



FOLT – Forum Open Language Tools • Filderbahnstr. 43 A • D-70567 Stuttgart-Möhringen • Germany
Telefon +49 (0) 711 166 46 0 • Fax +49 (0) 711 166 46 50 • E-Mail: folt@beo-doc.de • www.folt.de

Exposé

Translation Memory Open Source System

TMOSS

Version 1.0

Stuttgart, 28. Oktober 2007



Mitwirkende

Ulrike Baral, beo Gesellschaft für Sprachen und Technologie mbH
 Christine Bruckner, Referat SMD 1, Bundessprachenamt
 Lutz Fischer, Universität Mainz
 Sascha Hofmann, Universität Mainz
 Horst Liebscher, euroscript Deutschland GmbH
 Marc Mittag, transline GmbH
 Andrea Modersohn, oneword GmbH
 Michael Neuhäuser, euroscript Deutschland GmbH
 Michael Schneider, beo Gesellschaft für Sprachen und Technologie mbH
 Christian Taube, Matrix AG
 Thomas Wedde, euroscript Deutschland GmbH

	Organisation	Adresse
Urheber Rechtsinhaber	FOLT Forum Open Language Tools	Filderbahnstr. 43 A 70567 Stuttgart folt@beo-doc.de www.folt.de Telefon +49 (0) 711 166 46 0 Fax +49 (0) 711 166 46 50
Herausgeber	Linux Solutions Group (LiSoG) e. V. Geschäftsstelle Stuttgart	Breitscheidstr. 4 70174 Stuttgart info@lisog.org www.lisog.org

Dieser Inhalt ist unter einem Creative Commons Namensnennung - Keine Bearbeitung 2.0 Germany Lizenzvertrag lizenziert. Um die Lizenz anzusehen, gehen Sie bitte zu

<http://creativecommons.org/licenses/by-nc-nd/2.0/de/>

oder schicken Sie einen Brief an

Creative Commons, 559 Nathan Abbott Way, Stanford, California 94305, USA.

Creative Commons - Commons Deeds

Namensnennung - Keine kommerzielle Nutzung - Keine Bearbeitung 2.0 Deutschland

Sie dürfen den Inhalt vervielfältigen, verbreiten und öffentlich aufführen.

Zu den folgenden Bedingungen:

Namensnennung: Sie müssen den Namen des Autors/Rechtsinhabers nennen. Keine kommerzielle Nutzung: Dieses Werk darf nicht für kommerzielle Zwecke verwendet werden. Keine Bearbeitung: Der Inhalt darf nicht bearbeitet oder in anderer Weise verändert werden. Im Falle einer Verbreitung müssen Sie anderen die Lizenzbedingungen, unter die dieser Inhalt fällt, mitteilen. Jede dieser Bedingungen kann nach schriftlicher Einwilligung des Rechtsinhabers aufgehoben werden.

Die gesetzlichen Schranken des Urheberrechts bleiben hiervon unberührt. Das Commons Deed ist eine Zusammenfassung des Lizenzvertrags in allgemeinverständlicher Sprache.



Vorwort

Auf dem Markt für Translation Memory Systeme (TMS) gibt es wenige kommerzielle Systeme. Hinzu kommt, dass sie nicht kompatibel zueinander sind und keinen unproblematischen Datenaustausch ermöglichen. Dieser Umstand führt dazu, dass im Erstellungsprozess multilingualer Dokumente viele Einzellösungen von DMS über Übersetzungswerkzeuge bis hin zu DTP-Systemen entstanden sind mit der Folge, dass Fachübersetzer nach dem Einsatz des TMS und nicht mehr nach Fachgebiet ausgesucht werden.

Hinzu kommt, dass die Kosten für die bestehenden kommerziellen und vollständig proprietären Systeme wegen ihrer beherrschenden Stellung im Markt unverhältnismäßig hoch sind.

Auch im Bereich der universitären Ausbildung von Übersetzern wird die Lehre immer wieder mit diesen o. g. Sachzwängen konfrontiert. Insbesondere im Rahmen der Umstellung der Studiengänge auf BA und MA und einer damit einhergehenden Verkürzung der Studienzeit, ist eine Integration aller Systeme in die Curricula nicht mehr möglich.

Open Source ist zwar in aller Munde, aber bisher sind keine Bestrebungen zu erkennen, ein Open Source TMS zu entwickeln. Dies möchte das „Forum Open Language Tools“, kurz FOLT genannt ändern und legt mit diesem Dokument einen ersten Entwurf für die Entwicklung eines „Translation Memory Open Source System“ (TMOSS) vor.

FOLT ist ein Zusammenschluss und Arbeitskreis von Unternehmen, Organisationen und Hochschulen aus dem Bereich Übersetzung und der Dokumentation. Wesentliche Ziele von FOLT sind die Unterstützung von standardisierten Austauschformaten, nicht proprietäre Software und die Erprobung neuer Übersetzungstechnologien und -methoden. FOLT ist ein offenes Forum und wird nicht als Verein oder Gesellschaft geführt. Arbeitstreffen finden in regelmäßigen Abständen von 4-6 Wochen an unterschiedlichen Standorten bei den Mitgliedern statt. Eingeladen sind alle Interessierten, die über das Internet-Forum hinaus einer offenen Diskussion zu diesem Thema teilnehmen möchten. Anmeldungen sind über den Internetauftritt des Forums www.folt.de möglich.

FOLT ist ein Expertenverbund und versteht sich als Interessenvertretung unabhängiger Übersetzungsdienstleister, Übersetzer und Redakteure sowie der in der Übersetzerausbildung tätigen Hochschulen – und nicht zuletzt eines freien globalen Übersetzungsmarktes.

