CROSS-LINGUISTIC DATA FORMATS
WORKSHOP

14 - 15 DECEMBER 2023, MPI-EVA

Book of Abstracts

Workshop on Cross-Linguistic Data Formats 2023 –
Graphs and Text

Organized by

Robert Forkel
Johann-Mattis List
Annika Tjuka

CLLD apps as a tool for the construction of datasets

Sascha Alexeyenko <sascha.alexeyenko@uni-goettingen.de>
University of Göttingen

A common use case of the CLDF + CLLD architecture implies that there is an existing dataset which gets converted into the CLDF format such that it can then be optionally served by a suitably configured CLLD app. Differently from this, this talk will discuss an alternative scenario in which a CLLD app is created in order to be used as an organization and visualization tool *during* the construction of a dataset. More specifically, I will report some of the experiences with the CLDF + CLLD architecture in the process of creation of the *Nominal Modification Database* (NMDB), which will be described in some detail below. I will also discuss some potential changes in the setup of the default CLLD app, which may make it easier for linguists to use it for the construction of datasets.

The Nominal Modification Database, which is currently under construction and will become gradually available under https://nmdb.uni-goettingen.de, aims to provide systematic information about the types of adnominal modifiers existing in the languages of the world, the word order possibilities available to them, as well as some further related information (in total, there are 50 parameters at the moment). The database collects information about a wide range of nominal modifiers: non-deverbal modifiers (adjectives or adjective-like items), deverbal modifiers including clausal ones (relative clauses), and adpositional modifiers (adnominal PPs). In all cases, it pays special attention to the possibility for modifiers to be complex, i.e. it contains information as to whether the head of the modifier can take further dependents and if so which linear orders are permitted in this case. It is for this reason that the construction of this database is a long-term project: the information about complex adnominal modifiers is rarely available in reference grammars, which makes it necessary to collect this data from native speakers or researchers with access to native speakers. The goal is to get the relevant information about at least one language per language family according to the Glottolog classification (https://glottolog.org). In this context, the CLDF + CLLD architecture provides a great support for a long-term data collection as it provides a relational database structure with a comfortable visualization possibility (the app), which already in its default version is attuned to linguistic data. In the talk, I will also discuss some potential further adjustments, which may make it even easier for linguists to use the architecture for this purpose.

Extending CLDF to multilingual data

Jeff Good (jcgood@buffalo.edu)
University at Buffalo and Humboldt-Universität zu Berlin

Linguistic research typically prioritizes studies where data collection is conducted at the level of a "language" despite the fact that individual-level multilingualism is a widespread phenomenon that is also a foundational feature of the sociolinguistic dynamics of many communities. Moreover, data from language use is often multilingual in nature, which means that annotating its content properly requires a means of associating different stretches of language use for different linguistic varieties within a single text.

The linguistic literature has already documented many of the complexities involved in working with multilingual data. This includes the need for more detailed metadata about actors and events than is typically collected for projects focused on a single language and the fact that the annotation strategies required to analyze multilingual text data also need special attention (see, e.g., Good 2022 for an overview from the perspective of documentary linguistics). Multilingual data also raises potential difficulties in the domain of language identification since, in some cases, a morpheme or stretch of discourse may be ambiguous with respect to the language being used and be classifiable as belonging to multiple languages (see, e.g., Cobbinah et al. 2017). Issues like these are especially visible in work on the multilingualism of small-scale societies, the study of which is making increasing contributions to diversity linguistics (Pakendorf et al. 2021).

This presentation will provide an overview of the data encoding complications associated with multilingual data and consider what kinds of extensions to existing CLDF might be needed for it to encode such data effectively. General proposals will include: (i) adapting and extending existing metadata standards, such as IMDI (Broeder & Wittenburg 2006), so that they are expressible in CLDF and updated to reflect our current understanding of the metadata needs for multilingual data, (ii) developing a standardized means of defining new languoids that allows for flexibility in annotation of data for language choice while also linking it as accurately as possible to Glottolog languoids, perhaps using general purpose semantic vocabularies like SKOS (Miles & Bechofer 2009), and (iii) the need for a standard for encoding textual data that allows for the annotation of arbitrary stretches of text for the language being used, rather than assuming that these will correspond to the subsets of text that need to be identified for traditional morphosyntactic annotation. Achieving the last goal, however, is in tension with the principle that CLDF data should be editable by hand.

