RiCL Research in Corpus Linguistics

Review of Martín Arista, Javier and Ana Elvira Ojanguren López. 2024. *Structuring Lexical Data and Digitising Dictionaries: Grammatical Theory, Language Processing and Databases in Historical Linguistics*. Leiden: Brill. 412 pp. ISBN: 978-90-04-70266-0. https://doi.org/10.1163/9789004702660

Silvia Saporta Tarazona
University of La Rioja / Spain

In today's world, Artificial intelligence (AI) has revolutionised our understanding of technology, which has been primarily attributed to the emergence of Large Language Models (LLMs) and their exceptional ability to perform Natural Language Processing (NLP) tasks such as text generation, summarisation, sentiment analysis or machine translation. Considering this scenario, an effective organisation of the available linguistic data is deemed imperative, especially in the case of historical languages, as datasets and databases are extensively employed by NLP applications. To accomplish this, Martín Arista and Ojanguren López's proposal establishes the underpinnings of corpus compilation through historical lexicography and lexicology from a two-pronged approach, namely, the inclusion of research procedures arising from lexical databases and language processing, and the configuration of historical dictionaries for lexicography and corpus analysis. This book constitutes a contribution to the avenue of research in the processing of historical texts with computational and artificial intelligence approaches also pursued in recent publications such as Villa and Giarda (2023), Martín Arista (2024) and Martín Arista *et al*. (2025).

Whilst Chapter 1 discusses the current state of research and underlines the major contents of the book, the ensuing chapters, which are structured in two parts titled Lexical databases and language processing in digital historical lexicography" and "Structuring historical lexicons for lexicography and corpus analysis," delve into languages such as Old English (OE), Old Church Slavonic (OCS), Sanskrit or Greek and explore lexical processes including lemmatisation, encoding, semantic structure, lexical representation, dialectal lexicography or digitisation, among others. Thus, Chapter 2, "String similarity

Research in Corpus Linguistics 13/2: 232–240 (2025). ISSN 2243-4712. <https://ricl.aelinco.es>
Asociación Española de Lingüística de Corpus (AELINCO)
DOI 10.32714/ricl.13.02.09

measures for evaluating the lemmatisation in Old Church Slavonic," authored by Ilia Afanasev and Olga Lyashevskaya, seeks to evaluate the most significant metrics for the lemmatisation of an OCS corpus-based dictionary, as this language has not been efficiently digitised despite its limited number of sources. Having demonstrated that the implementation of accuracy score metrics on two distinct datasets is not able to reflect central erroneous patterns, special emphasis has been placed on string similarity measures, that is to say, the Levenshtein distance (Levenshtein 1966), the Damerau-Levenshtein distance (Damerau 1964) and the Jaro-Winkler distance (Jaro 1989; Winkler 1990), which aim for an efficient, stable and scalable language model tuning.

In Chapter 3, entitled "Encoding the specificities of encyclopedias," Alice Brenon underlines the need for a shift in the current encoding of encyclopedias, for the latter have proved to be dissimilar from conventional dictionaries as "not only do entries tend to be longer […], they often have a deeper structure." (p. 60). The impossibility to apply the dictionary module comprised in the encoding standard XML-TEI to works such as *La Grande Encyclopédie* has led the author to develop a new encoding scheme based on graph theory, which consistently adheres to XML-TEI, so as to fully represent the complexity entailed in encyclopedic content and ensure its accessibility to the scientific community. The subsequent chapter, "Challenges in the process of retro-digitisation of Croatian grammar books before Illyrism" by Marijana Horvat, Martina Kramarić and Ana Mihaljević, examines the main obstacles encountered in the retro-digitisation of a selection of pre-standard Croatian grammar books in an effort to design a model capable of such endeavour. Among these, encoding has emerged as a primary difficulty, for it demands multilevel annotation by means of TEI tags which are currently inexistent in the case of Illyrian grammar books. In view of this, Horvat, Kramarić and Mihaljević have elaborated a TEI Header along with a terminology index which provide "essential insights into the digitisation process of all Croatian historical documents" (p. 8) and pave the way for a comprehensive description of Croatian language history, thus becoming the most substantial contribution of this project. Ellert Thor Johannsson identifies in Chapter 5, entitled "The evolution of a dictionary of Old Norse Prose (ONP): from a collection of citations to a digital resource," the crucial aspects for the configuration of a digital dictionary that compiles the vocabulary found in Norwegian and Icelandic medieval manuscripts. To this end, the phases required for the transition from a collection of citations to an online digital resource are painstakingly noted, which not only include the organisation and registration of data in a lexical database that serves as the fundamental element of the digital dictionary, but also electronic applications that supplement its data with the aim of enabling users to interact with these resources and enhance their understanding of Old Norse language, literature and culture. In Chapter 6, "Agile lexicography: rapid dictionary prototyping with R Shiny, with examples from projects on Sanskrit and Tibetan," Ligeia Lugli illustrates the potential benefits identified in Shiny, an open-source framework which develops web applications by means of R programming language (Chang *et al*. 2021). Its extra functionality was exemplified in projects involving the compilation of historical dictionaries, namely, a dictionary and thesaurus of Buddhist Sanskrit and a diachronic corpus of Tibetan verb valency (Lugli 2019; Lugli *et al.* 2022; Pagel *et al*. 2021). The simplicity and flexibility of the aforementioned tool has revealed itself as particularly well-suited for data-intensive applications, and has also facilitated the creation of new functionalities, a creative design and prototyping, and a data-pipeline while adhering to time and budget constraints. Despite the fact of having minor shortcomings related to scalability, speed or latency, it is consistently regarded as a paramount framework for historical digital lexicography.