Linux Solutions Group e. V. (LiSoG) ist eine Kooperations-Plattform mit dem Ziel, den Einsatz von linuxbasierten Lösungen in Unternehmen zu fördern und die Marktakzeptanz zu erhöhen. Neben Großen der IT-Branche wird die Initiative auch von mittelständischen Unternehmen gefördert, die sich mit Linux und Open Source beschäftigen.

LiSoG fokussiert auf „Collaborative Innovation“, d. h. sie initiiert Projekte zu aktuellen marktrelevanten Open Source Themen, in denen Mitglieder und Interessierte gemeinsam Lösungen erarbeiten und treibt diese voran. Ihr Wirkungsbereich erstreckt sich auf den deutschsprachigen Raum. Zusätzlich bietet die LiSoG eine Netzwerk-Plattform für Projekte, Partner und Interessierte.

LiSoG unterstützt FOLT in dem Projekt „Translation Memory Open Source System“.

Versionsübersicht

Version	Datum	Autor	Status*	Kommentar
0.2	3. 9. 2007	M. Schneider, T. Wedde	erledigt	Ersterstellung
0.3	6. 9. 2007	M. Neuhäuser	erledigt	Formale und strukturelle Änderungen
0.4	5. 10. 2007	M. Schneider	offen	Übersetzung
1.0	22. 10. 2007	M. Neuhäuser	erledigt	Festlegung auf Version 1.0
1.0	28.10.2007	S. Hofmann	erledigt	Formale und inhaltliche Richtigstellungen

* erledigt/offen

Inhaltsverzeichnis

1	Über FOLT	7
2	Zweck dieses Dokumentes	7
3	Ausgangslage und Problemstellung	7
3.1	Einführung in die Thematik	8
3.2	Grundlegendes Prinzip	8
3.3	Segment	10
3.4	Dokument	10
3.5	Übersetzungsauftrag (Job)	10
3.6	Translation Memory	10
3.7	Match	11
3.8	Beschreibung eines Standard Übersetzungsvorganges	11
4	Modularer Aufbau von TMOSS	13
4.1	Translation Memory	13
4.2	Segmentierer & Segment-Matcher	13
4.3	Benutzeroberfläche	15
4.4	Filter	15
4.5	Schnittstellen	15
5	Entwicklungsablauf	15
5.1	Phase 1 – Prototyp mit Grundfunktionen	15
5.2	Phase 2 – Entwicklung komfortabler Handlingfunktionen	16
5.3	Phase 3 – Entwicklung einzelner Zusatzfunktionen	16
6	Zielbestimmung	16
6.1	Musskriterien	16
6.2	Wunschkriterien	16
6.3	Abgrenzungskriterien	16
7	Produkteinsatz	16
7.1	Einsatzbereich und grundlegende Architektur	17
7.2	Zielgruppen	17
7.3	Software	17
7.4	Hardware	17
7.5	Technische Schnittstellen	17
7.6	Produktschnittstellen	18
8	Produktfunktionen	18
8.1	Vor- und Nachbereitung	18



8.2	TM, Segmentierung, Matching	18
9	Produktdaten	19
10	Benutzeroberfläche	19
10.1	Übersetzer	19
10.2	Verwaltung	20
11	Produktleistungen	20
12	Nichtfunktionale Anforderungen	20
12.1	Schnittstellen	20
12.2	Plattformen	20
12.3	Datenbank	21
12.4	Skalierbarkeit	21
13	Thema Lizenz	21
14	Glossar	22



1 Über FOLT

Das Forum Open Language Tools (FOLT) ist ein Zusammenschluss und Arbeitskreis von Unternehmen und Organisationen aus den Bereichen Übersetzung und Dokumentation.

FOLT befasst sich mit dem gesamten Arbeitsablauf in der Fremdsprachendokumentation. Von der Erstellung des Ausgangstextes bis zur Produktion in den Fremdsprachen untersuchen wir unsere Prozesse auf Brüche und fehlende Standards.

Die wesentlichen Ziele von FOLT sind

- Erfahrungsaustausch in Bezug auf Prozesse bei branchenüblicher Software
- Erfahrungsaustausch über den Einsatz von Open Source Software
- Standardisierung von Austauschformaten
- Erprobung neuer und Verbesserung bestehender Open Source Technologien für den Übersetzungsmarkt
- Öffentliche Unterstützung nicht-proprietärer Software und Softwareentwicklung
- Unterstützung der Hochschulen bei der Integration eines Open Source Produkts für die Ausbildung von Übersetzern
- Veröffentlichung der Ziele und Ergebnisse

FOLT hat sich zu einem Sprachrohr entwickelt, das sich für die Belange aller Beteiligten am Übersetzungsprozess einsetzt, ganz gleich ob auf der Seite von Auftraggebern oder Dienstleistern.

2 Zweck dieses Dokumentes

Dieses Dokument enthält die wesentlichsten Anforderungen aus Sicht des Auftraggebers für ein Open Source Translation Memory System. Es dient als Ausschreibungsunterlage für Entwickler und wird vertraglicher Bestandteil eines zu erteilenden Auftrages. Es dient ebenfalls als Grundlage für ein vom Auftragnehmer zu erstellendes Pflichtenheft.

