Chapter 1

Übersetzungsdatenmanagement

XYZ Varga

Dieser Beitrag gibt einen Überblick über Strategien und Prozesse des Übersetzungsdatenmanagements. Die Reihenfolge der Kapitel orientiert sich an der Vorgehensweise bei der Durchführung von Translation-Memory-Harmonisierungs-/Zentralisierungsprojekten: von der Definition der Datenorganisation und der Identifizierung potenziell relevanter Datenquellen über deren Migration, Analyse, Bereinigung und Harmonisierung bis hin zu ihrer Implementierung in Produktionsprozessen und der kontinuierlichen Überwachung ihrer Qualität und Performance.

1 Einleitung

Die Übersetzungsindustrie hat in den letzten Jahrzehnten einen grundlegenden technologischen Wandel erfahren. Auf das Übersetzen als Tätigkeit an sich hatten dabei insbesondere drei Technologien einen massiven Einfluss: So wurde mit der Veröffentlichung der ersten kommerziellen CAT-Tools in den frühen 1990er Jahren zunächst die einfache Wiederverwendung bereits übersetzter Texte ermöglicht – wodurch sich nicht nur das Arbeiten von ÜbersetzerInnen, sondern auch Art und Weise, wie sie ihre Leistungen abrechnen, grundlegend änderte. Mit der neuronalen maschinellen Übersetzung (NMÜ) folgte im **Jahr2016** dann der nächste Technologiesprung, in dessen Folge maschinelle Übersetzung von der Nischen- zur Breitentechnologie wurde – auch hier sowohl mit Auswirkungen auf Arbeitsweisen und Abrechnungsmodalitäten. Mit der Entwicklung der adaptiven maschinellen Übersetzung im Jahr 2017, spätestens aber mit der Veröffentlichung der ersten Large Language Models (LLMs) **Ende2022** wurde es dann schnell einfacher, MÜ-Ergebnisse in Echtzeit – und ohne vorheriges aufwendiges Training – in die gewünschte Richtung zu lenken.

XYZ Varga. . To appear in: E. Di Tor & Herr Rausgeberin (ed.). Booktitle in localcommands.tex. Berlin: Language Science Press. [preliminary page numbering]

Was alle diese Technologien gemeinsam haben? Sie funktionieren auf Grundlage von Daten. Waren Übersetzungsdaten im Zeitalter der CAT-Tools noch eine hochspezialisierte Ressource, die außerhalb klassischer Übersetzungsprozesse kaum eine Rolle spielten, so gewannen sie mit der Verfügbarkeit trainierbarer bzw. customisierbarer NMÜ-Modelle zunehmend auch breitere Bedeutung (**Wang2016**, S. 12) – etwa für die Bereitstellung von maschineller Übersetzung als Self Service in Unternehmen; ein Trend, der sich mit dem schrittweisen Umschwenken vieler Anbieter auf LLMs nahtlos fortsetzt.

In erster Konsequenz lässt sich daraus folgende Beobachtung ableiten: In dem Maße, in dem Übersetzungsdaten breiter eingesetzt und die auf ihrer Grundlage erstellten Übersetzungen nicht systematisch von Menschen überprüft werden (können), steigen die Anforderungen an ihre Qualität. Diese lässt sich definieren anhand der Dimensionen Genauigkeit (*Accuracy*), Konsistenz (*Consistency*), Vollständigkeit (*Completeness*) und Aktualität (*Currency* und *Timeliness*) (Scannapieco/Missier/**Bat**
Batini/**Scannapieco2016**). Konkret sollten Übersetzungsdaten also:

- den an sie gestellten Qualitätsanforderungen entsprechen, etwa in Bezug auf die Einhaltung einer Corporate Language oder eines vordefinierten Metadatenschemas (*Accuracy*),
- konsistent sein in Bezug auf die Einhaltung sowohl sprachlicher Vorgaben als auch auf die Verwendung von Metadaten (*Consistency*),
- den jeweiligen Einsatzzweck abdecken (*Completeness*),
- in ihrem Inhalt dem aktuellen Stand des durch sie abgebildeten Sprachgebrauchs entsprechen (*Currency*), also z. B. keine veraltete Terminologie enthalten,
- durch geeignete Managementprozesse auf dem jeweils neuesten Stand gehalten werden (*Timeliness*) (vgl. Zielinski/**Varga2020**, S. 307–309).

Wie aus diesen Anforderungen deutlich wird, gehen erhöhte Anforderungen an die Übersetzungsdatenqualität Hand in Hand mit erhöhten Anforderungen an das Übersetzungsdatenmanagement.

Vor diesem Hintergrund geben die folgenden Kapitel einen Überblick über die wichtigsten Schritte und Ansätze für ein effizientes Management von Übersetzungsdaten in Form von Translation Memorys (TMs). Das Augenmerk liegt dabei nicht nur Aufbau und Pflege neuer Datenbestände, sondern auch und gerade auf der Zusammenführung von Daten aus unterschiedlichsten Quellen, wie sie im Rahmen von Zentralisierungs- und Harmonisierungsprojekten in Übersetzungsabteilungen erforderlich sind.

2 Datenorganisation definieren

Die Implementierung effizienter Managementprozesse für Übersetzungsdaten setzt zuallererst die Definition von Anforderungen und Zielen sowie einer darauf ausgerichteten Datenorganisation voraus.

2.1 Physische TM-Organisation

Dies betrifft zunächst die physische Organisation von TMs: Sollen (bzw. können!) je Übersetzungsrichtung alle Inhalte in ein und demselben TM gespeichert werden? Die Vorteile eines solchen Single-TM-Ansatzes liegen auf der Hand: je kleiner die Zahl an TMs, desto kleiner der Aufwand für ihre Implementierung, Verwaltung und Pflege.

Unterschiedliche TMs für Übersetzungen in derselben Sprachrichtung können jedoch etwa aus Vertraulichkeitsgründen erforderlich sein: Werden bspw. vertrauliche Inhalte von In-House-Übersetzern bearbeitet und andere Übersetzungen extern vergeben, kann über zwei separate TMs je Übersetzungsrichtung garantiert werden, dass ein Zugriff auf vertrauliche Inhalte durch externe Übersetzer – über Fuzzy-Matches oder Konkordanzsuchen – unmöglich ist.