Javier Martín Arista addresses in the seventh chapter, entitled "Interface of Old English dictionaries in database format: toward a knowledge base," the limited presence of OE in textual and lexical databases, which mainly stems from the lack of annotation and the incompatibility of its lexical resources. With this state of affairs, an interface able to merge information from different sources is introduced, in pursuit of providing full availability and access to its linguistic data and enhancing the compilation of Old English data. This undertaking entails the digitisation of lexicographical sources and the population of the relational database, as well as the development of knowledge graphs, giving rise to a knowledge base which allows multivariable queries and ensures compatibility with NLP and AI resources. The following contribution, "Bosworth-Toller's Anglo-Saxon Dictionary online" by Ondřej Tichý and Martin Roček (Chapter 8), sheds light on the digitisation process of the *Anglo-Saxon Dictionary*, "the only fairly comprehensive dictionary of Old English available to both experts and the public." (p. 184). Furthermore, its most significant obstacles are also underlined, devoting special attention to the disambiguation of references, the quality of its database and the accuracy of its digital representation, as well as future updates. Notwithstanding these remaining challenges, the digital Dictionary has effectively surpassed its printed version and constitutes a formidable lexicographical resource in the field, which will eventually develop into a high-quality dataset employed by a wide range of users.

Chapter 9, entitled "The adjective *gesælig* in Old English prose: towards the characterization of the lexical field of holiness in Old English" by Ondřej Fúsik and Alena Novotná, aims to provide a detailed portrayal of the OE lexical field of being holy, with a particular focus on the adjective *(ge)sælig*, so as to clarify its meaning and its existing relation with similar adjectives. To address said purpose, the lexicographic profile of this adjective has been examined by means of the *Thesaurus of Old English*, the *Bosworth-Toller Dictionary* and the *Dictionary of Old English*, although the *York-Toronto-Helsinki Parsed Corpus of Old Prose* has demonstrated to be key for the compilation of relevant data. Fúsik's (2018) prior contributions have also been deemed indispensable to enable meaningful comparisons with the adjective *halig*, which have evidenced that *(ge)sælig* might be more accurately translated as 'blessed', and that it is primarily used in the predicative function as opposed to the former, which is eminently attributive. Despite the fact that both adjectives are found in similar genres, the authors conclude that the holiness of *(ge)sælig* derives from good behaviour, whereas the blessedness of *halig* originates directly from God. In Chapter 10, entitled "Cultural labels as a means of organizing semantic structure of lexemes in an explanatory synchronic historical dictionary," Alenka Jelovšek thoroughly analyses the existing label classification for historical and specialised dictionaries (Hausmann 1989; Atkins and Rundell 2008) to present a universal typology based on their function. According to the author, encyclopedic, linguistic and register labels exhibit greater efficacy compared to encyclopedic notes, for it has been noted in the labelling of the Dictionary of the sixteenth-century Slovenian literary language, a synchronic historical dictionary detailing the period of the Slovenian Reformation. In light of this, it could be claimed that systematised labels constitute a plausible method which simplify the use of dictionaries, promote the standardisation of historical resources and provide specific meanings considering historical, ideological and textual dimensions.

