Localization Engineering für Übersetzer – mit Freeware- und OpenSource-Helferlein

Christine Bruckner

Translation Technology Consultant

Zusammenfassung des Kurzseminars und Publikation im Tagungsband der BDÜ-Konferenz „Übersetzen und Dolmetschen 4.0 – Neue Wege im digitalen Zeitalter“, Bonn, 22.-24.11.2019

1 Einführung

„Localization Engineering“ wird als eine der Zukunfts-/Alternativtätigkeiten für Übersetzer im Zeitalter von „Digitalisierung 4.0“ und der „Machtübernahme“ der neuronalen maschinellen Übersetzung genannt.

Das Berufsbild des Localization Engineers existiert jedoch schon seit Jahrzehnten, vorwiegend in unternehmensinternen Sprachendiensten und bei großen Übersetzungsdienstleistern. Hauptaufgabe ist in der Regel die Vor- und Nachbereitung von Übersetzungsressourcen für die Verwendung in Übersetzungstools (computergestützte Übersetzungssysteme/CAT-Tools, Translation Management-Systeme/TMS, maschinelle Übersetzung/MT, Terminologiesysteme).

Effiziente Vor- und Nachverarbeitung von Übersetzungsdateien, Translation Memories, Terminologiedaten u. ä. sind jedoch auch wichtig für Einzelübersetzer und kleine Übersetzergruppen und setzt – entgegen der weit verbreiteten Meinung – nicht unbedingt Programmierkenntnisse voraus.

Wer über Grundkenntnisse in XML (Extended Markup Language) und regulären Ausdrücken verfügt, kann kostenlose „Helferlein“ wie unter anderem Notepad++, OpenOffice, diverse TMX-Editoren sowie das Okapi Framework clever und Übersetzungstool-unabhängig nutzen, um ein- und mehrsprachige Dateien ins passende Format zu bringen und/oder inhaltlich zu optimieren.

Nach einer kurzen Vorstellung der Tools werden anhand von typischen Problemstellungen im Übersetzungskontext praktische Anwendungsbeispiele, Tipps und Tricks für die effiziente Anwendung dieser Tools gezeigt.

Bezüglich der Dateiformate liegt der Schwerpunkt auf textbasierten Formaten (*.txt*, *.html*, *.xml*/*.xliff*/*.tmx*/*.tbx* u. ä.); für proprietäre Formate wie MS Office, Adobe InDesign, Adobe Framemaker und weitere DTP-Formate eignen sich die vorgestellten Tools nicht bzw. nur sehr eingeschränkt.

2 Ziele und Grenzen

Ziel des Kurzseminars ist es, technisch versierten Übersetzern Anregungen zu geben, wie man mit Freeware- und OpenSource-Tools Arbeitsschritte bei der Vor- und Nachbereitung von Dateien vereinfachen oder teilweise automatisieren kann und Datenbestände (TMs, Terminologiebestände) effizienter pflegen und konvertieren kann.

Ein Grundverständnis über reguläre Ausdrücken und XML (insbesondere die übersetzungsrelevanten Austauschformate TMX, TBX und XLIFF) wird bei den Teilnehmenden vorausgesetzt; eine Einführung in diese Konzepte ist nicht Bestandteil des Kurzseminars.

Die vorgestellten Freeware- und OpenSource-Tools sind zwar kostenlos verfügbar (für Download-Links siehe Abschnitt „Bibliographische Angaben“), doch gelten auch hier Nutzungsbedingungen, die beim Einsatz zu beachten sind. Insbesondere sei darauf hingewiesen, dass einige der vorgestellten Tools die Installation von Java voraussetzen, für dessen Nutzung seit April 2019 geänderte Lizenzbestimmungen seitens Oracle gelten (siehe Oracle 2019).

