Martín Arista, Javier, ed., Sara Domínguez Barragán, Luisa Fidalgo Allo, Laura García Fernández, Yosra Hamdoun Bghiyel, Miguel Lacalle Palacios, Raquel Mateo Mendaza, Carmen Novo Urraca, Ana Elvira Ojanguren López, Esaúl Ruiz Narbona, Roberto Torre Alonso, Marta Tío Sáenz, and Raquel Vea Escarza. 2024. "Nerthusv5. Interface of Textual, Lexicographical and Secondary Sources of Old English." www.nerthusproject.com.

Reviewed by Ondřej Tichý 🆔
Charles University, Prague

Introduction

The Nerthus Project with its open-access lexical database of Old English has now been under development for almost twenty years and its recent fifth version has made some significant progress.[1] Perhaps the most notable update is its integration of and closer connection to the sister project: *An Open Access Annotated Parallel Corpus Old English-English* (*ParCorOE*).[2] With these updates, I have carefully sampled and reviewed both of the online resources,[3] as well as their documentation and some of the research based on them. Both of the projects are large, complex and ambitious and I recognize I may have missed or misunderstood some of their various features or datapoints.

Project description

The *Nerthus* database is currently described as an "Interface of textual and lexicographical sources of Old English," which hints at its plethora of diverse features available for each of its 32,812 entries—or dictionary headwords. It is at the same time a lexicographical resource: a kind of dictionary, a lexical database recording word-formation processes of Old English (OE) words, a collection of

---

[1] The first version is from 2007–9 and the fifth version was published in 2023–24.

[2] Its first version was published in 2018 and its current third version has been published in tandem with *Nerthus* in 2023–4.

[3] Using Chrome (v. 136.0.7103.114), Edge (v. 136.0.3240.92) and Firefox (v. 139.0.1) browsers on Windows 11.

related third-party resources such as dictionaries and research papers, and finally a channel to and the extension of the *ParCorOE* corpus.

As a dictionary, it provides its users with a number of different headword forms (to simplify search and sorting), alternative spellings, definitions, word class information (called lexical category) and morphological categories (such as nominal gender for nouns or class for verbs). The lexical information (perhaps its most unique feature) recorded in the database features a word-formation status (such as simplex, compound or derivation), the immediate affix, word-formation base, lexical prime and the status of ge-prefixation (attesting to the fact that *Clark-Hall* dictionary (1916) is the foundation of its lexicographical information). The linked resources provide the headword forms, definitions and attested forms from the major dictionaries of OE—in addition to *Clark-Hall* (1916), *Nerthus* provides the information from *Bosworth-Toller* (1921), Sweet (1967) and for the headwords starting with A-I also the information from the *Dictionary of Old English* (2018). Some headwords also feature a list of references to secondary research that has dealt with those lexical items. Finally, headwords attested in some form in the *ParCorOE* corpus feature the attested spellings and an OE example fragment (usually a sentence) containing a form of the headword and a Present-day English (PDE) translation of this fragment.

Nerthus (as well as *ParCorOE*) is running on the FileMaker database app by Claris Inc. and is presented to the users through its WebDirect interface (see fig. 1), which can be accessed online using modern web browsers.

