RAEL Revista Electrónica de Lingüística Aplicada

RAEL: Revista Electrónica de Lingüística Aplicada
Vol./Núm.: 24/1
Enero-diciembre 2025
Páginas: 221-225
Artículo recibido: 27/07/2025
Artículo aceptado: 03/10/2025
Artículo publicado: 31/01/2025
Url: https://rael.aesla.org.es/index.php/RAEL/article/view/763
Https://doi.org/ https://doi.org/10.58859/rael.v24i1.763

Martín Arista, J. and Ojanguren López, A. E. (2024). *Structuring Lexical Data and Digitising Dictionaries: Grammatical Theory, Language Processing and Databases in Historical Linguistics*. Leiden and Boston: Brill.

ISBN: 978-90-04-70265-3 (hbk), 978-90-04-70266-0 (ebk) (xiii+393 pages)
https://doi.org/10.1163/9789004702660

Marina Tamayo Campins
Universidad de La Rioja

1. INTRODUCTION

Structuring Lexical Data and Digitising Dictionaries: Grammatical Theory, Language Processing and Databases in Historical Linguistics (hereafter *SLDaDD*) is the 85th volume of the book series *Language and Computers* (Mair & Meyer, 1988). This series aims to provide comprehensive insight to theoretical, methodological, and empiric issues derived from the raise of technological advances as well as the pervasive digitization of textual material. For instance, the re-evaluation of corpus linguistics and how its methods progressively incorporate new technologies for visualizing and geolocating digital language data. As well as fulfilling the need for knowledge regarding the new ways of engaging with language in a world that continues to technologically evolve.

In turn, the edited collective volume at hand addresses the need for structured linguistic data, the incorporation of lexical databases and language processing, and the structuring of historical lexicons in terms of corpus analysis, emphasizing the integration of grammatical theories, language processing, and the development of databases in the evolving field of historical linguistics. The development of new Large Language Models (LLMs) has triggered the rapid evolution of Natural Language Processing (NLP). Nonetheless, these breakthroughs have yet to be applied in historical and minority languages, especially in their lexicology and lexicography, such as Old English. In the spirit of bridging said gap, this volume presents new methodologies for structuring linguistic data that allow for their integration within computational approaches.

Citar como: Tamayo Campins, M. (2025). Book review: Martín Arista, J. & Ojanguren López, A. E. (2024). *Structuring Lexical Data and Digitising Dictionaries: Grammatical Theory, Language Processing and Databases in Historical Linguistics*. *RAEL: Revista Electrónica de Lingüística Aplicada, 24*, 221-225. https://doi.org/10.58859/rael.v24i1.763

SLDaDD consists of fourteen chapters authored by internationally recognised scholars who offer theoretical approaches and applications for the organization of historical lexical data. It follows a twofold structure: "Part 1: Lexical Databases and Language Processing in Digital Historical Lexicography" (pp. 11-207) comprises chapters two to eight and focuses on the development and implementation of lexical databases and digitization processes; while "Part 2: Structuring Historical Lexicons for Lexicography and Corpus Analysis" (pp. 209- 393) covers chapters nine to fourteen and examines the organisation and analysis of lexical data from perspective that abides by semantic and grammatical principles.

2. PART 1: LEXICAL DATABASES AND LANGUAGE PROCESSING IN DIGITAL HISTORICAL LEXICOGRAPHY

Ilia Afanasev and Olga Lyashevskaya present in "String Similarity Measures for Evaluating the Lemmatisation in Old Church Slavonic" (pp. 13-35) a novel set of metrics for evaluating an Old Church Slavonic lemmatiser model. These metrics assess the model's neural network and are crucial for selecting the optimal Language Model for heterogeneous data.

Alice Brenon's "Encoding the Specificities of Encyclopedias" (pp. 36-62) highlights the contrasts between dictionaries and encyclopaedias while employing XML-TEI for encoding *La Grande Encyclopédie* (Morrissey & Roe), devising a new scheme that represents the complexities of encyclopaedias' entries.

Marijana Horvat, Martina Kramarić, and Ana Mihaljević introduce a project creating an open-access portal for historical Croatian grammar books, using TEI Header encoding and multilevel annotation of translated and transcribed material.