2.2 Unterstützende Ressourcen

Zusätzlich zu den TMs selbst müssen in diesem Schritt auch weitere Ressourcen erstellt bzw. geprüft werden, die sich direkt auf die Qualität der Übersetzungsdaten auswirken.

Dies betrifft zunächst die TM-Einstellungen. Je nach verwendetem Tool bieten diese unterschiedliche Optionen für die Verwendung von TMs in Übersetzungsprojekten, wie etwa Filter, über die TM-Treffer mit bestimmten Metadaten priorisiert oder aber ausgeschlossen werden können (siehe Kap. **??**).

Weitere Ressourcen, die direkten Einfluss auf die Qualität der im Übersetzungsprozess entstehenden Daten hat, sind Segmentierungsregeln und Dateiimporteinstellungen. Segmentierungsregeln definieren jeweils für eine spezifische Ausgangssprache, an welchen Stellen Texte in einzelne Segmente aufgesplittet werden sollen, also etwa nach Satzendzeichen, Zeilenumbrüchen oder bestimmten Sonderzeichen. Dateiimporteinstellungen wiederum definieren für ein bestimmtes Dateiformat (also etwa .docx oder .xml), wie Inhalte aus entsprechenden Dateien in den Übersetzungseditor importiert werden. Fehlerhafte Segmentierungsregeln und/oder Dateiimporteinstellungen wirken sich negativ auf Qualität und Wiederverwendbarkeit von Übersetzungseinheiten aus, indem sie Texte auseinanderreißen oder

aber zu grob segmentieren oder zum Import nichtübersetzbarer Inhalte führen (siehe Kap. **??**).

Je nach System besteht darüber hinaus u. U. die Möglichkeit, nichtübersetzbare Elemente, wie etwa Produktnamen und -nummern, zu hinterlegen und im Idealfall sogar zu schützen. Hier sind unterschiedliche Ansätze möglich: So bieten manche Tools die Möglichkeit, zu schützenden Text mithilfe regulärer Ausdrücke in Tags zu verwandeln.[1] Eine zusätzliche toolspezifische Option ist etwa die Erstellung sog. *Non-Translatables*-Listen in memoQ (siehe Abb. 1). Unabhängig von der konkreten Vorgehensweise können fehlende/veränderte Elemente in laufenden Projekten über QA-Routinen identifiziert werden, sodass fehlerhafte Segmente erst gar nicht in das Master-TM gelangen.

Figure 1: Regex-Tagging und Non-Translatables in memoQ

2.3 Berechtigungskonzepte

Neben der Frage, *wie* TMs bearbeitet werden sollen, ist vor allem auch die Frage wichtig, *wer* die entsprechenden Berechtigungen dazu erhalten soll. Diese Frage betrifft sowohl die Verwendung in laufenden Übersetzungsprojekten als auch die flankierende TM-Pflege.

In Bezug auf die Verwendung in Übersetzungsprojekten stellt sich etwa die Frage, ob Übersetzer nur Lese- oder auch Schreibzugriff auf bestimmte TMs erhalten sollen. Ein typischer Ansatz ist hier die Unterscheidung zwischen Master-TMs (die bereits übersetzte Texte enthalten) und Arbeits-TMs (die für jedes Projekt neu erstellt und in die neue Übersetzungseinheiten zunächst gespeichert werden). Indem man den Schreibzugriff der Übersetzer auf das Arbeits-TM beschränkt, kann verhindert werden, dass diese unerwünschte Änderungen in Master-TMs vornehmen.

2.4 Metadatenschema

Metadaten sind Daten über andere Daten. Konkret handelt es sich bei diesen „anderen“ Daten im Fall von TMs sowohl um das TM als Ganzes als auch um die darin enthaltenen Übersetzungseinheiten.

Die hier relevanten Metadaten zu deren Beschreibung lassen sich grob in drei Kategorien einteilen (vgl. **Pomerantz2015**; Zielinski/**Varga2020**):

- Administrative Metadaten

[1] Für einen Überblick über die Verwendung regulärer Ausdrücke in CAT-Tools siehe Rudd (2018).

- Beschreibende Metadaten
- Nutzungsbezogene Metadaten

2.4.1 Administrative Metadaten

Administrative Metadaten werden automatisch generiert, wenn ein TM erstellt, neue Übersetzungseinheiten darin gespeichert oder bereits darin vorhandene Übersetzungseinheiten geändert werden. Sie geben Aufschluss über die Erstellung und Herkunft von TM-Daten. Typische Beispiele hierfür sind etwa Name und Version des Tools, mit dem ein TM erstellt wurde, die IDs der Nutzer, die die darin enthaltenen Übersetzungseinheiten erstellt haben, sowie die Zeitstempel ihrer Erstellung und/oder Änderung.

Zu den administrativen Metadaten gehören auch sog. strukturelle Metadaten, die etwa Auskunft darüber geben, um welche Art von Text es sich bei einem Segment handelt (Überschrift, Listenelement etc.). Auch die in einer Übersetzungseinheit gespeicherten Kontextinformationen fallen in diese Kategorie.

2.4.2 Beschreibende Metadaten

Beschreibende Metadaten enthalten zusätzliche Informationen zu einem TM oder den darin enthaltenen Übersetzungseinheiten. Typische Beispiele für beschreibende Metadaten sind etwa die IDs von Kunden oder Projekten sowie Angaben zu Fachgebiet, Textsorte, Produkt(familie), Kommunikationskanal etc. Im Gegensatz zu administrativen Metadaten werden beschreibende Metadaten nicht automatisch generiert, sondern müssen von den NutzerInnen selbst definiert/ausgewählt werden. Gespeichert werden diese Informationen entweder in Metadatenfeldern, die standardmäßig von dem verwendeten TM-System bereitgestellt werden, oder aber in benutzerdefinierten Metadatenfeldern. Dies bietet einerseits Flexibilität bei der TM-Verwaltung, andererseits erschweren fehlende oder inkonsistente Metadaten die Zusammenführung von Übersetzungsdaten aus unterschiedlichen Quellen, selbst wenn sie alle aus demselben Tool stammen.