Die im Folgenden als „Helferlein“ bezeichneten Tools können oftmals fehlende Funktionen kommerzieller Übersetzungstools (CAT-Tools, Translation Management-Systeme, Terminologiesysteme) abdecken, aber eine 100 %-ige Kompatibilität mit dem verwendeten Übersetzungstool bzw. gewünschten Zielformat ist nicht immer gewährleistet. Zudem können auch sie nicht zaubern, wenn die Ausgangsformate suboptimal sind. Welche Zielsetzungen und Einschränkungen beim Einsatz dieser Helferlein gelten, und ggf. welche Tests und Prüfmaßnahmen empfehlenswert sind, wird im Rahmen der Anwendungsbereiche und -beispiele ebenfalls kurz skizziert.

Für bestimmte Zielsetzungen mag das verwendete kommerzielle Übersetzungstool bzw. ein entsprechendes Plugin oder Add-On die gewünschte Funktionalität benutzerfreundlicher oder zuverlässiger abdecken oder ist möglicherweise die Investition in ein kostenpflichtiges Tool (wie kommerzieller XML- oder Text-Editor) oder einen spezialisierten Dienstleister oder Inhouse-Programmierer/Localization Engineer eher ratsam.

Einige der vorgestellten Tools werden nicht mehr weiterentwickelt oder ihre Wartung liegt bei Entwicklern, die sich nebenberuflich oder in ihrer Freizeit um Support und Umsetzung von Feature-Wünsche kümmern. Man darf also nicht zu viel erwarten, aber sich gern positiv überraschen lassen, wie schnell schon mal der eine oder andere Feature-Wunsch umgesetzt wird oder man Unterstützung für seine Fragestellung in der jeweiligen Community erhält.

3 Localization Engineering – Begriffsdefinition und Verwendung

Im Kurzseminar wird der Begriff „Localization Engineering“ für Aufgaben und Tätigkeiten im Zusammenhang mit den vorgestellten Helferlein und Anwendungsbeispielen verwendet.

Localization Engineering ist in der Praxis meist nicht klar zu trennen von angrenzenden Tätigkeiten wie Übersetzungsprojektmanagement, Software-Entwicklung, Software-Testen usw. – oftmals, weil die Localization Engineering-Tätigkeiten Bestandteil der

jeweiligen Jobs sind, insbesondere in kleineren Übersetzungsagenturen und Inhouse-Sprachendiensten.

Betrachtet man die einschlägigen Stellenausschreibungen und Tätigkeitsdarstellungen, so zählen zu den typischen Aufgaben eines Localization Engineers:

- Schnittstellenfunktion zwischen Übersetzern, Projektmanagern, DTP-Team, Entwicklern, Endkunden
- Vor-/Nachbereitung bzw. Konvertierung und Rückkonvertierung von Dateien für die Bearbeitung durch Übersetzer
- Kontrolle der Integrität der Dateien nach erfolgter Übersetzung
- Erstellen von angepassten Datei(format)filter für CAT-Tools
- Sicherstellen, dass die übersetzten Textinhalte auch in der laufenden Software richtig dargestellt werden
- (formale) Pflege kundenspezifischer und übergreifender Terminologiedatenbanken und Translation Memories, Aufbereitung und Konvertierung für Training von MT-Engines
- Konzipieren, Analysieren und Optimieren von Übersetzungsprozessen unter Einsatz von Translation-Memory-Systemen, Terminologiesystemen und maschinellen Übersetzungssystemen
- Definieren von Best Practices im Übersetzungsprozess aus technischer Sicht
- technische Mitarbeit/Leitung bei der Einführung neuer Tools

Vereinfacht gesagt ist ein Localization Engineer also meist der „klassische“ CAT-Tool-Spezialist, mit Kenntnissen und Zuständigkeiten, die über die CAT-Tools hinausgehen.

(Teil-)Synonyme und verwandte Benennungen zum Localization Engineer sind neben dem CAT-Tool-Spezialisten unter anderem: *Translation Engineer*, *Translation/Language Technologist*, *Translation Support Engineer*, *Computerlinguist*, *Linguistic Data Scientist*.

