Translation Memory

Seminararbeit für das Virtueller Campus Seminar
"Mensch–Maschine–Interaktion in der Maschinellen Übersetzung"

Leitung: Dr. Folker Caroli und Prof. Claus R. Rollinger

vorgelegt von
Arno Erpenbeck, Daniela Hellmann, Tony Peters,
Frauke Schmeier, Timo Steffens, Annika Surrey und Joachim Wagner

WS 1999/2000

Inhalt

- Inhalt
- 1. Einleitung
  - 1.1 Zielsetzung
  - 1.2 Kapitelübersicht
- 2. Begriffe und Grundlagen
  - 2.1 Translation Memory
  - 2.2 Begriffe
  - 2.3 Funktionsweise
- 3. Vorhandene Programme
  - 3.1 Star Transit
    - 3.1.1 Module
    - 3.1.2 Arbeitsweise von Transit 3.0
  - 3.2 Trados TWB
    - 3.2.1 Module
    - 3.2.2 Arbeitsweise
  - 3.3 Atril Déjà Vu
    - 3.3.1 Komponenten des Systems
    - 3.3.2 Arbeiten mit Déjà Vu
    - 3.3.3 Features
  - 3.4 Weitere Produkte
- 4. Matchfunktion
  - 4.1 Bedeutung für den Übersetzungsvorgang
  - 4.2 Matchfunktion von Trados TWB
    - 4.2.1 Änderderung an einem einzelnen Nomen
    - 4.2.2 Änderderung an einem einzelnen Verb
    - 4.2.3 Linguistisches Wissen
    - 4.2.4 Verschiedene Satzlängen
    - 4.2.5 Interpunktion
    - 4.2.6 Vertauschungen von Wörtern innerhalb eines Satzes
    - 4.2.7 Einfügungen und Löschungen von Wörtern
    - 4.2.8 Ergebnis
  - 4.3 Fazit
- 5. Evaluationskriterien
  - 5.1 Zweck und Gründung der Kommission
  - 5.2 Methodik
  - 5.3 Anwendung auf Translation Memories
  - 5.4 Fazit
- 6. Abschließendes Fazit
- Literatur

1. Einleitung

Der Virtuelle Campus ist ein gemeinsames Projekt des Instituts für Angewandte Sprachwissenschaft (IAS) und des Zentrums für Fernstudium und Weiterbildung (ZFW) der Universität Hildesheim, des Instituts für Rechnergestützte Wissensverarbeitung der Universität Hannover und des Instituts für Semantische Informationsverarbeitung der Universität Osnabrück (ISIV).

Im Projekt Virtueller Campus Hildesheim–Osnabrück werden innerhalb von drei Jahren zwei Prototypen zum Logischen Programmieren und zur Mensch– Maschine–Interaktion in der maschinellen Übersetzung sowie die notwendige Infrastruktur für ein Studium im Netz entwickelt und erprobt.

Die zentrale Frage des diesjährigen Projekts (WS 1999/2000), das im Rahmen des Virtuellen Campus (Mensch–Maschine–Interaktion) durchgeführt wird, lautet:

> "Was können Systeme der maschinellen und maschinengestützten Übersetzung in realen Anwendungen wirklich leisten?"

Zu diesem Problem sollen von Projektgruppen, in denen jeweils Studierende aus allen drei Standorten zusammenarbeiten, Fallstudien zu realen Anwendungsbeispielen erarbeitet werden. Es soll untersucht werden, wie die Leistungsfähigkeit solcher Systeme beurteilt werden kann und wie sie den Arbeitsprozeß der Dokumenterstellung und der Übersetzung verändern.

Innerhalb dieses Projekts kamen drei Gruppen zustande, die folgende Themenbereiche bearbeiteten:

- Anwendungen von maschineller und maschinengestützter Übersetzung in der Europäischen Kommission.
- Erstellung und Verwaltung multilingualer technischer Dokumentation (MULTILINT und MULTIDOC)
- Vorstellung verschiedener Übersetzungssysteme mit deren Funktionsweisen, sowie Erstellung eines Evaluations–Leitfadens für Translation Memory Systeme.

Unsere Gruppe, die aus sieben Studenten der Studiengänge Computerlinguistik und Künstliche Intelligenz (CLKI), Internationales Informationsmanagement (IIM) und Internationale Fachkommunikation (IFK) besteht, bearbeitet das Thema Translation Memories.

Frauke Schmeier und Annika Surrey, 30. März 2000

Unterabschnitte

- 1.1 Zielsetzung
- 1.2 Kapitelübersicht

1.1 Zielsetzung

Zunächst stellen wir kommerzielle Translation–Memory–Produkte vor, die zur Zeit hoch gelobt werden. Seit ihrer Einführung in den frühen achtziger Jahren finden sie weite Verbreitung im professionellen Übersetzungsalltag. Wir werden in unserer Arbeit untersuchen, was diese Programme

wirklich leisten können. Dafür definieren wir zunächst den Begriff der Translation Memories. Darauf aufbauend stellen wir die gebräuchlichsten Systeme kurz vor, untersuchen deren Funktionsweise (insbesondere Fuzzy Matching) und erstellen schließlich einen Kriterienkatalog zur Evaluation von Translation Memory Systemen anhand des EAGLES−Reports.

Frauke Schmeier und Annika Surrey, 30. März 2000

1.2 Kapitelübersicht

In 1. Einleitung erläutern wir kurz die Art des Seminars, in dem diese Ausarbeitung entstand, und beschreiben unsere Zielsetzung. Anschließend werden in 2. Begriffe und Grundlagen wichtige Begriffe definiert, die im Text immer wieder benutzt werden, und es wird in die Funktionsweise eines Translation Memories eingeführt. Kapitel 3. Vorhandene Programme enthält eine Beschreibung von drei verschiedenen Programmen. Wir gehen jeweils auf die Programmkomponenten und die typische Arbeitsweise ein. In 4. Matchfunktion untersuchen wir die Matchfunktion, die die Auswahl der Übersetzungsvorschläge beeinflußt, näher. Die Versuchsergebnisse aus den Experimenten mit der Matchfunktion sind hier aufgelistet. In Kapitel 5. Evaluationskriterien gehen wir dann auf den EAGLES Report ein, der als Grundlage für die Evaluation von Translation−Memory−Systemen dienen kann. Schließlich fassen wir in 6. Abschließendes Fazit unsere Ergebnisse kurz zusammen.

Joachim Wagner, 12. April 2000

2. Begriffe und Grundlagen

Unterabschnitte

- 2.1 Translation Memory
- 2.2 Begriffe
- 2.3 Funktionsweise

2.1 Translation Memory

Die Idee eines Translation Memory ist im Grunde einfach. Als Hilfsmittel für den Übersetzer kommen sogenannte Translation Memories (TM's) zum Einsatz, die bereits übersetzte Satz− und Segmentpaare in der Ausgangs− und Zielsprache zur Verfügung stellen. Ist also ein ähnlicher oder identischer Satz in einem Dokument enthalten, braucht dieser nicht mehr übersetzt zu werden, sondern kann direkt übernommen oder noch einmal überarbeitet werden. Noch nicht bekannte Sätze überträgt der Übersetzer einmal in die Zielsprache. Diese werden in einer Datenbank abgelegt und stehen für weitere Anwendungen zur Verfügung. Translation Memories sind bildhaft gesprochen nicht mehr als zwei übereinstimmende Aktenschränke, in denen alle jemals angefertigten Übersetzungen jeweils in Ausgangssprache oder Zielsprache gespeichert sind.

Tony Peters, Frauke Schmeier und Annika Surrey, 29. März 2000

2.2 Begriffe

Um die Arbeitsweise eines Translation Memories zu verstehen, sollten zunächst einige Begriffe, die in diesem Zusammenhang auftauchen, näher erläutert werden.

Begriffe:

- Übersetzungseinheit / translation units: abgespeicherte Satz– und Segmentpaare
- Translationdatenbanken: Ansammlung von Übersetzungseinheiten
- Fuzzy matching: "Unscharfes Suchen" = ermöglicht das Auffinden von Daten, die nur eine gewisse Ähnlichkeit mit dem Suchargument aufweisen
- Exact matches : 100% übereinstimmende Daten
- Automatische Terminologieunterstützung: Das System sucht in der Terminologiedatenbank nach bereits übersetzten Einzelbegriffen, bzw. neu übersetzte Einzelbegriffe werden dort abgelegt.
- Alignment: nachträgliches Segmentieren von bereits vorhandenen nicht in TM erstellten Übersetzungen
- Tagger: Modul, das bestimmte Formate wie Framemaker oder Interleaf in das Format des Übersetzungsprogramms konvertiert.
- Filter: dienen zur Abgrenzung unterschiedlicher Terminologien (Projektbezogene Arbeit)
- TM: Translation Memory
- AT: Ausgangstext
- ZT: Zieltext
- AS: Ausgangssatz
- ZS: Zielsatz
- MÜ: maschinelles Übersetzen

Tony Peters, Frauke Schmeier und Annika Surrey, 29. März 2000

2.3 Funktionsweise

Wie funktioniert ein TM?

Ausgangspunkt der Arbeit mit Translation Memories sind die vom Auftraggeber der Übersetzung gelieferten Dateien.

Als erstes muß der Ausgangstext in einen zum Programm passenden Editor übernommen werden. Der Übersetzer kann nun direkt mit dieser Datei arbeiten. Der Text wird segmentiert, und die Software versucht nun durch einen programmgesteuerten Abgleich gleiche oder ähnliche Segmente zu finden. Sie nutzt dazu früher bereits übersetzte Texte, die in segmentierten Textdateien oder eigenen Datenbanken abgelegt sind. Ergebnis dieses Abgleichs sind je nach Übereinstimmungsgrad Exact Matches oder Fuzzy Matches. Ein Exact Match kann ohne weiteres Bearbeiten direkt in die Übersetzung übernommen werden. Ein Fuzzy Match liefert ähnliche Segmente, die Alternativen angeben. Das Zielsegment muß also noch in einem bestimmten Umfang bearbeitet werden.