Je nach Tool kann es unterschiedliche Optionen für das Speichern bestimmter beschreibender Metadaten geben. So verfügen etwa memoQ-TMs über eine Option zum automatischen Speichern des Namens oder Pfads der Datei, aus der eine Übersetzungseinheit stammt – in Trados Studio hingegen kann ein solcher Wert, wie alle beschreibenden Metadaten, standardmäßig nur manuell über die Projekteinstellungen eingegeben werden.

Generell gilt: Wo immer möglich, sollten die Metadatenwerte in beschreibenden Feldern als Auswahllisten hinterlegt werden, um Inkonsistenzen durch Schreibvarianten, Rechtschreibfehler etc. zu vermeiden. Freitextfelder bleiben typischerweise auf diejenigen Metadaten beschränkt, die sich von Projekt zu Projekt ändern, z. B. der Projektname.

2.4.3 Nutzungsbezogene Metadaten

Nutzungsbezogene Metadaten geben Aufschluss darüber, ob und wie oft, von wem und wann zuletzt individuelle Übersetzungseinheiten wiederverwendet wurden. Dies ermöglicht es etwa, alte/nicht (mehr) genutzte Übersetzungseinheiten zu löschen und so übermäßig große TMs gezielt zu bereinigen, um Performanceproblemen entgegenzuwirken. Darüber hinaus können entsprechende Informationen auch dazu genutzt werden, gezielte Qualitätssicherung für häufig wiederverwendete Übersetzungseinheiten zu betreiben.

3 Datenquellen identifizieren

Nachdem die Zielsituation definiert und die erforderlichen Ressourcen erstellt wurden, gilt es, Quellen für existierende Übersetzungsdaten zu identifizieren. Hier sind natürlich zuallererst vorhandene TMs relevant, die bereits in der eigenen Organisation oder von Übersetzungsdienstleistern verwendet und (idealerweise) gepflegt werden. Darüber hinaus können jedoch auch andere Daten von Interesse sein, insbesondere wenn vorhandene TMs unvollständig oder von schlechter Qualität sind. Alternative Möglichkeiten, um an Übersetzungsdaten zu gelangen, sind mehrsprachige Dokumente, XLIFF-Dateien aus vergangenen Übersetzungsprojekten oder auch mehrsprachige Exporte aus inhaltsführenden Systemen, wie etwa Content-Management- (CMS), Produktinformationsmanagement- (PIM) und Enterprise-Resource-Planning-Systemen (ERP).

4 Daten exportieren und migrieren

Nachdem alle relevanten Datenquellen identifiziert wurden, müssen die Daten im nächsten Schritt in ein einheitliches Format gebracht werden. Mit Blick auf die Verwendung von Übersetzungsdaten in CAT-Tools bedeutet dies konkret, dass diese in das proprietäre Format des jeweiligen Tools überführt werden müssen.

Das genaue Vorgehen dabei hängt von Art und Umfang der Übersetzungsdaten ab: So können etwa XLIFF-Dateien aus vergangenen Projekten in den CAT-Editor importiert und darüber in ein neues TM gespeichert werden. In mehreren

Sprachen verfügbare Dokumente, aber u. U. auch mehrsprachige Exporte aus inhaltsführenden Systemen erfordern hingegen ein aufwendigeres Alignment. Der vermeintlich einfachste Weg, bestehende Übersetzungsdaten zu migrieren, ist der Austausch über den Translation-Memory-eXchange-Standard (TMX). In den folgenden Unterkapiteln werden dieser Standard sowie seine Umsetzung durch unterschiedliche CAT-Tool-Hersteller und die daraus entstehenden Herausforderungen bei Datenmigrationen skizziert.

4.1 Der TMX-Standard

Der TMX-Standard hat die Aufgabe, den Datenaustausch zwischen Übersetzungssystemen unterschiedlicher Hersteller zu erleichtern. Entwickelt und veröffentlicht wurde er von der *Fachgruppe Open Standards for Container/Content Allowing Reuse* (OSCAR) der Localization Industry Standards Association (LISA). Nach einer ersten Version aus dem Jahr 1998 erfolgten mehrere Überarbeitungen, bis zur Veröffentlichung der aktuellen Version des Standards, TMX 1.4b, im **Jahr2005** (Localization Industry Standards **Association2005**).

Der TMX-Standard basiert auf der Extensible Markup Language (XML), einer Auszeichnungssprache zur Darstellung hierarchischer Datenstrukturen in Textform, die sowohl menschen- als auch maschinenlesbar ist.[2]

Eine TMX-Datei muss dabei mindestens die in Abb. 2 dargestellten Elemente enthalten:

- Die XML-Deklaration.
- Das Root-Element <*tmx*> inkl. der Version.
- Einen Header mit Metadaten über die TMX-Datei. Diese geben an:
  - mit welchem Tool (<*creationtool*>) in welcher Version (<*creationtoolversion*>) die Datei erstellt wurde,
  - welche Systemsprache in dem Tool eingestellt war (<*adminlang*>),
  - welches Dateiformat das Translation Memory hatte, auf dessen Grundlage die TMX-Datei erstellt wurde (<*o-tmf*>),
  - welche Art von Daten die TMX-Datei enthält (<*datatype*>),
  - wie die darin enthaltenen Übersetzungen segmentiert sind, bspw. satz- oder absatzbasiert (<*segtype*>),

[2] Für einen Überblick über den Einsatz von XML in Übersetzung in Lokalisierung siehe Savourel (2001).

  - welche Ausgangssprache sie hat (<*srclang*>).

- Das <*body*>-Element mit den Translation Units (<*tu*>), die die eigentlichen Übersetzungsdaten enthalten.

Figure 2: TMX-Beispieldokument mit verpflichtenden Elementen (in Anlehnung an ebd.)

Neben diesen verpflichtenden Elementen können TMX-Dateien eine ganze Reihe weiterer Elemente enthalten: darunter eine Reihe von Elementen, die im TMX-Standard definiert sind, wie etwa Angaben von wem und wann eine Übersetzungseinheit erstellt oder geändert wurde (<*creationid*> und <*creationdate*> bzw. <*changeid*> und <*changedate*>).

