UNIVERSITY OF MURCIA

International Journal of English Studies

IJES

http://revistas.um.es/ijes

Martín Arista, Javier, & Ojanguren López, Ana Elvira (Eds.) (2024). *Structuring lexical data and digitising dictionaries: Grammatical theory, language processing and databases in historical linguistics.* Brill. Pages: 412. ISBN: 978-9004702653

LUISA FIDALGO ALLO*

Universidad de La Rioja (Spain)

Received: 12/06/2025. Accepted: 27/03/2026

Recent developments in historical linguistics and digital lexicography reveal a decisive methodological shift, one that merges computational innovation with rigorous linguistic inquiry. Foundational studies such as Durkin (2019) and McGillivray and Tóth (2020) have laid the groundwork for this shift, highlighting the need for reliable digital frameworks to support the study of historical and low-resource languages. These works identify the difficulties of encoding complex textual features, ensuring interoperability across lexical resources, and achieving annotation accuracy through corpus-based methodologies. Concurrently, Adamska-Sałaciak (2019) examines conceptual and terminological challenges in defining lexicography as a science, arguing that clarifying key terms is essential to advance coherent theoretical development in the field.

Structuring lexical data and digitising dictionaries, edited by Javier Martín Arista and Ana Elvira Ojanguren López, emerges as a significant contribution in this context of ongoing academic change. Addressing the challenges previously outlined, the volume effectively integrates computational techniques with philological precision. Its twofold structure –Part 1 focusing on dictionary digitisation and Part 2 on lexicon structuring– resonates with the foundational objectives presented in O'Keeffe and McCarthy (2022), particularly the emphasis on creating robust, semantically sound corpora and the importance of developing interoperable linguistic resources that meet contemporary research needs. Importantly, the volume's attention to lesser-studied languages, such as Old Church Slavonic and Old English, extends the reach of digital humanities into typologically diverse territories, reinforcing the urgency of developing replicable, well-grounded approaches to lexicographic work across the digital humanities. Within the field of Old English studies, this book contributes to the area of research

Address for correspondence: Departamento de Filologías Modernas, Universidad de La Rioja. Edificio de Filologías, C/ San José de Calasanz, 33, 26004 Logroño (España); e-mail: luisa.fidalgoa@unirioja.es

Print ISSN: 1578-7044; Online ISSN: 1989-6131
doi: 10.6018/ijes.666751

in corpus linguistics and computational analysis of the Anglo-Saxon language pursued in Martín Arista (2012, 2018, 2022, 2024), Martín Arista et al. (2025), and Ojanguren López (2022, 2024, 2025).

Part 1 presents a wide-ranging set of computational approaches to the digitisation and annotation of historical linguistic resources. Afanasev and Lyashevskaya (Chapter 2) propose a hybrid lemmatization model for Old Church Slavonic that integrates neural sequence-to-sequence learning with edit-distance metrics (Levenshtein, Jaro-Winkler). Their adaptive evaluation framework effectively addresses the challenges of low-resource languages. Brenon (Chapter 3) explores 19th-century French encyclopedias using TEI XML encoding, illustrating how encoding frameworks can preserve structural complexity while enhancing machine readability. Similarly, Horvat et al. (Chapter 4) tackle the retro-digitisation of early Croatian grammars. By combining TEI with semantic infrastructures such as Jena, they ensure metadata preservation and conceptual consistency.

Johannsson (Chapter 5) describes the transformation of *A dictionary of Old Norse prose* (ONP Online), housed at the University of Copenhagen, from a traditional collection of citations to an online digital resource. As a case study, it documents the progressive transition from analogue data collection to a digital lexical database, illustrating how technological developments enable new forms of data structuring, access, and interaction in historical lexicography. This case study highlights key elements essential for establishing the project and outlines the methodology behind presenting technological solutions to historical lexicographers, emphasizing how this shift prioritizes accessibility. Lugli (Chapter 6) introduces *agile lexicography*, using *Shiny*, an R package, to develop dynamic dictionary environments for Sanskrit and Tibetan, making lexicography more adaptable to modern needs.

Building on this, Martín Arista (Chapter 7) proposes a graph-based lexical database for Old English that enhances Semantic Web integration by interlinking lexical, morphological, and textual data. Finally, Tichý and Roček (Chapter 8) conclude the section with a sophisticated reworking of the *Anglo-Saxon dictionary* (Bosworth & Toller, 1882–1898, 1921), incorporating TEI-structured annotations, Elasticsearch capabilities, and grammatical cross-referencing.

These chapters reflect and expand upon key ideas from foundational works in historical linguistics and computational lexicography. Afanasev and Lyashevskaya's hybrid model resonates with the hybrid architectures promoted by McGillivray and Tóth (2020) in addressing the challenges of low-resource languages through advanced computational models. Brenon and Horvat et al. echo Durkin's (2019) call for encoding frameworks that manage multilingualism and diachronic variation, as well as O'Keeffe and McCarthy's (2022) emphasis on ensuring corpus interoperability. Johannsson's work and approach align with Durkin's (2019) focus on user-centered, adaptable lexicographic tools. Similarly, Lugli follows Durkin's vision by introducing agile lexicography. Martín Arista supports the ideals discussed by Roberts and McConchie (2022) on interoperability in lexicographic tools. Lastly, Tichý and Roček reflect Adamska-Sałaciak's (2019) emphasis on academic rigor while foregrounding digital accessibility.

