Formulaic Language in Historical Linguistics: data, methods, tools, and theory, Helsinki, 2-3 June 2025

Keywords: formulaic language, historical linguistics, corpus linguistics, NLP, language technology, philology, repetition, genre

This is the first call for abstracts for a conference on formulaicity in the linguistic and philological research of historical language varieties. Please, mark the dates on your calendars.

The aim of the conference is to discuss the multiple roles formulaicity plays in historical language data, to examine the advances of other fields in the analysis and management of formulaic texts, and to evaluate how these advances can be applied to historical linguistic research settings.

Omnipresence of formulaic language

Many linguists, philologists, and language technologists, especially those who work on historical language varieties, are used to formulaic language and the challenges it poses to the interpretation of their research data.

Various source types that are important for linguistic and philological study, both qualitative and quantitative/computational, are characterized by formulaic language. In historical contexts lists, inventories, records, proceedings, contracts, official and private letters, dedications, prayers, and technical treatises, to name but a few, may be the only substantial sources available from a certain period, area, or social stratum. In some cases such texts are paradoxically also the only sources that reflect the diachronic development of the language in question while literary texts keep on perpetuating age-old grammatical and stylistic conventions, which enshroud most linguistic evolution. On the other hand, literary language may also be highly formulaic. Certain conventional literary devices and even some historical genres, such as oral poetry or traditional Japanese theatre, largely rely on the use of prefabricated linguistic building blocks.

There are also a plethora of modern text types that contain or are largely composed of formulaic elements: court records, dictated medical notes, buy and sell notices, weather forecasts, social media objects, etc., with some of them involving equally prefabricated multimodal elements. And if formulaicity is considered a continuum, including all types of phrasal lexical items, the major part of human language, if not all, is constituted of structures of varying degree of formulaicity: sociolinguistics and historical pragmatics emphasise the functional-communicative role of formulaic utterances in specific social contexts, while construction grammar posits that human language consists of constructions that are more or less schematic pairings of linguistic patterns with meanings.

Recent advances

Because of this omnipresence of formulaic language, formulaicity has emerged as an important theme in various fields in the past decades. Branches of applied linguistics, including language acquisition, have paid growing attention to formulaic expressions and repetitiveness in communication. There have been advances in the corpus-driven approaches to multi-word expressions in modern languages. Language-technological solutions have been developed to tackle formulaic data in various practical contexts. Social and political historians have become increasingly interested in the role of formulaic language in historical sources: how it should be defined, what additional information it carries in historical documents, how it affects their analysis, and how it can (or cannot) be identified in and extracted from text archives. This interest found a manifestation in the Formulaic Language in Historical Research and Data Extraction conference organized by the Resolutions Published in a Computational Environment (REPUBLIC) project at the Huygens Institute in Amsterdam on 7–9 February 2024 (see the Proceedings), of which conference this conference can be seen as a linguistic spin-off.

In spite of these advances, the phenomenon of *formulaic language* is still mainly approached pre-theoretically in many fields of language-related study. This is all the more acute in historical linguistics, where so many sources are highly formulaic and where the role of formulaicity is, perhaps, even more crucial to the correct interpretation of sources than it is in modern-day contexts familiar to us; for example, in epigraphy, the restoration of damaged inscriptions relies upon the identification of formulaic expressions, which are employed by experts to reconstruct the original text.

***

We invite proposals for presentations that are related to one or more of the following broad themes:

1) The **definition(s)** of formulae/formulaic language from a linguistic/philological point of view. Frequency counts, co-occurrence patterns of words and/or constructions, and fixed multi-word expressions as single processing units are central concepts in specific subfields of modern linguistics. How can they be applied to the conceptualization and analysis of formulaicity in historical language data? How do the linguistic definitions of formulae relate to the definitions proposed in other disciplines, historical and not, such as diplomatics, literary/poetry studies, cognition studies, communication studies, information extraction, and text reuse detection?

2) Formulaicity results in **repetition** in corpora that consist of several formulaic texts of the same type (e.g., epigraphical databases, documentary collections, scientific texts). This elicits the question if and, if so, to what extent such repetitive data can be used in (quantitative/statistical corpus-)linguistic research and whether there is something that can be done to avoid or mitigate the skewing effect of formulaicity-induced over-representation in the corpus-linguistic analysis of historical language data. How are repetitive research settings best operationalized for historical linguistics? Which NLP methods are best applicable to them?

