T1 ist bereits ein Satzarchivmodul „idioms“ enthalten, in dem als idiomatische Wendung unter anderem *it's raining cats and dogs* aufgelistet ist. Verwendet man nun bei der Übersetzung eines Textes, in dem beispielsweise die Sätze (3), (4) oder (5) auftreten, das Satzarchivmodul „idioms“, so wird für Satz (3) eine Ähnlichkeit von 60%, für die Sätze (4) und (5) eine Ähnlichkeit von jeweils 40% zu dem Referenzidiom *it's raining cats and dogs* berechnet. Der Prozentsatz von 60% Ähnlichkeit, der als Match-Wert für Satz (3) ermittelt wird, resultiert aus der Tatsache, dass in Satz (3) im Unterschied zur Eintragung im Satzarchiv der Satzanfang mit einem Großbuchstaben beginnt und der Satz darüber hinaus mit einem Satzzeichen (Punkt) beschlossen wird. Aufgrund der Konzeption des Match-Algorithmus (siehe den vorhergehenden Abschnitt) werden in den Sätzen (4) und (5) lediglich die Wörter *cats* und *and* als mit dem Referenzmaterial übereinstimmend identifiziert, so dass bei einer Satzlänge von 5 Wörtern ein Prozentsatz von lediglich 40% ermittelt wird.

(3) *It's raining cats and dogs.*

(4) *It rains cats and dogs.*

(5) *It rained cats and dogs.*

**translate: aString**

„Ermittelt den Match-Wert eines zu übersetzenden Satzes (testWords) relativ zu einem Referenzsatz (refWords).“

```
| testWords result refWords matchValue string index wordsFound |
index := aString indexOf: $. ifAbsent: [aString size].
string := aString copyFrom: 1 to: index.
testWords := T1Scanner new scanTokens: string.
result := Dictionary new.
Reference
keysDo:
    [:ref |
    refWords := T1Scanner new scanTokens: ref.
    wordsFound := 0.
    testWords do: [:word | (refWords includes: word)
    ifTrue: [wordsFound :=
    wordsFound + 1]].
    matchValue := (wordsFound / (testWords size
    max: refWords size)) asFloat * 100.0.
    matchValue >= self threshold ifTrue: [result at:
    ref put: matchValue]].
```

^result

*Abb. 1: Smalltalk-Implementierung der Funktionsweise des in T1 realisierten Match-Algorithmus: Gezählt werden die Wörter des zu übersetzenden Satzes (testWords), die in einem Referenzsatz (refWords) enthalten sind, und zwar unabhängig von ihrer Position und der Häufigkeit ihres Auftretens. Dieser Algorithmus, der syntaktische Strukturen völlig außer Acht lässt, liefert die in Tab. 1, 2 und 3 dargestellten Ergebnisse.*

Bei einer idiomatischen Wendung, die ein flektierendes Satzglied enthält, jedoch nur in einer Form im Satzarchiv abgelegt ist, liegt die Ähnlichkeit immer unterhalb von 100%. Ist die betreffende Wendung schließlich noch in einen Satz eingebettet, der im Normalfall über die idiomatische Wendung hinaus zusätzliche Wörter enthält, so sinkt der Match-Wert in Abhängigkeit von der Zahl der im Satz zusätzlich auftretenden Wörter. Entsprechend wird die Wendung *the most important thing*, für die im Archiv die Übersetzung „das A und O“ angegeben ist, in einem

Satz wie (6) nur mit einem Match-Wert von 60% identifiziert, obschon sie hier in derselben Form wie im Archiv auftritt.

(6) *He recognized **the most important thing** of the story.*

Berücksichtigt man, dass erfahrene Benutzer von TMs in der Regel Match-Werte empfehlen, die oberhalb von 70% liegen, um die benötigte Zeit für die Sichtung der vom System gefundenen Referenzsätze auf einen ökonomisch vertretbaren Rahmen zu beschränken, so sprechen die von T1 berechneten Werte für sich selbst.

Die genannten Beispiele belegen somit, dass das Satzarchiv mit dem in T1 implementierten Match-Algorithmus als Erweiterung des Wörterbuchs für Mehrwortausdrücke nicht geeignet ist.

