Lang*Reg: Collecting and processing comparable multi-lingual data for a variety of communicative interactions

Keywords: register variation, language comparison, normalization, syntactic annotations

One of the major challenges in cross-linguistic naturalistic data collection is the simultaneous monitoring of multiple dimensions of variation. We will present a data collection protocol used for the open-access Lang*Reg corpus (Adli et al. 2023), which provides an exemplary corpus incorporating intra-individual variation across multiple situations in five languages, including indigenous and minority languages, across three language families – German (Indo-European), Persian, Southern Kurdish (Indo-European); Yucatec Maya (Mayan); and Javanese (Austronesian). The language sample has been selected in order to investigate the role of cross-linguistic and cross-cultural aspects in register variation. Participants of each language traversed a course of recording situations in which they were asked to produce language following two tasks: a) talking freely with various kinds of interlocutors (friend, stranger, taxi driver or professor) and b) telling a story to a friend orally and graphically (Lehmann et al. 2025). The recorded situations aim at cross-cultural validity, meaning they occur naturally in a diverse set of cultures and languages, varying by the parameters (i) hierarchy vs. equality, (ii) distance vs. closeness, and also (iii) mode, and (iv) monologue vs. dialogue (see e. g. Halliday & Hasan 1989; Neumann 2014).

A further challenge concerns the processing of such multi-lingual data and ways to ensure maximum comparability, starting with decisions on a common script (e. g. Latin vs. Arabic for Persian and Kurdish). We will discuss the challenges and solutions that emerged during data processing and annotation of the Lang*Reg corpus. In a first step, all data received a close Latin script based transcription in collaboration with native speakers, using a basic syntactic segmentation, i. e. one matrix clause and all its dependent clauses, in ELAN (The Language Archive, 2022) with separate tiers per speaker. A syntactic segmentation supports later syntactic annotations such as dependencies as well as analyses but it requires a good understanding of the language syntax. Moreover, it depends on a rigorous documentation of decisions across languages and collaborators, for example with respect to the handling of repetitions, terminations, ellipses and apo koinou constructions.

A normalization layer was added on top of the close transcription to facilitate search and automatic annotation processes by adapting the orthography to the standard spelling or conventionalized practice, thereby reducing contractions (e. g. German *kannste* > *kannst du* 'can you') and phonological variants, for example not representing stem final coda weaking [l] ↛ [j] in Yucatec. Such a normalization is non-trivial as languages in Lang*Reg are at different stages of spelling standardizations, and none exists for the phonologically based Latin script used for Persian and Kurdish.

On the basis of the normalized layer, semi-automated processes were applied for some of the languages to obtain a gloss layer (using Fieldworks Explorer 2021), a part of speech layer and dependency annotations (using UDPipe 2016). Further, more specific annotations for concrete research questions are being added using Inception (Klie et al. 2018), with further challenges e. g. concerning annotating non-overt material and dealing with referential distance tracking across two speakers. All this also involves a series of file formats and conversions, for which we used inherent export functions and the conversion tool Annatto (Krause & Klotz 2024). We will share our workflows and report on some of the issues we encountered in this regard.

References

Adli, Aria & Verhoeven, Elisabeth & Lehmann, Nico & Mortezapour, Vahid & Vander Klok, Jozina (eds.). 2023. *Lang*Reg: A multi-lingual corpus of intra-speaker variation across situations. Version 0.1.0.* [Data set]. Zenodo. Berlin, Köln: Humboldt-Universität zu Berlin, Universität zu Köln. https://doi.org/10.5281/zenodo.7646320

FLEx. 2021. *Fieldworks explorer: Software tools for language and cultural data, with support for complex scripts. version 9 [computer software].* Dallas,TX: SIL International. https://software.sil.org/fieldworks/.

Halliday, Michael A. K. & Hasan, Ruqaiya. 1989. *Language, Context, and Text: Aspects of Language in a Social-Semiotic Perspective.* Oxford: Oxford University Press.

Klie, Jan-Christoph & Bugert, Michael & Boullosa, Beto & de Castilho, Richard Eckart & Gurevych, Iryna. 2018. The INCEpTION platform: Machine-assisted and knowledge-oriented interactive annotation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations.* 5–9. Association for Computational Linguistics.

Krause, Thomas & Klotz, Martin. 2024. Annatto. https://github.com/korpling/annatto/.

Lehmann, Nico & Mortezapour, Vahid & Klok, Jozina Vander & Farokhnejad, Zahra & Müller, David & Verhoeven, Elisabeth & Adli, Aria. 2025. Lang*Reg corpus: Documenting intra-speaker variation across languages and registers. *Language Documentation & Conservation* 19. 40–66. https://hdl.handle.net/10125/74810.

Neumann, Stella. 2014. *Contrastive Register Variation: A Quantitative Approach to the Comparison of English and German.* Berlin: De Gruyter Mouton.

Straka, Milan & Hajic, Jan & Straková, Jana. 2016. UDPipe: Trainable pipeline for processing CoNLL-U files performing tokenization, morphological analysis, POS tagging and parsing. In Calzolari, Nicoletta & Choukri, Khalid & Declerck, Thierry & Goggi, Sara & Grobelnik, Marko & Maegaard, Bente & Mariani, Joseph & Mazo, Helene & Moreno, Asuncion & Odijk, Jan & Piperidis, Stelios (eds.), *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016).* Paris, France: European Language Resources Association (ELRA).

The Language Archive, . 2022. *Elan.* Nijmegen: Max Planck Institute for Psycholinguistics. https://archive.mpi.nl/tla/elan.