The second part of the volume shifts the focus from tools and platforms to the theoretical and methodological foundations of historical lexicon construction. Across a range of case studies, the chapters address challenges related to semantic classification, syntactic structure, dialectal variation, and lexicographic consistency. While methodologically diverse, the contributions collectively aim to develop replicable models for organizing lexical data in ways that are both historically sensitive and computationally tractable.

Novotná and Fúsik (Chapter 9) take a close look at the Old English adjective *(ge)sælig* using the *York-Toronto-Helsinki parsed corpus* (Taylor et al., 2003) to track semantic drift and

Print ISSN: 1578-7044; Online ISSN: 1989-6131

syntactic tendencies over time. Their work suggests subtle shifts in usage and meaning, contributing to the understanding of broader lexical and grammatical developments. Similarly concerned with structure and consistency, Jelovšek (Chapter 10) proposes a three-part labeling system for the *Dictionary of 16th-century Slovenian* (Ahačič et al., 2021): encyclopedic, linguistic, and register-based. This model enhances internal consistency and illustrates how a taxonomic approach can address both descriptive needs and user expectations in historical lexicography.

Lacalle Palacios (Chapter 11) and Manolessou and Katsouda (Chapter 12) approach the challenge of lexical classification from different angles. Lacalle uses Role and Reference Grammar (RRG) to classify Old English verbs of DEPRIVING, such as *beniman*, *berēofan*, and *(ge)stelan*, which are classified as full members, prospects, and non-members, respectively. This approach offers a principled syntactic-semantic mapping that categorizes verbs into these groups. Manolessou and Katsouda tackle lemmatization in Greek dialectal dictionaries, contrasting methodologies from ILNE and DiCaDLand projects. Their efforts highlight the difficulty of maintaining both orthographic diversity and lexicographic clarity, a challenge in dialect-rich corpora.

Ojanguren López (Chapter 13), like Lacalle, also draws on RRG to analyze Old English deverbal nominalizations, revealing how nominal derivatives inherit logical and macrorole structures from their verbal sources. This paradigm-based methodology complements Lacalle's verb-focused study and reinforces the value of RRG in diachronic lexical analysis. Smith (Chapter 14) closes the section by questioning the *Oxford English dictionary*'s classification of *-some* adjectives as obsolete, presenting evidence from Google Books to suggest that several forms continue to be used, albeit in restricted contexts. This dynamic, data-driven approach underscores the importance of re-evaluating legacy classifications in light of corpus-based findings.

Taken together, these chapters offer a theoretical counterpoint to the digital tools presented in Part 1 while reflecting and extending key debates in contemporary historical lexicography. Novotná and Fúsik's stance and views align with O'Keeffe and McCarthy's (2022) insistence on empirical validation in diachronic studies. Jelovšek's labeling system reflects Durkin's (2019) emphasis on internal consistency and taxonomic structure in dictionary design. The RRG-based studies by Lacalle Palacios and Ojanguren López mirror McGillivray and Tóth's (2020) call for tighter integration between linguistic theory and lexicographic practice. Manolessou and Katsouda engage with the dialectal challenges raised by Roberts and McConchie (2022), particularly those of balancing representativeness and usability. Finally, Smith's re-evaluation of adjectival obsolescence exemplifies O'Keeffe and McCarthy's (2022) advocacy for corpus-based revision of received lexicographic classifications. The section as a whole demonstrates how theoretical models and empirical methods can inform one another in the structuring of historical lexicons.

One of the most compelling features of this volume lies in the way it brings together technical innovation and theoretical insight. The technical innovations outlined in Part 1, particularly those involving hybrid NLP architectures, graph-based databases, and agile lexicographic environments, complement the more theoretical contributions in Part 2, such as RRG-based syntax-semantic linking and new labeling systems. Both sections prioritize scalability, whether dealing with orthographic variation (Chapters 4, 12) or validating lexicographic labels across genres (Chapters 9, 14).

A major advancement here is the integration of natural language processing with philological expertise. Thus, Part 1's neural models for Old Church Slavonic lemmatization (Chapter 2) and Part 2's semantic analysis of *(ge)sælig* (Chapter 9) demonstrate how computational methods and historical studies can enrich each other. Similarly, the use of TEI

Print ISSN: 1578-7044; Online ISSN: 1989-6131

XML (Chapters 3, 8) and Semantic Web standards (Chapter 7) aligns with Durkin's (2019) vision of creating interoperable, long-lasting resources.

The volume also addresses lexicographic obsolescence, challenging fixed labels through quantitative corpus analysis (Chapter 14), reinforcing Roberts and McConchie's (2022) argument for diachronic flexibility in historical thesauri and ensuring lexicons reflect changing usage patterns.

