Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 3934–3941
Marseille, 20-25 June 2022

Towards the Construction of a WordNet for Old English

Background, Methods, Infrastructure

Anas Fahad Khan[1], Francisco J. Minaya Gómez[2], Rafael Cruz González[2], Harry Diakoff[3], Javier Diaz Vera[2], John P. McCrae[4], Ciara O'Loughlin[4], William Michael Short[5] Sander Stolk[6]
[1]Istituto di Linguistica Computazionale "A. Zampolli" (ILC-CNR) Pisa, Italy,
[2]University of Castilla-La Mancha, Ciudad Real, Spain, [3]The Alpheios Project,
[4]Data Science Institute, National University of Ireland Galway,
[5] University of Exeter, UK; [6]Leiden University, Leiden, the Netherlands,
[1]fahad.khan@ilc.cnr.it, [2]{francisco.minaya, rafael.cruz, javierenrique.diaz}@uclm.es
[3]harry.diakoff@gmail.com, [4]john@mccr.ae, ciara.oloughlin@insight-centre.org,
[5]w.short@exeter.ac.uk, [6]s.s.stolk@hum.leidenuniv.nl

Abstract

In this paper we will discuss our preliminary work towards the construction of a WordNet for Old English, taking our inspiration from other similar WN construction projects for ancient languages such as Ancient Greek, Latin and Sanskrit (on this overall endeavour, see now (Biagetti et al., 2021)). The Old English WordNet (OldEWN) will build upon this innovative work in a number of different ways which we articulate in the article, most importantly by treating figurative meaning as a 'first-class citizen' in the structuring of the semantic system. From a more practical perspective we will describe our plan to utilize a pre-existing lexicographic resource and the naisc system to automatically compile a provisional version of the WordNet which will then be checked and enriched by experts in Old English.

Keywords: WordNet, ancient, figurative, metaphor, metonymy, conceptual metaphor, embodiment

1. Introduction

In this paper we describe the proposed construction of a WordNet (WN) for the Old English language, the *Old English WordNet* (OldEWN). This resource will be based on lemmas and definitions extracted from a legacy Old English dictionary; these will be used to compile a list of candidate synsets and the relations between them using the Naisc linking system. Finally the candidate synsets will be subject a process of correction and enrichment by a team of expert annotators using a specially developed platform. We plan to publish the resulting resource, the OldEWN with an open license, both in the Global Wordnet LMF XML format[1] and as Linked Open Data using the RDF vocabulary OntoLex-Lemon. As we discuss below, this work is inspired by a number of previous efforts at creating historical language WN's, in particular in its strong focus on figurative language and semantic shift.

The rest of the article is structured as follows. We begin in Section 2.1 by giving some background on the Old English language itself and listing some of the language resources that currently exist for it. Next, in Section 2.3, we situate our proposals for a OldEWN in the context of other efforts in the construction of historic language WNs. Then in Section 3 we discuss the steps by which the OldEWN will be created using the Naisc system. Afterwards, in Section 4.1, we look at some sample entries from our proposed OldEWN using an enriched version of the Global WordNet Association WordNet LMF schema. Finally, we end with a discussion and conclusions in Section 5.

2. Background and Related Work

2.1. An Introduction to Old English

The term *Old English* (OE) refers to a set of West Germanic dialects spoken in Great Britain from the 5th century AD until the 12th. The predecessor of modern English, OE has a written corpus containing surviving texts dating from the period c. 650 to c. 1150 CE (when the use of the written language was abandoned in favour of French). Texts written in the language have been traditionally classified into four main dialectal areas, namely Northumbrian, Mercian, Kentish and West Saxon[2]. Moreover, we can distinguish four different sub-periods in the development of OE reflecting the four consecutive (but frequently overlapping) stages into which the history of Old English literature is sometimes organised (Magennis, 2011). These are: **OE1** (comprising texts written before 850); **OE2** (comprising texts written between 850 and 950); **OE3** (comprising texts written between 950 and 1050); and **OE4** (comprising texts written between 1050 and 1150).