#### **4.1 Bewertung des Suchalgorithmus des T1-Satzarchivs**

Der für das Satzarchiv von T1 realisierte Suchalgorithmus ist einfach und leicht nachvollziehbar. Er weist jedoch schwerwiegende Mängel auf, weil er beim Vergleich von Test- und Referenzsatz in zahlreichen Fällen zu niedrige, in anderen Fällen zu hohe Match-Werte liefert. Für den praktischen Einsatz zu niedrige Match-Werte werden an Stellen ermittelt, an denen im Testsatz ein Interpunktionszeichen auftritt, wo im Referenzsatz ein solches nicht vorhanden ist (vgl. Satz (3), Tab.1 oder Satz (10), Tab. 2). Da das Interpunktionszeichen als Bestandteil des ihm vorangehenden Wortes interpretiert wird, wirkt sich die unzureichende Segmentierung in besonderem Maße beim Abgleich kurzer Sätze aus (siehe Satz (3), Tab.1). Da die Abfolge der Elemente im Satz beim Abgleich mit dem Referenzmaterial nicht berücksichtigt wird, zudem auch keine Kontrolle über im Testsatz mehrfach auftretende Wörter erfolgt, werden für entsprechende Sätze (vgl. Satz (35) und (37) aus Tab. 3) zu hohe Match-Werte (100%) ermittelt, weshalb Übersetzungen von 100%-Matches unter keinen Umständen ungeprüft übernommen werden dürfen.

Bewertet man das Satzarchiv von T1 unter dem Gesichtspunkt von *Recall* und *Precision*, Parameter, die zur Bewertung von Retrievalsystemen herangezogen werden (siehe auch [Reinke99]), lässt sich die Leistungsfähigkeit des Suchalgorithmus in T1 wie folgt zusammenfassen:

Der Recall, d.h. die Zahl der vom Suchalgorithmus als Basis für die Übersetzung eines Testsatzes ermittelten Kandidaten im Verhältnis zu den tatsächlich im Referenzmaterial enthaltenen relevanten Kandidaten, ist zu gering, weil (a) Satzzeichen als Wortbestandteile interpretiert werden, (b) Teilsätze nicht identifiziert werden, da die Vergleichseinheit ein ganzer Satz ist, und schließlich (c) wegen des Ausklammerns einer linguistischen Analyse der Satzarchivdaten, mit der flektierende Bestandteile erkannt und mit anderen Flexionsformen desselben Paradigmas abgeglichen werden könnten. Gleichzeitig ist aber auch die Precision zu gering, und zwar in solchen Fällen, in denen sich die Vergleichsdaten relativ zum Referenzmaterial wie in Satz (33) aus Tab. 2 in der Abfolge oder der Häufigkeit der im Satz enthaltenen Wörter unterscheiden.

## 5 Der Suchalgorithmus von PT plus

Im Gegensatz zu T1 ist die Retrievalstrategie von PT plus weniger transparent und nachvollziehbar und um einiges komplexer, was sich auch in den Retrievalergebnissen (siehe Tabellen 1 und 2 im vorherigen Abschnitt) widerspiegelt. Beispielsweise werden Modifikationen der Interpunktion oder auch Groß- und Kleinschreibung „robuster“ gehandhabt als bei T1 und bei der prozentualen Bewertung ignoriert (vgl. Sätze (1) – (3) in Tab. 1). Hierbei ist allerdings zu erwähnen, dass sich dieses Verhalten bei unterschiedlichen Satzendezeichen ebenfalls verändert. Während das Hinzufügen oder Weglassen des ‘Punkts’ keinen Einfluss auf das tatsächliche Retrieval und die Angabe der Trefferquote hat, verändert sich sowohl das Retrievalergebnis als auch die vom System kalkulierte Trefferquote beispielsweise beim Hinzufügen eines Fragezeichens (7):

- (7) Referenzsatz:            *Chapter 2 Getting Started*  
      Testsatz:                *Chapter 2 Getting Started?*  
      Match-Wert des PT: 98%

