BMBF-Verbundprojekt
„**QUEST**: Quality – Established: Erprobung und Anwendung von Kurationskriterien und Qualitätsstandards für audiovisuelle, annotierte Sprachdaten“

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung

Curation of Interpreted Corpora on the example of ComInDat[1]

By Elena Arestau

[1] https://corpora.uni-hamburg.de/hzsk/de/islandora/object/spoken-corpus:comindat

Table of Contents

1. Interpreted Corpora on the example of ComInDat

On the one hand, this document provides an overview of the curation of interpreted corpora on the example of ComInDat and presents the main steps that I believe are important for the curation and reuse of such resources. On the other hand, this document serves as recommendation that should be considered by archives, data centres and people who need to manage such data.

Community Interpreting Corpora contain audio and/or video recordings of various types of community interpreted discourse (e.g. medical interpreting (doctor-patient communication, simulated doctor-patient communication) and legal interpreting (e.g. courtroom communication) conducted with the help of the consecutive or simultaneous interpreting or both.

1.1 Transcription, Annotation

When creating interpreted corpora, researchers requested a more precise description of the transcription conventions used and the documentation of them. This criterion is useful for further analysis of the corpus.

There are several transcription systems available to researchers with important theoretical variations. CHAT, HIAT and LIDES coding manual (LIPPS Group 2000) are well-known standards in the community.

The transcription in EXMARaLDA-Format can be checked, whether there is a particular transcription convention. Otherwise, this will be indicated as an error. The list of the extended conventions: CHAT, HIAT, LIDES, GAT.

Furthermore, the annotation applied to the corpus and the tier structure should be described including the annotation of the multilingual phenomena.

Regarding the language change phenomena, the following summary can be done:

- No agreed classification for the phenomena of language change. (different theoretical backgrounds)
- The process of annotation of language change is heterogenic: code switching can be annotated in many different modes: mostly as Code-Switching (further CS), cs, hybrid, mixed-language words, Wechsel, foreign und indigenous. This phenomena appears or in a commentary tier, language tier or in the separate tier for the cs.

To sum up, it can be said that there is no generally accepted annotation of code switching in the community. Moreover, different concepts have been used for the annotation of CS-tiers.

As far as the translation is concerned, there is also heterogeneity. Furthermore there are different concepts how languages can be defined in a corpus (cf. Myers-Scotton 2002 (14) and the Matrix-Language-Frame-Modell or the annotation model of Vaillant und Léglise (2014), where the „turn-taking“ and the „speech turns“ are the important elements for the transcription of the data). In this model, it being considered that „the utterance is multilingual and composed of several segments“(Léglise / Alby 2016: 7).It is important for the ComInDat Corpora that the data are consistently translated. The corpus should be at least translated in more widely accessible languages. The annotations in such a case should match source and target pairs. If the translation tier is named, it can be checked, whether this tier is consistent all over the corpus.

For the curation of ComInDat Corpora are also important the languages involved in the interaction. This criterion should describe according to common conventions in the research community the languages involved in the interaction.

All the languages in the corpus should be annotated. If the languages are annotated, it is possible to check the consistency of the annotation.

1.2 Metadata

Metadata of the ComInDat corpora depends on the corpus type, purpose and the public status of the given corpus. (cf. Meyer 2010)

Metadata describing the corpus

This set of metadata describes the event and discourse context / types, qualifications of interpreters and preparation of interpreters, spontaneity index.

Metadata describing speaker(s) and interpreters

This set of metadata gives information about speakers, the role in the communication, the status of the language(s), regional variety of the language,

For the interpreters: gender, level of expertise, native language, language combination.

Metadata describing speech event

This set of metadata describes the communication event, its location, languages, mode, and topic, the quality of sound and image, setting, speed of delivery.

Metadata describing translation status

The translation languages (original language and translation language, interpreted language) should be documented and the translation conventions status, translation modality and translation mode. When possible dialectal variation within one language should also be described.

Metadata describing language of an utterance

The status of the languages should be indicated: source and target language and information of the language affiliation of each utterance.

2. References

Bauer, Bruno; Ferus, Andreas; Gorraiz, Juan; Gründhammer, Veronika; Gumpenberger, Christian; Maly, Nikolaus; Mühlegger, Johannes Michael; Preza, José Luis; Sánchez Solís, Barbara; Schmidt, Nora; Steineder, Christian (2015): *Forschende und ihre Daten. Ergebnisse einer österreichweiten Befragung – Report 2015*. Version 1.2. DOI: 10.5281/zenodo.32043. Online auch unter: http://phaidra.univie. ac.at/o:407513.

