Computer-Assisted Language Comparison in Practice

Tutorials on Computational Approaches to the History and Diversity of Languages

Standardizing Phonetic Transcriptions for Kitchen et al.'s Comparative Wordlist on Semitic Languages with Language-Specific Orthography Profiles

Ben Sapirstein
Reichmann University

Comparative wordlists are a fundamental tool for tracing language history, allowing us to see how languages are related, much like biologists use DNA to infer phylogenies of species. When linguists compile data from different sources, scholars often code lexical data differently, using individual transcription systems that cannot be directly compared with each other. In order to make such data comparable, individual transcription systems must be unified in order to reflect a common standard. This study illustrates how such unification can be done by taking a particularly diverse dataset on Semitic languages as example and illustrating how transcriptions for individual language varieties can be harmonized as part of the general standardization workflow proposed by the Cross-Linguistic Data Formats initiative.

1 Introduction

The discipline of historical linguistics, which compares languages to reconstruct their history, offers a unique window into human prehistory, cultural evolution, and cognition. By analyzing linguistic data, researchers can shed light on ancient migrations and cultural contacts that occurred long before written records existed. This approach has proven particularly powerful for the Semitic language family, which is one of the world's least controversial and best-documented language groups (Bennett 1998). With a written history spanning nearly five millennia, the Semitic languages are associated with some of the world's earliest urban civilizations in Mesopotamia, the Levant, and the Horn of Africa, as well as major religious and literary traditions, making their history fundamental to our understanding of the ancient world (Bennett 1998).

Computer-Assisted Language Comparison in Practice
Volume 8, Number 2
URL: https://calc.hypotheses.org/8706 DOI: 10.15475/calcip.2025.2.5

Published on 10/11/2025

In recent decades, the quantitative turn in historical linguistics has led to the development of computational historical linguistics, a field that applies computational methods to tasks such as phonetic alignment, automated cognate detection, and phylogenetic reconstruction (Häuser and List 2025). These methods hold great potential for discovering new linguistic relationships that traditional scholarship may have overlooked. The Semitic languages offer an ideal test field for these computational approaches. The family's genetic relations are well-established through traditional scholarship, providing a solid foundation against which to evaluate automated methods. The family's complex morphological systems, characterized by triconsonantal roots and vocalic templates, present a unique challenge to standard computational models that often assume concatenative morphology. Furthermore, the family's temporal depth is key for understanding language change, spanning ancient languages preserved in historical texts (like Akkadian, attested from around 2400 BCE) and modern varieties spoken today.

While having been intensively investigated through traditional scholarship, only a few computational studies have investigated the Semitic languages so far. As a rare exception, Kitchen et al. (2009) provide a phylogenetic analysis of the language family, based on a short cognate-coded wordlist covering 25 language varieties. An advantage of this study is that the data are publicly shared and the cognates are coded in a transparent way that can be easily inspected. A disadvantage, however, is that the word forms are only represented in the original orthographies in which they were collected from the original studies. Due to the lack of standardized phonetic transcriptions, several interesting computational analyses cannot be carried out with the dataset. In order to address this problem, this study uses the Lexibank worfklow for the standardization of multilingual wordlists in order to harmonize the phonetic transcriptions in the data by Kitchen et al. (2009) with the help of orthography profiles (Moran and Cysouw 2018).

2 Background

As mentioned before, the dataset by Kitchen et al. (2009) was compiled for a Bayesian phylogenetic study of the Semitic language family. It consists of lexical data for 25 Semitic languages, covering 96 concepts from a modified Swadesh list and their cognacy labels. The dataset itself was not compiled from scratch but aggregates lexical items from two larger comparative studies on Ethio-Semitic languages (Bender 1971) and of several Semitic languages (Rabin 1975), as well as lexicographic materials for additional languages. The compilation drew from at least five distinct scholarly traditions, each with established transcription conventions, as shown in Table 1.