References

Broeder, Daan & Peter Wittenburg. 2006. The IMDI metadata framework, its current application and future direction. *International Journal of Metadata, Semantics and Ontologies* 1(2). 119–132.

Cobbinah, Alexander, Abbie Hantgan, Friederike Lüpke & Rachel Watson. 2017. Carrefour des langues, carrefour des paradigmes. In Margaret Bento Michelle Auzanneau & Malory Leclère (eds.), *Espaces, mobilités et éducation plurilingues: Éclairages d'afrique ou d'ailleurs*, 79–97. Paris: Édition des Archives Contemporaines.

Good, Jeff. 2022. Adapting methods of language documentation to multilingual settings. *Journal of Language Contact* 15:341–375. https://doi.org/10.1163/19552629-15020006.

Miles, Alistair & Sean Bechhofer. 2009. *SKOS: Simple Knowledge Organization System reference*. W3C Recommendation. http://www.w3.org/TR/skos-reference.

Pakendorf, Brigitte, Nina Dobrushina & Olesya Khanina. 2021. A typology of small-scale multilingualism. *International Journal of Bilingualism* 25. 835–859. https://doi.org/10.1177/ 13670069211023137.

Areal colexification and partial colexification in northern Australia

John Mansfield, University of Zurich
Ruth Singer, University of Melbourne

Recent years have seen exciting developments in areal semantics, that is the study of how semantic patterns characterise geographic and cultural areas, and can be distinguished from the shared inheritance of language families (Koptjevskaja-Tamm & Liljegren 2017; Schapper & Koptjevskaja-Tamm 2022; François 2022). This work has focused on lexical semantics, enabled by the colexification methodology proposed by François (2008) and subsequently implemented in the CLICS database (List et al. 2019). Australia is one region that has been suggested to have areal lexical semantic patterns (Dixon 1980), which in the north of the continent may spread across distantly related or unrelated languages (Evans 1992; Evans & Wilkins 2000; Schapper et al. 2016). In this presentation I outline some work-in-progress on northern Australian colexification, and raise questions about how best to implement this study in the CLDF data framework.

Northern Australian languages share colexifications to a degree that suggests areal diffusion. This includes some that are relatively rare, e.g. YEAR≈RAIN.SEASON, and some that have been shown to have an areal distribution, such as FIRE≈TREE (Schapper et al. 2016). Cultural diffusion, as opposed to shared lexical inheritence, is suggested by instances such as YEAR≈RAIN.SEASON being found in neighbouring languages with unrelated lexemes, for example *thangku* (Murrinhpatha) and *warri* (Marri Tjevin). These might also be considered 'partial colexifications', where only part of the lexical form is retained in a conceptual distinction. These languages make pervasive us of nominal classification systems, which allows one to further distinguish concepts such as *da thangku* 'time of the rainy season, one year', with a TIME classifier, versus *kura thangku* 'heavy rain (in the rainy season)' with a WATER classifier. Partial colexification has been identified as an important direction for further research in global semantics (List 2023), and may indeed constitute an iceberg of semantic relations with full colexification as the visible tip.

Table 1 shows some examples of how nominal classifiers are used to express concepts in four northern Austrlian languages, from the distinct subregions of Daly and Arnhem. Only the classifiers are shown here, and not the specific lexemes as in *ku murrurrbe* 'bird' (Murrinhpatha). Notice that the two Daly languages have perfect alignment of classifiers to concepts in this sample. The two Arnhem languages have some alignment, though less so, and they show little alignment with the Daly. This suggests that northern Australian lexical semantics may pattern according to several subregions. Further research will have to untangle areal and phylogenetic relationships. This will be especially challenging since the phylogenies are themselves elusive (Evans 2003), but focusing on cognacy among the classifier forms may make this more tractable.

Our classifier research follows interesting work on 'grammatical gender' alignment among Indo-European languages (McCarthy et al. 2020), which shows that these semantically bleached classification systems have a strongly phylogenetic distribution. We may hypothesise that more semantically robust nominal classification systems should show more areal diffusion, in line with the previous findings on areal semantics. This study will present several challenges for coding data in a way that is both practical and theoretically grounded. For example, will substring methods (List 2023) be sufficient to analyse classifier systems in the same CLDF format as colexification data? Or should we annotated linguistic structure in our lexical datapoints, thus departing from existing data conventions?