3 Ausgangslage und Problemstellung

Derzeit existieren kommerzielle TMS, die nicht kompatibel zueinander sind und keinen unproblematischen Datenaustausch ermöglichen. Dies hat insgesamt zu einem uneinheitlichen Zustand im Übersetzungsprozess mit TMS geführt. Übersetzer müssen sich

entweder auf ein TMS spezialisieren oder die verschiedenen Systeme beherrschen können. Da viele Auftraggeber irgendeines der kommerziellen Systeme einsetzen, erfolgt die Auswahl von Fachübersetzern nach ihrer Qualifikation auf ein bestimmtes TMS. Die Übersetzer sind deshalb zunehmend gezwungen, ihre Kernkompetenzen mit zusätzlichen Aufgaben zu belasten. Programmhandling, Schulung und Releasewechsel der kommerziellen Programme fordern immer mehr Zeitaufwand und Kosten, für die es keinen finanziellen Ausgleich gibt. Hinzu kommt, dass Anschaffungs- und Folgekosten für die gängigen kommerziellen Systeme unverhältnismäßig hoch sind.

Dementsprechend schwer fällt es auch den Ausbildungsinstituten für angehende Übersetzer oder technische Redakteure bereits im Studium an entsprechender Software auszubilden.

Der Wechsel von einem System und zu einem anderen wird durch Inkompatibilität der Systeme erschwert. Der Aufwand für Konvertierungen bei einem Systemwechsel ist hoch.

Ein Open Source TMS soll die Prozesse künftig vereinheitlichen und die bestehenden offenen Standards konsequenter umsetzen.

3.1 Einführung in die Thematik

3.2 Grundlegendes Prinzip

Translation Memory Systeme (TMS) oder Übersetzungsspeicher respektive Satzdatenbanken speichern die Übersetzungen von in der Regel menschlichen Übersetzern und bieten dem Übersetzer bereits vorhandene Übersetzungen später zur Wiederverwendung an. Hierbei können nicht nur identische, sondern auch ähnliche Ausgangstextsegmente erkannt werden. TMS dienen vor allem zur Sicherung der terminologischen, inhaltlichen, stilistischen und formalen Konsistenz von Übersetzungen. Ein weiterer Aspekt ist die Arbeitserleichterung und -beschleunigung und die damit verbundene Kostenreduzierung.

Besonders sinnvoll ist der Einsatz von TMS zur Übersetzung von Texten, in denen immer wiederkehrende identische oder ähnliche Formulierungen vorkommen, wie etwa Bedienungsanleitungen, aber auch Rechenschaftsberichte, Bilanzen, Werbetexte, und Kataloge. Grundsätzlich gilt: Je höher der Wiederholungsanteil im Textkörper ist, desto eher greift das Prinzip der TMS. Die Verwendung von Textbausteinen im redaktionellen Prozess bringt die Mächtigkeit von TMS besonders zur Geltung.



In der Praxis erfolgt die interaktive Arbeit mit dem TMS dergestalt, dass ein Übersetzer ein Segment zur Übersetzung auswählt. Das System sucht dann im Speicher nach gleichen oder ähnlichen Segmenten und bietet die vorhandenen Übersetzungen an. Diese Übersetzungsvorschläge können vom Übersetzer übernommen oder angepasst werden. Werden keine entsprechenden Segmente gefunden, gibt der Übersetzer eine eigene Übersetzung ein, die dann zusammen mit dem Ausgangssegment gespeichert wird und ab sofort beim Auftreten identischer oder ähnlicher Segmente zur Verfügung steht. Zusätzlich erhält der Übersetzer je nach System noch eine Reihe weiterer Informationen, die ihm die Übersetzung erleichtern sollen. Hierzu gehören:

- Benutzer, der die angebotene Übersetzung angelegt/geändert hat
- Datum der Anlage/Erstellung der Übersetzung
- Häufigkeit der Verwendung der Übersetzung
- Kontext der Übersetzung
- Weitere klassifizierende Informationen
- Angaben zur fach- oder nutzerspezifischen Terminologie

Neben diesem interaktiven Verfahren können die meisten TMS vor der eigentlichen Übersetzung eine vollautomatische Übersetzung („Vorübersetzung“) durchführen. Dabei vergleicht das System die Segmente in dem zu übersetzenden Dokument mit denen im Übersetzungsspeicher. Bei vollständiger Übereinstimmung wird das Segment durch die gespeicherte Übersetzung ersetzt. Der Benutzer muss sich anschließend mit den Sätzen/Segmenten beschäftigen, die nicht im Übersetzungsspeicher gefunden wurden, und die vorübersetzten Segmente auf Richtigkeit im jeweiligen Kontext untersuchen.

Diese Methodik ist um ein Vielfaches schneller als das herkömmliche Übersetzen. Ein TMS reproduziert zielsprachliche Inhalte und kann, entsprechende Schnittstellen und Prozessdefinitionen vorausgesetzt, die redaktionelle Erstellung des Ausgangssprachlichen Textes unterstützen. Diese Variante nennt sich Authoring Memory System (AMS).

TMS werden zur kommerziellen Verwendung seit über 20 Jahren entwickelt. Der Verbreitungsgrad ist hoch (> 70%). Der Markt wird von sehr wenigen kommerziellen Anbietern beherrscht.

3.3 Segment

Die einzelnen Einheiten der Datenbank werden Segmente genannt. Sie umfassen in der Regel je einen Satz oder Absatz. Der Zugriff auf und die Arbeit mit Übersetzungsspeichern erfolgt über das TMS.

Im Rahmen dieses Dokumentes sind Segmente einsprachig. In der Fachliteratur findet sich oft der Begriff „Translation Unit“ (TU). TU bezeichnet eine durch Tags dargestellte Struktur, die neben diversen möglichen Metadaten je ein ausgangs- und ein zugehöriges zielsprachliches Segment enthält.