Localization Engineers haben meist einen Ausbildungshintergrund in Sprachtechnologie, Computerlinguistik, Informatik/Softwareentwicklung – oder sind Übersetzer mit technischem Interesse und Fähigkeiten, die über Makro-Programmierung, Programmier- und Scripting-Sprachen, reguläre Ausdrücke und OpenSource/Freeware-Tools (wie die im Folgenden vorgestellten) ihre Übersetzungsarbeit und -abläufe technisch optimiert haben und so zu haupt- oder nebenberuflichen Localization Engineers geworden sind.

Im Kurzseminar geht es um diejenigen Localization Engineering-Tätigkeiten, die nicht unmittelbar das Arbeiten mit bzw. in (kommerziellen) CAT-/TMS-/Terminologie-Tools selbst betreffen, sondern diesen vor- oder nachgelagert sind bzw. sie ergänzen.

4 Localization Engineering-Tools für Übersetzer und typische Anwendungsbeispiele

4.1 Vorteile und Einsatzbereiche

Effiziente und (teil-)automatisierbare Vor- und Nachbereitung von Ausgangs- bzw. Zieldateien sowie Erstellung und Pflege von Übersetzungsressourcen wie Translation Memories, Terminologiedatenbanken, Referenzdateien, alternativen Übersetzungsquellen (z. B. Machine Translation-Engines) sparen nicht nur Zeit beim Übersetzen und Prüflesen, sondern helfen auch dabei, Qualität und Konsistenz zu verbessern. Technisch versierte Übersetzer, die über Localization Engineering-Kenntnisse und Erfahrungen mit entsprechenden frei verfügbaren bzw. Open-Source-Tools verfügen, können zudem auch Kunden bzw. deren technischen Ansprechpartnern Anregungen für entsprechende Format- und Prozessoptimierung geben.

4.2 Vorgestellte Helferlein im Überblick

Im Kurzseminar werden folgende Tools aus der Freeware- und OpenSource-Welt vorgestellt, die unabhängig vom eingesetzten kommerziellen Übersetzungstool clevere Funktionen für Vor- und Nachbereitung von Dateien und Übersetzungsressourcen bieten:

- *OpenOffice (Teilaspekte)*
- *Notepad++*
- *TMX-Editoren: Olifant, Heartsome TMX Editor*
- *Okapi Framework*: Komponenten *Rainbow* und *Checkmate*

Die Download-Links für diese lokal installierbaren Tools finden sich im Abschnitt „Bibliographische Angaben“. Die vorgestellten Helferlein zählen zu den populärsten Freeware- und OpenSource-Tools in Fachkreisen (siehe GILT Leaders Forum 2018). Es sei jedoch darauf hingewiesen, dass es weitere (kommerzielle oder kostenlose) Tools gibt, die ähnliche Funktionen bieten.

4.3 Anwendungsbereiche und -beispiele

4.3.1 Terminologie

Die meisten kommerziellen Terminologiesysteme bieten Funktionen und/oder ergänzende Komponenten für die Konvertierung von Wortlisten und Terminologiebeständen, meist auch über die Austauschformate TBX oder Martif oder über „einfache“ Excel- bzw. csv-Listen.

Die hier vorgestellten Helferlein bieten Zusatzfunktionen oder Alternativen unter anderem für folgende Aufgaben:

- Vor- und Aufbereiten von Glossaren im MS Excel-, csv- oder MS Access-Format (für den Import in Terminologiesysteme oder in Wörterbücher maschineller Übersetzungssysteme), auch mittels regulärer Ausdrücke beim Filtern, Suchen&Ersetzen

- Konvertieren von Terminologielisten in das TMX-Format zwecks Verwendung als Referenz-Translation Memories
- Terminologie-Extraktion

4.3.2 Translation Memory-Pflege und Konvertierung

4.3.2.1 Vorbemerkungen

Für Zwecke der Pflege und Konvertierung von Translation Memories (TMs) bieten die jeweiligen kommerziellen CAT-Tools, Translation Management-Systeme und auch Machine Translation-Systeme ebenfalls meist integrierte eigene Funktionen.