Darüber hinaus können jedoch auch sogenannte Properties (<*prop*>) enthalten sein, die nicht Teil des Standards sind, sondern vom jeweiligen Tool vorgegeben werden. Die Schwierigkeiten, die sich daraus für den Austausch von Übersetzungsdaten zwischen unterschiedlichen Tools ergeben, werden im folgenden Kapitel beschrieben.

4.2 TMX-Kompatibilität vs. -Interoperabilität

Idee und Ziel hinter der Entwicklung des TMX-Standards sind eindeutig: den Austausch von Übersetzungsdaten über verschiedene Installationen bzw. auch unterschiedliche Systeme hinweg zu ermöglichen. Betrachtet man die große Zahl an Übersetzungssystemen, die TMX implementiert haben,[3] so scheint dieses Ziel erreicht. Allerdings: So „nahtlos“, wie der Austausch häufig dargestellt wird (siehe etwa **Chan2015**; **Roturier2020**), ist er bei genauerem Hinsehen nicht. Tatsächlich basieren die Austauschformate vieler CAT-Tools zwar auf TMX, die Hersteller gehen bei der Umsetzung des Standards jedoch eigene Wege. Die dabei entstehenden TMX-„Dialekte“ sind mehr oder weniger interoperabel, jedoch bei weitem nicht voll miteinander kompatibel.[4]

Grob lässt sich hierzu Folgendes festhalten: Die in einem TM gespeicherten Übersetzungen lassen sich typischerweise – auch über verschiedene Systeme hinweg – problemlos(er) per TMX austauschen (wobei es auch hier Inkompatibilitäten geben kann, die zu Matchverlusten führen, etwa in Bezug auf die Auszeichnung von Tags!). Anders sieht es hingegen mit den Metadaten aus, die Informationen zu diesen Übersetzungen bereitstellen.

[3] Für eine Liste mit Beispielen siehe **Chan2015**.

[4] Toolspezifische Dialekte und die damit verbundenen Schwierigkeiten betreffen neben TMX auch andere Datenaustauschformate wie etwa TBX (TermBase eXchange) (siehe **Wright2018**).

Dies betrifft zunächst beschreibende Metadaten, die von unterschiedlichen Tools unterschiedlich umgesetzt werden. So gibt es:

- TM-Systeme, die eine feste Auswahl an Feldern für beschreibende Metadaten vorgeben. Dies trifft beispielsweise auf Phrase zu, wo es die Felder *Client*, *Business Unit*, *Domain*, *Subdomain* und *Note* gibt.
- TM-Systeme, die feste Felder für beschreibende Metadaten vorgeben, zusätzlich aber auch benutzerdefinierte Felder zulassen. Dies trifft etwa auf memoQ zu. Neben den Standardmetadatenfeldern *Client*, *Project*, *Domain* und *Subject* lassen sich in memoQ-TMs für eine feingliedrigere Auszeichnung benutzerdefinierte Felder anlegen.
- TM-Systeme, in denen für die Speicherung beschreibender Metadaten benutzerdefinierte Felder angelegt werden müssen. Dies ist beispielsweise bei Trados Studio der Fall.

Konkret bedeutet dies, dass bei der Konsolidierung von Übersetzungsdaten aus verschiedenen Systemen die beschreibenden Metadatenfelder aufeinander gemappt werden müssen, wobei es – etwa im Fall von Phrase – Einschränkungen in Bezug auf die Zahl verfügbarer Felder geben kann.

Ein weiteres eindrückliches Beispiel für durch TMX-Dialekte entstehende Herausforderungen ist die Speicherung von Fließtextkontexten in memoQ- und Trados-TMs: Während memoQ sowohl das vorgehende als auch das nachfolgende Segment als Kontext speichert (in den Property-Elementen des Typs *x-context-pre* bzw. *x-context-post*, siehe Abb. 3), speichert Trados Studio nur das vorhergehende Segment (in Property-Elementen des Typs *x-ContextContent*, siehe Abb. 4).

Figure 3: Speicherung von Fließtextkontext in memoQ-TMX-Dateien

Figure 4: Speicherung von Fließtextkontext in Trados-TMX-Dateien

Das bedeutet: Während (in dieser Hinsicht!) eine Migration von TM-Daten aus memoQ in Trados Studio keine besonderen Maßnahmen erfordert, führt im umgekehrten Fall ein einfacher Import von TMX-Dateien aus Trados Studio dazu, dass alle Segmente ihren Fließtextkontext verlieren. Da für Kontext-Matches in memoQ beide Kontexte im TM gespeichert sein müssen, würde dies bedeuten, dass alle potenziellen Kontext-Matches nur noch mit 100 % matchen – was zu erheblichen Mehrkosten führen kann (siehe Abb. 5 und 6)!

Figure 5: Match-Werte in memoQ mit unverändertem TMX-Import aus memoQ

Figure 6: Match-Werte in memoQ mit unverändertem TMX-Import aus Trados Studio

Dieses Beispiel zeigt bereits, dass ein einfaches Importieren von TMX-Dateien aus anderen CAT-Tools u. U. massive Matchverluste nach sich ziehen kann. Eine Migration von Übersetzungsdaten über verschiedene Systeme hinweg ist damit sinnvoll allein mit CAT-Tools nicht möglich. Vielmehr werden je nach beteiligten Systemen und Migrationsrichtung spezielle Migrationsskripte benötigt, die Inkompatibilitäten zwischen TMX-Dialekten so weit wie möglich ausgleichen.

5 Daten analysieren

Nachdem alle Übersetzungsdaten in das Format des Zielsystems gebracht wurden, kann die Analyse beginnen. Die in diesem Zusammenhang erforderlichen Analysen umfassen quantitative wie qualitative Aspekte, die in den folgenden Unterkapiteln vorgestellt werden. Eine nachhaltige, effiziente Nutzung und Pflege von Translation Memorys setzt nämlich nicht nur standardisierte Prozesse, sondern auch und vor allem explizit definierte Qualitätskriterien für die darin enthaltenen Daten voraus. Im Rahmen von Harmonisierungsprojekten ist eine Bewertung vorhandener Übersetzungsdaten anhand entsprechender Kriterien umso wichtiger, als diese Daten typischerweise aus unterschiedlichen Quellen stammen und unterschiedlich verwaltet wurden.

