Formale und inhaltliche Prüfschritte zur Korpusveröffentlichung

Reiner Konrad, Uta Salden

DOI (jeweils letzte Version): 10.25592/uhhfdm.838

Versionen:

Version 1	2017-12-31	Erste Fassung
Version 2	2018-07-02	div. Korrekturen
Version 3	2021-02-08	Konsolidierung mit der englischen Fassung

Abstract

Dieses Arbeitspapier beschreibt die Prüfschritte, die vor der Veröffentlichung der Videos und dazugehörigen Annotationen vorgenommen wurden mit dem Ziel der Qualitätssicherung. Diese Prüfschritte sind nach inhaltlichen Aspekten zusammengefasst.

Inhalt dieses Arbeitspapiers

Das Öffentliche DGS-Korpus[1] enthält ca. 50 Stunden annotierte DGS-Texte. Bei der Auswahl wurden vor allem Aufgaben (Gesprächsformate) aus der Erhebung berücksichtigt, die die Erfahrungen und Erlebnisse gehörloser Menschen und die Gehörlosenkultur widerspiegeln. Ebenso wurden Dialoge und freie Konversation bevorzugt. Neben den Videos werden die deutsche Übersetzung der DGS-Texte, die Lemmatisierung (in Form von Glossen) sowie die Mundbild-/Mundgestik-Annotationen veröffentlicht. Weiterhin sind die DGS-Texte mit Schlagwörtern annotiert, um einen thematischen Zugriff zu ermöglichen. Übersetzungen sowie Glossen und Schlagwörter sind ins Englische übersetzt. Die Daten sind auf verschiedenen Webseiten unterschiedlich aufbereitet, um den Erwartungen verschiedener Nutzergruppen besser gerecht zu werden (Näheres s. Jahn et al. 2018).

Zwischen 2015 bis 2017 wurden verschiedene Prüfschritte eingeführt, die die Qualität der zu veröffentlichenden Übersetzungen und Annotationen sicherstellen sollen. Diese Prüfschritte werden im Folgenden zu inhaltlichen Gruppen zusammengefasst und genauer beschrieben.

Prüfschritte

Nicht zu veröffentlichende Sequenzen (Schwärzungen)

Nach Abschluss der Erhebungen erhielten die Informanten ihre Videoaufnahmen mit der Bitte, diejenigen Stellen zu notieren, die sie nicht veröffentlicht sehen möchten.[2] Diese Sequenzen

[1] Siehe MEINE DGS (https://meine-dgs.de/ bzw. DOI 10.25592/dgs.meinedgs) und MEINE DGS – annotiert (https://ling.meine-dgs.de/ bzw. DOI 10.25592/dgs.corpus).

[2] Insgesamt wünschten 76 Informanten die Schwärzung bestimmter Stellen. Der Großteil lag unterhalb einer Minute. Die meisten Stellen waren für die Forschung freigegeben. Nur von zwei Informanten erhielten wir

wurden im Transkript getaggt. Weiterhin wurden die Übersetzungs-Tags entsprechend angepasst mit dem Ziel, durch die Schwärzung möglichst wenig Material für die Veröffentlichung zu verlieren, da alle Sequenzen, die ein Schwärzungs-Tag enthalten oder mit diesem überlappen, von der Veröffentlichung ausgeschlossen werden. Der Abschluss dieser Prüfung wird durch das Metadatum *TK_07: Schwärzungen* zu den für die Veröffentlichung ausgewählten Subtasks dokumentiert.
Weiterhin konnten Annotatoren und MitarbeiterInnen Stellen markieren, die sie für die Veröffentlichung als nicht geeignet erachteten. Diese Sequenzen wurden im Team diskutiert und über deren Schwärzung entschieden. In nur wenigen Fällen haben wir uns dazu entschlossen, den Informanten/die Informantin zu fragen. Passagen wurden vorsichtshalber geschwärzt, wenn deren Veröffentlichung möglicherweise negative Konsequenzen für einen Informanten oder die Sprachgemeinschaft haben könnte, wenn über Dritte unangemessen erzählt wird oder wenn Äußerungen ethisch fraglich interpretiert werden könnten. Bei allen Schwärzungen, egal ob durch die Informanten oder das Projekt vorgenommen, wurde überprüft, ob die Segmentierung der betroffenen Übersetzungs-Tags entsprechend angepasst werden kann, damit möglichst wenig Material geschwärzt werden muss.
Vormerkungs-Tags mit dem Wert „für Teilkorpus[3] irrelevant“ wurden von der Veröffentlichung ausgeschlossen. Das Prüfen dieser Sequenzen und das Anlegen von Schwärzungs-Tags wird mit dem Metadatum *TK_15: Vormerkungen* dokumentiert.