3) Formulaic language consists of prefabricated expressions of differing extent and rigidity indexed for particular conditions of use. Especially in formal texts, the formulae represent "**someone else's language**" which the writer adapts to their own language. Thus, the linguistic features of formulaic phrases do not necessarily reflect the linguistic competences of the writer; formulae may even contain vocabulary and grammar that is no longer or that has never been present in the language in which a specific formulaic text is written. This often provokes errors or hypercorrections. The question again arises whether and, if yes,

how the researcher can cope with such diachronic and/or stylistic diversity within a text and what consequences it has to (diachronic or socio)linguistic or philological analysis. How is variation in formulaic sequences to be understood and operationalized? To which extent is such variation consciously introduced?

4) Formulaic language always has a function, a role to play in a text. Such **discourse-organizational functions** vary from one communicative-pragmatic context to another. How do formulaic sequences operate within broader textual environments within which those sequences occur? What kinds of regularities are found between the use of formulaic language and genres/text types? How do formulaic sequences relate to discourse segmentation and to what extent is that standardized? Is there any visual marking at play (multimodality)? What is the relation of formulaic sequences to paratextual elements (broadly defined)?

Case studies on specific datasets, methods, or computational tools, as well as broader theoretical discussions, are welcome, providing that the presentations are founded on empirical evidence which, as well as its processing and analysis, is clearly and sufficiently explained. The presentations can be of 20 or 30 minutes followed by 10 minutes of discussion. The extra 10 minutes are reserved for presentations with a detailed explanation of the research data and how it is processed (something one does not usually have enough time for). We encourage this latter approach because we are confident that a more thorough description of the research process than usual helps others assess its validity and, if need be, apply it to their own datasets. Please, indicate in your abstract proposal whether and why you prefer to have the 10 extra minutes. The final decision rests with the scientific board.

The proposals of 250 to 500 words (excl. references), followed by a short academic bio, should be sent as docx or odt to timo.korkiakangas [ät] helsinki.fi by 21 October, 2024. The notifications of acceptance will be announced in November. The conference will be held in English. Coffee and some meals will be served. A few bursaries of 200 to 300 euros will be available for PhD students or other early career academics without travel funds, depending however on the final budget of the conference. Please, specify in your proposal if you apply for a bursary. Once at the conference, let us discuss together the possibility of publishing selected contributions as a special issue in some relevant journal or other platform.

The second call for abstracts will be sent in early September 2024.

The conference is organized by the Academy of Finland project “The learning of Latin in the 8th to 12th century: a linguistic approach to medieval Latin literacies” (PI Timo Korkiakangas) in collaboration with the Classical Philological Society of Finland. The venue is Tieteiden talo in the centre of Helsinki. The scientific board consists of

- Timo Korkiakangas (Academy of Finland/University of Helsinki)
- Marja Vierros (Professor of Classical Philology, University of Helsinki)
- Tommi Jauhiainen (PI of the project Automatic Classification and Analysis of Texts from Egyptian Antiquity, University of Helsinki)
- Margherita Fantoli (Assistant Professor of Digital Humanities, KU Leuven)
- Klaas Bentein (Research Professor, Ancient Greek, Department of Linguistics, Ghent University)
- Joanna Kopaczyk (Professor of Scots and English Philology, University of Glasgow)

Some bibliography

Please, see the Proceedings of the Formulaic Language in Historical Research and Data Extraction conference organized by the Resolutions Published in a Computational Environment (REPUBLIC) project at the Huygens Institute in Amsterdam on 7–9 February, 2024.

***

Bentein, Klaas. 2023. “A Typology of Variations in the Ancient Greek Epistolary Frame (I–III AD)”, in Giannakis, Georgios, Crespo, Emilio, de La Villa, Jesús & Filos, Panagiotis (eds), *Historical Linguistics and Classical Philology*, De Gruyter, 415–457.

Biber, Douglas. 2009. "A Corpus-Driven Approach to Formulaic Language in English: Multi-Word Patterns in Speech and Writing", in *International Journal of Corpus Linguistics* 14 (3): 275–311.

Buerki, Andreas. 2020. *Formulaic Language and Linguistic Change: A Data-Led Approach*, CUP.

Bybee, Joan L. & Torres Cacoullos, Rena. 2009. “The role of prefabs in grammaticization: How the particular and the general interact in language change”, in Corrigan, Roberta & al. (eds), *Formulaic Language, volume 1: Distribution and historical change*, Benjamins, 187–218.

Ceccarelli, Paola. 2018. “Letters and Decrees: Diplomatic Protocols in the Hellenistic Period”, in Ceccarelli, P., Doering, L., Fögen Th. & Gildenhard I. (eds), *Letters and Communities: Studies in the Socio-Political Dimensions of Ancient Epistolography*, OUP, 147–184.

