

# Fallbeispiel: Übersetzung eines Datenbankhandbuchs mit MTM

---

## Situation

**Zu übersetzende Dokumente** – Wir erhielten von unserem Kunden XML-Dateien mit über 300.000 Wörtern, die von Englisch nach Deutsch übersetzt werden sollten. Es handelte sich dabei um ein Datenbankhandbuch.

**Arbeitsumgebung** – Unser Kunde nutzt den Idiom Worldserver als Übersetzungsumgebung (Translation Memory und Workflow). Auf Grundlage unserer Erfahrung, der Sprachkombination und der Textart wählten wir Lucy LT, eine regelbasierte MT-Software.

Da der Kunde den Idiom Worldserver betreibt, hatten wir keinen Zugang zur API und konnten damit keine Worldserver-Funktionen zum Aufruf der MT-Software nutzen.

---

## Vorüberlegungen

**Einfache Anwendung** – Das Hinzufügen der MT-Software zur Übersetzungsumgebung soll das Leben der Übersetzer vereinfachen, nicht erschweren. Den Übersetzern sollten nur MT-erzeugte Vorschläge gemacht werden, wenn es keine guten TM-Matches gibt. Dabei sind 100%-Matches unverändert zu belassen.

**Memory Pollution** – Das TM wächst mit jeder Übersetzung und wird damit langsamer. Daher sollte das TM nicht durch nutzlose (niemals verwendete) Matches verunreinigt werden.

---

## Komplikationen

**Funktionswörter** – Der Text enthielt viele Funktionswörter, z. B. Befehle, die nicht verändert oder übersetzt werden dürfen. Da diese häufig mitten im Satz auftauchten, konnten wir sie nicht einfach vor der MT-Software „verbergen“, weil das die Satzanalyse verfälschen würde, sondern mussten sie als „Konstante“ markieren, die vom MT-System wie Eigennamen behandelt werden.

**Programmcode und Indexeinträge** – Neben den Funktionswörtern enthielt der Text zahlreiche Programmbeispiele, die von der MT-Bearbeitung ausgeschlossen werden mussten, und Indexeinträge, die in einzelne Segmente zu separieren waren, damit die MT-Software sie richtig übersetzen konnte.

---

## Lösung

**Standardansatz: Vorübersetzung und Import in TM** – Beim direkten (und üblichen) Ansatz werden alle Dateien zuerst durch eine Stapelübersetzung der MT Software geschickt. Dann werden die Quelldateien und die übersetzten Dateien segmentweise zugeordnet (Alignment) und in das Translation Memory importiert. Neben den Alignmentproblemen führt das zu zwei wesentlichen Nachteilen:

- ⬇ Das Memory wird durch unerwünschte „Matches“, die immer wieder hochkommen und die Übersetzer verwirren, unnötig aufgebläht.
- ⬇ Übersetzer erhalten auch dann „MT-Matches“, wenn ein 100%-Match vorhanden ist.

### **Statistische oder regelbasierte MT?**

Unabhängige Untersuchungen zeigen, dass die Ausgabequalität der beiden Systeme sich nicht wesentlich unterscheidet und der Nacharbeitung bedarf. Dafür wird eine Kombination mit einem TM genutzt (MTM). Es gibt keine allgemeine Empfehlung, welche Art von MT genutzt werden sollte, aber Auswahlkriterien:

- 1) Verfügen Sie über sehr große Corpora bilingualer Texte (TM) aus einem Fachgebiet?
- 2) Welche(s) Sprachpaar(e) benötigen Sie?
- 3) Welche Textarten übersetzen Sie?
- 4) Wer kodiert die Terminologie?

Im vorliegenden Fall führten das Sprachpaar Englisch-Deutsch und die viele Funktionswörter enthaltenden Texte zur Entscheidung, ein regelbasiertes System zu nutzen.

### **Stapelmodus oder auf Abruf?**

Maschinelle Übersetzung kann auf einmal im Voraus ausgeführt und dann zum Posteditieren bereitgestellt werden. Sie kann aber auch segmentweise vom TM angestoßen werden.