Chapter 11, "Organising the lexicon by means of grammatical behaviour: the verbal class of Deprive in Old English" by Miguel Lacalle Palacios, seeks to evaluate Old English verbal lexicon, in particular, those verbs denoting the meaning of depriving, along with the constructions and alternations typically associated with them. Thus, the Role and Reference Grammar (RRG) (Foley and Van Valin 1984; Van Valin and LaPolla

1997) and the framework of verbal classes and alternations (Levin 1993) have served as the foundation of the analysis. Equally significant for the extraction of data have been the diverse textual and lexicographical sources of the study, which comprise the *Dictionary of Old English Corpus*, the *York-Toronto-Helsinki Parsed Corpus of Old English Prose* or the *Thesaurus of Old English*. The findings of the undertaking have been able to illustrate the relationship between semantics and morpho-syntactic alternations, as well as proposing four alternations and two constructions that have enabled the evaluation of the aforementioned category of OE verbs. In Chapter 12, entitled "On lemmas and dilemmas again: problems in historical dialectal lexicography," Io Manolessou and Georgia Katsouda discuss the lemmatisation obstacles encountered in historical and dialectal lexicography, since headword selections may constitute a considerable challenge owing to the divergence of forms included in the same heading. The study has been conducted through specific instances observed in two main Greek lexicographic projects: the *Historical Dictionary of Modern Greek* (*ilne*) and the *Historical Dictionary of the Cappadocian Dialects*, which have provided the means for the compilation of specifically complex cases and their possible solutions. Furthermore, the criteria for headword selection has been established, although these are subjected to variation depending on the purpose of the dictionary, source availability or intended users.

Conversely, in Chapter 13, "Structuring the lexicon of Old English with syntactic principles: the role of deverbal nominalisations with aspectual and control verbs," Ana Elvira Ojanguren López explores the association between Old English semantic and syntax through the functions of deverbal nominalisations in aspectual and control verbs, which constitutes a substantial contribution, considering that the historical evolution of the English gerund relies on the acquisition of verbal properties by this type of nominalisations (Fischer 1992). To conduct such an endeavour, the theory of RRG (Van Valin and LaPolla 1997) has been employed as the theoretical framework of the study, whereas resources such as the *York-Toronto-Helsinki Parsed Corpus of Old English Prose* or the *Dictionary of Old English* have become pivotal for data retrieval. The results lead to claim that OE syntactic configurations, including derived constructions with deverbal nominalisations, are deemed a principle capable of structuring verbal lexicon. The closing chapter (Chapter 14), entitled "Assessing lexicographic obsolescence and historical frequency indicators in word entries in the OED: a corpus study of historical *-some* adjectival derivatives" by Chris Smith, implements a diachronic perspective to examine the field of obsolescence and low-frequency words in lexicographic databases such as the *Oxford English Dictionary* (OED). To this end, a systematic methodology is proposed to assess frequency labelling in a specific set of -some adjectival derivatives, which reveals the wide variability of obsolescence and relative frequency, and contributes to the study of lexicographic labelling and diachronic lexical competition.

Individual chapters have both strengths and limitations that deserve some comment. While the methodology of Chapter 2 is thorough, the study could benefit from more contextual examples illustrating lemmatisation problems. Moreover, some interpretation of the relationship of the metrics to general linguistic phenomena would strengthen the conclusions. The graph theory has revealed itself as an innovative approach, yet real-world examples of encoding problems and decisions would increase readability. Plus, the distinction between dictionaries and encyclopedias could be further clarified through structural diagrams, instead of relying on verbal descriptions only. Chapter 4 provides a strong contextual background, although the technical implementation of TEI headers draws more attention than the evaluation of the impact of the digitisation process. In spite of this, the grammar book comparison table is deemed excellent, although employing case studies to examine particular problematic elements could enhance comprehension among general audiences. Conversely, the author in Chapter 5 painstakingly illustrates database structure along with the evolution from analog to digital formats, but more in-depth critical examination of how lexicographical decisions affected the digitisation process would be highly beneficial. Additionally, more analysis of user experience with the final digital product could help to assess the impact of digitisation.