Gleichzeitig steht dem Übersetzer eine automatische Terminologieunterstützung zur Verfügung. Gefundene Übersetzungen von Einzelbegriffen in der Terminologiedatenbank werden angezeigt und/oder eingefügt. So können einzelne Wörter automatisch ersetzt werden, auch wenn der gesamte Satz vom früheren Text abweicht.

Weicht die angebotene Lösung stark vom Ausgangssatz ab, überträgt der Übersetzer den Satz gegebenenfalls komplett neu in die Zielsprache. Diese neue Übersetzung wird dann in das Translation Memory übernommen und steht beim nächsten zu übersetzenden Ausgangssatz wiederum als Exact Match oder als Fuzzy Match zur Verfügung.

Beginnt der Übersetzer seine Arbeit mit einem neuen TM, dann ist dieses zunächst noch leer und wird durch seine Einträge im Laufe der Zeit gefüllt, so daß die Bestände bei der Arbeit ständig wachsen. Es ist jedoch auch möglich, mit Hilfe eines Alignment–Tools bereits angefertigte Übersetzungen in das TM zu übernehmen. Dabei werden Ausgangstext und Zieltext in Segmente aufgesplittet, eindeutig zugeordnet und in das TM übernommen.

Für kunden– und projektbezogene Arbeiten kann der Übersetzer abhängig vom verwendeten Übersetzungsprogramm überdies verschiedene Filter setzen, um unterschiedliche Terminologien voneinander abzugrenzen und nur die jeweils gültige für die zu erstellende Übersetzung zu nutzen.

Vorteile:

Aus der vorhergehenden Beschreibung der Arbeit mit TMs, ergeben sich bestimmte Einsatzbereiche und Vorteile:

- Hat man eine TM–Datenbank erst einmal gefüllt, ergibt sich daraus eine Arbeits– und damit Zeit– und Kostenersparnis.
- Ein TM garantiert die Qualität und Konsistenz der Übersetzungen.
- TMs ermöglichen durch die Nutzung bestimmter Filter darüber hinaus das Bearbeiten von Formaten, die andernfalls nicht ohne weiteres importierbar oder exportierbar wären (z.B. Framemaker–Dateien)
- Der Einsatz von TM bringt die größten Vorteile, wenn ein hoher Wiederholungsgrad innerhalb eines Textes oder einen ganzen Reihe von Texten (z.B. bei Updates) gegeben ist.

Hardware:

In jedem Fall sollte man beim Einsatz dieser Systeme nicht bei der Rechnerausstattung sparen, da bei großen Textbeständen (dann lohnt sich der Einsatz von TM ja erst richtig) und zum Teil aufwendigen Abgleichprozeduren schnellere Prozessoren und Festplatten für ein komfortables Arbeiten erforderlich sind.

Tony Peters, Frauke Schmeier und Annika Surrey, 29. März 2000

3. Vorhandene Programme

Auf dem Markt werden verschiedene Translation–Memory–Tools angeboten:

- TRANSLATORS WORKBENCH, Fa. Trados, Stuttgart
  http://www.trados.com/
- TRANSIT, Fa. Star GmbH, Böblingen
  http://www.star–ag.ch/
- TRANSLATION MANAGER, Fa. IBM, Heidelberg
  http://www.ibm.com/software/ad/translat/tm/
- DEJA–VU, Fa. Atril
  http://www.atril.com/

- SDLX, Fa. SDL–International,
  http://www.sdlintl.com/
- OPTIMIZER, Fa. Eurolang, München

In folgenden Abschnitten werden wir einen Überblick über die Bestandteile und Arbeitsweise der drei Programme Translators Workbench, Transit und Déjà–Vu geben.

Tony Peters, Frauke Schmeier und Annika Surrey, 29. März 2000

Unterabschnitte

- 3.1 Star Transit
  - 3.1.1 Module
  - 3.1.2 Arbeitsweise von Transit 3.0
- 3.2 Trados TWB
  - 3.2.1 Module
  - 3.2.2 Arbeitsweise
- 3.3 Atril Déjà Vu
  - 3.3.1 Komponenten des Systems
  - 3.3.2 Arbeiten mit Déjà Vu
  - 3.3.3 Features
- 3.4 Weitere Produkte

3.1 Star Transit

Transit ist ein Produkt der Star Deutschland GmbH, das nicht an ein bestimmtes Textverarbeitungs–system gebunden ist, sondern über voll kompatible Schnittstellen zu gängigen Textverarbeitungs– und Desktop–Publishing–Systemen verfügt. Transit dient als Plattform und kombiniert zwei inte–grierte Komponenten, Translation Memory und Translation Editor, mit einem eigenständigen Mo–dul, der Terminologie–Datenbank TermStar.

Tony Peters, Frauke Schmeier und Annika Surrey, 29. März 2000

Unterabschnitte

- 3.1.1 Module
- 3.1.2 Arbeitsweise von Transit 3.0

3.1.1 Module

- Translation Memory
- Translation Editor
- TermStar
- Alignment Tool

Das **Translation Memory** speichert jede einmal überarbeitete Texteinheit, auf welche dann zu–

rückgegriffen werden kann. Transit segmentiert dabei beide Texte und numeriert die Segmente durch. Anhand der Segmentnummer sind diese mit der anderen Sprache verknüpft. Diese gelten als Sprachpaare. Transit nutzt diese Sprachpaare (Ausgangstext und Zieltext), die vom Benutzer als Referenzdateien angegeben werden. Das TM besteht also nicht aus einer Datenbank, sondern aus einem Assoziativen Netz, das aus Referenzdateien beim Start des Programms aufgebaut wird. Dadurch können identische Segmente in der Vorübersetzung übernommen werden. Das System liefert auch Übersetzungen ähnlicher Einheiten, Fuzzy Match, bei der Suche kann eine Mindestübereinstimmung in Prozent festgelegt werden. Unterschiede werden dabei farbig hervorgehoben, so daß sie dann angepaßt werden können.

Der **Translation Editor** bietet alle übersetzungsrelevanten Funktionen. Der Editor stellt eine automatische Synchronisierung der Fenster für Ausgangssprache, Zielsprache, Terminologie und Notizen bereit. Er weist mehrere übersetzungsspezifische Funktionen auf, wie die farbige Hervorhebung verschiedener Informationen und des Übersetzungsstatus. Außerdem ist ein Falten des Textes möglich, d.h. die ausschließliche Anzeige von Übersetzungseinheiten, die an besondere Bedingungen geknüpft sind.

TermStar nennt sich das eigenständige Terminologieverwaltungssystem. Es arbeitet im Hintergrund und sucht während der Übersetzung ständig nach bereits in der Datenbank vorhandenen Terminologie. Gefundenes wird dabei farbig hervorgehoben, und die Übersetzung wird dabei im Terminologiefenster eingeblendet. Die Sprachrichtung kann jederzeit geändert werden. Noch nicht erfaßte Termini können jederzeit per Mausklick/Tastendruck in eine Datenbank aufgenommen werden.

Transit ist also eine Arbeitsoberfläche, auf der die Möglichkeit geboten wird, mit drei integrierten Komponenten zu arbeiten. Der Übersetzungsprozeß findet dabei im integrierten Editor statt.

Hinzu kommt noch das Alignment Tool, das bereits angefertigte Übersetzungen formatiert, in Segmente zerlegt, durchnumeriert und als Referenzdateipaar nutzt.

Wenn man **in dem System arbeitet**, erstellt man zunächst ein eigenes Benutzerprofil. Dabei bestimmt man die Dialogsprache, die 1. Zielsprache, die 2. Zielsprache, die Fensteroptionen und die Editorattribute. Eine Projektdefinitionsdatei wird geladen und erstellt, in der die übersetzungsrelevanten Informationen gespeichert sind. Die zu übersetzenden Dateien müssen ins System importiert werden, dabei wird der Text in Segmente zerlegt. Das Translation–Memory–Programm Transit von Star konvertiert die zu übersetzenden Dateien zunächst in einem Importschritt in das Transit–Format. Dabei entstehen AS–Dateien und ZS–Dateien mit Dateinamenserweiterungen wie etwa "ger" für deutschsprachige und "eng" für englischsprachige Dateien. Hierzu ist zu beachten, daß Dateien sich nur durch ihre Namenserweiterung unterscheiden, vor dem Import umbenannt werden müssen. Während des Imports kann optional eine Vorübersetzung mit Exact Matches durchgeführt werden. Der zu übersetzende Text muß nun in den Editor geladen werden. Dabei legt Transit automatische ein Sprachpaar an, d.h. zusätzlich zu dem Ausgangstext wird ein Dokument für den Zielsprachentext geöffnet.

In der Arbeitsoberfläche sind beim Übersetzungsprozeß verschiedene Fenster möglich:

Das Ausgangsfenster enthält den Originaltext. Text und Steuerzeichen sind dabei schreibgeschützt. Im Wörterbuch gefundene Termini werden dort farbig angezeigt. Im Zieltextfenster wird der Übersetzungsprozeß durchgeführt. Wahlweise kann es sich dabei um eine Kopie des Ausgangstextes oder ein leeres Dokument sein. Auch hier sind die Termini des Wörterbuchs farbig hervorgehoben und die TAGs schreibgeschützt.