Doch mangelt es den integrierten TM-Editoren oft an Benutzerfreundlichkeit und Zusatzfunktionen, so dass in Fachkreisen – und manchmal von den CAT-Tool-/TMS-Herstellern selbst – auf kostenlose TM-Editoren wie *Olifant* oder *Heartsome TMX Editor* verwiesen wird (siehe auch GILT Leaders Forum 2018).

Im Hinblick auf die vorgestellten Funktionen der Helferlein ist vorab anzumerken, dass es sich bei *Olifant* um eine ältere Implementierung handelt, die unter Umständen mit großen Dateien nicht zuverlässig umgehen kann. In *Okapi Checkmate* sind die Terminologieprüffunktionen nur rudimentär bzw. noch in der Beta-Phase. Aber auch für *Heartsome TMX Editor* – und prinzipiell allen eingesetzten Tools – gilt: Backups anlegen, immer zunächst testen und Plausibilitätsprüfungen durchführen, damit man später kein böses Erwachen bezüglich fehlender oder beschädigter Übersetzungseinheiten oder Metadaten hat.

4.3.2.2 TM-Pflege für den computergestützten Übersetzungsprozess

Die vorgestellten Helferlein bieten neben Konvertierungsmöglichkeiten in das und aus dem TMX-Format Möglichkeiten zur formalen und sprachlichen Datenpflege, die über das reine Suchen/Ersetzen hinausgehen (z. B. mit regulären Ausdrücken oder SQL-Befehlen) sowie – im Falle von *Okapi Rainbow* – das Kombinieren und Wiederverwenden von Einzelschritten über so genannte *Pipelines*.

4.3.2.2.1 TM-Konvertierung und formale Bereinigung

In Zusammenhang mit der TM-Konvertierung und der formalen TM-Bereinigung bieten die Helferlein Lösungen und Unterstützung für die nachfolgenden Aufgaben:

- Konvertieren von TMX-Dateien zwecks Nutzung als Referenzdateien in MS Excel
- Anonymisierung von Benutzer-IDs in TMX-Dateien
- Konvertieren mehrsprachiger Listen und Bestände (aus Excel, SQL u. ä.) in das TMX-Austauschformat, inkl. Anwendung regulärer Ausdrücke zwecks Schützen/Taggen nicht zu übersetzender Zeichenfolgen
- Ermitteln leerer Ausgangs- oder Zielsegmente
- Validieren, Splitten und Zusammenführen von TMX-Dateien

Diese Schritte lassen sich sprachunabhängig und damit weitgehend ohne Kenntnisse in den bearbeiteten Sprachen durchführen.

4.3.2.2.2 Unterstützung für sprachliche TM-Pflege

Bei sprachlichen TM-Pflegemaßnahmen ist es in der Regel Best Practice, mithilfe von Zusatz-Tools QA-Berichte mit möglichen Fehlern und Problemen zu erstellen und diese an Übersetzer bzw. sprachkundige Linguisten weiterzugeben, die diese im TM-Editor des jeweiligen CAT-Tools bzw. TMS sichten und bereinigen. oder in einem zuvor für die jeweilige Aufgabe ausgiebig getesteten TMX Editor, mit anschließendem Import und Überschreiben der bearbeiteten Übersetzungseinheiten im „Live"-TM des CAT-Tools bzw. TMS.