Table 1. Examples of noun class (mis-)alignment in northern Australian languages

	Daly		**Arnhem**	
	Murrinhpatha	Marri Tjevin	Wubuy	Mawng
BIRD	*ku* (ANIM)	*awu* (ANIM)	*ngarra-* (NEUT)	*na-* (MASC)
FISH	*ku* (ANIM)	*awu* (ANIM)	*ngarra-* (NEUT)	*na-* (MASC)
WASP	*ku* (ANIM)	*awu* (ANIM)	*ana-* (FEM)	*niny-* (FEM)
FIREWOOD	*thungku* (FIRE)	*tjendji* (FIRE)	*ngarra-* (NEUT)	*ma-* (VEG)
FIRESTICK	*thungku* (FIRE)	*tjendji* (FIRE)	*ngarra-* (NEUT)	*niny-* (FEM)
ALCOHOL	*kura* (LIQ)	*wudi* (LIQ)	*ngarra-* (NEUT)	*nung-* (LAND)

References

Dixon, R.M.W. 1980. *The languages of Australia*. Cambridge: Cambridge University Press.

Evans, Nicholas. 1992. Multiple semiotic systems, hyperpolysemy, and the reconstruction of semantic change in Australian languages. In Kellermann, G. (ed.), *Diachrony within Synchrony: Language History and Cognition; papers from the International Symposium at the University of Duisburg 26-28 March 1990*, 475–508. Frankfurt: Peter Lang.

Evans, Nicholas. 2003. Comparative non-Pama-Nyungan and Australian historical linguistics. In Evans, Nicholas (ed.), *The non-Pama-Nyungan Languages of Northern Australia*, 3–25.

Evans, Nicholas & Wilkins, David. 2000. In the mind's ear: The semantic extensions of perception verbs in Australian languages. *Language*. Linguistic Society of America 76(3). 546–592. (doi:10.2307/417135)

François, Alexandre. 2008. Semantic maps and the typology of colexification: Intertwining polysemous networks across languages. In Vanhove, Martine (ed.), *From Polysemy to Semantic Change: Towards a typology of lexical semantic associations* (Studies in Language Companion Series), 163–215. John Benjamins Publishing Company. (doi:10.1075/slcs.106.09fra)

François, Alexandre. 2022. Lexical tectonics: Mapping structural change in patterns of lexification. *Zeitschrift für Sprachwissenschaft*. De Gruyter 41(1). 89–123. (doi:10.1515/zfs-2021-2041)

Koptjevskaja-Tamm, Maria & Liljegren, Henrik. 2017. Semantic patterns from an areal perspective. In Hickey, Raymond (ed.), *The Cambridge Handbook of Areal Linguistics* (Cambridge Handbooks in Language and Linguistics), 204–236. Cambridge: Cambridge University Press. (doi:10.1017/9781107279872.009)

List, Johann-Mattis. 2023. Inference of partial colexifications from multilingual wordlists. *Frontiers in Psychology* 14. (https://www.frontiersin.org/articles/10.3389/fpsyg.2023.1156540) (Accessed September 14, 2023.)

List, Johann-Mattis & Rzymski, Christoph & Tresoldi, Tiago & Greenhill, Simon & Forkel, Robert (eds.). 2019. *CLICS*^{3}. Jena: Max Planck Institute for the Science of Human History. (https://clics.clld.org/) (Accessed August 28, 2023.)

McCarthy, Arya D. & Williams, Adina & Liu, Shijia & Yarowsky, David & Cotterell, Ryan. 2020. Measuring the similarity of grammatical gender systems by comparing partitions. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 5664–5675. Online: Association for Computational Linguistics. (doi:10.18653/v1/2020.emnlp-main.456)

Schapper, Antoinette & Koptjevskaja-Tamm, Maria. 2022. Introduction to special issue on areal typology of lexico-semantics. *Linguistic Typology*. De Gruyter Mouton 26(2). 199–209. (doi:10.1515/lingty-2021-2087)

Schapper, Antoinette & San Roque, Lila & Hendery, Rachel. 2016. Tree, firewood and fire in the languages of Sahul. In Koptjevskaja-Tamm, M. & Juvonen, P. (eds.), *Lexico-typological approaches to semantic shifts and motivation patterns in the lexicon*, 355–422. Berlin: De Gruyter Mouton.