Broadly speaking, the texts in **OE1** represent the vernacular tradition of the Old English speaking populations (and consequently there is an under-representation of prose texts). The texts in **OE2**, how-

[1] `https://globalwordnet.github.io/schemas/#xml`

[2] Notwithstanding the fact that the bulk of surviving OE texts represent the West Saxon dialect, some of the oldest texts (most of them in verse) actually exemplify the Northumbrian or Mercian varieties of Old English

ever, are almost exclusively translations from Latin (mostly produced by Alfred the Great and his followers). In contrast to the latter, around, half of the texts contained in **OE3** were originally written in the vernacular; indeed most of them are attributed to just two Anglo-Saxon authors: Ælfric and Wulfstan. Lastly, **OE4**, a period characterised by a dramatic decrease in the use of written English as a consequence of the Norman Conquest, once again contains by a large number of translations from Latin. As this brief historical summary shows, Old English underwent a process of substantial change throughout its lifespan as a written language. Something that is manifested in the evolution of OE literature from the early epic poems and their echoes of pre-Christian cultures to the learned discourses associated with Anglo-Saxon Christian intellectual elites. Effectively this meant that while the lexicon of the language remained highly stable, the semantic scope of many Old English words began to be expanded (normally through the semantic extension of word senses via metonymization and metaphorization) in order to allow the integration of new concepts into the language from other cultures.

If we are to accurately reflect the surviving corpus for the OE language then our OE WordNet should take these aspects into full consideration. That is, it is desirable for many use cases that a resource of this kind make information on different OE dialectal variants accessible along with data on the distribution of senses across textual genres and time periods. Additionally and given OE's status as a historical language, one that is studied from a diachronic perspective, it would be very useful to be able to track the way that word senses in OE shift and evolve over time. We situate the current work in the context of previous and ongoing efforts at the construction of WNs for historical languages in Section 2.3.

2.2. Some Existing Lexical Resources for Old English

2.2.1. *The Dictionary of Old English*, Bosworth-Toller, and *the Thesaurus of Old English*

The Dictionary of Old English (DOE)[3] was first proposed in 1970 by Richard Venezky, Christoper Ball and Angus Cameron. The project was first released in 1981 under the supervision and leadership of Angus Cameron, Antonette diPaolo Healey and Ashley Crandell Amos in the Centre of Medieval Studies at the University of Toronto. The DOE is based on a corpus of Old English that contains at least one copy of every surviving OE text, in 2007 this digitized corpus was made available to the public to enable users to search for instances of words throughout the corpus or particular texts. The corpus contains approximately 3 million words and includes poetry, prose scientific texts, legal texts, brief inscriptions, glossaries and interlinear glossaries to Latin texts. Each text is assigned to Cameron number categories, e.g. *A - Poetry*, *B - PROSE*, *C- Interlinear Glossaries*.

So far, the DOE covers 'A to I' of the 24 letters of the Old English alphabet. However, DOE's 2020 progress report[4] detailed that they were working on the publication of the first two fascicles of 'L'. One of the aims of the DOE is to make it easier to use in comparison to other dictionaries of Old English (Cameron and diPaolo Healey, 1979). Spelling variants were grouped together under one single headword with this goal in mind. The aim of the DOE was also to improve upon the lexical resources that were available to academics, and in particular the Bosworth-Toller Anglo-Saxon dictionary (Bosworth, 1882).An online version of the latter work was constructed using digitized manuscripts of Bosworth's Anglo-Saxon dictionary and Toller's Supplement [5] by Sean Crist in 2001 as part of his German Lexicon Project[6] (GLP). The current version of the online Bosworth-Toller was created in 2010 using the data from the GLP. This resource is the largest and most complete OE dictionary available for use to date. Both of these resources were utilized to create *A Thesaurus of Old English*[7] (TOE). Unlike the DOE and Bosworth-Toller, but like the OldEWN, the TOE organises OE vocabulary into conceptual categories such as 'Life and Death' originally derived from Roget's Thesaurus[8]. The user must search a topic or subject rather than a specific headword or lemma. TOE uses words senses from Clark Hall (described below) and Bosworth-Toller and organizes them into 18 categories. These categories are further divided into subcategories, this creates a hierarchical structure in the classification of vocabulary that was devised for the Historical Thesaurus[9] of English. The TOE can be accessed via the *evoke* platform[10] (Stolk, 2021).