3.4 Dokument

Im Übersetzungskontext ist ein Dokument jede abgeschlossene, zusammen gehörende Ansammlung von zu übersetzenden Texten, die in einer Datei zusammengefasst werden. Beispiele für Dokumente in diesem Sinne sind:

- Einzelne Word- oder Open-Document-Dateien
- Ein Datenbank-Export
- Alle Oberflächentexte für ein Software-Modul

3.5 Übersetzungsauftrag (Job)

Ein Übersetzungsauftrag (Job) besteht aus ein oder mehreren Dokumenten, die zusammen mit allen weiteren Daten, die für die Bearbeitung notwendig sind, zusammengefasst werden. Diese können z. B. alle Daten sein, die ein Übersetzer zur Übersetzung eines Dokuments benötigt, inklusive des Dokumentes selbst.

3.6 Translation Memory

Das Translation Memory ist eine zentrale Datenbank von Translation Units. Die zu übersetzenden Dokumente werden in Segmente zerlegt. Für jedes dieser Segmente erfolgt eine Abfrage an die Datenbank, ob eine Übersetzung in der gewünschten Zielsprache vorhanden ist. Die gefundenen Treffer werden dem Übersetzer als Übersetzung vorgeschlagen.

Dabei werden nicht nur exakte Treffer gesucht, sondern auch die Übersetzungen „ähnlicher“ Segmente, die ebenfalls dem Übersetzer vorgeschlagen werden, zusammen mit der Information, wie „ähnlich“ das Segment ist.

Die Anwendung einer entsprechenden „Ähnlichkeitsfunktion“ wird im Allgemeinen als „Matching“ bezeichnet.

3.7 Match

Ein Match ist das Ergebnis einer unscharfen Suche (sog. fuzzy matching) des TMS nach Übereinstimmung bzw. Ähnlichkeit zwischen dem im TM befindlichen Ausgangssprachlichen und dem neu zu übersetzenden Segment. Die Ähnlichkeit wird in Prozent ausgedrückt. Zur besseren Übersicht und zur Verwaltung (Angebot, Rechnung sowie Transparenz in der Prozess- und Lieferantenkette) werden Matches in Matchklassen eingeteilt.

Eine typische Matchklassenverteilung sieht folgendermaßen aus:

- 100% Wiederholung
- 95% - 99% fuzzy match
- 85% - 94% fuzzy match
- 75% - 85% fuzzy match oder no match (je nach Einstellung)
- 50% - 74% fuzzy match oder no match (je nach Einstellung)
- < 50% no match

Die meisten TMS erlauben neben dem Ausdruck in Prozent eine Darstellung der absoluten Verteilung nach Menge der Segmente und Menge der Wörter.

3.8 Beschreibung eines Standard Übersetzungsvorganges

Übersetzungen folgen überwiegend einem ähnlichen Muster. Allerdings kann die Abweichung im Detail erheblich sein. Dies liegt insbesondere an der Qualität der Dokumente und Datenbanken, der Erfahrung und dem Wissen des Übersetzers, den Formaten und beigestellten terminologischen und anderweitigen Informationen. Redaktionelle Änderungsprozesse bzw. Korrekturwünsche über viele Zielsprachen hinweg und dies während einer bereits laufenden Übersetzung, können die Abläufe auf ein komplexes Niveau heben. Beispielsweise führen Änderungen an der Oberfläche einer Software zu entsprechenden Änderungen in abhängigen Textarten wie Hilfe, Handbuch, Marketing-Broschüre und Website.

Der prinzipielle Ablauf einer TMS gestützten Übersetzung sieht im wesentlichen wie folgt aus:

- Textextraktion der Ausgangsdokumente bzw. Trennen von Text und Struktur („filtern“)
- Segmentieren der Ausgangsdokumente
- Abgleich mit dem TM, um Wiederholungen und Alt-Übersetzungen zu finden
- Übersetzen der verbliebenen Segmente



- Austauschen der Quellsprachen-Segmente in den Ausgangsdokumenten durch ihre entsprechenden Übersetzungen (Resultat: Monolinguales ziel-sprachiges Dokument)

Weitere wichtige Arbeitsschritte sind TM-Pflege (Import von neuen Segmenten und ihrer Übersetzungen in das TM), Terminologiarbeit und QA-Maßnahmen (Prüfroutinen).

4 Modularer Aufbau von TMOSS

4.1 Translation Memory

Das eigentliche Translation Memory besteht aus einer Datenbank von Segmenten. Segmente sind Datenobjekte mit folgenden Eigenschaften:

- Inhalt (im wesentlichen Text, meistens ein Satz)
- Sprache
- Metadaten, wie z. B. Erstell- oder Änderungsdatum, Kommentare, etc.

Ein Segment in einer Quellsprache kann einem oder mehreren Segmenten in einer anderen Sprache zugeordnet sein (1-n Verknüpfung). Dies entspricht der Übersetzung des Segments in eine oder mehrere Zielsprachen. Eine Beschränkung des TM auf Sprachpaare ist zu vermeiden.

Es muss möglich sein, mehr als eine TM-Datenbank (hierarchisch gestaffelt) zu verwenden. Diese sollten sowohl lokal als auch remote vorgehaltene, von mehreren Übersetzern gemeinsam verwendete zentrale Datenbanken sein.