Rechtschreibung

Zur Unterstützung der Rechtschreibprüfung bei der Alignierung der deutschen Übersetzungen wurde eine automatische Rechtschreibprüfung in iLex integriert, so dass fehlerhafte oder fragliche deutsche Texte markiert werden und ggf. sofort geändert werden können. Ist die Alignierung abgeschlossen – erkennbar am Wert des Metadatums *Bearbeitungstand Übersetzung* beim Subtask [4] –, so werden die deutschen Übersetzungen auf formal korrekte Schreibung gegengelesen. Dieser Prüfschritt wird durch ein weiteres Metadatum dokumentiert (*TK_01: Rechtschreibprüfung deutsch*).
Ebenso gibt es für die englischen Texte eine automatische Rechtschreibprüfung, die bei der Eingabe fehlerhafte oder fragliche Schreibungen markiert. Nach Abschluss der englischen Übersetzungen werden auch diese Texte korrekturgelesen, der Subtask erhält danach das Metadatum *TK_02: Rechtschreibprüfung englisch* mit dem Wert „abgeschlossen“.
Die Mundbild-Annotationen werden in regelmäßigen Abständen korrekturgelesen. Dazu werden die ab einem bestimmten Zeitpunkt neu angelegten Mundbilder mithilfe einer SQL-Abfrage extrahiert und korrekturgelesen. Alle Fehler werden direkt in iLex wiederum über Abfragen korrigiert.
Die deutschen und englischen Glossennamen werden ebenfalls korrekturgelesen, neu hinzugekommene Glossen in regelmäßigen Abständen überprüft.

Thematischer Zugriff

Beim Alignieren der Übersetzungen legen die Annotatoren ein Tag mit Stichworten an, die die wesentlichen Inhalte des DGS-Textes des betreffenden Subtasks zusammenfassen. Diese Angaben bilden die Grundlage für eine systematische Verschlagwortung mithilfe eines kon-

---

insgesamt keine Freigabe zur Veröffentlichung der Videos, eine Informantin gab ihre Aufnahmen nur für die Forschung frei.

[3] Teilkorpus = Öffentliches DGS-Korpus.

[4] Als Subtask bezeichnen wir die DGS-Sequenz(en), in denen die beiden Informanten die ihnen gestellte Aufgabe (Task) bearbeiten, nachdem sie die Instruktionen erhalten und evtl. ein oder mehrere Stimuli ausgewählt haben. Eine gestellte Aufgabe, im öffentlichen DGS-Korpus auch Gesprächsformat genannt, kann einen oder mehrere Subtasks enthalten. Nur diese Sequenzen wurden übersetzt, annotiert und für die Veröffentlichung aufbereitet. Der Bearbeitungsstand der Übersetzung, der Transkription sowie verschiedener Prüfschritte wurde durch entsprechende Metadaten an diesen Subtasks dokumentiert.