Figure 1. User interface of the Nerthus Project showing the record of the headword *feormung*.
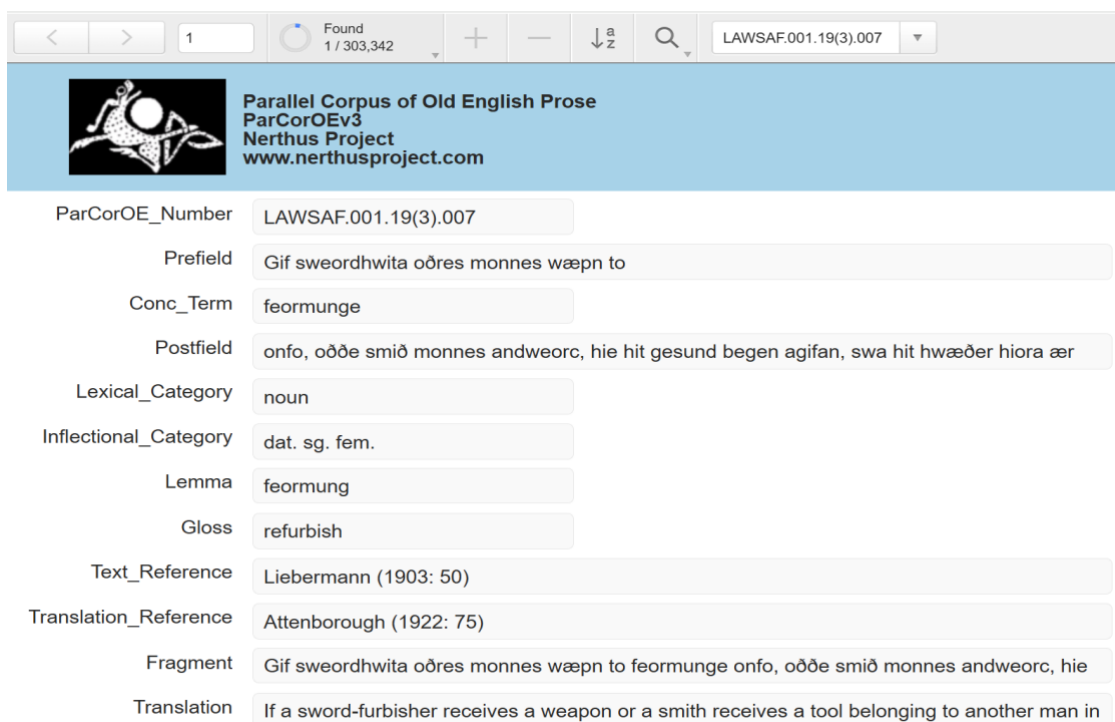
The default interface is structured like a form with multiple text fields in eight sections representing the information described above and a toolbar at the top of the screen allowing users to page through records, sort, filter or find them. Each page (record) represents one dictionary entry, or headword. The quick find function searches all the fields for a substring submitted by the user while the Find mode functions allows users to create more complex queries using wildcards and combining multiple text fields (such as the headword form and the lexical category). The search returns a filtered subset of records for the user to page through or refine the search/filter terms further.

The *ParCorOE* is a unique parallel corpus project that aligns the source language—Old English—with a present-day variety of the same language as a target. The two varieties are aligned on two levels. First, on the level of fragments (which seem to generally correspond to sentences), the original OE text is aligned with an already existing Present-day English (PDE) translation—often relatively dated[4]. Second, on the token level, each OE token is aligned with a PDE gloss. Usually, an OE token corresponds to a one-word PDE gloss, but sometimes a PDE gloss translates more than one OE token or vice versa. Given the differences between OE and PDE in morphology and syntax, the PDE glosses do not constitute a continuous translation text, but help users find the translational counterparts of individual words in the fragment level translation. In addition to the gloss, each OE token also features a unique ID, an OE lemma, a word-class information and a detailed inflectional categorization (such as "ind. pret. 3rd sg." for a verbal token or "dat. pl. masc." for a nominal token). The corpus is, therefore, not only glossed and translated, but also fully lemmatized and deeply morphologically tagged. The corpus is composed of 300.000 tokens in its third version and while this makes it rather small by corpus linguistic standards, it nevertheless represents c. a fifth of the *York-Toronto-Helsinki Parsed Corpus of Old English Prose* (2003), which is commonly used for corpus research in OE, and ca a tenth of the *DOE Corpus* (*DOEC*, 2009), which in a sense represents almost all extant OE material.

ParCorOE's user interface (see fig. 2) is similar in design and functions to the one used by *Nerthus*, since they are, as noted above, based on the same technology (FileMaker's WebDirect). In the *ParCorOE* interface, each page represents one token, a concordance term (Conc_Term), and users can click to browse to the previous or the following token. The page also displays the preceding and the following context in a given fragment, though context over the fragment boundary is only visible once tokens from the preceding/following fragments are displayed as concordance terms. For each token, the whole

---

[4] This seems to be for licensing reasons, the translations used by ParCorOE are all in public domain. Often these are translations made to accompany the editions of the OE texts compiled by the corpus, like Thorpe's *Charters* or *Apollonius of Tyre*.

fragment is also displayed both in the original OE and in the PDE translation (with respective references). Therefore, when the user browses forward, the concordance term keeps shifting by one token to the right, the preceding context (Prefield) keeps getting longer and the following context (Postfield) keeps getting shorter.

Figure 2. User interface of the ParCorOE corpus showing the full record corresponding to the illustrative quotation in Fig. 1.