Structuring Lexical data and digitising dictionaries represents a paradigm shift in historical lexicography. By blending computational innovation with rigorous linguistic theory, the volume not only addresses current methodological challenges but also anticipates the future needs of the field. Its contributions extend the legacy of foundational works such as Durkin (2019) and McGillivray and Tóth (2020) but, more importantly, offer a promising framework for creating historical lexicons, providing valuable insights into how digital tools can integrate with traditional methods to enhance semantic analysis. While the technical innovations are impressive, it is perhaps the book's thoughtful approach to balancing computational efficiency and linguistic depth that stands out.

The contributions of this volume, such as the use of neural networks for Old Church Slavonic lemmatization and the application of RRG to Old English, could potentially influence future directions in digital and historical lexicography. Of course, there may be areas where further refinement and development are needed, particularly as new technologies continue to evolve. However, the book sets a solid foundation for future research, encouraging exploration of AI-driven annotation systems and expanding its focus on low-resource languages.

The editors have done an impressive job in curating a collection that balances technical depth with a strong foundation in linguistic tradition. *Structuring lexical data and digitising dictionaries* offers a refreshing take on how computational methods can revitalize the study of historical languages. This volume provides a clear direction for lexicographers, corpus linguists, and digital humanists and stands out as an essential reading for scholars aiming to navigate the complexities of historical data in the age of AI. It sets a clear path for future research to expand these methodologies to non-Indo-European languages and refine AI-driven annotation frameworks. Ultimately, this volume lays the ground for a future in digital lexicography that is both innovative and adaptable, making it a must-read for anyone engaged in the advancement of historical linguistic scholarship.

ACKNOWLEDGEMENTS

This work is part of the project PID2023-149762NB-I00, funded by MICIU/AEI/10.13039/501100011033.

REFERENCES

Ahačič, K., Čepar, M., Jelovšek, A., Legan Ravnikar, A., Merše, M., Narat, J., Novak, F., & Premk, F. (2021). *Slovar slovenskega knjižnega jezika 16. stoletja. A–D*. Založba ZRC.

Adamska-Sałaciak, A. (2019). Lexicography and theory: Clearing the ground. *International Journal of Lexicography, 32*(1), 1–19.

Bosworth, J., & Toller, T. N. (1882–1898). *An Anglo-Saxon dictionary: Based on the manuscript collections of the late Joseph Bosworth, edited and enlarged by T. Northcote Toller*. Clarendon Press.

Bosworth, J., & Toller, T. N. (1921). *An Anglo-Saxon Dictionary: Based on the Manuscript Collections of Joseph Bosworth. Supplement by T. Northcote Toller*. Oxford University Press.

Durkin, P. (2019). *The Oxford handbook of lexicography*. Oxford University Press.

Print ISSN: 1578-7044; Online ISSN: 1989-6131

Martín Arista, J. (2012). Lexical database, derivational map and 3D representation. *RESLA: Revista Española de Lingüística Aplicada, Extra 1*, 119–144.

Martín Arista, J. (2018). The semantic poles of Old English. Toward the 3D representation of complex polysemy. *Digital Scholarship in the Humanities, 33*(1), 96–111.

Martín Arista, J. (2022). Toward the morpho-syntactic annotation of an Old English corpus with Universal Dependencies. *Revista de Lingüística y Lenguas Aplicadas, 17*, 85–97.

Martín Arista, J. (2024). Toward a Universal Dependencies treebank of Old English: Representing the morphological relatedness of un-derivatives. *Languages, 9*(3), 76.

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

McGillivray, B., & Tóth, G. M. (2020). *Applying language technology in humanities research: Design, application, and the underlying logic*. Palgrave Macmillan.

Ojanguren López, A. E. (2022). The morpho-syntactic alternations of Old English verbs of inaction. *International Journal of English Studies, 22*(2), 91–128.

Ojanguren López, A. E. (2024). *Predications in competition and the rise of serial verb constructions in English. The verbal and nominal complementation of Old English aspectual and manipulative verbs*. Peter Lang.

Ojanguren López, A. E. (2025). Old English perspectives on the complement shift: Toward the desententialisation of self-manipulative verbs. *Journal of Historical Linguistics, 15*(1), 44–77.

O'Keeffe, A., & McCarthy, M.J. (2022). *The Routledge handbook of corpus linguistics* (2nd ed.). Routledge.

ONP Project (n.d.). *ONP Online: Ordbog over det norrøne prosasprog / A dictionary of Old Norse prose*. University of Copenhagen. https://onp.ku.dk/

Oxford University Press. (n.d.). Oxford English Dictionary. https://www.oed.com/

Roberts, J., & McConchie, R. W. (Eds.). (2022). *Historical thesauri: Challenges and developments*. Edinburgh University Press.

Taylor, A., Warner, A., Pintzuk, S., & Beths, F. (2003). *The York-Toronto-Helsinki parsed corpus of Old English prose*. University of York.

Print ISSN: 1578-7044; Online ISSN: 1989-6131