5.1 Quantitative Analysen

Zu den quantitativen Aspekten bei der Analyse von Übersetzungsdaten zählen unter anderem folgende Frage:

- Wie viele Sprachenkombinationen gibt es?
- Wie viele TMs gibt es für diese Sprachenkombinationen jeweils?
- Wie viele Übersetzungseinheiten sind jeweils in diesen enthalten?
- Welche Inhalte sind in diesen TMs gespeichert?
- Welche beschreibenden Metadatenfelder sind in den TMs zu finden?

- Welche Metadatenwerte enthalten diese Felder?
- Wie viele Übersetzungseinheiten sind jeweils mit diesen Metadatenwerten ausgezeichnet?

Bereits hier zeigt sich eine der grundlegenden Hürden: So lassen sich Antworten auf grundlegende Fragen, wie die nach Sprachkombinationen, Anzahl und Größe der TMs, zwar mithilfe eines CAT-Tools ermitteln – viel weiter allerdings reicht deren Funktionsumfang typischerweise nicht!

So ist zwar etwa anhand der TM-Definition in memoQ oder Trados Studio ersichtlich, welche Metadatenfelder in einem TM vorhanden sind. Welche Werte aber in diesen Feldern existieren, ist hingegen bei Freitextfeldern mit diesen Tools ebenso wenig zu ermitteln wie die Zahl der Übersetzungseinheiten, die damit jeweils ausgezeichnet wurden.

Auch hier müssen also zusätzliche Tools eingesetzt werden – seien es XML-Editoren, in denen Werte manuell ermittelt und gezählt werden, oder dedizierte TM-Analyseskripte.

5.2 Qualitative Analysen

Neben rein quantitativen Analysen muss im Rahmen von Migrationsprojekten und TM-Pflege auch die Qualität der Übersetzungsdaten analysiert werden. Typische Fehler und Mängel lassen sich dabei in drei Kategorien einordnen (vgl. Zielinski/**Varga2020**, S. 301–306):

- Fehler/Inkonsistenzen auf Segmentebene
- Fehler/Inkonsistenzen auf Subsegmentebene
- Fehler/Inkonsistenzen in den Metadaten

5.2.1 Fehler und Inkonsistenzen auf Segmentebene

Auf Segmentebene können u.a. folgende Aspekte die Qualität und Performance von TMs negativ beeinflussen:

- unübersetzte Ausgangssegmente, die per Copy & Paste in die Zielsprache übertragen wurden,
- unvollständige Übersetzungen von Ausgangssegmenten (etwa bei nicht bearbeiteten Fragment-Matches),

- Dubletten,
- Segmente, die nur aus Tags, Sonderzeichen, Hex-Codes etc. bestehend, die durch fehlerhafte Dateiimporteinstellungen entstehen,
- Inhalte in abweichenden Sprachen, die etwa durch den fehlerhaften Import mehrsprachiger Projektdateien entstehen.

Problematisch sind auf Segmentebene weiterhin Inkonsistenzen, die potenziell zu einer Vielzahl fast identischer Segmente führen können. Moorkens (2015) unterscheidet in diesem Zusammenhang drei Szenarien, nämlich:

- inkonsistente Ausgangssegmente, die inkonsistent übersetzt werden (n:n),
- konsistente Ausgangssegmente, die inkonsistent übersetzt werden (1:n) und
- inkonsistente Ausgangssegmente, die konsistent übersetzt werden (n:1).

Mögliche Gründe für derartige Inkonsistenzen sind vielfältig: So wurden in der Vergangenheit vielleicht TM-Daten aus unterschiedlichen Quellen ohne entsprechende Aufarbeitung aggregiert. Weitere Ursachen können fehlerhafte TM-Update-Einstellungen oder Inkonsistenzen beim Import der Ausgangstexte sein. Abb. 7 verdeutlicht dieses Problem anhand des Beispiels einer Datei mit eingebetteten HTML-Inhalten. Je nachdem, wie die Ausgangsdateien für die Übersetzung importiert werden, ändern sich auch die dabei entstehenden Übersetzungseinheiten im TM:

- Im ersten Fall werden die HTML-Inhalte korrekt verarbeitet, sodass der Text korrekt segmentiert wird und keine überflüssigen Tags importiert werden.
- Im zweiten Fall werden die Tags mithilfe regulärer Ausdrücke geschützt, sodass sie importiert werden und keine korrekte Segmentierung erfolgen kann.
- Im dritten Fall wird der Text inklusive aller Tags als übersetzbare Inhalte importiert. Neben der fehlerhaften Segmentierung besteht hier das Risiko einer versehentlichen Bearbeitung der Tags.

Figure 7: Dubletten aufgrund abweichender Importeinstellungen bei Dokumenten mit eingebetteten HTML-Inhalten in memoQ

Die potenziellen Auswirkungen dieser und anderer Qualitätsprobleme auf Segmentebene hängen vom geplanten Verwendungszweck der Übersetzungsdaten ab: Bei der Verwendung als Übersetzungshilfe in CAT-Tools erhöhen TMs mit einer Vielzahl an Dubletten und nicht bzw. inkonsistent übersetzten Segmenten die kognitive Belastung bei Auswahl und Überprüfung von TM-Treffern und wirken sich negativ aus Usability, Produktivität und die Übersetzungsqualität insgesamt aus (vgl. O'**Brien2007**; **Wolff2016**). Aus technischer Perspektive können aufgeblähte TMs sich außerdem negativ auf Performance und Geschwindigkeit auswirken und den Wert der Übersetzungsdaten für Training und Customization maschineller Übersetzungsmodelle mindern (vgl. Khayrallah/**Koehn2018**; **OttEtAl2018**).