Challenges and insights from cross-linguistic word-meaning associations: A roadmap for the study of loose colexification

Thomas Brochhagen (thomas.brochhagen@upf.edu)
Universitat Pompeu Fabra, Barcelona, Spain

Large scale resources such as CLICS (Rzymski et al. 2019) are increasingly used to shed light on the organization of meaning across languages and its underlying principles. Some studies directly leverage the distribution of word-meaning associations to infer regularities across lexica in particular semantic domains (e.g., Jackson et al. 2019). Others study regularities across domains by focusing on the kind of relationship that meanings stand in (e.g., Xu et al. 2020, Brochhagen & Boleda 2022, Brochhagen et al. 2023). For instance, based on their similarity in associativity or affectiveness; their taxonomic distance; or their visual resemblance. This second line of research requires word-meaning associations to be enriched with more information, often from external resources.

This talk has two goals. First, I will discuss the challenges faced by such enrichments as well as possible solutions. Challenges include lack of resources (e.g., lack of psychometrics and models for non-Indo European languages); reliance on decontextualized semantic representations; and need for adjustments for geographic and phylogenetic imbalances (Guzmán Naranjo & Becker 2022). These challenges motivate the second goal, which is to present a working proposal for the scalable study of loose colexification.

By contrast to strict colexification (multiple meanings associated to the same form in a language), loose colexification (multiple meanings associated with overlapping forms) has received relatively little attention (List 2023). However, it is a semantically pervasive feature of many languages; and it is an empirical question to which extent it differs from full colexification –and why. The challenges that rear their head for the study of strict colexification are as, if not more, pressing for loose colexification. Typological adjustments are key; resources become even scarcer; and word segmentation is an added task. My proposal is to side-step some challenges in favor of a scalable analysis, using information-theoretic measures and compression algorithms (e.g., Gutierrez-Vasques et al. 2023) and more diverse resources (e.g., multilingual language models typologically diverse corpora) while –at least for now– not employing human-derived psychometrics. This will enable us to test a clear hypothesis about the forces that shape lexica, reflected by whether meanings tend to strictly or loosely colexify.

References

Brochhagen, T., & Boleda, G. (2022). When do languages use the same word for different meanings? The Goldilocks principle in colexification. Cognition, 226, 105179. 10.1016/j.cognition.2022.105179

Brochhagen, T., Boleda, G., Gualdoni, E., & Xu, Y. (2023). From language development to language evolution: A unified view of human lexical creativity. Science, 381(6656), 431–436. https://doi.org/10.1126/science.ade7981

Jackson, J., J. Watts, T. Henry, J.-M. List, P. Mucha, R. Forkel, S. Greenhill, R. Gray, and K. Lindquist (2019): Emotion semantics show both cultural variation and universal structure. Science 366.6472. 1517-1522.

Gutierrez-Vasques, X., Bentz, C., and Samardžić, T. (2023). Languages through the looking glass of BPE compression. Computational Linguistics.

Guzmán Naranjo, M, and Becker, L. (2022). Statistical bias control in typology. Linguistic Typology 26 (3), 605-670.

List, J.-M. (2023): Inference of partial colexifications from multilingual wordlists. Frontiers in Psychology 14.1156540. 1-10.

Rzymski, Christoph and Tresoldi, Tiago et al. 2019. The Database of Cross-Linguistic Colexifications, reproducible analysis of cross- linguistic polysemies.

Xu, Y., Duong, K., Malt, B. C., Jiang, S., & Srinivasan, M. (2020a). Conceptual relations predict colexification across languages. Cognition, 201, 104280. https://doi.org/10.1016/j.cognition.2020.104280

Representing Semantic Networks in Concepticon

Annika Tjuka[1], Johann-Mattis List[1,2]

[1] Department of Linguistic and Cultural Evolution, Max Planck Institute for Evolutionary Anthropology
[2] Chair for Multilingual Computational Linguistics, University of Passau

The Concepticon was established in 2015 and the first major release was in 2016 with 162 concept lists (Concepticon version 1.0: List et al. 2016a,b). The most recent major release with 413 concept lists was in 2023 (Concepticon version 3.0: Tjuka et al. 2023; List et al. 2022). The Concepticon began with the collection of concept lists from studies in historical linguistics that used cross-linguistic comparisons to create language family trees. These concept lists include basic vocabulary and cross-linguistically comparable concepts such as HAND, BAUM, YOU, or GIVE. The Concepticon was the first resource that contained various concept lists and made them comparable. In 2020, we launched a satellite project, the Database of Norms, Ratings, and Relations (NoRaRe), which builds on the established workflows in Concepticon and makes the available data from linguistics and psychology interoperable (Tjuka et al., 2022).