Both *Nerthus* and *ParCorOE* are rich and valuable resources for students as well as researchers of OE, linguists, literary scholars, philologists or translators. Both projects have an undeniable potential (especially when considered or used together), however, some of the potential seems currently to be hampered especially by the accessibility of the data. While both projects are open access and freely available online (which is to be recognized and appreciated), the underlying datasets are not offered for download.[5] Having the datasets available for download in a standardized format (like TEI-XML), with a detailed description and from a well infrastructurally connected repository would definitely help researchers and improve the general usability of the project results.

---

[5] They could, arguably, be scrapped, but that usually involves some information loss and there does not seem to be any note of license, beyond the general characterization of the projects as open access.

More importantly, from the point of view of the majority of users who want to take advantage of working directly with the online interface, the FileMaker's WebDirect does not seem to me to quite realize the full potential of the databases and sometimes even creates some obstacles for its users. Here are some specific comments on the interface in general, largely applicable to both projects:

- The interface does not seem to support clickable links, which greatly reduces the advantages of the two resources working in tandem—it is useful to have one illustrative quotation derived from *ParCorOE* displayed in *Nerthus* or a lemma for each token in *ParCorOE* that is also a headword in *Nerthus*, but it would be much more convenient to get all the quotations or all the headword information in one click. Copying from one database and pasting into the search function of the other is not only cumbersome, but also does not necessarily yield satisfactory results simply because there are plenty of homonymous forms and even lemmas in OE. Specifically for *Nerthus*, if it is to work as a lexicographical hub navigating users to external dictionaries and sources (and it very well could fulfil that role in the future), it needs to provide clickable links rather than just headword forms or short citation information.
- Viewing the search results, the search term employed to generate the results is not shown, which is very inconvenient when considering the results and possible refinements of the query.
- Search seems to be sometimes sensitive and sometimes insensitive to diacritics, but this may be down to user system settings or locale.
- Sometimes the search results appear to be simply wrong, e.g. search for *gear* as a base of derivation in *Nerthus* returns *geap*; or as a lexical prime returns *GEADOR*. I was unable to understand this behaviour.
- When text is longer than the size of its text-box, it is quite difficult to read the parts of the text that do not fit in the box—the user has to click inside the text box and then use arrow keys to scroll through the text. Moreover, there is no indication of either the fact that there is more text or of the controls needed to actually read it.
- Users are automatically signed out of the databases after a couple of minutes, which can get annoying relatively quickly.
- Last character of the search query is deleted in Find mode in Chrome and Edge (not in Firefox), e.g. when a user enters *gegaderode* as a Conc_Term in *ParCorOE* and clicks Perform, the final *-e* is deleted (which is difficult to notice, since the search query is not displayed, as noted above).[6]
- While the Find mode allows for wildcards and Boolean operators, standard regular expressions would probably be preferable for advanced users.

---

[6] This is probably a known bug (see TimAcri, 2025).

- Viewing results as a filtered series of pages with each result on its own page is not very user friendly for a lot of use cases. There is a View as List option, but that only puts all the forms on one long page. Search results in a lexicographical resource should ideally take the form of a clickable list. When looking for a particular word, a clickable list of headwords (possibly with snippets of more details) should be shown as a result of the search so that the user can decide which links to follow, but also immediately see how many results are actually relevant and whether the search needs to be further refined. Search results for a corpus should ideally take the form of the key word in context (KWIC)—with one token in context per line (rather than per page), possibly with additional metadata on the side or interlineally—to allow the user a quick overview of the resulting data and possibly facilitate refining the search terms.[7]
- Better search, clickable links and dynamic linking of the two resources in combination can encourage and help answer more complex and ambitious questions—such as, what is the token frequency of a particular word-formation base or process.
- It is also not possible to link directly to or report specific records or search results of either database, which is crucial both for researchers citing the resources as well as for drawing in more users in general. For *ParCorOE*, it is at least possible to copy, report and easily find again the ParCorOE_Number, which is a unique identifier of each token in the corpus. For *Nerthus*, however, no unique record identifier is shown, though it is possible to export an XML file via File, Save, Snapshot Link, Current record, that contains a NativeID attribute—possibly signifying a unique ID.