Subgroup	Languages	Primary Source
Ethiosemitic	Amharic, Argobba, Chaha, Gafat, Ge'ez, Geto, Harari, Innemor, Mesmes, Mesqan, Soddo, Tigre, Tigrinya, Walani, Zway	Bender 1971
Arabic varieties	Moroccan Arabic, Ogaden Arabic	Harrell and Sobelman 1963, Bender 1971
South Arabian	Soqotri, Jibbali, Harsusi, Mehri	Leslau 1938, Johnstone 1981, Johnstone 1977, Johnstone 1987
Semitic	Akkadian, Aramaic, Hebrew, Ugaritic	Rabin 1975, Gelb 1956

Table 1: Overview on the languages in the sample.

This diversity of sources created numerous encoding inconsistencies. Some were straightforward character mapping issues: Bender used <9> to represent the voiced pharyngeal fricative while Moroccan Arabic data used <ɛ> (inspired by Arabic ـعـ) and the other sources encoded as the usual <ʕ>. Other problems proved more complex, such as the ambiguous use of <h>, which served both as a distinctive phoneme /h/ and as a spirantization indicator (e.g., <bh> → /β/, <kh> → /χ/). Or apostrophe usage that varied between glottal stops, pharyngealization, ejectives and stress marks. In addition to transcriptional inconsistencies, the dataset contained typos and misplaced entries when compared directly against the original published sources.

Another notable gap emerged upon closer inspection of the concept list. While Kitchen et al. indicate that their dataset is based on a 96-item list, it actually provides data for only 95 concepts. A comparison with the primary sources they aggregated revealed the missing item: CLAW (ANIMAL). This concept is present in both Rabin's and Bender's wordlists. The concept was therefore added back into the dataset for all relevant languages, and its cognate labels were annotated based on the available lexical forms in the sources.

Finally, a conceptual mix-up surrounding the word 'fly' required correction. The original Swadesh list intends the verb FLY (v.), but Bender collected data for the noun FLY (n.) instead. Kitchen et al. adopted Bender's interpretation but also included Rabin's forms, which correctly correspond to the verb. To resolve this, I added a separate concept for FLY (v.) and mapped Rabin's forms accordingly.

3 Standardizing Phonetic Transcriptions

To make the dataset computationally tractable and comparable with other resources it was structured according to the standards of the Cross-Linguistic Data Formats (CLDF) initiative (Forkel et al. 2018), following the established Lexibank workflow (List et al. 2022, Blum et al. 2025).

Concepts were linked to the Concepticon reference catalog (List et al. 2025, https://concepticon.clld.org), languages were listed in the language table of the CLDF dataset with corresponding Glottolog identifiers (Hammarström et al. 2025, https://glottolog.org), and bibliographic sources were compiled in the sources accompanying all CLDF datasets, represented in BibTeX format. As in all wordlists that follow the Lexibank framework in CLDF, a Python script (`lexibank_kitchensemitic.py`) was used to handle the conversion, ensuring that the workflow remains transparent and reproducible.

While an earlier version of the CLDF dataset of Kitchen et al.'s (2009) Semitic dataset had already been created before, this first pass on the data lacked unified orthographies for all languages in the sample. As a result, the data was not included in the second installment of the Lexibank repository (Blum et al. 2025). As I was told by the Lexibank team, they had initially tried to standardize the orthographies by using the standard workflow to employ orthography profiles (Moran and Cysouw 2018) to convert the data to CLDF, but they had given up later, as they lacked time and even more importantly the detailed knowledge of Semitic languages to come to a satisfying solution.

The workflow for the standardization of the phonetic transcriptions that I will describe in the following consists of two major stages (see Forkel et al. 2024 for a first adaptation on languages from India and Miller and List 2024 for a followup on Pano and Tacana languages). In a first stage (§ 3.1), a general orthography profile is created, which does not need to be perfect but would at least cover all individual sounds that recur in the data. In a second stage (§ 3.2) language-specific orthography profiles are added for all individual language varieties in the sample. This workflow was accompanied by an intensive validation stage (§ 3.3) in which the resulting transcriptions were thoroughly investigated and adjusted further to minimize errors.