Das Wörterbuchfenster zeigt die Termini des aktuellen Segments an, das im TermStar Wörterbuch gefunden wurde. Zur Übernahme in den Zieltext stehen mehrere Möglichkeiten zur Verfügung.

Das Assoziativfenster zeigt Einträge des assoziativen Netzes an, die Ähnlichkeiten mit dem zu übersetzenden Segment haben. Dieses Fenster liegt standardgemäß im Hintergrund und wird per Tastendruck aufgerufen.

Nach der Durchführung der Übersetzung sind die Segmente in der Zieldatei entsprechend ihrem Status gekennzeichnet nach: vollständig übersetzt, teilweise übersetzt oder unvollständig übersetzt.

Der Befehl <Segment übersetzt> legt ein Segment als übersetzt fest, und es wird zum nächsten Segment übergegangen. Der Befehl <Datei übersetzt> beendet den Übersetzungsprozeß.

Transit bietet zusätzlich noch die Möglichkeit einer Rechtschreibprüfung und eines Terminologie–checks. Die fertige Übersetzung kann aus dem System in das ursprüngliche Format exportiert wer–den.

Tony Peters, Frauke Schmeier und Annika Surrey, 29. März 2000

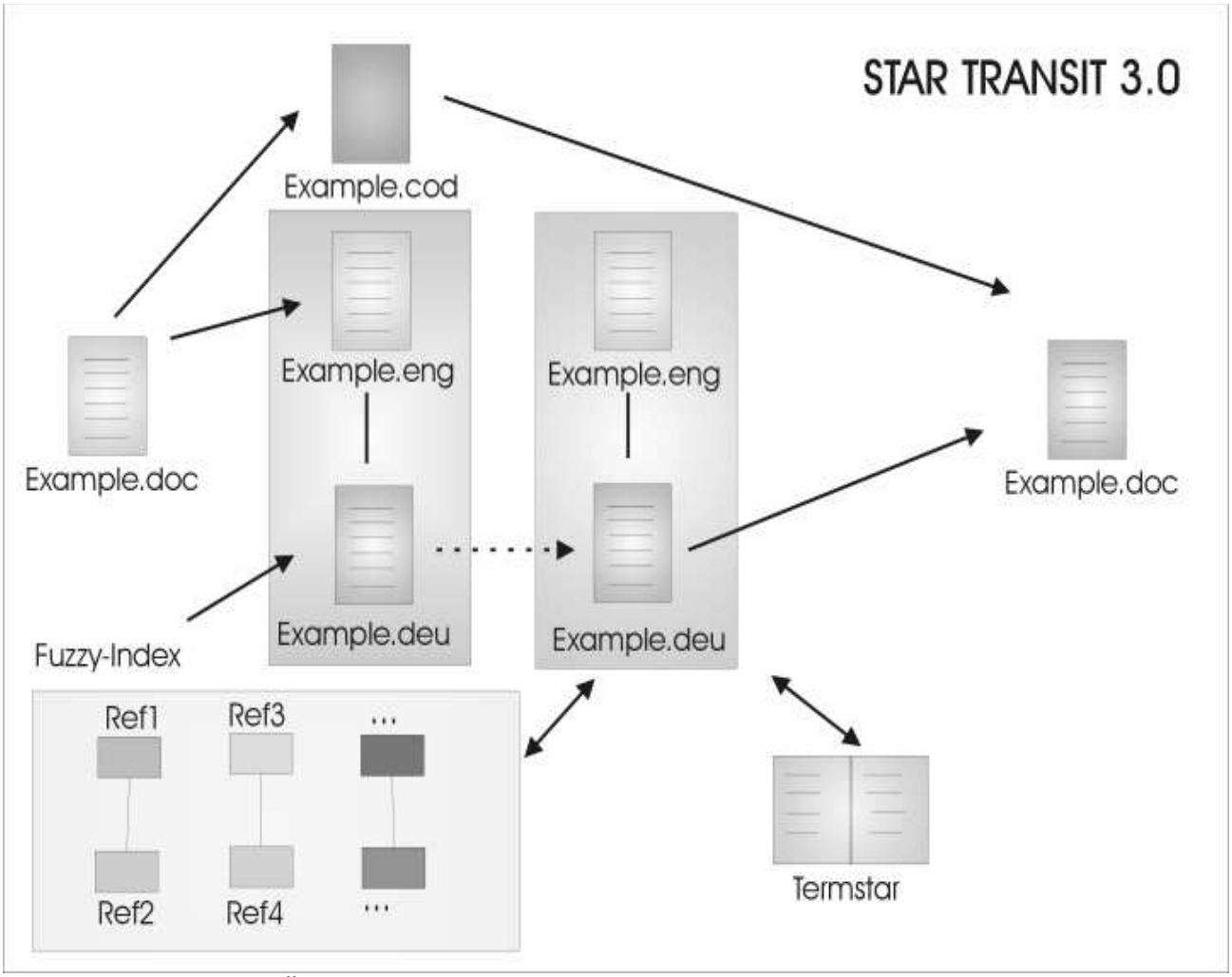
3.1.2 Arbeitsweise von Transit 3.0

Zu Projektbeginn hält der Übersetzer in einem Transit–Projekt mit Hilfe eines Assistenten die Pro–jektinformationen fest (Kundeninformationen, Sprachen, Referenzmaterial etc.).

Nachdem alle wichtigen Informationen eingegeben wurden, importiert der Anwender die zu über–setzende Datei und übersetzt sie im Editor. Transit extrahiert dabei die Textinformation in eine Textdatei und die Layoutinformation wie Grafiken, Zeichensätze usw. in eine eigene Datei. Wäh–rend des Übersetzens greift Transit auf das Referenzmaterial zu und fragt es nach bereits vorhande–nen Übersetzungen ab.

Nach der Übersetzung wird die übersetzte Datei exportiert und die Layout– und Textinformationen zusammengeführt. So erhält der Übersetzer automatisch die Datei wieder im Ausgangsformat.

Die nachstehende Grafik zeigt den Programmablauf bei Transit 3.0.



Beispielsituation: Ein Übersetzter soll einen englischen Text EXAMPLE.DOC in einen deutschen Text BEISPIEL.DOC übersetzen.

Er legt zunächst in Transit 3.0 mit Hilfe des Assistenten ein Projekt an und bestimmt die Referenz–dateien, die er in die Übersetzung einbeziehen will.

Transit extrahiert mit Hilfe eines Import–Filters aus der Datei EXAMPLE.DOC die Textinforma–tionen und erstellt eine schreibgeschützte Textdatei EXAMPLE.ENG, die keine Layout– oder Strukturinformationen enthält. Anhand des Suffixes .ENG wird die Ausgangssprache identifiziert. Zusätzlich teilt Transit den Text in Segmente ein.

Gleichzeitig erstellt Transit eine exakte Kopie der Textdatei EXAMPLE.ENG und nennt diese Ko–pie EXAMPLE.DEU. Der Suffix .DEU gibt die Zielsprache an. Diese Datei wird anschließend be–arbeitet.

Zusätzlich erzeugt Transit eine Datei EXAMPLE.COD, die sämtliche Layoutinformationen abspei–chert.

Transit baut anhand der vom Übersetzer gewählten Referenzdateien einen Fuzzy–Index auf. Dieser Fuzzy–Index stellt das Translation Memory dar, in dem bereits übersetzte Sätze zusammen mit den entsprechenden ausgangssprachlichen Sätzen als Sprachpaar abgespeichert sind.

Nachdem der Fuzzy–Index aufgebaut wurde, überprüft Transit die Datei EXAMPLE.DEU auf Sät–ze, die bereits in genau dieser Form im Fuzzy–Index abgespeichert sind. In diesem Fall ersetzt Transit die Sätze durch die Übersetzungen. Der Übersetzer hat nun eine vorbereitete Datei und kann

die Sätze, die Transit noch nicht vorübersetzt hat, übersetzen.

Die so übersetzten Sätze werden zusammen mit dem jeweiligen Ausgangssatz als Sprachpaar im Fuzzy–Index abgespeichert, der sich daraufhin ständig aktualisiert. Der Fuzzy–Index bietet dem Übersetzter während des Übersetzens Vorschläge an, sobald er "ähnliche" Sätze findet. Der Ähnlichkeitsgrad kann vom Übersetzer eingestellt werden.

Das Wörterbuch TermStar liefert dem Übersetzer Stichwörter, die der Übersetzer bereits abgespeichert hat. Der Übersetzer kann ebenso während des Übersetzens seine Terminologie im Wörterbuch einpflegen.

Hat der Übersetzer die Datei EXAMPLE.DEU komplett übersetzt, ruft Transit die Layoutinformationen aus der Datei EXAMPLE.COD ab und erstellt über den Filter die Datei EXAMPLE.DOC.

Daniela Hellmann, 29. März 2000

3.2 Trados TWB

Die Trados GmbH (www.trados.de) vertreibt das Produkt Translator´s Workbench.

Die Translator´s Workbench ist eine von der Firma Trados (Stuttgart) entwickelte Arbeitsumgebung für das rechnergestützte Übersetzen natürlicher Sprache. Sie umfaßt Module zur Abfrage von Terminologien und ganzen Sätzen sowie eine Schnittstelle zu einem maschinellen Übersetzungsprogramm. Der Zugriff auf diese Module ist in die Bedieneroberfläche einer Standard–Textverarbeitung integriert. Bei der mit der der Translator´s Workbench erfolgt kein Überschreiben des Quelltextes, sondern eine Formatierung als "Hidden Text", d.h. die Übersetzung fügt sich hinter dem Ausgangstext ein. Nach Fertigstellen der Übersetzung wird dann ein Clean–up durchgeführt, so daß der Ausgangstext verschwindet.