Auch bei der Veränderung von als Ziffern dargestellten Zahlen im Testmaterial vergibt PT plus großzügigere Trefferraten als T1 (vgl. (8)), so dass zu vermuten ist, dass auch bei der Modifikation von reinen Zifferndarstellungen bei PT plus andere Gewichtungen angesetzt werden, was allerdings durch spezifische Tests mit entsprechendem Datenmaterial noch zu belegen ist (siehe hierzu auch Kapitel 6):

- (8) Referenzsatz: *Chapter 2 Getting Started*  
 Testsatz: *Chapter 1 Getting Started*  
 Match-Wert des T1: 75%  
 Match-Wert des PT: 95%

Ähnlich hohe Bewertungen wie bei T1 erhalten auch bei PT plus Modifikationen, die syntaktisch unsinnig sind (siehe Satz (9)), so dass die Aussagefähigkeit der 100%-Bewertungen durch das System – wie dies auch schon bei der Beschreibung der Ergebnisse des T1 im vorherigen Kapitel festgestellt worden ist – in Frage gestellt werden muss:

- (9) Referenzsatz: *This chapter is an introduction to XII operations.*  
 Testsatz: *Chapter This an is to introduction operations XII.*  
 Match-Wert des PT: 91%

## 5.1 Mehrwortausdrücke im PT plus Satzarchiv

Im Gegensatz zu T1 erlaubt es die Wörterbuchschnittstelle von PT plus, Mehrwortausdrücke als benutzerdefinierte Lexikoneinträge für alle Wortarten zu definieren. Ein Eintrag für das obige Beispiel *rain cats and dogs* ist beispielsweise zulässig. Schwierigkeiten mit Mehrwortausdrücken treten bei PT plus in solchen Fällen auf, in denen ein Mehrwortausdruck neben dem verbalen ein nichtverbales Element enthält, das in Abhängigkeit vom Satzsubjekt variiert, wie z.B. *put one's oar in* (dt. *seinen Senf dazu geben*), das bei einem Subjekt in der 3. Person Singular Maskulinum *als he puts his oar in* und bei der 2. Person Singular *als you put your oar in* erscheint.

Für die Untersuchung der Satzarchivfunktionalität von PT plus mit idiomatischen Mehrwortausdrücken wurde daher ein mit den T1-Testdaten identisches Experiment durchgeführt. Hierzu mussten zunächst diese Testdaten in einer Archivdatei aufbereitet werden, da das idiomatische Wörterbuch des PT plus nur offline konsultiert werden kann und nicht online wie bei T1 als Referenzquelle verwendbar ist.

Bei der Betrachtung der Ergebnisse fällt auf, dass in einigen Fällen trotz teilweiser Übereinstimmung von modifizierter Eingabe und Satzarchivdaten der Match-Wert mit 0% angegeben und somit kein Übersetzungsvorschlag des be-

treffenden Fragments aus dem Archiv angezeigt wird (vgl. die Beispiele (10) bis (12)):

- (10) Referenzsatz: *read from cover to cover*  
Testsatz: *She reads the book from cover to cover.*  
Match-Wert des PT: 0%
- (11) Referenzsatz: *he told us everything*  
Testsatz: *He will tell us everything*  
Match-Wert des PT: 0%
- (12) Referenzsatz: *bask in the sun*  
Testsatz: *Yesterday I basked in the sun*  
Match-Wert des PT: 0%

Die Tatsache, dass die Modifikationen teilweise *innerhalb* der idiomatischen Wendungen aufgrund von Flexionsänderungen und Auslassungen oder Komplementierung auftreten, ist alleine keine ausreichende Erklärung für diese Ergebnisse. Gegenproben mit Testsatz (10), der abermals modifiziert wurde ((13) bis (17)), zeigen, dass noch andere Kriterien, wie z.B. die Zeichenlänge der Modifikation, eine Rolle spielen:

- (13) Referenzsatz: *read from cover to cover*  
Testsatz: *read XXX from cover to cover*  
Match-Wert des PT: 87%
- (14) Referenzsatz: *read from cover to cover*  
Testsatz: *read XXXX from cover to cover*  
Match-Wert des PT: 83%
- (15) Referenzsatz: *read from cover to cover*  
Testsatz: *I wanted to read XXXX from cover to cover*  
Match-Wert des PT: 67%