3.1 Master Orthography Profile Generation

The Lexibank workflow provides several commands that facilitate the handling of orthography profiles for the standardization of transcriptions, implemented as part of the PyLexibank Python library (Forkel et al. 2024b, https://pypi.org/project/pylexibank). With the help of PyLexibank and the CLDFBench

package (Forkel and List 2020, https://pypi.org/project/cldfbench), an initial orthography profile can be created from an existing CLDF dataset, as long as this provides lexical forms for each language variety. Thus, the following command will create an initial orthography profile for all word forms in the CLDF version of the dataset by Kitchen et al. (2009), trying to guess for the best way to convert original sound segments into the B(road)IPA system used in the Cross-Linguistic Transcription Systems (CLTS, Anderson et al. 2018, https://clts.clld.org) that is used as the standard transcription throughout the entire Lexibank repository.

```
$ cldfbench lexibank.init_profile lexibank_kitchensemitic.py --
context
```

As usually, this firts profile must be refined systematically. While this can often be done by refining the individual orthography profile created for all languages at once, the Semitic data by Kitchen et al. (2009) turned out to be too challenging. As a result, individual orthography profiles had to be compiled for all 25 language varieties.

3.2 Language-Specific Profile Creation

Language-specific orthography profiles in Lexibank are individualized orthographic profiles that provide the conversion pairs for individual language varieties in a dataset. In order to facilitate their creation from a basic profile, PyLexibank offers a dedicated command to transfer individual parts of the master profile to individual language profiles, listing only those character combinations observed for each particular variety.

```
$ cldfbench lexibank.language_profiles lexibank_kitchensemitic.py
```

Having created the language-specific profiles, each individual profile was cross-referenced with established phonological descriptions about the respective language variety to ensure accurate conversion to the IPA transcriptions employed by the CLTS reference catalog. Orthographic verification included checking the sources for their phonological tables and figuring out the matching (B)IPA character.

3.3 Lexical Validation

Beyond standardizing the transcriptions, I conducted a systematic lexical verification for every item in the dataset. This involved cross-referencing each word form with its original published source to correct inconsistencies. As an illustration, Table 2 shows the corrections applied to the lexical item KNOW and the resulted segmented transcribing.

Language	Original form	Fixed form	BIPA graphemes
Argobba	wçɔːnkˈa	wɔːnkˈa	w ɔː n kʼ a
Harari	7aˈkˈɛ	7aːkˈɛ	ʔ aː kˤ ɛ
Zway	caːlɛn	caːlɛⁿ	tʃ aː l ɛ̃
Gafat	salɛ	šalɛ	ʃ a l ɛ
Soddo	saːlɛ	šaːlɛ-	ʃ aː l ɛ
Mesmes	haˈroː-	haroː-	h a r oː
Hebrew	yadha	yaːðaʕ	j aː ð a ʕ
Ugaritic	y d '	ydʕ	j d ʕ
Aramaic	idha	iðaʕ	i ð a ʕ
Akkadian	idu	iduː	i d uː
Mɛhri	ɤaruub	ɣɛroːb	ɣ ɛ r oː b
Gibbali	yuɤɔreb	ɣarɔb	ɣ a r ɔ b
Soqotri	ʕeerob	ʕerob	ʕ e r o b
Moroccan Arabic	ɛref	ʕref	ʕ r e f

Table 2: Individual form fixes applied to the dataset.

We see here many examples for the corrections applied. fixing encodings (ɤ→ɣ), fixing normalized forms (s→š), fixing spacing, representing Hebrew and Aramaic fricative consonants with single graphemes, disambiguating apostrophes to the voiced pharyngeal ʕ, pharyngealization, ejective or stress marks.