Corrigan, Roberta, Moravcsik, Edith A., Ouali, Hamid & Wheatley, Kathleen (eds), *Formulaic Language, volume 1: Distribution and historical change*, Benjamins, 2009.

Corrigan, Roberta, Moravcsik, Edith A., Ouali, Hamid & Wheatley, Kathleen (eds), *Formulaic Language, volume 2: Acquisition, loss, psychological reality, and functional explanations*, Benjamins, 2009.

Culley, Robert. 1967. *Oral Formulaic Language in the Biblical Psalms*, University of Toronto Press.

Granger, Sylviane. 2018. “Formulaic sequences in learner corpora: Collocations and lexical bundles”, in Siyanova-Chanturia, A. & Pellicer-Sanchez, A. (eds), *Understanding Formulaic Language: A Second Language Acquisition Perspective*, Routledge, 228–247.

Gruber, M. Catherine. 2009. “Accepting responsibility at defendants' sentencing hearings: No formulas for success”, in Corrigan, Roberta & al. (eds), *Formulaic Language, volume 2: Acquisition, loss, psychological reality, and functional explanations*, Benjamins, 545–566.

Kiparsky, Paul. 1976. “Oral poetry: some linguistic and typological considerations”, in Stolz, Benjamin A. & Stoll Shannon, Richard (eds), *Oral Literature and the Formula*, Ann Arbor, 73–106.

Koolen, Marijn & Hoekstra, Rik. 2022. "Detecting formulaic language use in historical administrative corpora", in *Proceedings of the Computational Humanities Research Conference 2022, Antwerp, Belgium, December 12–14, 2022*, 127–151. https://ceur-ws.org/Vol-3290/long_paper5740.pdf

Kopaczyk, Joanna. 2013. *The Legal Language of Scottish Burghs: Standardization and Lexical Bundles (1380–1560)*, OUP.

Korkiakangas, Timo. 2022. “From memory or formulary: how were medieval documentary formulae reproduced?”, in *Mirator* 22, 4–24. https://doi.org/10.54334/mirator.v22i1.119760

Korkiakangas, Timo. 2023. “Documentary formulae as text reuse templates: *constat* and *manifestus* clauses in early medieval Latin charters”, in *Digital Medievalist* 16, 1–44. https://doi.org/10.16995/dm.8195

Kuiper, Koenraad. 2004. “Formulaic performance in conventionalised varieties of speech”, in Schmitt, Norbert (ed.), *Formulaic Sequences: Acquisition, Processing, and Use*, Benjamins, 37–54.

Kuiper, Koenraad. 2009. *Formulaic Genres*, Palgrave.

Netz, Reviel. 2003. *The Shaping of Deduction in Greek Mathematics: A Study in Cognitive History*, CUP.

Rio, Alice. 2009. *Legal Practice and the Written Word in the Early Middle Ages: Frankish Formulae, c.500–1000*, CUP.

Sabatini, Francesco. 1965. “Esigenze di realismo e dislocazione morfologica in testi preromanzi”, in *Rivista di Cultura Classica e Medievale* 7, 972–998.

Sahala, Aleksi & Lindén, Krister. 2020. “Improving Word Association Measures in Repetitive Corpora with Context Similarity Weighting”, in Fred, A. L. N. & Filipe, J. (eds), *Proceedings of the 12th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management, IC3K 2020, Volume 1: KDIR, Budapest, Hungary, November 2–4, 2020*. SCITEPRESS, 48–58.

Scafuro, Adele. 2014. “Decrees for Foreign Judges: Judging Conventions – Or Epigraphic Habits?”, in Gagarin, M. & Lanni, A. (eds), *Symposion 2013*, Austrian Academy of Sciences Press, 365–395.

Schmitt, Norbert (ed.). 2004. *Formulaic Sequences: Acquisition, Processing, and Use*, Benjamins.

Stefanowitsch, Anatol & Gries, Stefan. 2003. “Collostructions: Investigating the interaction of words and constructions”, in *International Journal of Corpus Linguistics* 8(2), 209–243.

Vierros, Marja. 2018. “Copying practices in Ptolemaic Egypt: A discussion based on agoranomic contracts from Pathyris”, in *Tyche* 33, 207–230 (+ Tafel 11). https://tyche.univie.ac.at/index.php/tyche/issue/view/386

Vierros, Marja. 2018. “The Greek of the Petra Papyri” in Arjava, A., Frösén, J. & Kaimio, J. (eds), *The Petra Papyri V*, American Center of Oriental Research, 8–34.

Wray, Alison. 2008. *Formulaic Language: Pushing the Boundaries*, OUP.