Tony Peters, Frauke Schmeier und Annika Surrey, 29. März 2000

Unterabschnitte

- 3.2.1 Module
- 3.2.2 Arbeitsweise

3.2.1 Module

Die Workbench besteht aus:

- Translation Memory
- MultiTerm
- Schnittstelle zur Textverarbeitung
- Transcend
- WinAlign

Das **Translation Memory** ist als wichtigste Komponente der TW eine mehrsprachige Satzdaten–

bank, in der ausgangs– und zielsprachliche Sätze paarweise einander zugeordnet sind. Die Satzpaare werden als Übersetzungseinheiten bezeichnet und zusammen mit bestimmten Zusatzdaten abgespeichert. Bei der Übersetzung eines neuen Textes wird das Translation Memory abgefragt und zeigt (Teil–) Übereinstimmungen mit vorhandenen Sätzen an. Die Abfrage erfolgt fuzzy oder unscharf und führt zu den als Matches bezeichneten Treffern. Eine fuzzy Suche führt zum Auffinden ähnlicher Sätze im Translation Memory, wobei ein als Neuronales Netz bezeichnetes Abgleichverfahren verwendet wird. Das TM erlaubt auch die Suche nach beliebigen Zeichenketten (Wörtern, Wortteilen, Wendungen usw.) in der sogenannten Konkordanzsuche.

MultiTerm ist ein Terminologieverwaltungsprogramm, das als Einzelanwendung oder Teilprogramm der TW genutzt werden kann, um einen sachgebietsbezogenen Spezialwortschatz zu pflegen. In MultiTerm sind Termini mehrsprachig zusammen mit Zusatzinformationen wie Quellenangaben, Definitionen, Sachgebieten, Grammatikangaben u.ä. als strukturierte Begriffseinträge abgelegt. Die Datenstruktur in MultiTerm ist in Systemfelder, Indexfelder, Textfelder und Attributfelder gegliedert. Indexfelder dienen zur Aufnahme/Abfrage der Schlagwörter, Textfelder für die frei formulierten Zusatzinformationen(z.B. Definitionen, Kontexte, Quellenangaben) und Attributfelder für feste Werte einer Auswahlliste (z.B. Sachgebiet, Grammatikangaben). Wie beim Translation Memory dienen diese Felder zur Anzeige und gefilterten Abfrage. Eine Abfrage des Translation Memories bewirkt bei entsprechender Voreinstellung eine automatische Abfrage von MultiTerm und eine Anzeige der Ergebnisse (aktive Terminologieerkennung). Die Abfrage in MultiTerm erfolgt fuzzy, wobei zuvor ein Fuzzy–Index aktiviert werden muß.

Die Abfrage des TM erfolgt aus der Textverarbeitung heraus über eine DDE–Schnittstelle (DDE=Dynamic Data Exchange, ein Verfahren zur Verknüpfung mehrerer MS–Windows–Anwendungen). Die über das Makro (Dokumentenvorlage) in die Textverarbeitung eingeblendeten Menüoptionen bzw. Symbole stellen den einzigen Anwenderzugang zum TM dar. Die Arbeitsweise ist interaktiv. Das TW segmentiert den Quelltext während der Übersetzung nach vordefinierten Regeln und durchsucht die Datenbank nach bereits übersetzten Teilen. Ein Abfragesatz wird in der Textverarbeitung automatisch farbig unterlegt und in das gleichzeitig geöffnete TM–Fenster übernommen. Ein ggf. passender Match erscheint mit Angabe der prozentualen Übereinstimmung, und die Übersetzung des TM–Satzes wird in ein Farbfeld der Textverarbeitung übernommen, wo sie weiter editiert werden kann. Sobald ein neues Segment gespeichert ist, erhält es einen Eintrag im TM und wird beim nächsten ähnlichen Segment zur Auswahl gestellt. Zusätzlich werden im TM–Fenster die Treffer aus MultiTerm markiert und in einem eigenen Fensterbereich angezeigt. Eine Konkordanzsuche kann aus der Textverarbeitung oder unter der TM–Oberfläche durchgeführt werden. Als Ergebnis werden alle Übersetzungseinheiten mit der Suchmenge (ebenfalls fuzzy) in einem eigenen Fenster dargestellt. Nicht zu übersetzende Satzteile wie Grafiksymbole oder Feldinformationen (Datum, Seitenzahl) werden unverändert in den ZT–Satz eingefügt.

Transcend ist eine PC–Software zur vollautomatischen Übersetzung natürlichsprachlicher Texte ins Englische und aus dem Englischen. Transcend liefert keine Wort–für–Wort – Übersetzung, sondern analysiert morphologische und syntaktische Strukturen sowie semantische Merkmale und versucht, diese korrekt in die Zielsprache zu übertragen. Transcend erstellt jedoch, wie alle MÜ–Systeme, Rohübersetzungen, die überarbeitet werden müssen. Transcend ist in die MS–Windows–Umgebung und insbesondere in die TW integriert und "interagiert" mit den Modulen Translation Memory und MultiTerm sowie mit der Textverarbeitung MS Word oder Corel WordPerfect für Windows über ein DDE–Makro. Transcend arbeitet mit eigenen, komplex aufgebauten Wörterbüchern, und zwar einem nicht zugänglichen Hauptwörterbuch sowie mit erweiterbaren Benutzerwörterbüchern. Letztere, von denen immer nur eines aktiv sein kann, werden vor dem Hauptwörterbuch abgefragt und haben eine höhere Priorität.

Mit **WinAlign** bietet Trados ein spezielles Tool an, mit dem aus schon vorhandenen Dokumenten und ihren Übersetzungen Translation Memories erstellt werden können. Es handelt sich hierbei um

ein visuelles, interaktives, sehr flexibles und vielseitiges Tool, das eine Fülle von Optionen zur Optimierung der Ergebnisse des Alignments bietet. Das Programm muß vor allem Satzgrenzen richtig erkennen und Probleme bewältigen, die dadurch entstehen, das ein AT–Satz mit mehreren ZT–Sätzen (oder umgekehrt) übersetzt wurde. Zur Analyse verwendet WinAlign eine Kombination aus statistischen Parametern (mittlere Satz– und Wortlängen) und Texthinweise wie in Sätzen enthaltene Eigennamen, Akronyme, Zahlen, Tags von Markup–Sprachen. Das noch zu überarbeitende Ergebnis wird anschließend in ein TM importiert und als Alignmentergebnis gekennzeichnet.

Fazit: Die Translator's Workbench kann als rechnergestützte Arbeitsumgebung nur unter bestimmten Bedingungen nutzbringend eingesetzt werden. Die MÜ–Komponente eignet sich nur für Texte mit einfachen syntaktischen Strukturen, die Wörterbuchpflege ist relativ zeitaufwendig, die Kodierung der Einträge nicht immer leicht nachvollziehbar, die Interaktion mit der Terminologieverwaltung bringt zwar einen Gewinn für die Terminologie, stört aber die syntaktische Analyse von Transcend. Das Translation Memory ist durch seine Einbettung in zwei bekannte Textverarbeitung bedienerfreundlich, wozu auch die "optische Führung" durch farbige Unterlegung der gerade zu übersetzenden Sätze zählt. Sein Nutzen zeigt sich aber vor allem bei sehr ähnlichen Texten, wobei auch eine thematische und textsortenmäßige Nähe nicht ausreicht. Die Konkordanzsuche ist für das Auffinden von Einzelbegriffen oder Wendungen sehr hilfreich. Die aktive Terminologieerkennung ist sehr nützlich, bedingt jedoch den vorherigen Aufbau spezifischer Terminologiedatenbanken.

Tony Peters, Frauke Schmeier und Annika Surrey, 29. März 2000

3.2.2 Arbeitsweise

Der Übersetzer startet zuerst das Programm Trados Translator´s Workbench und die Textverarbeitung. Unterstützt werden die Programme MS Word und Corel WordPerfect.

Im Translator´s Workbench öffnet der Übersetzer eine bereits bestehende Translation Memory Datenbank oder erstellt eine neue. In der Datenbank werden sämtliche Satzpaare aus früheren Übersetzungen gespeichert. Daneben enthält die Datenbank Projektinformationen. Diese können als Filter benutzt werden, um bestimmte Übersetzungen, die z.B. zum Sachgebiet passen, zu bevorzugen.

Das zu übersetzende Dokument wird in der Textverarbeitung geöffnet. Soll es erhalten bleiben, muß der Benutzer das Dokument unter einen neuen Namen abspeichern. Mit einer Tastenkombination oder Mausklick wird die Übersetzung gestartet. Im Textverarbeitungsfenster werden der aktuell zu übersetzende Satz und der Übersetzungsvorschlag aus dem Translation Memory farblich unterlegt. Der Benutzer kann diesen Vorschlag direkt in der Textverarbeitung wie gewohnt editieren oder in das Translator´s Workbench Fenster wechseln, um dort alternative Übersetzungsvorschläge auszuwählen. Mit einer weiteren Tastenkombination oder Mausklick wird der übersetze Satz zusammen mit dem Ausgangssatz in der Datenbank abgelegt und zum nächsten Satz weitergegangen.

Neben dieser iterativen Vorgehensweise bietet Translator´s Workbench auch die Möglichkeit, vorab alle Sätze mit 100% Übereinstimmung automatisch übersetzen zu lassen.

Außer der Translation Memory Funktion enthält das Produkt das Programm MultiTerm zur Pflege der Terminologie. Es muß ebenfalls extra gestartet werden.

Joachim Wagner, 25. März 2000

3.3 Atril Déjà Vu

Die Firma Atril (www.atril.com) mit Sitz in Madrid, Spanien, stellte ihr Übersetzungssystem Déjà Vu erstmals am 20. November 1999 offiziell vor.