Perhaps most challengingly were the three South Arabian languages (Mehri, Jibbali, Harsusi) that lacked clear source attribution, requiring detective work to trace them to Johnstone's dictionary series through systematic lexical comparison. When discrepancies emerged between the dataset and lexicographic sources, I would compare original form against lexicon entry, apply lexicon form when clearly superior (correct the transcription when similar). When multiple values are possible, I would consider expanding the forms.

The process of choosing between multiple plausible words for a single concept is fraught with challenges and has significant implications for any downstream analysis. As Chaim Rabin noted regarding Morris Swadesh's lexicostatistical lists, "Swadesh insists, rightly, that in each case only one word should be considered, namely, the most commonly used one". However, Rabin immediately highlights the practical difficulty of this rule: "We have no frequency lists for any of the languages studied, and must go by guess-work; besides, frequency in a certain group or type of literary text may have been quite different from frequency in common speech".

This "guess-work" remains a persistent issue. A linguist's choice can be influenced by how well a word matches entries from other languages, but this can lead to misinterpretations (Bennett 1998). A wordlist creator might select a less common word simply because it appears to be a cognate with words in other languages, thus skewing

the data towards a particular hypothesis. As recent work by Snee et al. (2025) shows, such decisions lead to a great deal of variation in multilngual wordlists.

4 Usage Examples

The curated dataset comprises 25 language varieties linked to Glottocodes and 97 concepts tied to Concepticon sets, accumulating a total of 2,396 word forms, which resolve into 2,150 cognates distributed across 665 cognate sets. The word forms themselves are built from a total of 110 different phonemes that are all referenced in the CLTS reference catalog for speech sounds. On average, the language varieties in the sample consist of 37 distinct speech sounds. In this form, the dataset qualifies for the inclusion in future versions of the Lexibank repository, given that with the latest version (2.0), Lexibank only includes datasets for which standardized phonetic transcriptions are available (Blum et al. 2025).

The utility of the standardized transcriptions is immediately apparent when analyzing the data further with the help of the EDICTOR tool (List et al. 2025b, https://edictor.org), that allows us to inspect individual cognate sets in phonetically aligned form. In order to do so, one only has to convert the CLDF data to the internal wordlist format used by EDICTOR. This can be done by using a dedicated terminal command provided by the EDICTOR package that takes a CLDF dataset as input and returns a TSV-file that can be directly edited with the help of EDICTOR.

After installing EDICTOR with `pip` (`pip install edictor`) on the commandline, we first define a small code snippet in the file `preprocessing.py`. This makes sure that the cognate sets, which are not globally defined in the original data, but rather per concept slot, are converted into numerical values.

```
def run(wordlist):
    wordlist.add_entries(
        "cog",
        "cognacy,concept", l
        ambda x, y: x[y[0]] + "-" + x[y[1]])
    wordlist.renumber("cog")
    return wordlist
```

To convert the CLDF data , we download the latest version of the kitchensemitic dataset from the Lexibank repository with the help of GIT in the same folder in which we have stored the preprocessing file. We can then use the wordlist sub-command from the EDICTOR package to convert the data to EDICTOR's required TSV format and open the app directly in a local server.

```
$ git clone https://github.com/lexibank/kitchensemitic --depth=1 --
version=v2.1
$ edictor wordlist --dataset=kitchensemitic/cldf/cldf-metadata.json
--preprocessing=preprocessing.py --name=kitchensemitic.tsv --
addon=cogid_cognateseid:cognacy
$ edictor server
```

We can now open the EDICTOR app at the URL https://localhost:9999 in the webbrowser and open the file `kitchensemitic.tsv` in the FILES tab of the application. Having done this, we can add a new column ALIGNMENT and compute alignments for the dataset (the HELP tab will offer more detailed information on how to do this in the app). This allows us now to inspect individual entries in their automatically aligned form.