Fandrych, Christian / Elena Frick / Hanna Hedeland / Anna Iliash / Daniel Jettka / Cordula Meißner / Thomas Schmidt / Franziska Wallner / Kathrin Weigert / Swantje Westpfahl (2016), „User, whoartthou? User Profiling for Oral Corpus Platforms“. In: Nicoletta Calzolari / Khalid Choukri / Thierry Declerck / Sara Goggi / Marko Grobelnik (Hrsg.): *Proceedings of the Tenth International Conference on Language Resources and Evaluation* (LREC 2016), Portorož, Slovenia. Paris: European Language Resources Association (ELRA), 280-287.

Gardner-Chloros P., Moyer M., Sebba M. (2007) Coding and Analysing Multilingual Data: The LIDES Project. In: Beal J.C., Corrigan K.P., Moisl H.L. (eds) Creating and Digitizing Language Corpora. Palgrave Macmillan, London. https://doi.org/10.1057/9780230223936_5

Neuroth, Heike / Ortgiese, Michael (eds.)(2018): *Umfrage zum Forschungsdatenmanagement an der FH Potsdam. Projektbericht* Potsdam: Verlag der Fachhochschule Potsdam.

Angermeyer, P. / Meyer, B. / Schmidt, T. (2012). Sharing Community Interpreting Corpora: A pilot study, in: Schmidt, T. / Wörner, K. (eds.) *Multilingual Corpora and Multilingual Corpus Analysis*. Amsterdam: Benjamins, 275-294.

Bergmann, Anka / Caspers, Olga /Stadler, Wolfgang (Hg.) (2018): *Didaktik der slawischen Sprachen – Beiträge zum 1. Arbeitskreis in Berlin* (12.–14.9.2016) Innsbruck: University press.

Bührig, Kristin / Kliche, Ortrun / Meyer, Bernd &Pawlack, Birte (2012): The corpus “Interpreting in Hospitals” – possible applications for research and communication training, in: Schmidt, Thomas / Wörner, Kai (eds): *Multilingual Corpora and multilingual corpus analysis*, Hamburg Studies on Multilingualism, volume 14. Amsterdam: Benjamins.

Deuchar, M. / P. Davies / J. Herring / M. ParafitaCouto / D. Carter (2014). Building bilingual corpora, in: E. M. Thomas / I. Mennen (Eds.): *Advances in the Study of Bilingualism*, Bristol: Multilingual Matters.

Empfehlungen des DFG-Fachkollegiums 104 “S*prachwissenschaften*„ Stand: 31 Oktober 2019. <https://www.dfg.de/download/pdf/foerderung/grundlagen_dfg_foerderung/informationen_fachwissenschaften/geisteswissenschaften/standards_sprachkorpora.pdf>

Fandrych, Christian /Meißner, Cordula/ Wallner, Franziska (Hg.) (2017): *Gesprochene Wissenschaftssprache – digital. Verfahren zur Annotation und Analyse mündlicher Korpora.* Tübingen: Stauffenburg.

Gusev, Valentin / Klooster, Tiina (2018): “INEL Kamas Corpus.” Version 0.1. Publication date 2018-12-31. http://hdl.handle.net/11022/0000-0007-CAE6-2. Archived in Hamburger ZentrumfürSprachkorpora, in: Wagner-Nagy, Beáta / Arkhipov, Alexandre / Ferger, Anne / Jettka, Daniel / Lehmberg, Timm (eds.): *The INEL corpora of indigenous Northern Eurasian languages*.

Hedeland, Hanna / Lehmberg, Timm / Schmidt, Thomas / Wörner, Kai (2014): Multilingual Corpora at the Hamburg Centre for Language Corpora, in: *Best Practices for Speech Corpora in Linguistic Research*. Cambridge Scholars Publishing.

Heegård Petersen, Jan / GertFoget Hansen / Jacob Thøgersen / Karoline Kühl. (Ahead of print). *Linguistic proficiency in immigrant and heritage speakers of Danish in Argentina and North America: A quantitative approach*. Corpus Linguistics and Linguistic Theory.

Johannessen, Janne Bondi (2015): The Corpus of American Norwegian Speech (CANS), in: BéataMegyesi (ed.): *Proceedings of the 20th Nordic Conference of Computational Linguistics,* NODALIDA 2015, May 11-13, 2015, Vilnius, Lithuania. NEALT Proceedings Series 23.