Ebenso wie die oben beschriebenen Systeme Translator's Workbench von Trados und Transit von Star ist Déjà Vu ein System mit Translation−Memory−Komponente. Es eignet sich besonders für den professionellen Einsatz, da es sich um ein sehr leistungsfähiges und zugleich benutzerfreundliches, weil individuell anpaßbares Übersetzungswerkzeug (auch Übersetzungstool genannt) für computergestützte Übersetzung (CAT) handelt.

Tony Peters, Frauke Schmeier und Annika Surrey, 29. März 2000

Unterabschnitte

- 3.3.1 Komponenten des Systems
- 3.3.2 Arbeiten mit Déjà Vu
- 3.3.3 Features

3.3.1 Komponenten des Systems

Déjà Vu besteht aus mehreren Programmen:

- Déjà Vu Interactive (DVI)
- Database Maintenance
- Terminology Maintenance
- SGML Filter Maintenance
- Term Watch
- Database Conversion Wizard
- Online Hilfe

Déjà Vu Interactive (DVI) ist die zentrale Arbeitsoberfläche für den Benutzer und gleichzeitig Hauptprogramm des Übersetzungssystems. Hier werden alle Übersetzungsprojekte erstellt und verwaltet sowie Schreib− und Lesezugriff auf die Memory−Datenbanken (Memory Maintenance) und die Terminologie−Datenbank (Terminology Maintenance) ermöglicht. Es wird als Hauptprogramm des Systems bezeichnet, weil man sich bei der Übersetzung darauf beschränken kann, lediglich mit diesem Programm zu arbeiten. DVI stellt alle grundlegenden Funktionen bereit, die für den Übersetzungsprozeß notwendig sind.

Database Maintenance ist eine Arbeitsoberfläche, die der Pflege von Memory−Datenbanken dient. Es bietet alle dazu benötigten Funktionen wie Import/Export, Datenabgleich, Reparaturen, manuelles Hinzufügen und Entfernen von Satzpaaren usw.

Das Programm Terminology Maintenance ist das Äquivalent zur Database Maintenance. Sein Funktionsumfang entspricht dem der Database Maintenance, bezieht sich aber auf die Pflege der Terminologiedatenbanken.

SGML Filter Maintenance dient zur Erstellung von Filtern für Dateien, die den SGML−Standard

benutzen.

TermWatch ist ein Programm, das es ermöglicht, aus so gut wie jeder MS−Windows−Anwendung auf die Terminologiedatenbanken zuzugreifen. Der Benutzer kann Tastenkombinationen so definieren, daß das Programm bei deren Betätigung die Terminologiedatenbank nach dem ausgewählten Begriff durchsucht.

Der Database Conversion Wizard dient lediglich dazu, Memory−Databanken, die mit früheren Déjà Vu−Versionen erstellt wurden in das jetzige Format zu konvertieren.

Hinzu kommt eine Online−Hilfe, die der Benutzer bei jeglicher Art von Schwierigkeiten um Rat fragen kann. Sollte das Programm nicht in der Lage sein, bei Problemen zu helfen, steht die Firma Atril seinen Kunden per e−mail immer zur Verfügung. Es gibt für Atril−Kunden aller europäischen Länder, in denen Déjà Vu vertrieben wird, einen eigenen Support/Ansprechpartner, der häufig schon nach wenigen Stunden die Fragen der Benutzer beantwortet hat. Ein weiterer Pluspunkt für Déjà Vu.

Tony Peters, Frauke Schmeier und Annika Surrey, 29. März 2000

3.3.2 Arbeiten mit Déjà Vu

Als Erstes muß ein Projekt angelegt werden. Hier wird zunächst das Dateiformat bestimmt, denn das Übersetzungstool ist durch die Vielzahl von Filtern in der Lage, mit Word, RTF, Powerpoint, FrameMaker, QuarkXPress, Interleaf, IBM TM, RC, C/C++, HTML und Nur−Text−Dateien zu arbeiten.

Als zweiter Schritt sind die Eigenschaften des Projekts festzulegen, d.h. Thema, Kunde, Fachgebiet, Sprachkombination etc. Für die ausgewählte Sprachkombination, z.B. Übersetzungsrichtung Englisch −> Deutsch, müssen entsprechende Datenbanken angegeben oder neu angelegt werden. Hier bietet das System eine sehr praktische Funktion an, die es ermöglicht, die Sprachrichtung einer Datenbank vom Maintenance Programm umkehren zu lassen.

Im dritten Schritt wird der zu übersetzende Text, der Quelltext, in das Translation Memory importiert. Je nach Format können bestimmte Elemente wahlweise auch unterdrückt und vom Import und der Übersetzung ausgenommen werden. Nun vergleicht das System den importierten Text mit dem Übersetzungsarchiv, das zu Beginn der Arbeit mit dem TM angelegt wird und mit jedem Übersetzungsprojekt wächst. Die Effektivität des TM steigert sich also um so mehr, je länger es im Einsatz ist. Durch den sogenannten Alignment−Prozeß läßt sich der Import bereits vorhandener Übersetzungen durchführen und somit die Brauchbarkeit des Translation Memories schnell und leicht verbessern. Von Alignment spricht man, weil Ausgangs− und Zielsprache in zwei nebeneinander liegenden Fenstern aneinander ausgerichtet werden müssen, so daß dem ausgangssprachlichen Satz oder Segment das korrekte zielsprachliche Segment zugeordnet ist.

Nachdem der Quelltext nun importiert ist und die notwendigen Einstellungen durchgeführt wurden, beginnt die eigentliche Arbeit des Systems. Es führt eine Vorübersetzung (Pretranslation) durch, indem es den zu übersetzenden Text prüft und hinsichtlich jedes einzelnen darin enthaltenen Satzes die Memory−Datenbank nach früher übersetzten Sätzen mit ähnlichem Wortlaut durchsucht. Déjà Vu nimmt den Satz mit der höchsten Ähnlichkeit und fügt ihn an der entsprechenden Stelle der Übersetzung ein. Außerdem unterlegt es alle Wörter des Quelltextes farbig, deren Entsprechung es im Übersetzungsarchiv gefunden hat. 100%ige Treffer (Full Matches) werden grün, alle Fuzzy Matches in der Farbe Magenta unterlegt. Außerdem werden die Fuzzy Matches noch mit einer Pro−

zentzahl versehen, die den Grad der Übereinstimmung zwischen dem Wort aus dem Quelltext und der Datenbank angibt. Eine Mindestprozentzahl kann vom Benutzer von vorne herein festgelegt werden, so daß beispielsweise nur Treffer mit mindestens 75% Übereinstimmung ausgegeben werden.

Des weiteren gibt es noch blau markierte Treffer, die Assembled Matches und Multiple Exact Matches, also Treffer, für die mehr als nur ein Äquivalent in der Zielsprache vorliegt, die in dunkelgrün hervorgehoben werden. Déjà Vu ist im übrigen das derzeit einzige System, das diese "Multi–Match–Technology" anbietet.

Nach der Vorübersetzung hat der Benutzer verschiedene Möglichkeiten mit dem Übersetzungsprozeß fortzufahren. Er kann beispielsweise alle Wörter, die nicht farbig unterlegt sind, selbst übersetzen und per Hand eingeben oder mit einer der vielen Funktionen fortfahren, die Déjà Vu dem Übersetzer zur Verfügung stellt. Im folgenden Abschnitt wird eine kleine Auswahl an Features vorgestellt.

Tony Peters, Frauke Schmeier und Annika Surrey, 29. März 2000

3.3.3 Features

Während der Arbeit mit Déjà Vu Interactive können eine Reihe von Features im Hintergrund ablaufen, je nachdem ob sie aktiviert wurden oder nicht.

So kann AutoAssemble z.B. nach Übereinstimmungen mit eingegebenen Zahlen und Akronymen suchen, die bei der Vorübersetzung übergangen wurden.

Die Funktion AutoSearch ermöglicht eine automatische Suche nach dem aktuellen Satz im Projektlexikon und den Memory– und Terminologiedatenbanken.

AutoPropagate übergibt die neue Übersetzung eines Satzes, die bei der Vorübersetzung noch nicht zur Verfügung stand, an den Rest der zu übersetzenden Datei oder alle Dateien, die zu dem Projekt gehören.

Für das automatische Verschicken gerade übersetzter Segmente oder Sätze an die Memory – Datenbanken ist die Funktion AutoSend zuständig.

Damit sind die Möglichkeiten von Déjà Vu aber keineswegs erschöpft. Einem Vergleich mit anderen Übersetzungswerkzeugen kann das leistungsfähige Tool ohne weiteres standhalten, vor allem weil es als Komplettsystem zu einem attraktiven Preis angeboten wird, so daß ein Nachkauf von Erweiterungen und Zusatztools entfällt. Eine sehr positive Eigenschaft, die die meisten anderen Hersteller nicht bieten.

Tony Peters, Frauke Schmeier und Annika Surrey, 29. März 2000

3.4 Weitere Produkte

Neben den in 3. Vorhandene Programme genannten Produkten SDLX und OPTIMIZER sind uns noch zwei weitere Produkte aufgefallen:

- Die linguatec Sprachtechnologien GmbH (www.linguatec.de) bietet ein Übersetzungsprogramm Personal Translator 2000 an, das Translation Memory einsetzt.
- Die Firma MorphoLogic (www.morphologic.hu) entwickelt nach eigenen Angaben ein Translation Memory, das nicht einfaches Oberflächen–Stringmatching sondern linguistische Strukturen benutzt. Derzeit bietet sie das Translation Memory Tool MoBiMem an.