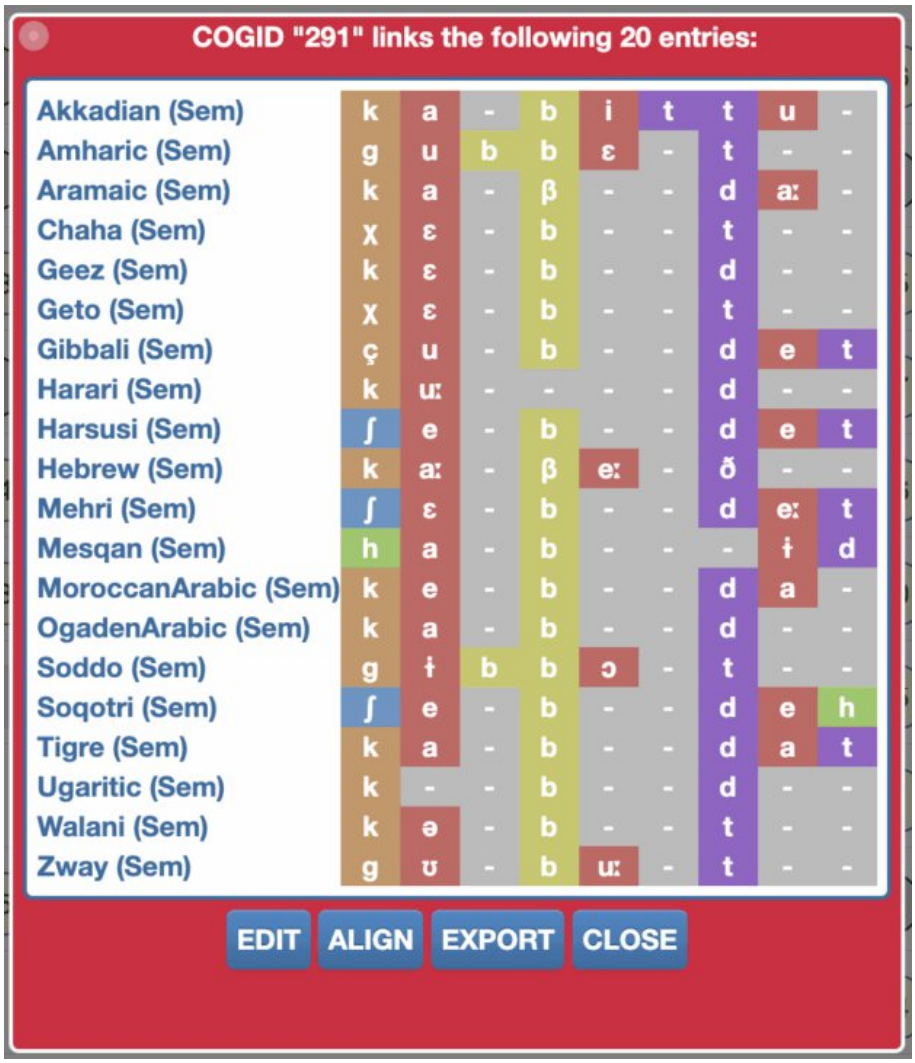

Figure 1: Phonetic alignment of 20 entries for “liver” in the Semitic languages in the sample.

For instance, the Proto-Semitic term **kabid-*, meaning ‘LIVER’ (Kogan 2015), was retained across a vast majority of the Semitic languages surveyed, providing a perfect

lens through which to inspect phonetic evolutionary phenomena. From the alignment, shown in a screenshot in Figure 1, we can observe several consonant changes when comparing the root form with the reflexes, including voicing and devoicing (k > g, d > t), instances of palatalization (k > ʃ), sound lengthening (b > bb, t > tt), fricativization (b > β, k > χ, d > ð), and retraction (χ > h).