Léglise, Isabelle / Alby, Sophie (2016): *Plurilingual corpora and polylanguaging, where corpus linguistics meetscontact linguistics*. Sociolinguistic Studies, 10 (3).

Lleó, Conxita (2011): "*ALCEBLA*." Archived in Hamburger Zentrum für Sprachkorpora. Version 0.1. Publicationdate 2011-06-30.

Lüdi, Georges (2011): Neue Herausforderungen an eine Migrationslinguistik im Zeichen der Globalisierung, in: Stehl, Thomas (eds.): *Sprachen in mobilisierten Kulturen; Aspekte der Migrationslinguistik*, Universitätsverlag. Potsdam. (=MobilisierteKulturen 2).

MacWhinney, Brian (2000): *The CHILDES Project: Tools for Analyzing Talk*. Volume 1: Transcription format and programs. NJ: Lawrence Erlbaum Associates.

Meyer, Bernd (2010): "*Consecutive and Simultaneous Interpreting (CoSi)*." Archived in Hamburger ZentrumfürSprachkorpora. Version 1.1. Publication date 2010-02-26.

Myers-Scotton, Carol (2002): *Contact Linguistics: Bilingual Encounters and Grammatical Outcomes*. Oxford, UK: Oxford University Press.

Rehbein, Jochen / Schmidt, Thomas / Meyer, Bernd / Watzke, Franziska &Herkenrath, Annette (2004): Handbuch für das computergestützte Transkribieren nach HIAT, in: *Arbeiten zur Mehrsprachigkeit*, Folge B 56. Hamburg: Universität Hamburg.

Rehbein, I. / Schalowski, S. / Wiese, H. (2014): The KiezDeutschKorpus (KiDKo) Release 1.0., in: *Proceedings of the 9th International Conference on Language Resources and Evaluation* (LREC), May 24-31, 2014. Reykjavik, Iceland.

Rehbein, J.; Schmidt, T.; Meyer, B.; Watzke, F. &Herkenrath, A. (2004). Handbuch für das computergestützte Transkribieren nach HIAT. In: Arbeiten zur Mehrsprachigkeit, Folge B 56. https://talkbank.org/manuals/CHAT.pdf

Schmidt, Thomas / Wörner, Kai (eds.) (2012): Multilingual Corpora and Multilingual Corpus Analysis. Amsterdam: Benjamins.Schmidt, Thomas / Wörner, Kai / Hedeland, Hanna /Lehmberg, Timm (2013): *Leitfaden zur Beurteilung von Aufbereitungsaufwand und Nachnutzbarkeit von Korpora gesprochener Sprache*, Mannheim, Institut für Deutsche Sprache.

Schmidt, Thomas / Schütte, Wilfried /Winterscheid, Jenny (2015): *cGAT. Konventionen für das computergestützte Transkribieren in Anlehnung an das Gesprächsanalytische Transkriptionssystem 2 (GAT2),* Mannheim: Institut für Deutsche Sprache.

Schütte, Wilfried (2013): Metadaten für Gesprächsdatenbanken: ein Überblick und ihre Verwaltung in der IDS-Datenbank Gesprochenes Deutsch (DGD), in: Kratochvílová, Iva/Wolf, Norbert Richard (Hrsg.): *Grundlagen einer sprachwissenschaftlichen Quellenkunde.*Tübingen: Narr (Studien zur Deutschen Sprache 66).

Selting, Margret et al. (2009): Gesprächsanalytisches Transkriptionssystem 2 (GAT 2), in: *Gesprächsforschung – Online-Zeitschrift zur verbalen Interaktion* 10

The LIPPS Group (2000): *The LIDES Coding Manual*. A document for preparing and analyzing language interaction data. Version 1.1 - July, 1999. In: International Journal of Bilingualism, 4, Nr. 2

Thomason, S (2010): Contact Explanations in Linguistics, in: Hickey, R. (ed.): *The Handbook of Language Contact* 31-47. Wiley-Blackwell.

Vaillant, P. / Léglise, I. (2014): *A la croisée des langues: Annotation etfouille de corpus plurilingues,* RNTI Revue des Nouvelles Technologies de l'Information. Hermann: Paris. 81-100.

Vila, M. / González, S. / Martí, M. A. / Llisterri, J. / Machuca, M.J. (2010). '*ClInt: a Bilingual Spanish-Catalan Spoken Corpus of Clinical Interviews*'. Pendent de publicació a Procesamiento del Lenguaje Natural,45. pp. 105-111. València (Spain).