2.2.2. Clark Hall's *A Concise Anglo-Saxon Dictionary*

In addition to the lexicographic resources mentioned above we can also list Clark Hall's *A Concise Anglo-Saxon Dictionary* (CAS) subtitled *For the Use of Students*. First published in 1894 by the the Victorian barrister and scholar of Old English John R. Clark Hall the dictionary is still currently in print (and is now in its

[3] https://doe.artsci.utoronto.ca/

[4] https://doe.artsci.utoronto.ca/wp-content/uploads/2021/04/report.pdf

[5] https://bosworthtoller.com/

[6] http://www.germanic-lexicon-project.org/etc/aa_texts.html

[7] https://oldenglishthesaurus.arts.gla.ac.uk/

[8] Unlike the WordNet organisational approach, these categories have not been widely adopted for structuring computational language resources, a limiting factor on the interoperability of the resource.

[9] https://ht.ac.uk/about/

[10] http://evoke.ullet.net/app/#/view?source=toe

fourth edition). The second edition of the work, published in 1916, has been made available on the Gutenberg website in HTML and textfile versions[11]. This edition of the CAS contains 36744 lemmas although a large number of these are orthographic variants of other lemmas. It also contains information on citations, in the case of hapaxes, and provides genre information when senses are only found in poetic texts; there is also a limited amount of etymological information. The brevity of the entries in the dictionary (and the lack of nesting of senses typical of the Bosworth-Toller) makes it easier to work with using text processing and analysis tools, and so we made the decision to construct the first provisional stage of the OldEWN using the CAS. In particular we take its lemma list and sense definitions as the basis for the OldEWN's synsets. However we plan to enrich the WordNet in subsequent stages through the work of manual annotators/validators and also, potentially, by adding information from other public domain OE resources.

2.3. WordNets for Historic Languages and OldEWN

Our work has been been strongly informed by previous attempts at the creation of ancient or historical language WordNets including a Latin WordNet (Section 2.3.1) and a WordNet for Ancient Greek (Section 2.3.2). More generally, our project for the creation of a OldEWN is associated with a recent initiative towards the publication and enrichment of a family of interoperable WordNets for ancient Indo-European languages (discussed in Section 2.3.3). Since several of the themes we raised in Section 2.1 and a lot of the technical issues we will discuss in the rest of this paper have already been touched upon in these other initiatives we will dedicate a section to each of them with an emphasis on the commonalities they share with the OldEWN project. In what follows wel assume the reader is already familiar with some of the basics of WordNets, if not they are invited to consult (Miller, 1998).

2.3.1. Latin WordNet

As part of the Fondazione Bruno Kessler's MultiWordNet project (Pianta et al., 2002), Stefano Minozzi succeeded in 2008 in assigning, through an automatic process of sense-matching via English and Italian dictionary glosses, around 9000 lemmas to synsets defined for English and available from the Princeton WordNet (PWN)(Minozzi, 2017). Prior work on Italian had also made available several synsets specific to Italian, designated using an identification schema similar to that of PWN but including a new language marker within the offset tag; these appear sporadically among Minozzi's synset assignments, but in recent work have largely been discounted because in the creation of the Italian MultiWordNet lemmas they were often (mistakenly) treated as constituting synsets *per se*. This was the genesis of the Latin WordNet (LWN).

In 2018, Minozzi's abandoned project was taken up by William Short at the University of Exeter, one of the co-authors of the current work. The Latin WordNet was consequently expanded[12] to include more than 40,000 lemmas, and a new process of synset matching (utilizing multiple available lexical resources) created a foundation of semantic assignation for manual curation by scholars. In addition to the new lemmas, new morphological and etymological information was added to the LWN, along with data structures for capturing diachronic and generic information in relation to word usage. More radically, the LWN has introduced mechanisms for describing figurative meanings, both at the lexical level (that is, within the semantic structures of individual words) and at the conceptual level (in the form of large-scale metaphorical and metonymic mappings). Chiara Fedriani's companion project on conceptual metaphors of emotion in Latin has now seeded the LWN with data on the conceptual mappings involved in Latin's construal of ANGER, LOVE, and FEAR[13]. At the same time, some effort has been given to revising the sometimes wildly inaccurate synset assignations of Minozzi's 9000 lemmas, within the Marco Passaroti's *Linking Latin* project, see (Franzini et al., 2019).