- (16) Referenzsatz: *read from cover to cover*  
 Testsatz: *She wanted to read XXXX from cover to cover*  
 Match-Wert des PT: 0%
- (17) Referenzsatz: *read from cover to cover*  
 Testsatz: *She wants to read XXXX from cover to cover*  
 Match-Wert des PT: 65%

Anhand der Match-Werte, die beim Abgleich von Testsatz (10) bis (17) mit im Satzarchiv enthaltenen idiomatischen Wendungen ermittelt wurden, zeigt sich, dass auch das Satzarchiv von PT plus nur in eingeschränktem Maße als Erweiterung des Wörterbuchs brauchbar ist. Hier kann es lohnenswert sein, entsprechende Wendungen, die mit hoher Frequenz auftreten, in ein spezifisches Benutzerwörterbuch einzutragen. Die Frage nach dem Kosten-Nutzen-Effekt einer solchen Aktivität im Vergleich zur Verwendung des Satzarchivs (Vor- und Nachbereitungsaufwand) erfordert weitergehende Untersuchungen und Testprozeduren, die nicht Gegenstand der hier beschriebenen Arbeiten waren.

## 5.1 Bewertung des Suchalgorithmus des PT plus Satzarchivs

Wie die im vorhergehenden Kapitel illustrierten Beispiele zeigen, ist der Suchalgorithmus des PT plus nicht ohne weiteres nachvollziehbar. Vor allem Testbeispiel (10) und seine Modifikationen (13) bis (17) machen deutlich, dass offensichtlich unterschiedliche Faktoren für die Berechnung der Match-Werte eine Rolle spielen, die sich wiederum je nach Gewichtung gegenseitig zu beeinflussen scheinen. Während die prozentualen Angaben zur Retrievalleistung in (13) und (17) eine scheinbar logische Abhängigkeit von der Länge der Modifikation sowohl innerhalb als auch außerhalb der archivierten Wendung reflektieren, weichen (10) und (16) von dieser Kalkulation ab. Schließlich zeigt auch die unterschiedliche Bewertung von (16), für das der Suchalgorithmus einen Match-Wert von 0% ermittelt, im Vergleich zu (15), das einen Match-Wert von 67% erhält, obschon sich die beiden Sätze lediglich durch das Personalpronomen in Subjektposition – *she* versus *I* – unterscheiden, dass die Berechnung der Ähnlichkeit zwischen Test- und Referenzmaterial häufig nicht plausibel ist.<sup>7</sup>

Somit bleibt festzustellen, dass die Match-Werte des Satzarchivs von PT plus unabhängig von der Struktur der archivierten Daten häufig sehr willkürlich wirken. Die Brauchbarkeit der Satzarchivfunktion hinsichtlich einer Wörterbucherweiterung für idiomatische Mehrwortausdrücke ist – ähnlich wie bei T1 – beschränkt.

## 6 Leistungsfähigkeit der Übersetzungsspeicher-Module

Vergleicht man abschließend die Match-Werte der beiden hier untersuchten Systeme beim Abgleich der Testdaten, so ergeben sich bei der Bewertung der Ähnlichkeit Unterschiede bis zu 50% zwischen T1 und PT plus (vgl. die Beispiele (18) bis (20)).

- |                    |   |
|--------------------|---|
| (18) Referenzsatz: | <i>Chapter 2 Getting Started</i>                                      |
| Testsatz:          | <i>Chapter 2</i>  |
| Match-Wert (T1):   | 50%   |
| Match-Wert (PT):   | 0%  |
| (19) Referenzsatz: | <i>Chapter 2 Getting Started</i>                                      |
| Testsatz:          | <i>Chapter 2 Getting Started.</i>                                     |
| Match-Wert des T1: | 75%   |
| Match-Wert des PT: | 100%  |
| (20) Referenzsatz: | <i>This chapter is an introduction to X11 operations.</i>             |
| Testsatz:          | <i>This chapter is <b>a brief</b> introduction to X11 operations.</i> |
| Match-Wert des T1: | 77%   |
| Match-Wert des PT: | 89%   |