Chapter 6 possesses a strong practical focus and provides convincing examples, although more emphasis should be made on the limitations of Shiny for production environments. The comparison between different development approaches is particularly valuable, yet more quantitative metrics on performance would strengthen the argument. Chapter 7 constitutes a significant contribution to Old English lexicography through the development of the *Interface of Old English Dictionaries* (IOED), a component of the *Knowledge Base of Old English* (KBOE). Moreover, the methodology of converting relational databases to knowledge graphs represents an innovative approach to historical lexicography, whereas the combination of type analysis (lemmas) with token analysis (inflectional forms) provides a comprehensive resource for both historical corpus linguists and lexicographers. The chapter excels in demonstrating how inconsistencies across dictionaries (regarding headword spelling, vowel quantity, etc.) can be resolved through standardisation and normalisation in a database format. In addition to this, the graphs generated from edge lists demonstrate the potential for more efficient complex queries. Chapter 8 effectively documents the transformation process from print to digital dictionary by incorporating an excellent historical context. Nonetheless, alternatives to the chosen encoding approaches might have been discussed, and the influence of user needs on technical decisions could have received more focus.

Chapter 9 presents a thorough corpus-based investigation of the Old English adjective *(ge)sælig* and its position within the lexical field of holiness. While the study acknowledges limitations due to the restricted corpus (mainly religious texts), it makes a convincing case that translating *(ge)sælig* as ‘blessed’ rather than ‘happy’ may be more appropriate in many contexts. Chapter 10 proposes a three-class labeling system (encyclopedic, linguistic, register) and therefore contributes to bridging the gaps in previous classifications. However, the chapter could benefit from more concrete examples demonstrating how the proposed structure improves digital interoperability beyond theoretical frameworks. Chapter 11 applies the framework of verb classes and alternations with Role and Reference Grammar to analyse Old English verbs of depriving and convincingly demonstrates that grammatical behaviour constitutes a sounder basis for lexical organisation than meaning definitions alone. Plus, the chapter succeeds in presenting a rigorous approach that could be applied to other verbal classes in historical English. As it can be noted, Chapter 12 presents a strong analysis of lemmatisation issues in Greek lexicography with excellent discussion of methodology. Whilst the examples are detailed and well-chosen, a visual depiction of the decision tree for selecting headwords would enhance readability.

Chapter 13 analyses deverbal nominalisations with aspectual and control verbs in Old English. The methodology combining lexical database analysis with corpus linguistics offers insights into both synchronic and diachronic aspects of the language and demonstrates that nominal linked predications display semantic and syntactic configurations parallel to verbal linked predications. Furthermore, the analysis of the macrorole transitivity of nominal linked predications represents a particularly original contribution to the field. The author convincingly argues that the syntactic configurations of Old English can serve as a principled basis for structuring the verbal lexicon. Chapter

14 constitutes an impressive quantitative analysis of *-some* adjectives, although the methodology section could have connected the corpus testing techniques with the theoretical framework more clearly. In spite of the visual presentation of data in figures supporting the arguments about frequency patterns, more discussion of the implications for ongoing OED revisions would help general audiences to understand the connection between lexicographical theory and practice.

Overall, *Structuring Lexical Data and Digitising Dictionaries* represents a significant advance in historical lexicography and computational linguistics, particularly for diachronic English studies. The interdisciplinary approach of the volume, which combines traditional philological methods with cutting-edge computational techniques, offers innovative solutions to long-standing problems met when accessing structuring, and analysing historical lexical data. Across thirteen chapters, the contributors demonstrate how relational databases, knowledge graphs, corpus-based semantic analysis, and syntactic frameworks can organise lexical information from various sources. The book particularly excels in addressing practical problems of standardisation across dictionaries and corpora, given that it proposes methodologies for semantic field analysis, and establishes principled approaches to verb classification that underline the relationship between semantics and syntax. While some theoretical frameworks might be applied too rigidly to historical data, and certain computational methodologies would benefit from more rigorous evaluation, the book as a whole significantly improves our understanding of how to transform traditional lexicographical resources into structured digital datasets. This volume will undoubtedly become a reference work for future projects in digital historical lexicography.

