First workshop on Data Models, Citation, Access, and Re-usability impacting Historical Linguistic Datasets

Hugh Paterson III & Oksana Zavalina

The role of library models (e.g., IFLA-LRM: Riva, Le Bœuf, and Žumer 2017) and archival practice (e.g., lifecycle management: Higgins 2012) is under-explored in relation to the construction and reuse of *Historical Linguistic Information Sources*. This workshop proposes to provide a forum to discuss the structures and models of information resources in historical-comparative linguistic research outputs through the integration of informatic models from library science and archivy.

Significant advances have been made in historical linguistics through the use of large compiled datasets (e.g., Kamholz et al. 2024; Tresoldi 2023; Arora et al. 2023; Dellert et al. 2020; Greenhill 2015; Segerer and Flavier 2013; Mielke 2008; Greenhill, Blust, and Gray 2008). While not precluding the contributions of single historical manuscripts and traditional manuscript consultation methods, the use of and creation of datasets (including corpora) has become the defacto way of generating new hypotheses (Wichmann and Saunders 2007; Steiner, Cysouw, and Stadler 2011; Segerer 2015). Datasets in historical linguistics generally do two things: (1) record critical researcher-created information such as **reconstructed forms**, **cognacy judgments**, **confidence levels,** along with **contextual notes**; and (2) contain foundational content from sources not created by the dataset compiler. Such source material often include historically published and unpublished resources including: **maps** (Hessle and Kirk 2020), language specific **lexicons** and **published reconstructions** (Kamholz et al. 2024), **wordlists** (Forkel et al. 2024; Segerer and Flavier 2013), **transcriptions of manuscripts and texts** (Weber et al. 2023; Genee and Junker 2018; Kytö 2011), and even **reconstructions by other scholars**, etc.

Interactional platform-tools such as *RefLex* (Segerer and Flavier 2013) or *OUTOFPAPUA* (Kamholz et al. 2024) allow users to create custom datasets based on specific selected resources available to the platform. They do this without requiring users to interact with the complete set of underlying resources and/or the platforms allow users to create new derivative aggregate collections (reconstructed forms and cognacy relations) independent of other platform users. Citing, referencing, and redistributing these custom datasets is challenging and impacts the verifiability of claims.

It is broadly accepted across linguistic research that scholarly work—including evidence— should be citable, accessable, and reusable (Bird and Simons 2003). Together these issues impact reproducibility, an important tenet in scholarship often overlooked in linguistics (Berez-Kroeker et al. 2018). However, it is also well acknowledged that the citation and reference of original source material for linguistic evidence is lacking across the field (Gawne et al. 2017). More specifically in historical-comparative linguistics, the context of citation and referencing of the evidentiary record along with current dataset assemblage and

distribution practices generally do not support fine-grained or Work-oriented citation and referencing. This often means that specific and necessary details in comparative linguistics are not retrievable. Therefore, the data models embedded within historical comparative datasets become all the more important for the reproducibility of work and the testing, verification, and refinement of hypotheses (Bakro-Nagy 2010).

With the exception of leading work around Cross-Linguistic Data Formats (CLDF) use with historical-comparative data (Forkel et al. 2018; Forkel, Swanson, and Moran 2024) and approaches using linked data in linguistics (Kesäniemi et al. 2018; Tittel, Gillis-Webber, and Nannini 2020), the literature has been silent about the storage formats for historical-comparative data. Undiscussed are the information categories represented in historical comparative linguistic datasets. The informatic arrangement and description of compiled datasets has generally been ad-hoc and served the needs of individually-funded projects. This has resulted in a proliferation of divergent data categories mitigating against ease-of-reuse.

We set out to ignite discussion around compilations of manuscripts, wordlists, and other derivative resources which have become mainstream tools in hypothesis generation related to the language evolution. We explore the heretofore unapproached contribution that models such as Work-Expression-Manifestation-Item (WEMI), illustrated in figure 1, from library and information science (Coyle 2023; Riva, Le Bœuf, and Žumer 2017; IFLA, 1998) can offer those who compile, and cite/reference aggregate linguistic resources. Specifically, clarifying linking relationships between the literature and datasets, including dataset portions.

We invite papers describing the information models used for assembling large corpora (including wordlists) used in historical linguistics, highlighting assumptions for citation, referencing, segmentation, and reusability of the assembled collection of texts and their digital surrogates. We encourage papers which present typologies of use cases, categories of tracked information, provenance of data content, citability of aggregate content, and the identifiers-for and permanence-of user-generated datasets on research platforms.