2.3.2. Ancient Greek WordNet

In its original incarnation, the Ancient Greek WordNet (AGWN) arose as a collaboration between the Alpheios Project[14], the Perseus Project[15], the University of Pisa and the Istitute of Computational Linguistics (Bizzoni et al., 2014). Preliminary work on the AGWN was motivated by an interest in developing Ancient Greek semantic resources that could complement the Ancient Greek Treebank (previously developed by the Perseus Project and sponsored by Alpheios) and with a view to permitting concept-based querying in corpora of ancient languages[16]. The object was thus to optimize recall rather than precision, tolerating an increase in retrieval "noise" in the interests of more comprehensive recall of potentially relevant material, so the approach was simply to merge the definitions of all the

[11] https://www.gutenberg.org/ebooks/31543

[12] https://latinwordnet.exeter.ac.uk/

[13] see https://latinwordnet.exeter.ac.uk/lexicon

[14] https://alpheios.net/

[15] http://www.perseus.tufts.edu/hopper/

[16] Several of the principals involved in Alpheios had previously been involved in utilizing the UMLS medical hierarchical thesaurus developed by the National Library of Medicine to expand online search queries and rank the results, and saw the potential of WordNets for doing something similar when it came to querying ancient language corpora. William Short's *Cylleneus* project, taking advantage of the Latin, Greek, and Sanskrit, now affords this possibility for any corpus of these languages

machine-readable dictionaries available at the time and associate the corresponding Greek word with each of the English equivalents that reported by one or more of the dictionaries. No complete definitions of any words were retained and the addition of glosses to the resulting synsets was left for manual curation. Like practically all WordNets based on gloss matching, the resulting resource suffered in particular from transferring the polysemy of its pivot language (namely, English) to Greek, producing excessive "noise" and often unacceptable sense definitions. For some purposes this may be irrelevant, for others crippling. The experience of attempting the epuration of AGWN is described in (Bizzoni et al., 2014). Note that similarly inappropriate synset assignation can readily be found in Minozzi's LWN, for example.

In 2019, the AGWN project having formerly been abandoned, was revitalized through a joint intiative of the University of Exeter and Harvard's Centre for Hellenic Studies. Similarly to the LWN, the WordNet was expanded on the basis of newly available lexical resources and glosses of lemmas (over 100,000) were newly processed to determine sense definitions. The updated version of the Ancient Greek WordNet can be found at `https://greekwordnet.chs.harvard.edu/`.

2.3.3. Towards a Family of Ancient Indo-European WordNets

The collaboration described in the last section led to an ongoing international project under the direction of William Short with the aim of creating a network of similarly structured and fully interoperable WordNets for ancient Indo-European (IE) languages. This project, described in (Biagetti et al., 2021) and which we will refer to as the IEWN project, is being jointly developed by scholars at the University of Exeter, the University of Pavia, the Center for Hellenic Studies at Harvard University, and the Alpheios Project. As well as including the updated Latin and Ancient Greek WN's mentioned above the project is also building a WordNet for Sanskrit from scratch[17].

The IEWN project aims to explore the potential of lexico-semantic resources adhering to the same organising principles, in this case the principles of the WordNet family of lexical resources, for carrying out comparisons of semantic structures across languages, here additionally exploiting the pool of sense designations (synsets) available from the Princeton WordNet as the core of semantic description. In particular the idea is to leverage this semantic interoperability in order to compare ancient Indo-European languages belonging to societies with significant linguistic and cultural similarities and differences. Notably the project is guided by insights from Cognitive Linguistics and with a strong emphasis on the annotation of figurative meaning. We share IEWN's view of figurative meaning as a 'first-class citizen' in the structuring of the semantic system (an illustration of our cognitive linguistic based approach to describing figurative language can be found in the example entries which we present in Section 4.1.1). Indeed our intention is to fully align the OldEWN with all of the IEWN WordNets. This would enable us to link together, for instance, the OldEWN and the IEWN's Latin WordNet, something which is highly desirable given the abundance of conceptual borrowings recorded in OE translations and re-elaborations of Latin texts. At the same time, as shown by (Geeraerts and Gevaert, 2008) and (Díaz-Vera and Manrique-Antón, 2014) among others, there are numerous figurative expressions in Old English, and in other ancient Germanic languages, the origins of which are to be found not in the classical Mediterranean cultures but, rather, in the system of beliefs and in the folk psychologies of the ancient Germanic peoples (Lockett, 2018). The OldEWN is intended to make information about both kinds of linguistic phenomena more accessible to users of the resource. In order to help ensure a high level of interoperability between the OldEWN and the IEWN WordNets we will follow the strategies adopted by the IEWN project for capturing polysemy and for lumping together or splitting apart senses, all of which is laid out in detail in (Biagetti et al., 2021). In addition, and once again taking our cue from the IEWN family of WNs, we will include both antonymy and morphological relations between lexical units in addition to using a compatible system of morph-syntactic tags for OE. Moreover we also intend to tag for historical periods (currently those mentioned in Section 2.1) and genres and include etymological information, aligning our annotation strategies with those proposed by IEWN as far as is possible.