trollierten Vokabulars (insgesamt 564 Schlagworte). Diese Schlagworte sind 35 Oberbegriffen zugeordnet, wobei auch Mehrfachzuordnungen vorgenommen wurden. So ist z.B. das Schlagwort *Rente* sowohl dem Oberbegriff *Arbeit und Beruf* als auch *Finanzen* zugeordnet. Die Oberbegriffe enthalten die 26 Sachgebiete, die bei der Erhebung verwendet wurden: *Arbeit und Beruf, Behörden, DDR, Energie und Umwelt, Essen und Trinken, Familie und Verwandte, Fest – Feier – Party, Finanzen, Gefühle, Haus und Wohnung, Kino – Theater – Museum – Kunst, Kleidung und Mode, Kommunikation, Körperpflege – Gesundheit – Krankheit, Natur, Partnerschaft und Beziehung – Liebe und Sexualität, Politik, Recht und Gesetz, Religion, Schule und Unterricht, Sport und Spiel, Stadt und Land, Urlaub –Freizeit – Reise, Verkehr, Wetter, Wirtschaft.* Diese Liste wurde um 9 Begriffe erweitert, um alle Schlagwörter abdecken zu können: *Do-It-Yourself, Gebärdensprache, Gehörlosenkultur, Gesellschaft, Information und Unterhaltung, Lebensabschnitt, Schwangerschaft und Geburt, Terror –Krieg – Katastrophen, Wissenschaft und Technik.* Anhand dieser Oberbegriffe können im Öffentlichen DGS-Korpus (https://meine-dgs.de/) Texte nach Themen ausgewählt werden. Die Stichwörter werden bei der Auswahl eines DGS-Textes unter der Rubrik *Themen* aufgelistet, wobei der Oberbegriff vorangestellt ist wie z. B. *Gebärdensprache: Fingeralphabet, Gebärdensprachdozent*.
Die Schlagwörter und Oberbegriffe wurden ins Englische übersetzt und korrekturgelesen. Ist die systematische Zuordnung der inhaltlichen Stichwörter zu Schlagwörtern abgeschlossen, wird das Metadatum *TK_05: Verschlagwortung* angelegt, danach werden diese Schlagwörter verzeitet. Unterhalten sich die Informanten z.B. über eine gewisse Zeit ausschließlich über ein Thema, dann wird dieser Abschnitt mit dem entsprechenden Schlagwort versehen. Springen die Informanten aber zwischen verschiedenen Themen, dann wird der gesamte Abschnitt mit mehreren Schlagwörtern versehen. Für den Abschluss dieses Arbeitsschritts gibt es das Metadatum *TK_06: Schlagwortverzeitung*. Ein zusätzlicher Kontrollschritt ist die Überprüfung, dass Schlagwort-Tags nur innerhalb eines Subtasks liegen und diesen vollständig abdecken.

Inhaltliche Überprüfung und Überarbeitung

Im normalen Arbeitsablauf findet eine Konsistenzprüfung zwischen Transkription und deutscher Übersetzung statt, bevor mit der englischen Übersetzung begonnen wird. Dabei wird die deutsche Übersetzung mit der Lemmatisierung und den annotierten Mundbildern abgeglichen (Metadatum *TK_04: Abgleich Lemmatisierung – Übersetzung*). Dieser Prüfschritt ist entscheidend für die inhaltliche Konsistenz der Annotationen und besonders wichtig, wenn viele verschiedene Personen daran beteiligt sind.

Anonymisierung

- Im Zuge der Anonymisierung müssen ggf. Referenzen auf dritte Personen im Video, in der Lemmatisierung und in den Übersetzungen entfernt werden. Folgende Kriterien wurden dabei angewandt:
  - Namen von Dritten werden anonymisiert, wenn über den privaten Bereich gesprochen wird.
  - Nicht anonymisiert werden hingegen Namen von Dritten, wenn sie eine Funktion innehaben (z.B. im Verein oder Theater) und wenn positiv über sie gesprochen wird.
  - Nicht anonymisiert werden Namen von bekannten Personen, auch wenn negativ über sie gesprochen wird. Eine Person gilt dann als bekannt, wenn sie über die Bundeslandgrenzen hinaus bekannt ist.
  - Anonymisiert werden ferner Ortsnamen, wenn aus dem Ort (z.B. kleines Dorf) direkt auf eine dritte Person geschlossen werden kann.