Ellert Thor Johannsson's chapter provides an account of the evolution of *A Dictionary of Old Norse Prose* into an online interactive dictionary, creating a lexical database with curated information presented in novel ways, including gamification elements.

Ligeia Luigli offers two case studies where R Shiny has proven valuable for developing lexicographical resources of historical and minority languages: a Buddhist Sanskrit dictionary and a Tibetan verb valency dictionary, highlighting Shiny's advantages for rapid prototyping and financial independence.

Javier Martín Arista presents an interface that compares and relates four dictionaries and three corpora of Old English, normalizing headwords and inflections to ensure comparability across sources and creating a knowledge base capable of answering extensive multivariable queries.

Ondřej Tichý and Martin Roček comment on the digitisation of Bosworth and Toller's *Anglo-Saxon Dictionary*, addressing the challenges of creating an open-access lexical database that overcomes the limitations of the print edition while maintaining fidelity to the source.

3. PART 2: STRUCTURING HISTORICAL LEXICONS FOR LEXICOGRAPHY AND CORPUS ANALYSIS

Ondřej Fúsik and Alena Novotná examine the meaning of the Old English adjective *(ge)sælig* through quantitative analysis, determining it likely meant "blessed" rather than "happy" and comparing its concept of holiness with that of *halig*.

Alenka Jelovšek assesses different sets of labels used in synchronic dictionaries to describe semantic structure, proposing a three-fold universal typology based on reference and function, emphasizing their compatibility with electronic resources.

Miguel Lacalle Palacios analyses the verbal class of deprive in Old English using Role and Reference Grammar (Foley & Van Valin, 1984; Van Valin & LaPolla, 1997; Van Valin, 2005) and Levin's (1993) framework, describing four alternations and two constructions, and advocating for organization of verbal lexicon by grammatical behavior.

Io Manolessou and Georgia Katsouda explore headword spelling in historical dictionaries, focusing on lexicographical projects about modern Greek and Cappadocian dialects, offering solutions that vary based on the specific problem, dictionary aim, source availability, and target audience.

Ana Elvira Ojanguren López explores the lexicon-syntactic relation in Old English between derivation bases, deverbal nominals, and syntactic constructions, showing that deverbal nominalisation acts like linked verbal predication, sharing principles, categories, and functions.

Chris A. Smith addresses obsolescence and low frequency in the *Oxford English Dictionary*, devising a methodology for assessing obsolescence through a lexicographical database of -some adjectival derivatives, revealing the continued availability of this derivational marker despite its low productivity.

4. EVALUATION

This collective volume represents a remarkable contribution to the fields of computational linguistics, historical lexicography, and corpus linguistics as it creates a dialogue between traditionally philological perspectives and state-of-the-art computational methods. Thus, the interdisciplinarity of this work allows for the discerning of the complex challenges inherent to the structuration and digitization of historical lexical data.

One of the most salient characteristics of *SLDaDD* is its diversity. Its chapters cover a variety of languages and periods which not only elicit the convergence of problems that historical lexicographers face when devising new digital resources, but also language-specific issues that highlight the unique nature of each project. Furthermore, the multiplicity of methods employed on the course of the book showcases the different approaches that researchers can pursue in their topics. Thus, highlighting the multifaceted nature of digital humanities. Moreover, the volume presents novel tools and methodologies that contribute to the development of ongoing projects regarding historical lexicography, and corpus linguistics. For instance, new models for lexical database design, metrics for the evaluation of lemmatization accuracy, lexical organization frameworks, and diverse approaches to encode complex historical sources.

It is important to mention that the volume is intended for linguists, lexicographers, and users of dictionaries and repositories of linguistic data. While the book is generally cohesive, the technical complexity of certain chapters –especially those that rely on computational methods or specialized grammar theories– may be challenging for some readers. Nevertheless, this complexity only reinforces the revolutionary nature of the projects at hand and the rapid development of research in the field.