4.2 Segmentierer & Segment-Matcher

Der Segmentierer regelt die Segmentgrenzen. Die am häufigsten vorkommende Segmentgrenze ist der Satzpunkt. Jeder Satz kann ein Segment sein, allerdings ist längst nicht jedes Segment ein Satz. Es gilt zu unterscheiden zwischen allgemeinen Segmentgrenzen, z. B. die Absatzmarke, und besonderen Segmentgrenzen, z. B. ein Steuerzeichen aus dem Quellcode einer Maschinensteuerung. Aus diesen Gründen muss der Segmentierer konfigurierbar sein, d. h. die Regeln, die die Segmentgrenzen beschreiben, müssen anpassbar sein. Eine kontextabhängige Segmentierung wird zunächst nicht gefordert ist aber für eine Weiterentwicklung zu berücksichtigen.

Der Segment-Matcher ist das Modul, das die einzelnen zu übersetzenden Segmente mit den Inhalten der Datenbank vergleicht und gleiche oder möglichst ähnliche Segmente findet. Dabei wird eine „Match-Qualität“ (meist in Prozentanteil) ermittelt und dem Treffer zugeordnet. Die gefundenen Treffer sowie die ermittelte Qualität wird mit XLIFF der Übersetzer-Oberfläche oder der Darstellungsform des Analyseergebnisses übermittelt.

Hierbei ist der Kontext zu beachten. Nichtbeachtung von Kontext führt immer wieder zu vorgebliehen 100% Matches, die faktisch keine sind, da das zielsprachliche Segment nur

in einem bestimmten Kontext gültig ist. Ein anschauliches Beispiel ist das Wort „Bank“, das kontextabhängig eine Sitzbank oder eine Geldbank sein kann.

Je länger eine Datenbank im Einsatz ist, desto höher ist die Wahrscheinlichkeit, dass solche Unschärfen zunehmen. TMOSS wird diesem Umstand Rechnung tragen müssen und die optionale Funktion anbieten, eine zu definierende Menge an Segmenten vor und nach dem zu übersetzenden, aktiven Segment zu berücksichtigen und in das Matching einzubeziehen.

In diesem Zusammenhang ist ein weiteres Phänomen wichtig: Die Match-Klassen bewirken, dass die Qualität der Übersetzungsvorschläge abnimmt bis zu dem Punkt, an dem die Überarbeitung eines schlechten Übersetzungsvorschlages aufwändiger ist als das Segment von Grund auf neu zu übersetzen. Dennoch ist es nicht ungewöhnlich, dass dem Übersetzer einzelne Phrasen oder Teilsätze bekannt vorkommen und er diese mit einer Konkordanzsuche tatsächlich in der Datenbank findet, wenn er gezielt danach sucht. Vorstellbar wäre ein Matching auf der Ebene so genannter Subsegmente. Erkennt das TMS einen "no match", wird das aktive Segment, soweit möglich und im Rahmen der Segmentierungsregeln zulässig, in Subsegmente zerlegt, die erneut gegen die Datenbank abgefragt werden. Diese Funktion bewirkt einen signifikanten Fortschritt in der Arbeitsgeschwindigkeit und schöpft bereits vorhandene Übersetzungen wesentlich effizienter aus.

Der Segment-Matcher soll austauschbar sein, um verschiedene Matching-Algorithmen parallel implementieren zu können. Notwendig sind Funktionen, die die Segmentierung erweitern oder einschränken können, beispielsweise wenn das Semikolon als Segmentgrenze definiert ist oder wenn mehrere Sätze in einem Segment abgelegt wurden.

Dieses Phänomen kann auch im normalen Übersetzungsprozess nur mit dem TMS auftreten. Solche Multisatz-Segmente treten häufig beim Alignment auf. Dies ist ein Vorgang, bei dem ein Tool ein Ausgangssprachliches Dokument mit einem Zielsprachlichen Dokument Segment für Segment zu TU verschränkt und in der TM-Datenbank ablegt.

Beispiele hierfür sind lange Sätze, die in der Übersetzung in mehrere kurze Sätzen umgewandelt wurden. Hier wird eine erwartete 1:1-Beziehung aufgelöst und es entstehen 1:n-Beziehungen, die eine Entsprechung in den TU finden müssen. Grundsätzlich sind auch n:1 sowie seltene n:n Beziehungen möglich.

Der Segment-Matcher muss alle diese Phänomene berücksichtigen.

4.3 Benutzeroberfläche

Die Benutzeroberfläche von TMOSS gliedert sich in eine Übersetzungs- und eine Verwaltungsoberfläche, die in einem Browser abgebildet wird. Die Browserabbildung verzichtet auf jegliche Funktionalitäten proprietärer Run-Time-Umgebungen sowie auf sicherheitskritische Browsereinstellungen, die üblicherweise von Firewalls ausgefiltert werden (z. B. bei Bundesbehörden). Hierzu gehören Flash, ActiveX, Cookies und die lokale Speicherung von Session-IDs (vgl. ZDv 54/100).

In der Übersetzungsoberfläche sind Darstellungen für Quell- und Zieltext enthalten, Vorschläge aus dem TM und der Terminologie, Schalter für funktionale Aufrufe sowie Zusatzinformationen z. B. für Kontext. Die Verwaltungsoberfläche enthält Konfigurationsoptionen (z. B. für die Oberfläche selbst oder für Segmentregeln), Projektinformationen, Analysefunktionen und Funktionen für den Import bzw. Export von Dateien.

4.4 Filter

Es werden zunächst keine eigenen Dokumentenfilter implementiert. Stattdessen werden bestehende Filter aus anderen Open Source Projekten über die XLIFF-Schnittstelle angebunden.