5.3 Fehler und Inkonsistenzen auf Subsegmentebene

Inkonsistenzen auf Segmentebene sind häufig auf Qualitätsmängel in den Segmenten selbst zurückzuführen, etwa in Bezug auf:

- Rechtschreibung, Grammatik und Zeichensetzung
- die Verwendung inkonsistenter oder veralteter Terminologie
- Verstöße gegen Styleguide-Regeln
- Sprach- oder Rechtschreibvarianten (belgisches vs. französisches Französisch, Genitivvarianten im Deutschen (*-es* vs. *-s*)
- Inlineformatierung und Tags
- nichtdruckbare Zeichen (einfaches vs. geschütztes vs. schmales Leerzeichen etc.)
- abweichende Formatierung von Zahlen, Daten, Zeit- und Maßangaben
- Auslassungen oder Hinzufügungen

Hinzu kommt, dass über lange Zeiträume hinweg genutzte TMs häufig Übersetzungseinheiten mit an sich veralteten Inhalten enthalten.

Aus dieser Liste möglicher Qualitätsprobleme – sowohl auf Segment- als auch auf Subsegmentebene – wird schnell deutlich, welche Herausforderungen mit der Analyse von Übersetzungsdaten verbunden sind. Es zeigen sich außerdem bereits hier deutlich die Unzulänglichkeiten von CAT-Tools, die nur über sehr eingeschränkte TM-Management-Funktionen verfügen.

6 Daten bereinigen

Die qualitative Analyse von Übersetzungsdaten geht Hand in Hand mit ihrer Bereinigung. Es gibt dabei je nach Anforderungen und zu erwartendem Aufwand unterschiedliche Möglichkeiten, fehlerhafte TM-Einträge zu korrigieren.

6.1 TM-Bereinigungstools und -funktionen

Clientbasierte CAT-Tools wie memoQ, Trados Studio oder Across bieten integrierte TM-Editoren, in denen die Bearbeitung von Übersetzungsdaten im jeweils eigenen proprietären Format möglich ist. Der Funktionsumfang dieser Lösungen ist jedoch typischerweise sehr gering, sodass diese eher für die punktuelle Korrektur bekannter Fehler als für Qualitätsanalysen über große Datenmengen hinweg nutzbar sind.

So bietet etwa der integrierte TM-Editor der aktuellen memoQ-Version 11.1 folgende Optionen:

- Filtern von TM-Einträgen (auf Grundlage von Volltextsuchen in Ausgangs- und/oder Zielsprache, administrativen/beschreibenden Metadaten oder Inlinetags)
- Auf-/absteigende Sortierung der Ergebnisse (alphabetisch oder auf Grundlage der Segment-ID oder des Datums der letzten Bearbeitung)
- Bearbeitung von Metadaten im Batch-Verfahren
- Suchen und Ersetzen von Zeichenfolgen in Ausgangs- oder Zieltext (an beliebiger Stelle oder nur als ganze Wörter bzw. ganze Segmente)
- Löschen aller/bestimmter Tags

Wie diese Liste zeigt, zielen diese Optionen eher darauf ab, bestimmte Segmente gezielt herauszufiltern und zu bearbeiten als umfängliche Qualitätssicherung zu betreiben. Für umfangreichere Überprüfungen, etwa von Terminologie- und Rechtschreibung, sind somit externe Tools erforderlich. Hier gibt es unterschiedliche Möglichkeiten:

- Clienbasierte QA-Tools mit Unterstützung für TM-Dateien wie ErrorSpy, ApSIC Xbench und QA Distiller
- Cloudbasierte Tools wie myproof (http://www.glossa.de/de/dienstleistungen/glossa-myproof.html#myproofplatform)

- TM-Bereinigungsskripte in Python oder anderen Programmiersprachen (siehe etwa **Barbu2015**; **BarbuEtAl2016**; **NegriEtAl2017**; **SabetEtAl2016**)

Dezidierte QA-Anwendungen wie diese verfügen typischerweise über einen deutlich höheren Funktionsumfang im Vergleich zu integrierten TM-Editoren in CAT-Tools. So bietet etwa ApSIC Xbench in seiner aktuellen Version 3.0 u. a. Funktionen zum Auffinden von nicht oder inkonsistent übersetzten Segmenten, Terminologiefehlern, Differenzen zwischen Zahlen, Tags, Sonderzeichen etc. in Ausgangs- und Zielsegment und Wortwiederholungen.

6.2 TM-Bereinigung mit Large Language Models

Large Language Models (LLMs) bieten heute eine Reihe neuer Möglichkeiten bei der Pflege und Bereinigung von Übersetzungsdaten. Ein Anwendungsfall, der dies eindrücklich zeigt, ist die Bereinigung von Terminologiefehlern. Klassische Terminologieprüfungen dienen lediglich dem Auffinden von Segmenten mit potenziellen Terminologiefehlern. Eine automatische Korrektur (oder auch nur eine halbautomatische mittels Suchen und Ersetzen) war bei Terminologiefehlern jedoch bislang in den allermeisten Fällen nicht möglich. Problematisch waren in diesem Zusammenhang sowohl flektierte Formen der Benennung als auch abhängige Wörter wie Adjektive, Artikel, Pronomen etc. Eine einfache Ersetzung von Benennungen hätte damit in der Vergangenheit unweigerlich zu Folgefehlern geführt.

LLMs bieten nun erstmals die Möglichkeit, nicht nur inkorrekte Benennungen zu ersetzen, sondern auch alle erforderlichen grammatischen Anpassungen in deren Umfeld vorzunehmen – also etwa a) Artikel, b) Flexionsendungen und c) Pronomen zu ändern (siehe Abb. 8).

Figure 8: Terminologiekorrektur mit ChatGPT (Zielinski/**Varga2024**)

Ob LLMs dabei als Ersatz oder Ergänzung klassischer Qualitätsprüfungen eingesetzt werden, hängt nicht zuletzt von wirtschaftlichen Überlegungen ab: Je nach Kosten des verwendeten Modells und Menge der zu prüfenden Daten kann es so etwa vorteilhaft sein, zunächst mit klassischen Methoden problematische Segmente zu identifizieren und nur diese dann gezielt mithilfe von LLMs korrigieren zu lassen.