In general, given the complexity of the projects, it would be beneficial to have more information on their usage, structure but also editorial decisions available for the users—beyond and above the otherwise very useful search tips on the entrance page of the databases, the project/database descriptions and the corpus manuals. There are a couple of specific pieces of information and clarifications that may be especially helpful to users:

- Some *Nerthus* fields have surprising and unexplained punctuation or abbreviations, possibly derived from the original sources, but difficult to interpret out of context, e.g. SW_meaning "(!) ineffable" for *untōsprecendlic*.
- Lists and explanation of labels used in certain text boxes would be really helpful (you have to know what you're looking for to find it), e.g. the morphological_status attribute in *Nerthus* contains some more cryptic labels,

---

[7] Interestingly, a similar view was reported in Martín Arista (2021, 82, Fig. 1), but it is not available in the current version of *ParCorOE*.

such as "ZD_EXPLICIT_INFLECTION." Using select boxes rather than text-boxes for some of the features might also help.

- Some non-obvious editorial decisions should be explained e.g. in a research paper/status report – and they possibly are, but if so, they are not clearly referenced for the project users. An example of such a decision may be the diachronic depth of the word-formation processes reported. In some cases, derivation is definitely not synchronic in OE, but goes at least as far back as Proto-Germanic (like the ablaut derivations reported often as V>N).
- How complete is the *Nerthus* data with respect to the other linked dictionaries? This is important, if *Nerthus* is to be a general lexicographical hub for OE. Note e.g. the headword *gegadere*, it features no headword links or inflectional forms from *Bosworth-Toller*, *DOE* and *DOEC*, but it does show the *Bosworth-Toller* meaning and one of the alternative spellings listed (*gegædere*) is the headword form for both the *Bosworth-Toller* and *DOE*.[8]
- Some *Nerthus* entries (e.g. *abbot*) feature alternative spellings and meanings from *Bosworth-Toller* or *Clark-Hall*, but do not have associated headword forms listed for those same sources.
- The overall quality of the data has undoubtedly been improving over the years—e.g. all the major problems noted by Zimmermann (2013) appear to have been remedied, but it is possible to still encounter some apparent errors— as can be expected in projects of this magnitude and complexity.
- While some *Nerthus* data may be missing, as noted above, some data seem duplicated, e.g. *ēarmǽlum* is listed three times, with the second and third instances missing the lexical_category and ge_prefix values (nativeID 1, 32815 and 32816).
- Some glosses in *ParCorOE* are in Latin and some of them are incorrect, e.g. *Hǽlendum* is glossed as *perpetuum* (CHAR.016.001.004) probably due to an error in alignment where the Latin exposition (originally in both the OE text and its translation by Thorpe) was removed from the original, but not from the translation.
- Some lemmatization also seems off, e.g. the token at CHAR.016.001.012 *untosprecendlican* is lemmatized as *untōlȳsendlic* (but its gloss follows the original meaning of "ineffable," though in multiple words "most ineffable Creator"). This error is then propagated to Nerthus, where *untosprecendlican* is listed as an inflectional of *untōlȳsendlic*. It might be useful to calculate something like an edit distance between headwords, alternative spellings and inflectional forms and check the extreme values manually.
- As noted above, there are a lot of homonymous forms as well as lemmas in OE and this may have caused problems with some automatic processing, so that

---

[8] It is also interesting to note that *Nerthus* does not include *DOE* meanings, while it does include *DOE* headword and *DOEC* spellings. Is this for licensing reasons?

in Nerthus, *dǣl* (part) and *dæl* (dale) share the dame SW_meaning_definition, though the rest of the semantic information is correctly differentiated.

There are some decisions the authors of the projects made that I personally do not agree with, but these may very well be just matters of opinion or differences of approach. For example, the decision to merge all the eth and thorn spellings in *ParCorOE* under eth makes searching easier and would be fine with me in case of *Nerthus* (since it is a secondary resource), but in a primary resource like a language corpus, this precludes certain types of research and the ease of searching for just one spelling variant could have been achieved by either standardizing the spelling in lemmas (that are also secondary) only, or by other technical means.

While the notes above understandably address mostly problems I encountered testing and reviewing the two resources, the time I have spent with them has very clearly shown me their considerable potential. It is already indisputable that the two projects are playing an important role in the scholarship of OE—they have been used for original research by the project team, including some very interesting recent plans for further development of the project, e.g. Martín Arista (2024a, 2024b) as well as by other scholars, such as Novo Urraca (2016), Kharlamenko (2020) or García Fernández (2021).