4.5 Schnittstellen

Folgende Schnittstellen werden im Export/Import verwendet:

- TTX, TMX, XLIFF
- Terminologie (MARTIF gemäß ISO 12200, TBX, OLIF, CSV)
- Metadaten, wie SRX und GMX sowie und XLIFF
- Einbinden externer Tools zur Qualitätssicherung

5 Entwicklungsablauf

Die Entwicklung von TMOSS soll in drei Phasen erfolgen. Die Spezifikation der Phasen 1 und 2 erfolgt später.

5.1 Phase 1 – Prototyp mit Grundfunktionen

Im ersten Schritt entsteht ein Kern-System, mit dem ein einzelner Übersetzer schon Übersetzungen anfertigen kann.

Dazu müssen die folgenden Funktionen implementiert sein:

- Vor- und Nachbereitung der zu übersetzenden Dokumente (Filter)



- Segmentierer
- Translation Memory Datenbank mit Matching
- Oberflächen für Übersetzer und für einfache Verwaltungsaufgaben
- Statistische Auswertungen und Analysen
- Rechtschreibprüfung
- Schnittstelle zu Terminologiesystemen

5.2 Phase 2 – Entwicklung komfortabler Handlingfunktionen

z. B. Erweiterung der Benutzeroberfläche, Terminologieerkennung, Persistenz

5.3 Phase 3 – Entwicklung einzelner Zusatzfunktionen

z. B. Anbindung an Workflow Management Systeme

6 Zielbestimmung

Das Hauptziel für die Entwicklung von TMOSS ist es, eine Alternative zu den kommerziellen TMS zu bieten, bei der alle Anwender mitbestimmen können. Dies wird durch die Entwicklung als Open Source Programm gewährleistet werden. Bestehende Prozesse im TMS-gestützten Übersetzen werden hierbei vereinheitlicht und die offenen Standards werden konsequent umgesetzt.

6.1 Musskriterien

TMOSS ist vollständig XML-fähig , d. h. es beherrscht den XML-Standard und die für den Übersetzungsprozess abgeleiteten XML-Derivate.

Alle für die Phase 1 erforderlichen Funktionen sind Musskriterien.

Alle internen Datenformate und Verarbeitungswege aller Module werden vollständig Unicode-basiert implementiert.

6.2 Wunschkriterien

Wird später spezifiziert.

6.3 Abgrenzungskriterien

Wird später spezifiziert.

7 Produkteinsatz

7.1 Einsatzbereich und grundlegende Architektur

TMOSS wird als Webapplikation realisiert werden, wobei ein Webbrowser als Thin Client fungiert. Clientseitig werden außer dem Browser keine weiteren Installationen erforderlich sein. Aus Sicherheitsgründen wird auf jegliche Funktionalitäten proprietärer Runtime-Umgebungen sowie auf sicherheitskritische Browsereinstellungen, die üblicherweise von Firewalls ausgefiltert werden (z. B. bei Bundesbehörden) verzichtet. Hierzu gehören Flash, ActiveX, Cookies und die lokale Speicherung von Session-IDs (vgl. ZDv 54/100).

Serverseitig soll das System skalierbar sein, d. h. die Funktionalitäten müssen auf verschiedene Server verteilt werden können.

7.2 Zielgruppen

TMOSS wird von folgenden Anwendergruppen benutzt:

- Einzelübersetzer und Übersetzungsdienstleister
- Mehrere Übersetzer, die gleichzeitig an einem Projekt arbeiten
- Dokumentationsabteilungen von Firmen, Organisationen und Behörden mit Übersetzungsbedarf
- Hochschulen und andere Bildungseinrichtungen

7.3 Software

Die Konzeption und Konfiguration von TMOSS ist an keine speziellen Softwaresysteme gebunden, gleichwohl steht die Verwendung von Open Source Technologien im Vordergrund und wird besonders gefördert.

7.4 Hardware

Clientseitig ist keine spezielle Hardwareanforderung erforderlich, da nur die Installation eines Webbrowsers notwendig ist. Serverseitig soll Standardhardware zum Einsatz kommen.

7.5 Technische Schnittstellen

Eine Anbindung an technische Organisationseinheiten ist für spätere Einsätze vorgesehen und zählt zu den Wunschkriterien. Solche technischen Organisationseinheiten sind beispielsweise Workflow Management Systeme oder Rechnungssysteme. Zum gegenwärtigen Zeitpunkt liegen hierfür keine Spezifikationen vor.

7.6 Produktschnittstellen

Austausch zu anderen Produkten wird in der Phase 1 durch Import oder Export über TMX oder XLIFF realisiert.

8 Produktfunktionen

8.1 Vor- und Nachbereitung

Unter „Vor- und Nachbereitung“ wird im wesentlichen die Extraktion der zu übersetzenden Texte aus Standard-Dokumentenformaten (das „Filtern“) bzw. das Wiedereinfügen der übersetzten Texte in die Ursprungsdokumente verstanden. Dies soll zunächst schon vorhandenen Filtern wie z. B. dem Okapi-Filter-Framework überlassen werden. Bei diesem Vorgang wird der Text von Format- und Strukturinformationen getrennt, die als Tags dargestellt werden und während der Übersetzung mitlaufen. Allerdings muss es möglich sein, diese Formatinformationen mitzunehmen, um sie korrekt zu platzieren. Beispielsweise ist ein hochgestelltes Copyright-Zeichen – in dem Ausgangssprachlichen Segment auf Position vier – als Tag geschützt und für den Übersetzer frei im Zielsprachlichen Segment platzierbar. Denkbar ist eine länderspezifische automatische Ersetzung von Auszeichnungen, beispielsweise der Fettformatierung durch All-Caps für den US-Markt.