In fact, the editors, Javier Martín Arista and Ana Elvira Ojanguren López, have worked extensively on the incorporation of lexical databases and computational approaches to historical linguistics, particularly to the study and research of Old English. Martín Arista's current research delves on Universal Dependencies annotation (Martín Arista, 2024; Martín Arista et al. 2025a; 2025b), as well as on the incorporation of artificial intelligence to the study of historical

languages (Martín Arista, 2025); and it informs the volume's first part on structuring lexical data. In turn, Ojanguren López's research currently focuses on verb complementation (Ojanguren López, 2024a; 2024b; 2024c) sets the theoretical base for understanding the syntactic principles present across different chapters of the second part of the volume.

5. CONCLUSION

SLDaDD constitutes a major contribution to the developing field of computational historical linguistics as it represents a state-of-the art overview of theoretical insights and new technological applications from projects on multiple languages, as well as a roadmap for future research on the field. Historical linguists, digital lexicographers, and scholars working on computational approaches to language may find these methodologies and frameworks valuable tools to deepen their research and inform their projects. The book opens up new possibilities for studying historical language data through the combination of language processing, grammatical theory, and database design. In a time shaped by LLMs and NLP, structured lexical data and digitized resources are becoming valuable means for the preservation and research of our linguistic heritage.

REFERENCES

Bosworth Toller's Anglo-Saxon Dictionary online. Retrieved from: https://bosworthtoller.com/

Foley, W., & Van Valin, R. (1984). *Functional Syntax and Universal Grammar.* Cambridge: Cambridge University Press.

Levin, B. (1993). *English verb classes and alternations.* Chicago: University of Chicago Press.

Mair, C., & Meyer, C. F. (Eds.). (1988). "Language and Computers: Studies in Digital Linguistics". In *Language and Computers*. Leiden: Brill. Retrieved from: https://brill.com/view/serial/LC

Martín Arista, J. (2024). Toward a Universal Dependencies Treebank of Old English: Representing the Morphological Relatedness of Un-Derivatives. *Languages*, *9*(3). https://doi.org/10.3390/languages9030076

Martín Arista, J. (2025). The Computational Study of Old English. *Encyclopedia, 5*(3), 137. https://doi.org/10.3390/encyclopedia5030137

Martín Arista, J., Ojanguren López, A. E., & Domínguez Barragán, S. (2025a). Universal Dependencies annotation of Old English with spaCy and MobileBERT. Evaluation and perspectives. *Procesamiento del language natural*, *74*, 253-262.

Martín Arista, J., Ojanguren López, A. E., & Domínguez Barragán, S. (2025b). Parsing Old English with Universal Dependencies. The impact of model architectures and dataset sizes. *Big Data and Cognitive Computing*, *9*(8), 199. https://doi.org/10.3390/bdcc9080199

Morrissey, R., & Roe, G. (Eds.), *Encyclopédie, ou dictionnaire raisonné des sciences, des arts et des métiers, etc*. University of Chicago: ARTFL Encyclopédie Project. Retrieved from https://artflsrv04.uchicago.edu/philologic4.7/encyclopedie0922/

Ojanguren López, A. E. (2024a). Are There Serial Verb Constructions in Old English? A New Perspective on the Changes in Verbal Complementation. *Philologica canariensia*, *30*, 393-423. https://doi.org/10.20420/phil.can.2024.683

Ojanguren López, A. E. (2024b). Competition in the Complementation of Old English Control Verbs with Oblique Marking: A Corpus Analysis. *Languages*, *9*(3). https://doi.org/10.3390/languages9030086

Ojanguren López, A. E. (2024c). *Predications in competition and the rise of serial verb constructions in English: The verbal and nominal complementation of Old English aspectual and manipulative verbs.* Lausanne, Berlin, Bruxelles, Chennai, New York, Oxford: Peter Lang. https://doi.org/10.3726/b21357

Ordbog over det norrøne prosasprog / A Dictionary of Old Norse Prose Online. Retrieved from: http://onp.ku.dk

Oxford English Dictionary online. Retrieved from: https://www.oed.com/

Van Valin, R. (2005). *Exploring the syntax-semantics interface.* Cambridge: Cambridge University Press.

Van Valin, R., & LaPolla, R. (1997). *Syntax: Structure, meaning and function*. Cambridge: Cambridge University Press.