As noted above, making the underlying data available in a standard repository would definitely foster further research, as well as bolster the long-term preservation of the projects. To fulfil its promise as a major online resource for the majority of users, however, the projects should focus some effort towards making the interface more user friendly and more in line with the standard lexicographical and corpus resources. Further data verification and cleaning should also be considered. Overall, the projects are indisputable contributions to scholarship and scholars of OE with a lot of promise in their future instalments.

References

Bosworth, Joseph, and Thomas N. Toller. 1921. *An Anglo-Saxon Dictionary: Based on the Manuscript Collections of Joseph Bosworth*. Supplement by Thomas N. Toller. Oxford: Clarendon Press.

Cameron, Angus, Ashley C. Amos, Antonette diPaolo Healey, Roy Liuzza, and Haruko Momma, eds. 2018. *Dictionary of Old English: A to I Online*. Dictionary of Old English Project. https://tapor.library.utoronto.ca/doe/.

Clark Hall, John R. 1916. *A Concise Anglo-Saxon Dictionary for the Use of Students*. 2nd ed., rev. and enl. New York: The Macmillan Company.

diPaolo Healey, Antonette, John P. Wilkin, and Xin Xiang. eds. 2009. *Dictionary of Old English Web Corpus*. Dictionary of Old English Project. http://tapor.library.utoronto.ca/doecorpus/.

García Fernández, Laura. 2021. "Insights into Old English Lexicography: Lemmatisation of *gĀn* and Its Prefix-Formations Using a Corpus-Based Database." *International Journal of Lexicography* 34 (4): 493–508.

Kharlamenko, Oxana. 2020. "The Expression of Non-Individual in Some Old English Nouns." In *Of ye Olde Englisch. Language and Textes: New Perspectives on Old and Middle English Language and Literature*, edited by Rodrigo Pérez Lorido, Carlos Prado-Alonso, and Paula Rodríguez-Puente, 25–50. Berlin: Peter Lang.

Martín Arista, Javier. 2021. "Word Alignment in a Parallel Corpus of Old English Prose." In *Corpora in Translation and Contrastive Research in the Digital Age*, edited by Julia Lavid-López, Carmen Maíz-Arévalo, and Juan R. Zamorano-Mansilla, 75–100. Amsterdam: John Benjamins.

Martín Arista, Javier. 2024a. "Toward a Universal Dependencies Treebank of Old English: Representing the Morphological Relatedness of Un-Derivatives." *Languages* 9 (3): 76. https://doi.org/10.3390/languages 9030076.

Martín Arista, Javier. 2024b. "Interface of Old English Dictionaries in Database Format: Toward a Knowledge Base." In *Structuring Lexical Data and Digitising Dictionaries*, edited by Javier Martín Arista and Ana E. Ojanguren López, 150–83. Leiden: Brill.

Nerthus. n.d. "Nerthus Project." University of La Rioja. https://www.nerthusproject.com/.

Novo Urraca, Carmen. 2016. "Morphological Relatedness and the Typology of Adjectival Formations in Old English." *Studia Neophilologica* 88 (1): 43–55. https://doi.org/10.1080/00393274.2016.1150788.

ParCorOE. n.d. "ParCorOEv3. An Open Access Annotated Parallel Corpus Old English-English." University of La Rioja. https://www.nerthusproject.com/search-parcoroe.

Sweet, Henry. 1967. *Anglo-Saxon Primer*. Oxford: Clarendon Press.

Taylor, Ann, Anthony Warner, Susan Pintzuk, and Frank Beths. 2003. *The York-Toronto-Helsinki Parsed Corpus of Old English Prose*. Electronic Texts and Manuals Available from the Oxford Text Archive. http://www.ling.upenn.edu/histcorpora/PCME2RELEASE 4/index.html.

TimAcri. 2025. "Webdirect is Removing the Last Character When Tabbing or Clicking out of a Field" *Claris Community* (blog). February 19. https://community.claris.com/en/s/question/0D5Vy00000bbdzfKAA/webd irect-is-removing-the-last-character-when-tabbing-or-clicking-out-of-a-field.

Zimmermann, Richard. 2013. "The Nerthus Proyect" A New Old English Language Online Database. Old Engli.Sh: The Portal to the Language of the

Anglo-Saxons (blog). https://old-engli.sh/news.php?ID=news-Nerthus-review.

(Received 03/06/2024)