Unabhängig davon stellt sich die Frage, wie automatisch per LLM korrigierte Übersetzungseinheiten in Produktionsprozesse integriert werden. Unter bestimmten

Voraussetzungen ist eine manuelle Überprüfung der angepassten Übersetzungseinheiten denkbar, bevor diese eingesetzt werden: dann etwa, wenn die die erforderlichen Ressourcen vorhanden, die Datenmengen überschaubar und ihre Kritikalität hoch ist. Eine weitere Möglichkeit ist die Auszeichnung der geänderten Segmente mit entsprechenden Metadatenwerten, die dann in den TM-Einstellungen mit einer Penalty belegt werden können. Dadurch ist gewährleistet, dass automatisch geänderte Übersetzungseinheiten in Übersetzungsprojekten nicht als Kontextmatches erscheinen, sondern auf jeden Fall vom Übersetzungsdienstleister geprüft und ggf. bearbeitet werden.

7 Daten annotieren und harmonisieren

Nach Analyse und Bereinigung der Übersetzungsdaten müssen diese anhand des neuen Metadatenschemas annotiert und harmonisiert werden. Ein wichtiger Schritt ist in diesem Zusammenhang die Ergänzung fehlender beschreibender Metadatenwerte. Diese können mithilfe unterschiedlicher Ansätze ermittelt werden, entweder anhand vorhandener Metadaten oder aber auf Grundlage der in den Übersetzungseinheiten enthaltenen Texten selbst.

Die Kreuzung von Metadatenwerten kann dabei auf unterschiedliche Weise erfolgen. So lassen etwa beschreibende Metadaten wie Projekt- oder Dateinamen potenziell Rückschlüsse auf Content-Typen oder Fachgebiete zu. Gleiches gilt u. U. auch für administrative Metadaten, wenn z. B. ein bestimmter Content-Typ ausschließlich von einem festen Übersetzungsdienstleister bearbeitet wurde.

Letztendlich sind die Möglichkeiten in diesem Bereich durch die vorhandenen Metadaten vorgegeben. Sind aus diesen keine Rückschlüsse auf die Inhalte von Übersetzungsdaten möglich, so besteht unter bestimmten Umständen auch die Möglichkeit, die Segmente selbst auf ihre Zugehörigkeit zu relevanten Kategorien hin zu untersuchen, und zwar anhand der der darin verwendeten Terminologie. Hierzu ein einfaches Beispiel: Liegen TMs mit Übersetzungen vor, die von verschiedenen Abteilungen – z. B. Marketing und Technische Dokumentation – in Auftrag gegeben wurden, können diese potenziell durch die darin vorkommenden Benennungen voneinander unterschieden werden. Voraussetzung hierfür ist natürlich zunächst, dass es zwischen den Inhalten beider Abteilungen terminologische Unterschiede gibt; darüber hinaus muss eine Terminologiedatenbank mit entsprechend gelabelten Einträgen existieren (bzw. angelegt werden!).

8 Daten anreichern

Wie in den bisherigen Kapiteln deutlich wurden, befasst sich Übersetzungsdatenmanagement nicht nur mit der Verwaltung und Pflege von TMs, sondern auch mit flankierenden Ressourcen und Prozessen, die sich unmittelbar auf die Nutzbarkeit der Übersetzungsdaten auswirken – also etwa Segmentierungsregeln und Non-Translatables. Eine weitere zentrale Tätigkeit ist in diesem Zusammenhang die Generierung von Übersetzungsdaten. Ob und in welchem Umfang eine solche erforderlich sein kann, hängt vom Zweck und den Zielen ab, zu denen die Übersetzungsdaten eingesetzt werden sollen.

In klassischen Übersetzungsworkflows etwa ist der Wert solcher synthetischer Daten gegenüber „organischen" TM-Treffern eher gering einzustufen: Einzelne Fuzzy-Matches genügen hier, um eine konsistente Übersetzung neuer Inhalte zu ermöglichen. Anders sieht es jedoch im Bereich MÜ-Training/-Customization bzw. MÜ-Output-Customization aus. Hier können automatisch generierte Übersetzungseinheiten die Qualität des MÜ-Outputs verbessern. Interessant ist diese Möglichkeit vor allem für Anwendungsfälle, in denen authentische Trainingsdaten nicht in ausreichender Menge vorhanden sind – was aus Sicht von Organisationen mit spezifischer Corporate Language eher die Regel als die Ausnahme darstellt.

Es gibt eine Reihe von Möglichkeiten, synthetische Trainingsdaten zu generieren, von denen an dieser Stelle zur Illustration zwei kurz genannt werden sollen: die sogenannte Back-Translation und die Umformulierung existierender Übersetzungseinheiten.

Back-Translation bezeichnet die Generierung zweisprachiger Trainingsdaten durch die maschinelle Übersetzung zielsprachlicher Inhalte in die gewünschte Ausgangssprache (**EdunovEtAl2018**; Sennrich/Haddow/**Birch2016**). Das Ergebnis sind korrekte Zielsegmente mit einem oder mehreren (potenziell fehlerhaften!) maschinell übersetzten Ausgangssegmenten (Imamura/**Sumita2019**). In diesem Zusammenhang können nicht nur monolinguale Daten (also nicht übersetzte Texte) verwendet werden, sondern auch die Zielsegmente bereits existierender Übersetzungseinheiten, zu denen alternative Ausgangssegmente generiert werden.

Die Umformulierung bestehender Übersetzungseinheiten wiederum kann sowohl auf Ausgangs- als auch auf Zielsegmente angewendet werden. Die konkreten Möglichkeiten sind hier vielfältig: So können etwa negierte Varianten von Sätzen erstellt werden (Wetzel/**Bond2012**). Ein weiterer Ansatz besteht darin, Entitäten, wie etwa Personen- und Städtenamen, durch andere Entitäten derselben Kategorie zu ersetzen und so verschiedene Varianten derselben Übersetzungsein-

heit zu erstellen (Winter/**Zielinski2020**). Auch hier ermöglichen LLMs unterschiedlichste Transformationen, ohne dass hierfür aufwendige Methoden und Skripte entwickelt werden müssen – im einfachsten Fall mithilfe eines simplen Prompts, der bestehende Segmente rein syntaktisch variiert.