5 Outlook

This work on the Semitic dataset demonstrates how the standardization of phonetic transcriptions can enable meaningful computational analyses, that may in turn help to obtain new insights into linguistic relationships. The dataset with standardized phonetic transcriptions is now available as part of the Lexibank repository (https://github.com/lexibank/kitchensemitic), and has been archived with Zenodo (Version 2.1, https://doi.org/10.5281/zenodo.17510140). In this form, it may be used for additional comparative studies of Semitic languages. More importantly, however, I would hope that it might serve as a starting point for the creation of larger integrated datasets on Semitic languages that are amenable for computational analysis.

References

Anderson, Cormac, Tiago Tresoldi, Thiago Chacon, Anne-Maria Fehn, Mary Walworth, Robert Forkel, and Johann-Mattis List. 2018. “A Cross-Linguistic Database of Phonetic Transcription Systems.” *Yearbook of the Poznan Linguistic Meeting* 4 (1): 21–53. https://doi.org/10.2478/yplm-2018-0002

Blum, F., C. Barrientos, J. Englisch, R. Forkel, S. Greenhill, C. Rzymski, and J.-M. List (2025): Lexibank 2: pre-computed features for large-scale lexical data [version 2; peer review: 3 approved]. *Open Research Europe* 5.126. 1-19. https://doi.org/10.12688/openreseurope.20216.2

Bender, Marvin Lionel. 1971. “The Languages of Ethiopia: A New Lexicostatistic Classification and Some Problems of Diffusion.” *Anthropological Linguistics*, 165–288. https://www.jstor.org/stable/30029540

Bennett, Patrick R. 1998. *Comparative Semitic Linguistics: A Manual*. Eisenbrauns.

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): 1–10. https://doi.org/10.1038/sdata.2018.205

Forkel, R., J.-M. List, C. Rzymski, and G. 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). 10578-10583. https://aclanthology.org/2024.lrec-main.925

Forkel, Robert, Simon J. Greenhill, Hans-Jörg Bibiko, Christoph Rzymski, Tiago Tresoldi, and Johann-Mattis List. 2021b. *PyLexibank. The Python Curation Library for Lexibank [Software Library, Version 2.8.2]*. Geneva: Zenodo. https://doi.org/10.5281/zenodo.2630582

Forkel, Robert, and Johann-Mattis List. 2020. “CLDFBench. Give Your Cross-Linguistic Data a Lift.” In *Proceedings of the Twelfth International Conference on Language Resources and Evaluation*, 6997–7004. Luxembourg: European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2020/pdf/2020.lrec-1.864.pdf

Gelb, Ignace Jay, Benno Landsberger, A. Leo Oppenheim, Erica Reiner, and Verlag JJ Augustin. 1956. *Assyrian Dictionary*. Vol. 1. 1. The Institute.

Harrell, Richard S. and Harvey Sobelman (editors) (2007): “A Dictionary of Moroccan Arabic Moroccan-English, English-Moroccan.” Waschington: Georgetown University Press.

Hammarström, Harald, Martin Haspelmath, Robert Forkel, and Sebastian Bank. 2025. *Glottolog [Dataset, Version 5.2.1]*. Leipzig: Max Planck Institute for Evolutionary Anthropology. https://glottolog.org

Häuser, Luise, and Johann-Mattis List. 2025. “Lexibench: Towards an Improved Collection of Benchmark Data for Computational Historical Linguistics.” *Computer-Assisted Language Comparison in Practice: Tutorials on Computational Approaches to the History and Diversity of Languages* 8 (1): 9–16. https://doi.org/10.15475/calcip.2025.1.2

Johnstone, T. M., University of London School of Oriental, and African Studies. 1977. *Ḥarsūsi Lexicon and English-Ḥarsūsi Word-List*. School of Oriental and African Studies. Oxford: Oxford University Press.

———. 1981. *Jibbāli Lexicon*. School of Oriental and African Studies. Oxford: Oxford University Press.