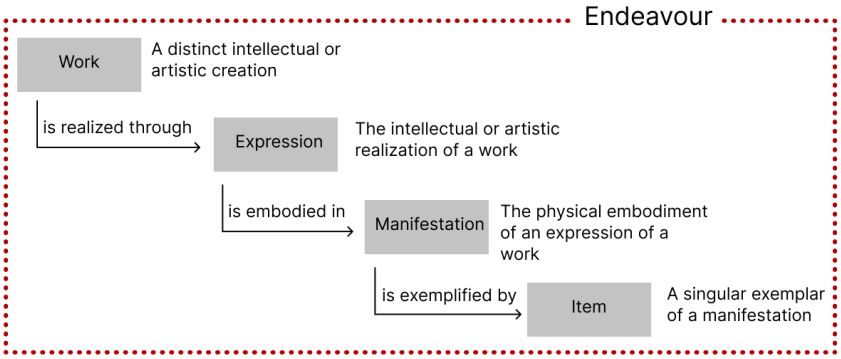

Figure 1. The basic WEMI model.

References

Arora, Aryaman, Adam Farris, Samopriya Basu, and Suresh Kolichala. 2023. “Jambu: A Historical Linguistic Database for South Asian Languages.” arXiv. https://doi.org/10.48550/arXiv.2306.02514.

Bakro-Nagy, Marianne. 2010. “Data in Historical Linguistics: On Utterances, Sources, and Reliability.” *Sprachtheorie Und Germanistische Linguistik 20.2: 133-195.*, January. https://www.academia.edu/3629841/Data_in_historical_linguistics_On_utterances_sources_and_reliability.

Berez-Kroeker, Andrea L., Lauren Gawne, Susan Smythe Kung, Barbara F. Kelly, Tyler Heston, Gary Holton, Peter Pulsifer, et al. 2018. “Reproducible Research in Linguistics: A Position Statement on Data Citation and Attribution in Our Field.” *Linguistics* 56 (1): 1–18. https://doi.org/10.1515/ling-2017-0032.

Bird, Steven, and Gary F. Simons. 2003. “Seven Dimensions of Portability for Language Documentation and Description.” *Language* 79 (3): 557–82. https://doi.org/10.1353/lan.2003.0149.

Coyle, Karen. 2023. “openWEMI.” In *Proceedings of the International Conference on Dublin Core and Metadata Applications*. Dublin, Ohio: Dublin Core Metadata Initiative. https://doi.org/10.23106/DCMI.953115290.

Dellert, Johannes, Thora Daneyko, Alla Münch, Alina Ladygina, Armin Buch, Natalie Clarius, Ilja Grigorjew, et al. 2020. “NorthEuraLex: A Wide-Coverage Lexical Database of Northern Eurasia.” *Language Resources and Evaluation* 54 (1): 273–301. https://doi.org/10.1007/s10579-019-09480-6.

Forkel, Robert, Johann-Mattis List, Simon J. Greenhill, Christoph Rzymski, Sebastian Bank, Michael Cysouw, Harald Hammarström, Martin Haspelmath, Gereon A. Kaiping, and Russell D. Gray. 2018. “Cross-Linguistic Data Formats, Advancing Data Sharing and Re-Use in Comparative Linguistics.” *Scientific Data* 5 (1): 180205. https://doi.org/10.1038/sdata.2018.205.

Forkel, Robert, Johann-Mattis List, Christoph Rzymski, and Guillaume Segerer. 2024. “Linguistic Survey of India and Polyglotta Africana: Two Retrostandardized Digital Editions of Large Historical Collections of Multilingual Wordlists.” In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, edited by Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue, 10578–83. Torino, Italia: ELRA and ICCL. https://aclanthology.org/2024.lrec-main.925.

Forkel, Robert, Daniel G. Swanson, and Steven Moran. 2024. “Converting Legacy Data to CLDF: A FAIR Exit Strategy for Linguistic Web Apps.” In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, edited by Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue, 3978–82. Torino, Italia: ELRA and ICCL. https://aclanthology.org/2024.lrec-main.353.

Gawne, Lauren, Barbara F. Kelley, Andrea L. Berez-Kroeker, and Tyler Heston. 2017. “Putting Practice into Words: The State of Data and Methods Transparency in Grammatical Descriptions.” *Language Documentation & Description* 11:157–89. http://hdl.handle.net/10125/24731.

Genee, Inge, and Marie-Odile Junker. 2018. “The Blackfoot Language Resources and Digital Dictionary Project: Creating Integrated Web Resources for Language Documentation and Revitalization.” *Language Documentation & Conservation* 12 (June):274–314. http://hdl.handle.net/10125/24770.

Greenhill, Simon J. 2015. “TransNewGuinea.Org: An Online Database of New Guinea Languages.” *PLOS ONE* 10 (10): e0141563. https://doi.org/10.1371/journal.pone.0141563.