Joachim Wagner, 30. März 2000

4. Matchfunktion

Unterabschnitte

- 4.1 Bedeutung für den Übersetzungsvorgang
- 4.2 Matchfunktion von Trados TWB
  - 4.2.1 Änderderung an einem einzelnen Nomen
  - 4.2.2 Änderderung an einem einzelnen Verb
  - 4.2.3 Linguistisches Wissen
  - 4.2.4 Verschiedene Satzlängen
  - 4.2.5 Interpunktion
  - 4.2.6 Vertauschungen von Wörtern innerhalb eines Satzes
  - 4.2.7 Einfügungen und Löschungen von Wörtern
  - 4.2.8 Ergebnis
- 4.3 Fazit

4.1 Bedeutung für den Übersetzungsvorgang

Exact Matches im Translation Memory sind nur selten die Regel, z.B. bei der erneuten Übersetzung eines leicht geänderten Ausgangstextes. Translation Memories bieten daher zusätzlich die Möglichkeit, ungenaue Übereinstimmungen zu verwenden. So werden Sätze gefunden, die nur eine gewisse Ähnlichkeit mit dem zu übersetzenden Satz haben. Diese Ähnlichkeit wird von der Matchfunktion bestimmt. Im Falle von Exact Matches ist sie natürlich 100%, sonst kleiner. Da dem Übersetzer die Treffer sinnvollerweise sortiert nach der Ähnlichkeit präsentiert werden, damit er schnell eine Übersetzung findet, die nur leicht modifiziert werden muß oder sogar direkt übernommen werden kann, ist ein angemessenes Verhalten der Matchfunktion von großer Bedeutung, um produktiv mit dem Translation Memory arbeiten zu können.

In den folgenden Abschnitten werden wir daher die Matchfunktion genauer untersuchen.

Joachim Wagner, 30. März 2000

4.2 Matchfunktion von Trados TWB

Die Matchfunktion ist für das Translation Memory entscheidend. Sie bestimmt, wie ähnlich ein neuer Satz zu einem im TM abgelegten Übersetzungspaar ist. Sätze, die wortwörtlich übereinstimmen, werden natürlich für die Übersetzung sofort vorgeschlagen. Stimmen die Sätze nicht direkt überein, werden sie mit der Matchfunktion bewertet und erhalten so einen Ähnlichkeitswert.

In diesen Abschnitt wollen wir untersuchen, inwiefern diese Werte geeignet sind, um übersetzungs–relevante Vorschläge zu erlauben. Die Vorgehensweise, die wir dazu wählen, ist, speziell gewählte Beispielsätze in das TM zu stellen, um diese dann mit weiteren Sätzen zu vergleichen. Hierzu er–stellen wir jeweils ein Dokument, das alle Beispielsätze und Testsätze enthält. Bei der Übersetzung der Beispielsätze werden diese von Trados TWB automatisch ins TM übertragen. Erreichen wir bei der Übersetzung die Testsätze, wechseln wir jeweils zum Trados TWB Fenster, in dem die Matches mit dem Grad der Übereinstimmung abgelesen werden können.

Zwar wird im Trados TWB Fenster angezeigt, welches Satzpaar aus dem TM verwendet wird. Wir haben aber jeden Versuch unter gleichen Bedingungen durchgeführt, indem wir immer eine neue, leere Datenbank für das TM benutzt haben.

Timo Steffens und Joachim Wagner, 29. März 2000

Unterabschnitte

- 4.2.1 Änderderung an einem einzelnen Nomen
- 4.2.2 Änderderung an einem einzelnen Verb
- 4.2.3 Linguistisches Wissen
- 4.2.4 Verschiedene Satzlängen
- 4.2.5 Interpunktion
- 4.2.6 Vertauschungen von Wörtern innerhalb eines Satzes
- 4.2.7 Einfügungen und Löschungen von Wörtern
- 4.2.8 Ergebnis

4.2.1 Änderderung an einem einzelnen Nomen

Es werden verschiedene Varianten vom Ausgangssatz "Peter goes to Hamburg" miteinander vergli–chen, die sich nur im ersten Wort unterscheiden. Folgende Tabelle zeigt die Ergebnisse.

im TM	im zu übersetzenden Fragment	Übereinstimmung
Peter	Petera, OPeter, Petei, PeaPeter	90
Peter	Hans, OttoPeter, Kohl, Peterabi, Weter	80
Hans	Kohl, Peterabi, Petera, OPeter, Petei, Weter	80

Offensichtlich werden kurze Vorsilben (Pea– und O–) und kleine Änderungen am Wortende (Petera und Petei) bevorzugt bewertet. Größere Veränderungen (z.B. Peterabi und OttoPeter) und ebenso Änderungen am Wortstamm (Peter zu Weter oder Hans zu Kohl) führen zu einer niedrigeren Be–wertung.

Versuchsdurchführung: Timo Steffens und Joachim Wagner, 04. März 2000
Protokollaufbereitung: Joachim Wagner, 29. März 2000

4.2.2 Änderderung an einem einzelnen Verb

Wieder haben wir den Satz "Peter goes to Hamburg" variiert. Dies soll zeigen, ob Verben von Trados TWB anders als die oben untersuchten Nomen behandelt werden.

im TM	im zu übersetzenden Fragment	Übereinstimmung
goesed	goesxy	94
goes	goez, goesu, gogees, goesed, goesxy, goez, gogexyes	90
goesu	goesabi, goez	90
goxyes	gogexyes	90
goes	drives, foes, goesabi, ugoes, goxyes, entgoes, abigoes	80
foes	goesabi	80
entgoes	abigoes	80

Aus diesen Ergebnissen können wir keine zwangsläufigen Unterschiede zu der Behandlung der Nomen ableiten. Einzelne Änderungen am Wortanfang (goes zu foes) und längere Vor– oder Nachsilben (–abi–) führen wie bisher zu einer Abwertung der Ähnlichkeit. Ein interessanter Aspekt ist allerdings, daß das Einfügen von "ge" in der Mitte des Wortes auch zu nur geringen Abwertungen führt, während "xy" an gleicher Stelle zu größere Abwertung führt. Dagegen führt ein Anhängen von "–ed" zu der gleich geringen Abwertung wie ein "–xy". Vermutlich ist in die Match–Funktion die Eigenheit des Deutschen eingeflossen, daß "ge" die einzige Silbe ist, die regelmäßig innerhalb eines Wortes eingefügt wird. In den Versuchen ist Englisch als Ausgangssprache eingestellt.

Es werden also verschiedene Arten von Änderungen differenziert. Dies geschieht anscheint sprachunabhängig und ohne Berücksichtigung der Wortart (Nomen oder Verb). Eine bevorzugte Bewertung kurzer Vorsilben und kleiner Änderungen am Wortende ist klar erkennbar. Berücksichtigt man, daß das TM morphologischen Informationen z.B. zum Wortstamm von Verben zur Verfügung hat, scheint uns dieses Vorgehen, soweit möglich, linguistisch angemessen zu sein

---

Versuchsdurchführung: Timo Steffens und Joachim Wagner, 04. März 2000
Protokollaufbereitung: Joachim Wagner, 29. März 2000

4.2.3 Linguistisches Wissen

Hier möchten wir Änderungen am Subjekt direkt Änderungen am Prädikat gegenüberstellen. Mit linguistischem Wissen zumindest über die Grammatik sollte die Matchfunktion erstere Änderungen geringer gewichten.

TM	zu übersetzendes Fragment	Übereinstimmung
Hans goes to Hamburg.	Peter goes to Hamburg.	80%
Peter moves to Hamburg.	Peter goes to Hamburg.	80%

Wie man sieht, unterscheidet Trados TWB nicht zwischen solchen Änderungen.

Versuchsdurchführung: Timo Steffens und Joachim Wagner, 04. März 2000
Protokollaufbereitung: Joachim Wagner, 29. März 2000

4.2.4 Verschiedene Satzlängen

In diesem Abschnitt werden wir untersuchen, wie weit sich die Länge des Satzes auf die Bewertung einzelner Wortänderungen auswirkt. Wir ersetzen jeweils das Subjekt durch Hans bzw. OPeter.

Ausgangssatz	Wörter	bei großer Änderung	bei kleiner Änderung
Peter goes to Hamburg.	4	80%	90%
Every morning Peter goes to Hamburg	6	86%	93%
In the middle of the day Peter sleeps well.	9	90%	95%
Unfortunately, Peter goes to Hamburg in the evening.	8	90%	95%

Betrachtet man die Differenz der Bewertung zu 100%, so fällt auf, daß sie den gerundeten Werten von 1/5, 1/7 bzw. 1/10 entsprechen. Zählt man die Satzzeichen mit, sind die Nenner genau die Satzlängen. Man kann also vermuten, daß ein Bewertungsabzug von 1/(n+s) bzw. 0.5/(n+s) für Einzelwortänderungen benutzt wird, wobei n die Anzahl der Wörter und s die Anzahl der Interpunktionen im Satz ist.

Versuchsdurchführung: Timo Steffens und Joachim Wagner, 04. März 2000
Protokollaufbereitung: Joachim Wagner, 29. März 2000

4.2.5 Interpunktion

Wenn Satzzeichen als Wort gezählt werden, stellt sich die Frage, ob Unterschiede im Satzzeichen anders behandelt werden als Änderungen an Wörtern. Um diese Frage zu klären, nehmen wir folgende Ersetzen vor:

Experiment	Satzlänge	Änderung	Bewertung
Änderung am Satzzeichen: "Peter goes to Hamburg."	4 Wörter	Punkt zu Fragezeichen	94%
Änderung an einem Wort mit Länge eines Satzzeichens: "P is crazy."	3 Wörter	P zu Q	75%
Wird Komma als Wort gezählt? "However, Peter goes to Hamburg in the evening."	8 Wörter + Komma	However zu Unfortunately	90%

Änderungen an einem Satzzeichen führen also zu einem geringen Bewertungsabzug als Änderung an Wörtern. Das zweite Experiment zeigt, daß dies nicht darin begründet ist, daß Satzzeichen im Gegensatz zu den bisher untersuchten Wörtern nur ein Zeichen lang sind.