The majority of lists in Concepticon and NoRaRe are discrete lists that represent a gloss or a value for a gloss in one row so that they can be conveniently mapped to a single Concepticon concept set. For example, the gloss *tree* with the frequency value 3.52 (Brysbaert & New, 2009) is mapped to 906 TREE. In recent years, an increasing number of lists containing data on concept pairs have been published (e.g. similarity ratings: Hill et al. 2015; Vulić et al. 2020, or semantic relations: Urban 2011; Maciejewski & Klepousniotou 2016). These lists are of great interest to Concepticon also in connection with another satellite project, the Database of Cross-Linguistic Colexifications (CLICS), which provides semantic relations in the form of polysemy frequencies between concepts (Rzymski et al., 2020). However, adding lists of concept pairs to Concepticon and NoRaRe is challenging, especially when they span multiple languages (List, 2021). By improving our workflows and adopting a consistent representation of network-like lists, we are able to add the lists and, more importantly, check the data for accuracy. The improved workflow was particularly useful in resolving consistency errors in the lists from Urban (2011) and the extended list that builds on Urban's publication in Winter & Srinivasan (2022).

In this talk, we present our efforts and give a first insight into the representation of networks in Concepticon using a concrete example, namely the inclusion of partial colexification provided in List (2023). We will also show how the data can be visualized in the form of graphs.

References

Brysbaert, Marc & Boris New. 2009. Moving Beyond Kučera and Francis: A Critical Evaluation of Current Word Frequency Norms and the Introduction of a New and Improved Word Frequency Measure for American English. *Behavior Research Methods* 41(4). 977–990. doi:10.3758/BRM.41.4.977.

Hill, Felix, Roi Reichart & Anna Korhonen. 2015. SimLex-999: Evaluating Semantic Models with (genuine) Similarity Estimation. *Computational Linguistics* 41(4). 665–695. doi:10.1162/COLI_a_00237.

List, Johann-Mattis. 2021. Mapping Multi-SimLex to Concepticon. *Computer-Assisted Language Comparison in Practice* 4(3). 1–8. `https://calc.hypotheses.org/2684`.

List, Johann-Mattis. 2023. Inference of Partial Colexifications from Multilingual Wordlists. *Frontiers in Psychology* 14. 1–10. doi:10.3389/fpsyg.2023.1156540.

List, Johann-Mattis, Michael Cysouw & Robert Forkel. 2016a. Concepticon: A Resource for the Linking of Concept Lists. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk & Stelios Piperidis (eds.), *Proceedings of the 10th International Conference on Language Resources and Evaluation*, 2393–2400. Portorož, Slovenia: European Language Resources Association. `https://aclanthology.org/L16-1379/`.

List, Johann-Mattis, Michael Cysouw & Robert Forkel. 2016b. *Concepticon. A Resource for the Linking of Concept Lists (Version 1.0)*. Jena: Max Planck Institute for the Science of Human History. doi:10.5281/zenodo.47143. `https://concepticon.clld.org/`.

List, Johann-Mattis, Annika Tjuka, Christoph Rzymski, Simon J. Greenhill & Robert Forkel. 2022. *Concepticon. A Resource for the Linking of Concept Lists (Version 3.0)*. Leipzig: Max Planck Institute for Evolutionary Anthropology. doi:10.5281/zenodo.7296458. `https://concepticon.clld.org/`.

Maciejewski, Greg & Ekaterini Klepousniotou. 2016. Relative Meaning Frequencies for 100 Homonyms: British Edom Norms. *Journal of Open Psychology Data* 4(1). 1–5. doi:10.5334/jopd.28.