Zwecks Analyse möglicher sprachlicher Inkonsistenzen in TMs und Filterung bzw. Export der ermittelten problematischen Übersetzungseinheiten bieten die Helferlein u. a. die folgenden Funktionen:

- Ermitteln und ggf. Normalisieren diverser Arten von Dubletten
- Ermitteln von verdächtigen Übersetzungseinheiten anhand von mit definierten Längenunterschieden für Quell- und Zielsegmente
- Filtern nach bestimmten Wörtern und Wortfolgen in Quell- und/oder Zielsegmenten, mithilfe von Wortlisten oder regulären Ausdrücken bzw. SQL-Befehlen

4.3.2.3 Ergänzende TM-Aufbereitungsschritte für MT-Training

Translation Memories und auch Terminologiebestände stellen in den datengestützten maschinellen Übersetzungsansätzen (statistische maschinelle Übersetzung, *SMT*, und neuronale maschinelle Übersetzung, *NMT*) die wichtigste Ressourcen für das Training bzw. die Anpassung (*Customization*) der entsprechenden MT-Engines dar.

Inzwischen bieten zunehmend auch die NMT-Anbieter *Do-it-Yourself*-Möglichkeiten für das Trainieren eigener MT-Engines mittels Hochladen von TMX-Dateien. Jedoch kann sich durch „naives" Hochladen von nicht weiter aufbereiteten TMX-Dateien oder Terminologiebeständen die Qualität des MT-Outputs verschlechtern. Denn für das MT-Training sind saubere Daten erforderlich (für NMT sogar *sehr* saubere Daten); Fehler in den Trainingsdaten können zu unerwarteten Ergebnissen führen, deren Ursache sich nicht oder schwer nachvollziehen bzw. nur sehr aufwändig ausmerzen ist. Daher ist es nach wie vor meist effizienter und führt zu qualitativ besseren Ergebnissen, wenn die Aufbereitung der TMX-Dateien und das Training der MT-Engines durch den MT-Hersteller bzw. entsprechende Spezialisten erfolgt (Bruckner 2018: 7).

Neben den proprietären Funktionen von MT-Systemen sowie eigenentwickelten Tools zählen zu den von MT-Spezialisten genannten Tools auch einige der hier vorgestellten Helferlein (Fernández Rowda 2015; GILT Leaders Forum 2018).

Neben den bereits im Zusammenhang mit TM-Pflege genannten formalen und sprachlichen Schritten werden u. a. folgende Funktionen der Helferlein im Hinblick auf MT-Vorbereitung genutzt:

- Entfernen von überlangen Übersetzungseinheiten
- Entfernen von zu kurzen Übersetzungseinheiten
- Ermitteln und Reparieren abweichender Zeichensätze und ungültiger Zeichen

4.3.3 Vor- und Nachbereitung von Übersetzungsdateienformaten

Die Vor- und Nachbereitung von zu übersetzenden Dateien, die im Text- oder XML-Format vorliegen, ist ein weites Feld: Hinter *.txt, .csv, .yml, .json, .xml, .xliff* und sonstigen Text- oder XML-basierten Formaten verbirgt sich eine Vielzahl von Varianten und man ist in der Regel nicht gut beraten, in seinem CAT-Tool ohne weitere Anpassung die generischen XML- bzw. txt-Dateifilter für die Verarbeitung solcher Formate zu verwenden.

Das bevorzugte Helferlein zur Sichtung, aber auch zur Validierung, Weiterverarbeitung und ggf. Prüfung solcher Formate ist der kostenlos verfügbare *Notepad++* mit seiner Vielzahl an ergänzenden Plugins.

Sowohl in *Notepad++* als auch in *Okapi Rainbow* lassen sich Bearbeitungsschritte auf mehreren Dateien gleichzeitig durchführen. Wie auch bei der TM-Pflege bereits angeführt, lassen sich mit *Okapi Rainbow* zusätzlich Einzelschritte über so genannte *Pipelines* kombinieren, wiederverwenden und damit teilautomatisieren.