Hierzu werden die Übersetzungen geprüft, die per Named Entity Recognition (s. Bleicken et al. 2016) dafür markiert worden sind, dass sie Namen enthalten, und von Hand entschieden, ob die automatisch vorgeschlagene Anonymisierung angewandt werden soll oder nicht. Entsprechend wird geprüft, ob auch die Lemmatisierung anonymisiert werden muss. Damit

werden zu schwärzende Bereiche im Video in ihrer zeitlichen Ausdehnung festgelegt, für die automatisch ein Schwärzungsbereich in Bildkoordinaten berechnet wird (wie für aus anderen Gründen – s.o. – zur Schwärzung markierte Passagen).
Ferner werden die Verwendungen von Namensgebärden daraufhin überprüft, ob sie durch die eben beschriebene Vorgehensweise schon anonymisiert sind. Ggf. wird dies nachgeholt.
Das Metadatum *TK_08: Namen* zeigt an, dass die manuelle Sichtung der Anonymisierungsvorschläge abgeschlossen ist.
Über eine Liste werden in iLex die automatisch erstellten Schwärzungsbereiche zur manuellen Überprüfung angeboten. Diese werden direkt im Datensatz als geprüft markiert.

Tag-Grenzen

Die automatisch eingelesenen Übersetzungen werden typischerweise in iLex auf mehrere Tags aufgeteilt. Dabei wird häufig die zeitliche Ausdehnung korrigiert, jedoch nicht so fein, dass die Übersetzungen mitten in einer Gebärde enden könnten. Weitere Prüfschritte bestehen deshalb darin, Übersetzungen gegen Subtask-Grenzen, Token- oder Mundbild-Tags zu testen, um die Übersetzungstag-Grenzen geeignet anzupassen. Ferner wird geprüft, ob sich Übersetzungs-Tags der beiden Informanten überlappen, ohne dass tatsächlich beide gleichzeitig gebärden. Auch in diesem Fall werden die Grenzen geeignet verschoben.
Der Abschluss der entsprechenden Arbeitsschritte wird durch Vergabe der Metadaten *TK_16: Tag-Grenzen: Subtask-Übersetzung*, *TK_17: Tag-Grenzen: Übersetzung-Token/MB* bzw. *TK_18: Überlappungen Übersetzungs-Tags* markiert.

Konsistenz

Mithilfe von SQL-Abfragen werden Passagen ermittelt, bei denen die Längenverhältnisse deutsche Übersetzung vs. englische Übersetzung und deutsche Übersetzung vs. Anzahl der Gebärden auffällig sind. Diese werden daraufhin überprüft, ob die Übersetzung bzw. die Lemmatisierung vollständig ist.
Weitere Abfragen werden benutzt, um vor der Produktion die Einhaltung der Glossierungskonventionen zu überprüfen. (Nicht alle in der Datenbank verwendeten Glossen sollen in veröffentlichten Annotationen erscheinen.)

Technische Qualität

Auf Basis der Trims (die zur Veröffentlichung vorgesehenen Videoabschnitte der Subtasks) wird geprüft, ob zu veröffentlichendes Material Bildstörungen etc. enthält. Betroffene Passagen werden, sofern Kürzungen etc. nicht in Frage kommen, als solche markiert.

Literatur

Jahn, Elena / Konrad, Reiner / Langer, Gabriele / Wagner, Sven / Hanke, Thomas (2018). Publishing DGS Corpus Data: Different Formats for Different Needs. In Bono, Mayumi et al. (Hrsg.): Proceedings of the LREC2018 8th Workshop on the Representation and Processing of Sign Languages: Involving the Language Community. 11th International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan, 12.5.2018. Paris, Frankreich: European Language Resources Association (ELRA), S. 83–90. URL: https://www.sign-lang.uni-hamburg.de/lrec/pub/18018.html.