Rzymski, Christoph, Tiago Tresoldi, Simon J. Greenhill, Mei-Shin Wu, Nathanael Schweikhard, Maria Koptjevskaja-Tamm, Volker Gast, Timotheus A. Bodt, Abbie Hantgan, Gereon A. Kaiping, Sophie Chang, Yunfan Lai, Natalia Morozova, Heini Arjava, Nataliia Hübler, Ezequiel Koile, Steve Pepper, Mariann Proos, Briana Van Epps, Ingrid Blanco, Carolin Hundt, Sergei Monakhov, Kristina Pianykh, Sallona Ramesh, Russell D. Gray, Robert Forkel & Johann-Mattis List. 2020. The Database of Cross-Linguistic Colexifications, Reproducible Analysis of Cross-Linguistic Polysemies. *Scientific Data* 7(1). 1–12. doi:10.1038/s41597-019-0341-x.

Tjuka, Annika, Robert Forkel & Johann-Mattis List. 2022. Linking Norms, Ratings, and Relations of Words and Concepts Across Multiple Language Varieties. *Behavior Research Methods* 54. 864–884. doi:10.3758/s13428-021-01650-1. `https://norare.clld.org/`.

Tjuka, Annika, Robert Forkel & Johann-Mattis List. 2023. Curating and Extending Data for Language Comparison in Concepticon and NoRaRe. *Open Research Europe* 2(141). 1–13. doi:10.12688/openreseurope.15380.3.

Urban, Matthias. 2011. Asymmetries in Overt Marking and Directionality in Semantic Change. *Journal of Historical Linguistics* 1(1). 3–47. doi:10.1075/jhl.1.1.02urb.

Vulić, Ivan, Simon Baker, Edoardo Maria Ponti, Ulla Petti, Ira Leviant, Kelly Wing, Olga Majewska, Eden Bar, Matt Malone, Thierry Poibeau, Roi Reichart & Anna Korhonen. 2020. Multi-SimLex: A Large-Scale Evaluation of Multilingual and Cross-Lingual Lexical Semantic Similarity. *Computational Linguistics* 46(4). 1–51. doi:10.1162/coli_a_00391.

Winter, Bodo & Mahesh Srinivasan. 2022. Why Is Semantic Change Asymmetric? The Role of Concreteness and Word Frequency and Metaphor and Metonymy. *Metaphor and Symbol* 37(1). 39–54. doi:10.1080/10926488.2021.1945419.

Collecting Character Sequences for Paleolithic Signs and Written Languages

Christian Bentz
University of Tübingen

Introduction "Language does not fossilize". This statement is often found in textbooks and articles on language evolution. The lack of empirical data is one of the major reasons for why explaining language evolution is considered one of the hardest problems in science. However, what might have fossilized after all are structural aspects of language, for instance, *symbolic combinatoriality*. First traces of so-called *geometric signs* are found deep into the paleolithic – several ten thousand or even hundred thousand years ago. These have recently been argued to constitute the first *artificial memory systems* (D'Errico *et al.* 2017) which have "complexified" over time. Can we measure this complexification? Moreover, can we statistically distinguish sequences of geometric signs from later visual information encoding, i.e. ancient and modern writing? If yes, what are the distinctive structural features?

Data Bases To start tackling these questions, two data bases are presented: a) *SignBase* (Dutkiewicz *et al.* 2020), a collection of geometric signs on mobile objects of the paleolithic; b) the *TeDDi* sample (Moran *et al.* 2022), a collection of currently c. 20k texts written in 89 typologically diverse languages and 15 writing systems. The *TeDDi* sample is available in CLDF (Forkel *et al.* 2018), and can be exported to a range of file formats (*Rdata*, *CSV*, *JSON*). *SignBase* is currently managed via *Django*, and available as CSV and XLSX download.

Challenges and Opportunities In this talk, the challenges and problems with building these data bases are discussed. This ranges from issues with data entries (e.g. how to systematically store geometric signs applied to three dimensional objects in sequential format), to problems of finding data output formats which are interoperable, but still human readable. Furthermore, some opportunities for not only using cross-linguistic but "cross-semiotic" data to analyse languages in comparison to other sign sequences are discussed.