8.2 TM, Segmentierung, Matching

Das Translation Memory stellt das zentrale Repositorium für Segmente in Quell- und Zielsprachen dar.

Der Segmentierer soll den zu übersetzenden Text in Segmente zerlegen. Da, je nach Anwendungsfall die Definition eines „Segments“ sehr unterschiedlich sein kann, muss der Segmentierer konfigurierbar sein.

Matching ist das Suchen nach identischen oder ähnlichen Segmenten im TM. Dabei wird die „Ähnlichkeit“ eines Segment in Prozenten ausgedrückt. Diese „Ähnlichkeit“ wird auch Match-Qualität oder Match-Klasse genannt.

8.2.1 Der Segment-Matcher soll

8.2.1.1 Ziffern, Satz- und Sonderzeichen berücksichtigen (oder, je nach Einstellung, ignorieren), hinterlegte Listen beispielsweise mit Eigennamen berücksichtigen.

8.2.1.2 Informationen der Dokumentenfilter verwenden, um z. B. Inhalte von Tabellen-Zellen als einzelne Segmente zu erkennen.

8.2.1.3 Die Match-Qualität an die weiteren Module weitergeben.

8.2.1.4 Die Match-Qualität in Prozenten lässt sich z. B. einfach als die Anzahl der identischen

Zeichen zweier Segmente definieren.

- 8.2.2 Auswertungen, Statistik: Für jedes zu übersetzende Dokument und jeden Übersetzungsauftrag sind folgende Werte zu ermitteln
 - 8.2.2.1 Gesamtanzahl Wörter
 - 8.2.2.2 Gesamtanzahl Segmente
 - 8.2.2.3 Gesamtanzahl Zeilen und Zeichen etc. Als Standardzeile sind 55 Zeichen inkl. Leerzeichen gesetzt. Die Zeilenlänge muss aber konfigurierbar sein.
 - 8.2.2.4 Anzahl und % Segmente und Wörter, die sich innerhalb des Dokumentes/Auftrages wiederholen („Repetitions“)
 - 8.2.2.5 Anzahl und % Segmente und Wörter, die in den TM gefunden wurden, aufgeschlüsselt nach „Match-Klasse“
- 8.2.3 Filter

Außer XML werden zunächst keine eigenen Dokumentenfilter implementiert. Jedoch müssen bestehende externe Filter aus anderen Open Source Projekten über die XLIFF-Schnittstelle angebunden werden.
- 8.2.4 Schnittstellen und Schnittstellenformate

Folgende Schnittstellen werden implementiert und jeweils im Export/Import verwendet:

 - 8.2.4.1 Übersetzungsinhalte (in Phase 1 XLIFF und TMX, um zwischen TM und Übersetzer-Oberfläche Daten zu tauschen, später auch TTX)
 - 8.2.4.2 Terminologie (TBX, CSV in Phase 1, später MARTIF gemäß ISO 12200, OLIF, HTML)
 - 8.2.4.3 Metadaten, wie z. B. Anlege-/Änderungsdatum, Kommentare etc. Austauschformate für Metadaten sind z. B. SRX (= Segmentierungsregeln), GMX (= Globalisierungsregel für den Wordcount) und XLIFF (Workflowdaten mit Infos bis auf Segmentebene).
 - 8.2.4.4 Einbinden externer Tools zur Qualitätssicherung (Rechtschreibprüfung und andere Prüfroutinen, Validatoren, Parser etc.)

9 Produktdaten

Wird noch spezifiziert.

10 Benutzeroberfläche

10.1 Übersetzer

Die Benutzeroberfläche für den Übersetzer ist immens wichtig für die Akzeptanz dieser neuen Lösung. Übersetzer bevorzugen, abhängig von den Abläufen, bestimmte Aufteilungen. Die Übersetzung von Texten in mindestens einer Zielsprache muss mindestens die folgenden Spezifikationen erfüllen.

- 10.1.1 Anzeige der ausgangs- und zielsprachlichen Segmente unmittelbar untereinander.



- 10.1.2 Anzeige des Segments in der Quellsprache, das gerade bearbeitet wird.
- 10.1.3 Eingabemöglichkeit für den Text der Übersetzung.
- 10.1.4 Anzeige der Treffer aus dem TM zur Auswahl. Dabei steht der vom TMS als am ähnlichsten vermutete Match an erster Stelle und ist direkt sichtbar bzw. wird automatisch in dem zielsprachlichen Segment dargestellt.
- 10.1.5 Kontextreferenz durch Anzeige eines oder mehrerer Segmente vor und nach dem zu bearbeitenden Segment, um den Sinnzusammenhang herzustellen.
- 10.1.6 QA-Funktionen: Rechtschreibung, Segment-Prüfer, Validität bei getaggten Texten.

10.2 Verwaltung

- 10.2.1 Analyse der zu übersetzenden Texte hinsichtlich Textwiederholungen und TM Treffern.
- 10.2.2 Zusammenstellung aller Informationen und Einstellungen, die zu einem Job gehören, als „Übersetzungsprojekt“.
- 10.2.3 Exportieren und Importieren von Übersetzungsprojekten.
- 10.2.4 Erstellen von Statistiken nach Häufigkeit und Matchklassen
- 10.2.5 Anpassen der Segmentregeln
- 10.2.6 GUI Konfiguration

11 Produktleistungen

- 11.1.1 Durchgängige Verwendung von Unicode
- 11.1.2 Spezielle Anforderungen an das Datenbankmodell hinsichtlich der Auswahl von Quellsprache und Sprachumschaltung
- 11.1.3 Skalierbarkeit zur Aufgaben- und Lastverteilung
- 11.1.4 In der letzten Ausbaustufe wird TMOSS mandantenfähig sein.