References

Atkins, B. T. Sue and Michael Rundell. 2008. *The Oxford Guide to Practical Lexicography*. Oxford: Oxford University Press.

Chang, Winston, Joe Cheng, J. J. Allaire, Carson Sievert, Barret Schloerke, Yihui Xie, Jeff Allen, Jonathan McPherson, Alan Dipert and Barbara Borges. 2021. *Shiny*: https://CRAN.R -project.org/package=shiny

Damerau, Frederick J. 1964. A technique for computer detection and correction of spelling errors. *Communications of the ACM* 7/3: 171–176.

Fischer, Olga. 1992. Syntax. In Norman Blake ed. *The Cambridge History of the English Language II. 1066- 1476*. Cambridge: Cambridge University Press, 207–407.

Foley, William and Robert Van Valin. 1984. *Functional Syntax and Universal Grammar*. Cambridge: Cambridge University Press.

Fúsik, Ondřej. 2018. *Old English Prose Adjectives Meaning 'holy': Towards a Characterization of a Lexical Field*. Prague: Charles University dissertation.

Hausmann, Franz Josef. 1989. Die Markierung in einem allgemeinen einsprachigen Wörterbuch: eine Übersicht. In Franz Josef Hausmann, Oskar Reichmann, Herbert Ernst Wiegand and Ladislav Zgusta, eds. *Wörterbücher: Ein Internationales Handbuch zur Lexikographie*. Berlin: Walter de Gruyter, 649–657.

Jaro, Matthew A. 1989. Advances in record linkage methodology as applied to the 1985 Census of Tampa Florida. *Journal of the American Statistical Association* 84: 414–420.

Levenshtein, Vladimir I. 1966. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady* 10/8: 707–710.

Levin, Beth. 1993. *English Verb Classes and Alternations*. Chicago: University of Chicago Press.

Lugli, Ligeia. 2019. Smart lexicography for low-resource languages: Lessons learned from Sanskrit and Tibetan. In Iztok Kosem, Tanara Zingano Kuhn, Margarita Correia, José Pedro Ferreira, Maarten Jansen, Isabel Pereira, Jelena Kallas, Miloš Jakubíček, Simon Krek and Carole Tiberius eds. *Electronic Lexicography in the 21st Century: Smart Lexicography. Proceedings of the eLex 2019 Conference*. Sintra, 198–212.

Lugli, Ligeia, Matej Martinc, Andraž Pelicon and Senja Pollak. 2022. Embeddings models for Buddhist Sanskrit. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Jan Odijk and Stelios Piperidis eds. *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. Marseille: European Language Resources Association, 3861–3871.

Martín Arista, Javier. 2024. Toward a Universal Dependencies Treebank of Old English: Representing the morphological relatedness of un-derivatives. *Languages* 9/3: 76.

Martín Arista, Javier, Ana Elvira Ojanguren López and Sara Domínguez Barragán. 2025. Universal Dependencies annotation of Old English with spaCy and MobileBERT: Evaluation and perspectives. *Procesamiento del Lenguaje Natural* 75: 253–262.

Pagel, Ulrich, Edward Garrett, Ligeia Lugli and Christian Faggionato. 2021. *A Visual Dictionary of Tibetan Verb Valency*. Mangalam Research Institute. https://doi.org/10.5281/zenodo.5596064

Van Valin, Robert and Randy J. LaPolla. 1997. *Syntax: Structure, Meaning and Function*. Cambridge: Cambridge University Press.

Villa, Luca Brigada and Martina Giarda. 2023. Using modern languages to parse ancient ones: A test on Old English. *Proceedings of the 5th Workshop on Research in Computational Linguistic Typology and Multilingual NLP (SIGTYP 2023)*. Dubrovnik: Association for Computational Linguistics, 30–41.

Winkler, William Erwin. 1990. String comparator metrics and enhanced decision rules in the Fellegi-Sunter model of record linkage. In *Proceedings of the Section on Survey Research Methods*. American Statistical Association: 354–359.

Reviewed by
Silvia Saporta Tarazona
University of La Rioja
Department of Modern Philologies
C/ San José de Calasanz, 33
E-26004, Logroño, La Rioja
Spain
e-mail: silvia.saporta@unirioja.es