References • **D'Errico, F., Doyon, L., Colagé, I., Queffelec, A., Le Vraux, E., Giacobini, G., Vandermeersch, B., and Maureille, B.** (2017). From number sense to number symbols. an archaeological perspective. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **373**(1740), 20160518. • **Dutkiewicz, E., Russo, G., Lee, S., and Bentz, C.** (2020). Signbase, a collection of geometric signs on mobile objects in the paleolithic. *Scientific data*, **7**(1), 364. • **Forkel, R., List, J.-M., Greenhill, S. J., Rzymski, C., Bank, S., Cysouw, M., Hammarström, H., Haspelmath, M., Kaiping, G. A., and Gray, R. D.** (2018). Cross-linguistic data formats, advancing data sharing and re-use in comparative linguistics. *Scientific data*, **5**(1), 1–10. • **Moran, S., Bentz, C., Gutierrez-Vasques, X., Pelloni, O., and Samardzic, T.** (2022). TeDDi sample: Text data diversity sample for language comparison and multilingual NLP. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pp. 1150–1158. European Language Resources Association.

Generating CLDF from heterogenous input in the Open Text Collections project: Input from FLEx, ELAN, tex

Sebastian Nordhoff

This talk will present discuss experiences from setting up a publishing project using CLDF as a backend for storing interlinear glossed texts. It will focus on routines for extracting IGT from ELAN files and writing the content to CLDF and highlight fundamental conceptual issues as well as issues arising from the current practices by ELAN users “in the wild“.

The project Open Text Collections will establish a presitigious high-quality platform for the publication of text collections. A text collection will consist on a cultural/linguistic/anthropological introduction and a set of 10+ interlinearized texts. The platform shall be easy to use for linguists, but have clearly defined and highly automated workflows. Together, these requirements mean that we must allow linguists to submit their texts in the format they were collected in. As of today, this is typically *flextext by Fieldworks Language Explorer or *eaf as saved by ELAN. This format must then be converted to a backend format for further processing. This backend format is CLDF. From there, a variety of output formats can be generated (JSON, HTML, PDF via tex). In this talk, however, I will focus on eaf ingestion and the storage of this information as CLDF.

ELAN uses a well-specified XML-based format defining various “tiers“, which can enter into a variety of relationships (time subdivision, symbolic subdivision, association). The relevant tiers for our purposes are the tiers covering transcription, interlinearization, and translation. Users are free to name and arrange the tiers as they see fit. In a set of 20k ELAN files, we find over 2000 different arrangements of tiers (already disregarding names), and the most frequent arrangement only makes up for 8% of all cases (von Prince & Nordhoff 2020). This means that we cannot base ourselves on a fixed structure; rather, we must employ heuristics to identify the relevant tiers in order to extract the information we need. I will briefly sketch methods for the development of “tier type induction” used in the eldpy python library and their problems. This can inform requirements for the specifications of a CLDF format for running texts. Some of the issues encountered in the ELAN files relate to users being “creative” and putting annotations like “#” or “***” or “???” or “(also found in XYZ)” in one of linguistically relevant tiers. Other issues are implicit conventions for the use of white space and morpheme separators (-,=), the existence of multiple speakers or multiple glosses/translations and mixing of implicit morpheme breaks via '-' inside a string and explicit morpheme breaks via an annotation boundary. I will briefly touch upon models like XIGT (Goodman et al. 2015) and Generalized Glossing Guidelines (Mortensen et al. 2023) which aim at modelling these issues.

I will then discuss the representation of the extracted IGT content in CLDF, where there are issues of unique identification of sentences, linearity, and dereferenceability.

Finally, I will address output formats like HTML, pdfs and printed books. These formats require different types of information (eg presence/absence of interlinear line, or of a special orthography/morphophonology line). This information must be present in CLDF in an accessible way, but the question is which pieces are obligatory/optional/primary and how users can agree on naming conventions for the relevant columns.

References

Goodman, Michael W., Joshua Crowgey, Fei Xia & Emily M. Bender. 2015. Xigt: Extensible interlinear glossed text for natural language processing. *Language Resources And Evaluation* 49(2). 455–485. DOI: `10.1007/s10579-014-9276-1`.