Unabhängig vom gewählten Ansatz müssen synthetische Daten in nachgelagerten Prozessen eindeutig identifizierbar bleiben, indem sie etwa in separaten TMs gespeichert, mindestens aber mit entsprechenden Metadaten ausgezeichnet und mit einer Penalty belegt werden. Ob und wie sie in klassische Übersetzungsprozesse eingebunden werden, hängt von der eingesetzten MÜ-Technologie ab: Bei Training/Customization von MÜ-Modellen sind sie nicht als Ressource in die eigentliche Übersetzung eingebunden. In den letzten Jahren zeichnet sich für viele Anwendungsfälle allerdings ein anderer Trend ab: Anstatt aufwendig MÜ-Modelle zu trainieren oder zu customizen, wird mithilfe sogenannter „adaptiver" MÜ der Output des Modells selbst auf Grundlage der zur Verfügung stehenden Referenzübersetzungen angepasst (mittlerweile typischerweise mithilfe von LLMs!).

9 Daten implementieren

Nach Abschluss ihrer Aufbereitung können die Übersetzungsdaten dann in die relevanten Systeme und Prozesse integriert werden, im einfachsten Fall etwa durch Import in lokale bzw. server- oder cloudbasierte TMs. Werden die Daten hingegen (auch) an anderer Stelle benötigt, bspw. für MÜ-Training/-Customization, müssen ggf. vorab geeignete Austauschformate und -routinen definiert werden.

10 Daten überwachen und pflegen

Übersetzungsdatenmanagement ist ein kontinuierlicher, aktiver Prozess. Nach ihrer initialen Aufbereitung und Bereitstellung müssen Qualität und Performance der Daten überwacht werden. Hierfür bieten sich APIs an, über die ein kontinuierliches Monitoring durchgeführt werden kann, ohne dass die Daten aus laufenden Prozessen genommen werden müssen. Prinzipiell kommen auch hier etwa die in Kap. **??** vorgestellten Analysedimensionen zur kontinuierlichen Qualitätssicherung in Frage.

Grundsätzlich sind jedoch Qualitätssicherungsmaßnahmen an der Quelle, also in laufenden Übersetzungsprojekten, die effektivsten Garanten für die Qualität von Übersetzungsdaten. Das Augenmerk sollte also auch und vor allem auf der Fehlervermeidung in den Übersetzungsprozessen selbst liegen – u. a. durch die in

Kap. ?? genannten Maßnahmen. Dadurch werden nicht nur Qualitätsprobleme in den Übersetzungsdaten, sondern auch in den eigentlichen mehrsprachigen Inhalten vermieden – schließlich ist ein Fehler in den Übersetzungsdaten immer auch ein Fehler an anderer Stelle!

11 Ausblick

In den letzten Jahren haben Übersetzungsdaten zunehmend an Aufmerksamkeit gewonnen. Schließlich sind sie nicht mehr nur reine Übersetzungshilfen in von außen typischerweise unsichtbaren Prozessen, sondern können auch für andere Zwecke eingesetzt werden.[5] Wie bei anderen Daten auch hängt ihre Eignung für diese intendierten Einsatzzwecke – ihre *fitness for use* (**Wang1998**) – dabei von explizit zu definierenden Qualitätskriterien ab, für deren Einhaltung ein aktives Datenmanagement unerlässlich ist. Ein solches setzt, wie aus den vorangegangenen Kapiteln deutlich wurde, eine ganze Reihe von Maßnahmen und entsprechenden Kompetenzen voraus, sei es auf technischer, organisatorischer oder prozessualer Ebene.

Die Ausführungen hierzu bezogen sich in diesem Kapitel auf den aktuell vorherrschenden technischen Standard für die Speicherung von Übersetzungsdaten, nämlich in Form fein segmentierter Übersetzungseinheiten in Translation Memorys. Dieser Ansatz, Inhalte auf möglichst kleine, wiederverwendbare Einheiten herunterzubrechen, ist auch in der Content-Erstellung seit Jahren Standard, wie sich etwa an Component-Content-Management-Systemen (CCMS) in der technischen Redaktion deutlich erkennen lässt. Dabei handelt es sich aber keineswegs um den einzigen Ansatz: So halten Tools wie STAR Transit Übersetzungsdaten in Form von Referenzdokumenten ohne Feinsegmentierung vor. War dieser Ansatz bislang die Ausnahme, so könnten die aktuellen technologischen Entwicklungen ihn in Zukunft aus seinem Nischendasein holen: Die neuesten Generationen von Übersetzungssystemen sind mittlerweile nicht nur in der Lage, wesentlich umfassendere Kontexte einzubeziehen, als diese durch klassische Segmentgrenzen vorgegeben sind – sie liefern auch potenziell bessere Ergebnisse, wenn ihnen mehr Kontext zur Verfügung steht! Dies ist der Fall bei LLMs, aber auch bei NMÜ. Dies wird zunehmend zu Nachteilen führen, wenn etwa Texte aus CAT-Tools weiterhin segmentweise statt im Kontext maschinell übersetzt werden (siehe Abb. 9 und 10):

[5] Die Erwartungen an ihr Monetisierungspotenzial, die im Zuge der Verbreitung trainierbarer NMÜ-Modelle aufkamen, scheinen sich jedoch – zumindest bislang – nicht zu bewahrheiten. So wurde etwa der von der Translation Automation User Society (TAUS) im **Jahr2020** geschaffene Data Marketplace, auf dem Übersetzungsdaten gehandelt werden konnten, **Anfang2024** wieder eingestellt.

Figure 9: Maschinelle Übersetzung segmentierter Texte in memoQ mit DeepL (Zielinski/**Varga2024**)

Figure 10: Maschinelle Übersetzung unsegmentierter Texte im DeepL-Webeditor (ebd.)

Unberührt von einem möglicherweise bevorstehenden Paradigmenwechsel bleibt die Erkenntnis, dass datengetriebene Anwendungen und Prozesse nur so gute Ergebnisse liefern können, wie die Daten es zulassen. Für Hochschulen in der Übersetzerausbildung scheint also eine stärkere Einbeziehung von Aspekten wie *Data Literacy* (**Krüger2022**), Engineering und (Daten-)Prozessmanagement geboten, damit ihre Studierende auch lernen, mit Datenmengen umzugehen, die mit klassischen Übersetzungsstrategien nicht zu bewältigen sind.