Bei der Vor- und Nachbereitung von textbasierten und XML-Dateien machen sich die Helferlein insbesondere bei den folgenden Aufgaben nützlich:

- Ermitteln des Dateiformats; Syntax-Highlighting, Identifizieren des Trennzeichen- und Zeilenumbruchtyps
- Prüfen (Wohlgeformtheit/Validierung) von XML-Dateien
- Ermitteln und Reparieren von „kaputten“ Zeichen(-kodierungen) in Dateien; Ergänzen/Entfernen des Byte-Order-Marks (BOM)
- Suchen und Ersetzen in Dateien und Ordnern mit regulären Ausdrücken sowie Auffinden der Treffer über Ergebnislisten
- Zählen/Plausibilitätsprüfungen nach Änderungen
- Vergleichen von Dateien
- Automatisches Ermitteln vorhandener Elemente/Attribute in XML-Dateien
- Automatisiertes Befüllen zielsprachlicher Einheiten mit ausgangssprachlichem Text in XML-/XLIFF-Dateien
- ID-basiertes Alignment

Insbesondere bei der Bearbeitung großer Dateien und in Zusammenhang mit umfangreichen Ersetzungsoperationen ist Vorsicht geboten, da Inkonsistenzen oder sogar Datenverlust/-beschädigung auftreten können, insbesondere wenn nicht mit leistungsstarker Hardware gearbeitet wird. Backups, Roundtrips, Pseudo-Übersetzung im eingesetzten CAT-Tool bzw. TMS sowie Prüfungen in der Zielanwendung bzw. in einem WYSIWYG-ähnlichen Modus sind daher in allen Fällen ratsam, um böse Überraschungen zu vermeiden.

5 Zusammenfassung

Technisches Verständnis, analytische Denk- und Vorgehensweise und Experimentierfreude sind wichtige Voraussetzungen bei der Nutzung der vorgestellten Tools für Localization Engineering-Aufgaben, aber auch Übung, ein gesundes Maß an Skepsis und Einschätzung der Grenzen des Machbaren. Mit dieser Grundeinstellung können Übersetzer die vorgestellten Tools aus der Freeware- und OpenSource-Welt zu ihren Helferlein für technische Bearbeitungsschritte im Rahmen der Vor- und Nachbereitung von Übersetzungsressourcen machen und sich damit gegebenenfalls ein zweites – oder erstes – Standbein als Localization Engineer aufbauen.

6 Bibliographische Angaben

Apache OpenOffice: https://www.openoffice.org/de (zuletzt aufgerufen am 21.07.2019).

Bruckner, C. et. al. (2011). *Okapi Framework Evaluation - Conclusions for Phase I*. (SAP AG, unveröffentlicht).

Bruckner, C. (2018): „Terminologie und maschinelle Übersetzung - Herausforderungen und Möglichkeiten einer Integration“, in: *edition - Fachzeitschrift für Terminologie, Ausgabe 02/2018*, S. 5–11.

Fernández Rowda, J. M. (2015). *5 Tools to Build Your Basic Machine Translation Toolkit – Part I.* https://www.linkedin.com/pulse/5-tools-build-your-basic-machine-translation-toolkit-fern%C3%A1ndez-rowda?trk=pulse_spock-articles (zuletzt aufgerufen am 21.07.2019).

GILT Leaders Forum (2018). *Best Practices in Translation Memory Management v.2.1*. https://github.com/GILT-Forum/TM-Mgmt-Best-Practices/blob/master/best-practices.md#summary-table-of-recommended-tasks-for-tms (zuletzt aufgerufen am 21.07.2019).

Heartsome TMX Editor 8: https://github.com/heartsome/tmxeditor8 (zuletzt aufgerufen am 21.07.2019).

Notepad++: https://notepad-plus-plus.org (zuletzt aufgerufen am 21.07.2019).

Okapi Framework: http://okapiframework.org (zuletzt aufgerufen am 21.07.2019).

Olifant: http://okapi.sourceforge.net/Release/Olifant/Help (zuletzt aufgerufen am 21.07.2019).

Oracle (2019). *Java FAQ: Lizenzierung und Bereitstellung* https://www.java.com/de/download/faq/distribution.xml (zuletzt aufgerufen am 28.07.2019).

Pawelec, M. (2017): „Okapi Framework Localization Swiss Army Knife“, in: *tekom-Jahrestagung 2017 – Tagungsband*, S. 274 ff.