Mortensen, David R., Ela Gulsen, Taiqi He, Nathaniel Robinson, Jonathan Amith, Lindia Tjuatja & Lori Levin. 2023. Generalized glossing guidelines: An explicit, human- and machine-readable, item-and-process convention for morphological annotation. In Garrett Nicolai, Eleanor Chodroff, Frederic Mailhot & Çağrı Çöltekin (eds.), *Proceedings of the 20th sigmorphon workshop on computational research in phonetics, phonology, and morphology*, 58–67. Toronto, Canada: Association for Computational Linguistics. DOI: `10.18653/v1/2023.sigmorphon-1.7`. `https://aclanthology.org/2023.sigmorphon-1.7`.

von Prince, Kilu & Sebastian Nordhoff. 2020. An empirical evaluation of annotation practices in corpora from language documentation. English. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk & Stelios Piperidis (eds.), *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 2778–2787. Marseille, France: European Language Resources Association. `https://aclanthology.org/2020.lrec-1.338`.

The morpho-syntax of Archaic Chinese verbs: Loss of morphology as trigger for the emergence of analytic structures

Barbara Meisterernst
University Stuttgart
bmeisterernst@gmail.com

In my new project on the diachronic morpho-syntax of Chinese, I am focusing on the reconstructed morphology of Chinese verbs. The corpus to be established for this study will contain all verbs which have the falling tone (*qusheng*) in Middle Chinese and for which a suffix *-*s* has been reconstructed. The latter is the most frequently attested, and the most uncontroversial and best studied affix of Chinese. However, there is still some debate about the precise morpho-syntactic functions of this suffix. In my project, I am particularly interested in possible resultative readings of the *-*s* suffix.

My hypothesis is that the suffix *-*s*, and the verbal morphology in Chinese in general, was derivational and not grammatical. This means that it rather served to derive aktionsart types than to express grammatical aspect (contra Unger 1983, Jin Lixin 2006); the expression of aktionsart types (lexical aspect) is typical for derivational morphology (Kiefer 2010). I propose that the derivational morphology of Archaic Chinese was hosted in a split VP (following Ramchand 2008), and that its loss was one of the triggers for a change of Chinese from a more synthetic to an analytic language. In previous studies (2019, 2023), I tentatively proposed that the *-*s* suffix may have functioned as an overt res head in the sense of Ramchand (2008) with both unaccusative/intransitive and causative/transitive verbs, uniting the two major functions proposed for the suffix in the literature (e.g. Schuessler 2007). When at the end of the Archaic period the verbal morphology increasingly lost its transparency, new structures such as disyllabification of verbs and resultative constructions, including the source structures of the Modern Mandarin aspectual suffixes, emerged in order to replace the old morphology. In order to provide evidence for my hypothesis, all verbs in the corpus will be subjected to the tests established cross-linguistically for the determination of the event structure of verbs.

As basis for the classification of the verbs, the respective reconstructions proposed in the relevant literature (mostly Pulleyblank, Baxter&Sagart, Unger, Jin Lixin) and the glosses provided therein will be collected and analysed. The resulting list will be connected to and compared with the phonological entries in the *Jingdian shiwen* by Lu Deming (6th – 7th c. CE), a commentary on the Classical Chinese literature. Example sentences from the *Jingdian shiwen* and from the Archaic Chinese literature will be added on which the cross-linguistically established syntactic tests will be conducted.

Selected References

Kiefer, Ferenc. 2010. Areal-typological aspects of word-formation: The case of aktionsart-formation in German, Hungarian, Slavic, Baltic, Romani and Yiddish. In Franz Rainer, Wolfgang U. Dressler, Dieter Kastovsky & Hans Christian Luschützky (eds.), *Variation and change in morphology: Selected papers from the 13th International Morphology Meeting,* Vienna, February 2008, 129–148. Amsterdam: John Benjamins.

McFadden, Thomas. 2015. Preverbal ge- in Old and Middle English. In André Meinunger (ed.), *Byproducts and side effects: Nebenprodukte und Nebeneffekte (ZAS Papers in Linguistics* 58), 15–48. Berlin: ZAS.

Meisterernst, Barbara. 2023. The loss of morphology and the emergence of analytic structures in Chinese. *Journal of Historical Syntax,* 7, artice 15, 1-55.

Ramchand, Gillian C. 2008. *Verb meaning and the lexicon: A first phase syntax.* Cambridge: Cambridge University Press.

Schuessler, Axel. 2007. *ABC etymological dictionary of old Chinese*. Honolulu Hawaii: University of Hawai'i Press.