12 Nichtfunktionale Anforderungen

12.1 Schnittstellen

- 12.1.1 TTX, TMX, XLIFF
- 12.1.2 Terminologie (MARTIF gemäß ISO 12200, TBX, OLIF, CSV, HTML)
- 12.1.3 Metadaten, wie SRX und GMX sowie und XLIFF
- 12.1.4 Einbinden externer Tools zur Qualitätssicherung

12.2 Plattformen

- 12.2.1 Es wird keine Betriebssystem-Plattform bevorzugt oder vorgegeben. Vielmehr ist gefordert, dass das TMS auf Windows, MacOSX und Unix bzw. Unix ähnlichen (Linux, BSD etc.) portierbar ist und die Interoperabilität zwischen den unterstützten Plattformen gewährleistet ist.

12.3 Datenbank

12.3.1 Die eigentliche TM-Datenbank muss ein Open Source Produkt sein wie z. B. MySQL, PostgreSQL oder ein ähnlich weit verbreitetes, frei und offen verfügbares Datenbanksystem. Da das gesamte System für Einzelbenutzer komplett auf einem Rechner installierbar sein wird, muss das Datenbanksystem alle vorgesehenen Plattformen unterstützen.

12.3.2 Die verwendete Datenbank muss Unicode UTF16 unterstützen.

Die strukturellen Anforderungen an das Datenbankmodell für ein multilinguales Datenbank-Konzept müssen folgendes berücksichtigen:

12.3.2.1 Es gibt keine fixen Sprachpaare

12.3.2.2 Es gibt keine bevorzugten Sprachen (keine Quelle-Ziel-Zuordnung)

12.3.2.3 Die Sprachrichtung ist jederzeit umschaltbar

12.4 Skalierbarkeit

Langfristig soll das System in unterschiedlichen Versionen bezüglich Funktionalität und TM-Backend vorliegen und entsprechend funktional skalierbar sein. Folgende Ausführungen müssen möglich sein

12.4.1 Testversion mit Minimalausstattung

12.4.2 Einzelplatzversion mit definierten Funktionalitäten

12.4.3 Voll ausgestattete Version

12.4.4 Optionale mandantenfähige Version

13 Thema Lizenz

Alle neu zu entwickelnde Software ist nach GPL Version 3 zu lizenzieren. Separate Lizenzen auf einzelne Module sind denkbar.

14 Glossar

Alignment	Aufbereitung vorhandener Übersetzungen für einen Translation Memory. Dadurch kann das TM auch Übersetzungen verwenden, die zuvor ohne dessen Mithilfe erstellt wurden oder für die keine entsprechenden Daten vorliegen. Hierzu wird ein Ausgangstext Segment für Segment (meist Satz für Satz) den entsprechenden Segmenten des übersetzten Zieltextes zugeordnet.
DMS	Document Management System
Dokumentenfilter	Eine Funktion zur Trennung von Struktur und Inhalt der zu übersetzenden Dokumente, wobei dass die spätere Zusammenführung von übersetztem Inhalt und ursprünglicher Struktur sichergestellt ist.
FOLT	Forum Open Language Tools
Fuzzy-Match	Ergebnis einer Ähnlichkeitssuche (siehe auch Match)
GMX	Metadaten zu Globalisierungsregeln für den Wordcount
QA	Quality Assurance
LiSoG	Linux Solution Group e. V.
Match	Ergebnis einer unscharfen Suche (fuzzy matching) des TMS nach Übereinstimmung bzw. Ähnlichkeit zwischen dem im TM befindlichen ausgangssprachlichen und dem neu zu übersetzenden Segment.
Multisatz-Segmente	Quellsprachliches Segment mit einer 1-n oder n-1 Beziehung
Segment	Ein Segment ist das Ergebnis eines Segmentierungsprozesses.
Segmentieren	Ist ein Prozess, bei dem nach definierten Regeln Segmentgrenzen festgelegt werden.
Segmentierungsprozess	siehe segmentieren



SRX	Metadaten zu Segmentierungsregeln
TMOSS	Translation Memory Open Source System
TM	Translation Memory ist ein Textarchiv, das multilinguale Texte enthält, die segmentiert, zugeordnet, analysiert und klassifiziert sind. Es erlaubt die Speicherung von multilingualen Texten und den Zugriff auf diese unter Verwendung unterschiedlicher Suchanfragen.
TMS	Translation Memory System
TU	Translation Unit; ein im Translation Memory abgelegtes Sprachpaar
XLIFF	XML Localization Interchange File Format. Die Format bildet die vielen unterschiedlich zu lokalisierenden Daten auf ein einheitliches Format ab. XLIFF-Dokumente können nicht nur die zu übersetzenden Daten, sondern auch andere für den Workflow benötigte Daten enthalten. Damit könnten XLIFF-Dokumente den Übersetzungsprozess auch steuern.
ZDv	Zentrale Dienstvorschriften der Bundeswehr, z. B. ZDv 54/110 VS-NfD „Behandlung und Einsatz von Kryptomitteln“; ZDv 54/100 VS-NfD „IT-Sicherheit in der Bundeswehr“