Versuchsdurchführung: Timo Steffens und Joachim Wagner, 04. März 2000
Protokollaufbereitung: Joachim Wagner, 29. März 2000

4.2.6 Vertauschungen von Wörtern innerhalb eines Satzes

Ausgangsbasis ist wieder der Satz "Peter goes to Hamburg", den wir in einigen Permutationen vergleichen.

Vergleich von	Permutation*	Verschiebung**	Bewertung
Peter goes to Hamburg. –> Peter to Hamburg goes.	(2,3)(2,4)	[2,4]	94%
Peter goes to Hamburg. –> Peter to goes Hamburg.	(2,3)	[3,2]	94%
Peter to Hamburg goes. –> Peter to goes Hamburg.	(3,4)	[4,3]	94%
Peter goes to Hamburg. –> Peter Hamburg goes to.	(2,3)(3,4)	[4,2]	87%
Peter to Hamburg goes. –> Peter Hamburg goes to.	(2,3)(2,4)	[2,4]	94%
Peter to goes Hamburg. –> Peter Hamburg goes to.	(2,4)	[2,4][3,4]	87%
Peter Hamburg goes to. –> Hamburg goes to Peter.	(2,3)(1,2)(1,4)	[1,4]	94%
Peter goes to Hamburg. –> Hamburg goes to Peter.	(1,4)	[3,1][1,4]	87%
Peter to goes Hamburg. –> Hamburg goes to Peter.	(2,3)(1,4)	[3,1][2,1][1,4]	80%

* Die Permutationen sind als Verkettung von elementaren Vertauschungen von rechts nach links angegeben.
** Zusätzlich haben wir die Permutationen als Folge von elementaren Verschiebungen notiert: [a, b] verschiebt das a–te Wort zur b–ten Position.

Einen klaren Zusammenhang können wir hier nicht erkennen. Zeile 4 und 5 werden ohne erkennbaren Grund verschieden bewertet. Man muß aber berücksichtigen, daß laut Trados TWB Handbuch ein Neuronales Netz die Match–Funktion berechnet. In einem solchen Netz wird die Bewertungsvorschrift nicht in klaren "Wenn–Dann"–Regeln kodiert, wie dies in den meisten Computerprogrammen geschieht, sondern durch Training mit Bewertungsbeispielen im Netz aufgebaut. Vermutlich können die Regeln für Permutationen so nicht exakt nachgebildet werden.

Versuchsdurchführung: Timo Steffens und Joachim Wagner, 04. März 2000
Protokollaufbereitung: Joachim Wagner, 29. März 2000

4.2.7 Einfügungen und Löschungen von Wörtern

Bisher haben wir nur gleich lange Sätze miteinander verglichen. In den Vierwortsatz "Peter goes to Hamburg" fügen wir im folgenden an unterschiedlichen Stellen Wörter ein und/oder nehmen Löschungen vor.

Vergleich von "Peter goes to Hamburg" mit	Operation	Bewertung
Sometimes Peter goes to Hamburg.	vorn 1 Wort eingefügt	85%
In the middle of the day Peter goes to Hamburg.	vorn 6 Wörter eingefügt	0%
Peter goes Hamburg.	drittes Wort gelöscht	75%
Peter goes to Hamburg Hauptbahnhof.	hinten 1 Wort eingefügt	85%

Peter goes directly to Hamburg.	in der Mitte 1 Wort eingefügt	85%
Peter goes very often to Hamburg.*	in der Mitte 2 Wörter eingefügt	73%
Peter often goes to Hamburg Hauptbahnhof.	verteilt 2 Wörter eingefügt	73%
unbekannt / nicht mehr nachvollziehbar	1 gelöscht und 1 eingefügt	72%

* Einige Sätze entsprechen nicht der englischen Grammatik.

Die Position der Einfügung ist anscheinend unerheblich. (Zeile 1, 4 und 5; oder Zeile 6 und 7) Folglich ist hier kein linguistisches Wissen zur Grammatik investiert worden. Weiter fällt auf, daß eine Löschung stärker gewichtet als eine Einfügung ist.

Versuchsdurchführung: Timo Steffens und Joachim Wagner, 04. März 2000
Protokollaufbereitung: Joachim Wagner, 29. März 2000

4.2.8 Ergebnis

Trados TWB setzt linguistischen Wissen nur sehr begrenzt, nämlich auf Wortebene, ein. Dies folgern wir aus der Tatsache, daß zum einen unterschieden wird zwischen Änderungen am Wortstamm und Änderungen an Vor– und Endsilben, zum anderen Einfügen von –ge– toleriert wird. Auf grammatikalischer Basis wird allerdings nicht geprüft. Es findet ein einfacher, wortweiser Vergleich statt.

Timo Steffens und Joachim Wagner, 29. März 2000

4.3 Fazit

Es wird tatsächlich eine einfache Oberflächenanalyse eingesetzt. Die Grammatik der Sprache oder auch nur statistische Information aus Wörterbüchern werden anscheinend nicht eingesetzt. Trotzdem werden diese Produkte produktiv eingesetzt. Die einfache Idee des Translation Memory und die guten Anbindung an vorhandene Textverarbeitungssysteme bzw. deren Dateiformate macht also dieses Übersetzungs–Werkzeug aus. Daher dürfte spannend sein, ob das angekündigte Produkt von MorphoLogic (siehe 3.4 Weitere Produkte) mehr leisten wird.

Joachim Wagner, 30. März 2000

5. Evaluationskriterien

Eine Evaluation von Translation–Memory–Systemen ist sehr aufwendig und ohne eigens dafür erstellte Textkorpora nicht durchführbar. Glücklicherweise gibt es bereits ein Framework der EAGLES–Kommission für diese Aufgabe, so daß wir auf dieses zurückgreifen können. Im folgenden werden wir dieses Framework eingehend beleuchten.

Arno Erpenbeck und Timo Steffens, 29. März 2000

Unterabschnitte

- 5.1 Zweck und Gründung der Kommission
- 5.2 Methodik
- 5.3 Anwendung auf Translation Memories
- 5.4 Fazit

5.1 Zweck und Gründung der Kommission

Der EAGLES Report ist ein Bericht der "Expert Advisory Group on Language Engineering Systems". Diese besteht aus einer Anzahl von renommierten Linguisten und Computerfachleuten und hat sich gebildet, um Evaluierungskriterien für Software im Bereich Language Engineering zu finden. Ziel der Kommission war es auch, Ordnung in den mittlerweile unübersichtlichen Markt von Software im Übersetzungsbereich zu schaffen.

Die erste Phase von EAGLES begann 1993. Während dieser Phase beschäftigte sich die Kommission in erster Linie mit der Definition von Evaluationskriterien und –methoden. Dabei baute die Gruppe auf bestehenden Standards auf, insbesondere der ISO/IEC 9126:1991. Dieser Standard der "International Organization for Standardization" setzt Maßstäbe für Software–Produkt–Evaluation, Qualitätsmerkmale und ihre Umsetzung. Die zweite Phase von 1997 bis 1999 erweiterte dieses Modell und führte verschiedene Fallstudien durch.

Um valide und reproduzierbare Ergebnisse erhalten, ist ein allgemeines Framework nötig. Zum einen sollen vollkommen verschiedene Ansätze miteinander in Beziehung gesetzt werden, zum anderen aber auch Kriterien für einzelne Systeme erstellt werden. Die Evaluation erfolgt in beiden Fällen nach demselben Schema. Mit Hilfe dieses Frameworks wurden Fallstudien in den folgenden Bereichen untersucht:

- Grammar Checker
- Spelling Checker
- Translation Memories
- Auskunftssysteme

Die Evaluation Working Group hat 1996 einen Zwischenbericht veröffentlicht (http://issco–www.unige.ch/projects/ewg96/ewg96.html) und mittlerweile auch einen Abschlußbericht vorgelegt (http://issco–www.unige.ch/projects/eagles/ewg99/index.html).

Arno Erpenbeck und Timo Steffens, 29. März 2000

5.2 Methodik

Aufgrund der engen Zusammenarbeit der Evaluation Working Group mit der ISO entstand eine relativ detaillierte Beschreibung des Evaluationsvorgangs. Die Evaluation soll in drei Schritten vorgenommen werden und kann sowohl für komplexe System als auch einzelne Komponenten durchgeführt werden.

- Definition der Qualitätsanforderungen:
  Da keine absoluten Maßstäbe herangezogen werden sollen, muß jeder Kunde selbst für sich entscheiden, welche Anforderungen er an das System stellt.
- Vorbereitung der Evaluation:

Typischerweise beinhaltet die Vorbereitung drei Teile:
  - Auswahl von Qualitätsmaßstäben (Metriken), d.h. wie werden Eigenschaften quantitativ bewertet
  - Festlegen, wie sich die Bewertungsklassen auf den gesamten Bereich der Metrik verteilen
  - Methoden finden, wie Einzelergebnisse zusammenzufassen sind
- Durchführung der Evaluation

Dabei sollen die Systeme nach folgenden Kriterien begutachtet werden:

- Funktionalität (Functionality):
  Fähigkeit der Software, Methoden und Techniken zu gegebenen Anforderungen des Benutzers bereitzustellen; Fähigkeit der Software, korrekte bzw. genaue Ergebnisse zu liefern
- Zuverlässigkeit (Reliability):
  Ist das System fehlerfrei? Wie werden Störungen behandelt?
- Benutzbarkeit (Usability):
  Ist das System einfach/verständlich zu bedienen? Wie groß ist die Einarbeitungszeit bzw. Umgewöhnungszeit bei Systemwechsel?
- Effizienz (Efficiency):
  Ausnutzung von Ressourcen; Reaktionszeit des Systems
- Wartbarkeit (Maintainability):
  Welche Handlungen sind zur Pflege des Systems erforderlich? Welchen Support bietet der Hersteller?
- Portierbarkeit (Portability):
  Welche Möglichkeiten zur Zusammenarbeit mit anderen Systemen bieten sich? Welche Anforderungen an vorhandene Hard– bzw. Software werden gestellt?