**Automatisierung:** Die hier beschriebenen Automatisierungsschritte (und viele andere) kann man mit dem Pattern Matcher von Lucy Software oder mit einem Editor erstellen, der Suchvorgänge mit Hilfe von regulären Ausdrücken beherrscht (z. B. MS Word, Notepad++). Die Erstellung ist für jeden Texttyp nur einmal nötig und erfordert überraschend wenig Aufwand.

**Terminologiekodierung** war früher der entscheidende Kostenfaktor. Lucy Software reduziert diesen erheblich durch das sogenannte Defaulting: die Software bestimmt sehr erfolgreich Standardwerte der grammatikalischen Informationen. Für dieses Projekt mussten wir über 4000 Einträge kodieren, was ohne das Defaulting nicht möglich gewesen wäre.

**Einsparungen:** Die erzielten Einsparungen (E) werden wie folgt berechnet:  
E = ZE - KD - AK - AM  
ZE = Zeiteinsparung durch Posteditieren statt Übersetzen multipliziert mit dem Stundensatz  
KD = Kosten der Terminologiekodierung  
AK = Kosten der Automatisierung  
AM = Amortisation der MT-Software

Eule Lokalisierung GmbH  
Holstenstrasse 104  
24103 Kiel  
+49-431-99042-0  
www.eule2005.de  
info@eule2005.de



## Unser Ansatz: Schutz vorhandener Übersetzungen –

Bevor wir die Dateien an die MT-Software übergaben, bereiteten wir sie sorgfältig vor, um vorhandene Übersetzungen zu schützen und das Übersetzen von Programmcode und Funktionswörtern zu verhindern.

### Einzelschritte:

- 1 Alle Quelldateien mit vorhandenen Übersetzungen als bilinguale Dateien, d.h. Dateien mit englisch/deutschen Satzpaaren, exportieren.  
Bemerkung: Die „deutschen“ Sätze waren nur deutsch, wenn es bereits eine Übersetzung gab, sonst waren sie eine Kopie des englischen Originals.
- 2 Alle vorhandenen Übersetzungen schützen.
- 3 Funktionswörter als nicht zu übersetzende „Konstante“ markieren.
- 4 Programmcode vor dem MT-System verbergen.
- 5 Indexeinträge in Einzelsegmente auftrennen.

All diese Schritte waren automatisiert und benötigten wenig Aufwand.

Danach ließen wir die MT-Software die vorbereiteten Dateien übersetzen. Dabei wurden nur die „deutschen“ Sätze übersetzt, die gar nicht deutsch waren (siehe obige Bemerkung).

Nach der maschinellen Übersetzung, die nur wenige Sekunden dauerte, entfernten wir die Schutzmarkierungen (wiederum automatisch) und importierten die vorübersetzten Dateien in Idiom Worldserver.

---

## Was haben wir erreicht?

- ✓ Vor dem eigentlichen Übersetzungs-/Posteditierschritt fügten wir nichts in das TM ein. Damit enthält das Memory nur überprüfte Segmentübersetzungen. Es gab KEINE Memory Pollution (keinen Unsinn in der Datenbank).
- ✓ Wir erhielten keine dummen „Übersetzungen“ der Funktionswörter. Ein Befehl wie `SELECT * FROM TABLE` blieb unverändert und wurde nicht als `WÄHLE * VOM TISCH` übersetzt, was für die Datenbank unverständlich gewesen wäre.
- ✓ Und wir hatten die perfekte Umgebung für Übersetzer: 100%-Matches waren immer noch 100%-Matches (blau markiert), Fuzzy Matches aus dem TM wurden wie üblich gezeigt (braun) und MT-Übersetzungen waren bereits als „manuelle Übersetzungen“ (grün markiert) eingesetzt. Damit konnten unsere Übersetzer auf einen Blick entscheiden, ob Sie die Übersetzung so stehen ließen oder den Fuzzy Match bzw. die MT-Ausgabe bearbeiteten.
- ✓ **Einsparungen:** Wir konnten unserem Kunden einen 20 %igen Rabatt auf alle MT-übersetzten Segment gewähren. Mit anderen Worten, der Kunde musste niemals den vollen Preis für „neue Worte“ bezahlen.