Johnstone, T. M., and G. R. Smith. 1987. *Mehri Lexicon and English-Mehri Word-List*. School of Oriental; African Studies, University of London.

Kitchen, Andrew, Christopher Ehret, Shiferaw Assefa, and Connie J. Mulligan. 2009. “Bayesian Phylogenetic Analysis of Semitic Languages Identifies an Early Bronze Age Origin of Semitic in the Near East“. *Proceedings of the Royal Society B: Biological Sciences* 276 (1668): 2703–10. https://doi.org/10.1098/rspb.2009.0408

Kitchen, Andrew, Christopher Ehret, Sheferaw Assefa, and Connie J. Mulligan. 2025. “CLDF dataset derived from Kitchen et al.’s “Bayesian Phylogenetic Analsyis of Semitic Languages” from 2009“. Leipzig: Max Planck Institute for Evolutionary Anthropology. https://doi.org/10.5281/zenodo.17510140

Kogan, Leonid. 2015. *Genealogical Classification of Semitic: The Lexical Isoglosses*. Berlin and New York: Walter de Gruyter.

Leslau, W. 1938. *Lexique Soqoṭri: (Sudarabique Moderne) Avec Comparaisons Et Explications Étymologiques*. Collection Linguistique. C. Klincksieck.

List, Johann-Mattis, Annika Tjuka, Frederic Blum, Alžběta Kučerová, Carlos Barrientos Ugarte, Christoph Rzymski, Simon J. Greenhill, and Robert Forkel. 2025. “CLLD Concepticon [Dataset, Version 3.4.0]“. Leipzig: Max Planck Institute for Evolutionary Anthropology. https://concepticon.clld.org

List, Johann-Mattis, Kellen Parker van Dam, and Frederic Blum. 2025b. “EDICTOR 3. An Interactive Tool for Computer-Assisted Language Comparison [Software Tool, Version 3.1]“. Passau: MCL Chair at the University of Passau. https://edictor.org

List, Johann-Mattis, Robert Forkel, Simon J. Greenhill, Christoph Rzymski, Johannes Englisch, and Russell D. Gray. 2022. “Lexibank, a Public Repository of Standardized Wordlists with Computed Phonological and Lexical Features.” *Scientific Data* 9 (1): 316. https://doi.org/10.1038/s41597-022-01432-0

Miller, John and List, Johann-Mattis. 2025. Adding Standardized Transcriptions to Panoan and Tacanan Languages in the Intercontinental Dictionary Series. *Computer-Assisted Language Comparison in Practice*, 7.2: 69-77. https://doi.org/10.15475/calcip.2024.2.3

Moran, Steven, and Michael Cysouw. 2018. *The Unicode Cookbook for Linguists: Managing Writing Systems Using Orthography Profiles*. Berlin: Language Science Press. https://langsci-press.org/catalog/book/176

Rabin, Chaim. 1975. “Lexicostatistics and the Internal Divisions of Semitic.” *Hamito-Semitica. The Hague, Paris*, 85–99.

Snee, D., L. Ciucci, A. Rubehn, K. van Dam, and J.-M. List (2025): Unstable Grounds for Beautiful Trees? Testing the Robustness of Concept Translations in the Compilation of Multilingual Wordlists. In: Proceedings of the 7th Workshop on Research in Computational Linguistic Typology and Multilingual NLP (SIGTYP 2025). 16-28. https://aclanthology.org/2025.sigtyp-1.3/

Swadesh, Morris. 1952. “Lexico-Statistic Dating of Prehistoric Ethnic Contacts: With Special Reference to North American Indians and Eskimos.” *Proceedings of the American Philosophical Society* 96 (4): 452–63. https://www.jstor.org/stable/3143802

Supplementary Materials
The dataset presented in this study is curated on GitHub (https://github.com/lexibank/kitchensemitic, Version 2.1) and archived with Zenodo (https://doi.org/10.5281/zenodo.17510140).