Hauptaugenmerk liegt vor allem auf der Frage, ob das, was der Kunde braucht und wünscht, auch gewährleistet wird. EAGLES nimmt also Evaluierungen aus Sicht des Kunden vor und hat daher einen eher pragmatischen als wissenschaftlich–theoretischen Ansatz.

Es wurden Richtlinien erstellt, auf die verschiedenen Systemattribute aufgeteilt und Methoden definiert, um diese Richtlinien zu testen. Durch diese Operationalisierungen werden diesen Aspekten Zahlen aufgrund einer Metrik zugeordnet.

Man spricht von den "7 Schritten einer Evaluation":

1. Was genau und wieso wird evaluiert?
2. Was wird mit dem System gemacht und wer benutzt es?
3. Welche Features des Systems sind die wichtigsten und müssen evaluiert werden?
4. Davon ausgehend, detaillierte Anforderungen an die Komponenten stellen.
5. Metrik für die Komponenten finden
6. Design der Evaluation: Testmaterial festlegen
7. Durchführung

Arno Erpenbeck und Timo Steffens, 29. März 2000

5.3 Anwendung auf Translation Memories

Wir wollen uns hier auf die Gruppe der Translation Memories konzentrieren. Für diesen Begriff benutzt die Kommission folgende Definition:

"A translation memory is a multilingual text archive containing (segmented, aligned, parsed and classified) multilingual texts, allowing storage and retrieval of aligned multilingual text segments against various search conditions."

Dies ist eine sehr allgemeine Definition, die daher alle gängigen Produkte, die sich Translation Memory nennen, umfaßt.

Im Gegensatz zu anderen Systemen werden mit Translation–Memory–Programmen keine Daten bzw. Datenbanken mitgeliefert, sondern es handelt sich bei dem Produkt hauptsächlich um eine Arbeitsumgebung. Daher muß sich die Evaluation auf die Funktionen und Features konzentrieren.

Folgende Fragen sollen untersucht werden:

Zunächst einmal muß man sich darüber klar werden, ob es überhaupt zwingend notwendig ist, ein TM zu benutzen, oder eine andere Methode erfolgversprechender ist. Dazu muß man sich überlegen, ob der Sourcetext viele Wiederholungen aufweist und reich an Terminologie ist. Ist Konsistenz innerhalb des Dokuments nötig?

Hat man entschieden, daß man tatsächlich ein TM benutzen möchte, stellt sich die Frage, welches TM–System zu benutzen ist. Ein Kriterium ist die benutzte Sprache. Obwohl TMs KEIN linguistisches Wissen benutzen, müssen beispielsweise notwendige Zeichensätze unterstützt werden. Dafür gibt es noch keinen Standard. Weiterhin muß man untersuchen, welche Formate das jeweilige System unterstützt.

Falls Texte häufig aktualisiert werden müssen, hat dies Einfluß auf die Entscheidung, ob ein Datenbank– oder Datei–Speicherung–orientiertes System vorteilhafter ist.

Auch der Übersetzer (Endnutzer) ist in die Planung einzubeziehen. Welche Software und Hardware ist nötig? Womit hat der Benutzer Erfahrung?

Handelt es sich um einen einzelnen Übersetzer oder um ein Team? Falls letzteres der Fall ist, ist von Bedeutung, ob alle lokal in der Nähe und greifbar oder verstreut an verschiedenen Orten sind. Müssen Arbeiten übers Netzwerk möglich sein? Arbeiten Leute u.U. gleichzeitig am selben Dokument? Kann man in dem System individuelle Lese–/Schreibrechte vergeben? Ist Exportieren von TMs möglich, um sie beispielsweise auszugliedern? Projektspezifische TMs möglich?

Die letzte Frage muß berücksichtigen, daß die Anschaffung eines TMs den Umstand der Teamarbeit/Nähe ändern kann.

Wie man sieht, kann man also keine generellen Evaluationen durchführen, sondern es ist stets vom Kunden und seinen Bedürfnissen abhängig, wie geeignet ein Produkt ist.

5.4 Fazit

Es ist der EAGLES–Kommission durchaus gelungen, Richtlinien bzw. ein Fundament für Evaluationen im Bereich Sprachsoftware zu erstellen. Obwohl das Gebiet sehr diffus und weit gefächert ist, haben sie Merkmale ausmachen können, die in allen Systemen auffindbar sind, und haben zusätzlich für die einzelnen Produktkategorien spezielle Merkmale herausgearbeitet. Nun ist es am Kunden, daraus individuelle Evaluationen durchzuführen und darauf aufbauend Entscheidungen zu treffen.

Aufgrund seiner Detailliertheit und seines formalen Aufbaus sind die auf ihm aufbauenden Evaluationen eindeutig und daher untereinander replizierbar. Genau dies war das Ziel und der Anspruch der Kommission.

Allerdings braucht man ein gewisses Vorwissen und Erfahrung, um die teilweise sehr speziellen Features anwenden zu können. Daher ist er für Privatpersonen und kleinere Betriebe schwer handzuhaben. Diese müssen sich dann auf fremde Evaluationen verlassen, die unter Umständen nicht direkt die tatsächlichen Anforderungen des eigenen Betriebs berücksichtigen.

Arno Erpenbeck und Timo Steffens, 29. März 2000

6. Abschließendes Fazit

Translation Memory Werkzeuge verbreiten sich stark. Es hat sich ein breiter Markt an Produkten gebildet, der dem Kunden erlaubt, die Software nach seinen speziellen Wüschen zu wählen.

Die eingesetzte TM–Technik der Programme unterscheidet sich nicht wesentlich und ist von der Grundidee her einfach. Die Leistungen der Produkte liegen in der Integration in die Arbeitsumgebung des Übersetzers, die Verknüpfung mit Datenbanken und anderen Übersetzungswerkzeugen und der Projektverwaltung. Hier finden sich auch die Unterschiede der Programme, die teilweise völlig verschiedene Ansätze verfolgen.

Zur Auswahl des richtigen Produkt bietet der EAGLES Report hilfreiche Ansätze zur individuellen Evaluation des Marktangebots. Leider ist der Aufwand zur Durchführung einer solchen Evaluation recht groß, so daß sich gerade kleine Betriebe auf Werbeversprechen verlassen müssen.

Joachim Wagner, 30. März 2000

Literatur

1 Trados GmbH: *Handbuch zu Trados Translator´s Workbench 2*

2 Lynn E. Webb (1999): *Advantages and Disadvantages of Translation Memory: A Cost/Benefit Analysis*, master thesis, San Francisco State University

3 MorphoLogic:, http://www.morphologic.hu/e_phare.htm

4 *Translation Memory Technologie und Maschinelle Übersetzung im Vergleich*. www.heeg.de/uta/AK–Protokoll–TM–1.html

5 Bachmann, R./ Heizmann, S./ Schmitz, K.–D.: *EDV–Unterstützung für Übersetzer.*

Einsatzmöglichkeiten und Grenzen. In: Technische Dokumentation optimieren. Oktober 1994.

6 *Im Test: TRANSIT 2.6 der Fa. STAR GmbH, Böblingen "Das Übersetzungssystem für Fachübersetzer* In: MDÜ. Mitteilungsblatt für Dolmetscher und Übersetzer. 2/98

7 Script Uni Hildesheim

8 *Translation Memory*. Bulletin of the Institute of Translation and Interpreting. 5/99

9 *Translation Memories: Übersetzungshelfer mit Gedächtnis.* www.doculine.com/zeitsch/9910/transmem.htm

Da bis zum Redaktionsschluß nur zwei Autoren Literaturlisten eingereicht haben, sind die Angaben wahrscheinlich unvollständig.

Joachim Wagner, 10. Mai 2000

Über dieses Dokument ...

Um die von den einzelnen Autoren geschriebenen Abschnitte zusammenzufügen, wurden folgende Schritte durchgeführt:

1. Die einzelnen Beiträge werden ins HTML–Format konvertiert oder bereits in diesem Format geschrieben.
2. Alle unerwünschten Formatierungen werden entfernt.
3. Ein LaTeX–Dokument wird erstellt, das alle Überschriften und Platzhalter für die Inhalte enthält.
4. Der LaTeX2HTML Konverter von Nikos Drakos, University of Leeds, und Ross Moore, Macquarie University, Sydney, erzeugt zu jedem Abschnitt eine HTML–Datei mit fertigen Navigationselementen und Überschriften.
5. Die Beiträge werden entsprechend zerschnitten und sortiert. Die so entstehenden Abschnitte werden genauso numeriert, wie die von LaTeX2HTML erstellten Dateien.
6. Ein Shell–Skript ersetzt die Platzhalter in den HTML–Dateien durch die Inhalte der entsprechenden Abschnittsdateien.

Hinweis: Namentlich gekennzeichnete Beiträge geben nicht immer die Meinung der gesamten Arbeitsgruppe wieder. Dies betrifft insbesondere die Produktwertungen, die eingentlich nicht Gegenstand unserer Ausarbeitung sein sollten.

Joachim Wagner, 10. Mai 2000