Greenhill, Simon J., Robert Blust, and Russell D. Gray. 2008. “The Austronesian Basic Vocabulary Database: From Bioinformatics to Lexomics.” *Evolutionary Bioinformatics* 4 (January):EBO.S893. https://doi.org/10.4137/EBO.S893.

Hessle, Christian, and John Kirk. 2020. “Digitising Collections of Historical Linguistic Data: The Example of The Linguistic Atlas of Scotland.” *Journal of Data Mining & Digital Humanities* Special issue on Visualisations in Historical Linguistics (December). https://doi.org/10.46298/jdmdh.5611.

Higgins, Sarah. 2012. “The Lifecycle of Data Managment.” In *Managing Research Data*, edited by Graham Pryor, 17–46. London, UK: Facet Publishing.

IFLA Study Group on the Functional Requirements for Bibliographic Records and Plassard, Marie-France. 1998. “Functional Requirements for Bibliographic Records: Final Report.” 19. 2nd ed. [UBCIM Publications, New Series] IFLA Series on Bibliographic Control. Munich, Germany: K.G. Saur. http://www.ifla.org/VII/s13/frbr/.

Kamholz, David, Anne van Schie, Allahverdi Verdizade, Maria Zielenbach, and Antoinette Schapper. 2024. “OUTOFPAPUA.” Database. 2024. https://outofpapua.com/.

Kesäniemi, Joonas, Turo Vartiainen, Tanja Säily, and Terttu Nevalainen. 2018. “Exploring Meta-Analysis for Historical Corpus Linguistics Based on Linked Data.” *Journal of Research Design and Statistics in Linguistics and Communication Science* 5 (1–2): 4–47. https://doi.org/10.1558/jrds.36709.

Kytö, Merja. 2011. “Corpora and Historical Linguistics.” *Revista Brasileira de Linguística Aplicada* 11 (2): 417–57. https://doi.org/10.1590/S1984-63982011000200007.

Mielke, Jeff. 2008. *The Emergence of Distinctive Features*. Oxford, England: Oxford University Press.

Riva, Pat, Patrick Le Bœuf, and Maja Žumer, eds. 2017. *IFLA Library Reference Model: A Conceptual Model for Bibliographic Information*. December 2017. Den Haag, Netherlands: International Federation of Library Associations and Institutions (IFLA). https://www.ifla.org/publications/node/11412.

Segerer, Guillaume. 2015. “How Databases Shape Research: Labial-Velars Distribution in Africa.” In *8th World Congress of African Linguistics (WOCAL8)*. Kyoto, Japan. https://inria.hal.science/halshs-01251122/.

Segerer, Guillaume, and Sébastien Flavier. 2013. “The RefLex Project: Documenting and Exploring Lexical Resources in Africa.” Oral Presentation presented at the Research, records and responsibility: Ten years of the Pacific and Regional Archive for Digital Sources in Endangered Cultures, Sydney, Australia. http://hdl.handle.net/2123/9854.

Steiner, Lydia, Michael Cysouw, and Peter Stadler. 2011. “A Pipeline for Computational Historical Linguistics,” January. https://doi.org/10.1163/221058211X570358.

Tittel, Sabine, Frances Gillis-Webber, and Alessandro A. Nannini. 2020. “Towards an Ontology Based on Hallig-Wartburg’s Begriffssystem for Historical Linguistic Linked Data.” In *Proceedings of the 7th Workshop on Linked Data in Linguistics (LDL-2020)*, edited by Maxim Ionov, John P. McCrae, Christian Chiarcos, Thierry Declerck, Julia Bosque-Gil, and Jorge Gracia, 1–10. Marseille, France: European Language Resources Association. https://aclanthology.org/2020.ldl-1.1.

Tresoldi, Tiago. 2023. “A Global Lexical Database (GLED) for Computational Historical Linguistics.” *Journal of Open Humanities Data* 9 (1): Article 2. https://doi.org/10.5334/johd.96.

Weber, Natalie, Tyler Brown, Joshua Celli, McKenzie Denham, Hailey Dykstra, Rodrigo Hernandez-Merlin, Evan Hochstein, et al. 2023. “Blackfoot Words: A Database of Blackfoot Lexical Forms.” *Language Resources and Evaluation* 57 (3): 1207–62. https://doi.org/10.1007/s10579-022-09631-2.

Wichmann, Søren, and Arpiar Saunders. 2007. “How to Use Typological Databases in Historical Linguistic Research.” *Diachronica* 24 (2): 373–404. https://doi.org/10.1075/dia.24.2.06wic.