Die Unterschiede in der Bewertung der Ähnlichkeit zwischen Testsatz und im Übersetzungsarchiv enthaltenem Referenzsatz machen deutlich, dass weitergehende systematische Erhebungen erforderlich sind, um fundierte Einschätzungen über die Aussagekraft von Match-Werten einzelner Systeme machen zu können. Hierzu sind neben einem größer angelegten Testkorpus, das systematisch

Modifikationen zwischen Testsatz und Referenzsatz an unterschiedlichen Positionen erfasst, auch authentische Testmaterialien erforderlich, die idealerweise in verschiedenen Textversionen vorliegen sollten. Solange solche Daten nicht vorliegen, muss eine Einschätzung der Leistungsfähigkeit der einzelnen Übersetzungsspeicher und somit auch die hier vorliegende Bewertung vorläufigen Charakter haben.

## Literatur

- [Nübel/Seewald-Heeg98] Nübel, R. und Seewald-Heeg, U. (eds.) (1998): *Evaluation of the Linguistic Performance of Machine Translation Systems*. Proceedings of the KONVENS-98 Workshop in Bonn. St. Augustin: Gardez! Verlag.
- [Andrés-Lange98] Andrés-Lange, C. (1998): 'Tying the Knot. How Baan wed machine translation to translation memory – and survived the honeymoon', in: *Language International* 10/5, 34-36.
- [Reinke94] Reinke, U. (1994): 'Zur Leistungsfähigkeit integrierter Übersetzungssysteme', in: *Lebende Sprachen* 3, 97-104.
- [Reinke99] Reinke, U. (1999): 'Evaluierung der linguistischen Leistungsfähigkeit von Translation Memory-Systemen - Ein Erfahrungsbericht', in: *LDV Forum*, vorliegendes Heft.
- [Twents97] Twents, A. (1997): *Terminologieerkennung in integrierten Übersetzungssystemen am Beispiel des Französischen*. Saarbrücker Studien zu Sprachdatenverarbeitung und Übersetzen. FR 8.6, Universität des Saarlandes, Saarbrücken.

## ANMERKUNGEN

- <sup>1</sup> Unter den kommerziell verfügbaren vollautomatischen Systemen zählen dazu Logos, Personal Translator plus von Linguattec, Power Translator Pro von Globalink, Systran PROfessional von Systran, T1 Professional von Langenscheidt oder Transcend von HEI-Soft, die im Rahmen einer von Nübel und Seewald-Heeg [Nübel/Seewald-Heeg98] dokumentierten Evaluation auf ihre linguistische Leistungsfähigkeit hin untersucht wurden.

- 
- 2 Die Zahl der auf dem Markt angebotenen TM-Systeme ist in den vergangenen Jahren bedeutend gestiegen. Produkte wie SDLX von SDL, Déjà Vu von Atril, Eurolang Optimizer, Joust von AlpNet, Transit von STAR, Translation Manager von IBM, Translator's Workbench von Trados oder ZERESTRANS von Zeres gehören in diese Systemklasse.
  - 3 Eine hiervon abweichende weitergefasste Definition des Terminus „integrierte Systeme“ liefert [Reinke94], auf die sich auch [Twents97] bezieht.
  - 4 Eine entsprechende Systemkonfiguration ist z.B. bei der Firma Baan im Einsatz. Vgl. [Andres-Lange98].
  - 5 Allerdings wird für den Abgleich von Eingabe- und Referenzmaterial die Suche jeweils bis zur Satzgrenze durchgeführt, unabhängig von der Größe der archivierten Übersetzung.
  - 6 Abb. 1 zeigt den Kern des Retrieval-Programmes von T1, wie er anhand der Match-Ergebnisse beim Abgleich des Testkorpus mit dem Referenzmaterial (siehe Tab. 1) mittels der objektorientierten Programmiersprache Smalltalk (VisualWorks 5i) nachgestellt werden konnte.
  - 7 Inwieweit diese Ergebnisse eine Konsequenz der in Retrieval-Software vielfach berücksichtigten Übereinstimmung von Trigrammen (hier: Buchstabenfolgen der Länge '3') oder anderen n-Grammen sind, die in unterschiedlicher Gewichtung in die Berechnung von Match-Werten eingehen, kann im Rahmen der vorliegenden Arbeit nicht beantwortet werden.