3. Constructing the OldEWN

As mentioned above, the OldEWN is being bootstrapped using the Concise Anglo-Saxon Dictionary (CAS) (Clark Hall, 1916)[18]. In addition to this automatic bootstrapping of OE synsets however, our intention is also to incorporate semantic information from a dataset compiled by Díaz-Vera as part of a program of research in figurative terms used to refer to emotion in Old English, as described in works such (Díaz-Vera and Manrique-Antón, 2014); an attempt to model this dataset as linked data is described in (Khan et al., 2020). This dataset describes the emotion lexicon of

[17] The project WordNets are accessible at: Latin `https://latinwordnet.exeter.ac.uk/`, Ancient Greek `https://greekwordnet.chs.harvard.edu/` Sanskrit `https://sanskritwordnet.unipv.it/` and can be queried via a unified RESTful API.

[18] We have previously alluded to the use of pre-existing lexicographic resources as a part of the process of constructing or refining the Latin and Ancient Greek WordNets; this has also been done in the case of modern language WN's, such as for instance the Turkish language WN *KeNet* (Bakay et al., 2021))

Old English and features an organisation of this lexicon into synsets; it also includes data on metonynmic and metaphoric sense shifts in polysemic OE emotional terms along with other kinds of etymological information, as well as data on the distribution of senses across genres and historical periods. Our intention is to manually merge the synsets from this dataset (along with the additional lexico-semantic information which it contains) with the other synsets which we will derive from CAS.

3.1. Bootstrapping the OldEWN from CAS

Lemmas have been assigned sense descriptors using the modern English glosses of CAS, which have been matched to the glosses of synsets in the Princeton WordNet. As mention the glosses for our provisional OE synsets will be taken from the CAS; though these are subject for subsequent revision by exerts. We will develop the data extracted from the CAS into a WordNet by linking the lemmas from the former to Modern English lemmas from the Open English WordNet project (McCrae et al., 2019; McCrae et al., 2020), which is based on the Princeton WordNet (Fellbaum, 2010). This will be performed based on the algorithms implemented in the Naisc system (McCrae and Buitelaar, 2018)[19]. The Naisc system implements several features for linking, and has previously been applied for monolingual linking by means of measuring the textual similarity of definitions (McCrae et al., 2021a), however in the case of the glosses in CAS, these are not full dictionary definitions with genus and differentiae, but instead they are translations of the Old English lemma into Modern English. For this reason, we will take an approach based on the graph-based feature extraction implemented in Naisc. In particular, we will build a graph based on Old English corpora consisting of the collocations of terms in CAS, this graph will then be compared to existing network of relations found in Open English WordNet. This will use link prediction techniques to rank the candidate links from CAS, by means of the similarity between the collocation graph and the Open English WordNet graph. We will then create the initial version of the OldEWN based on these links using the *expand* approach (Bond and Foster, 2013), in which we copy the relationships from the modern English WordNet and replace the lemmas with those from the target language, in this case Old English.

In the next stage of the construction of the OldEWN expert annotators, in addition to checking the candidate synsets for correctness, will be able to expand this WordNet with new synsets that are specific to the target languages. As in the IEWN project they will tag senses and entries for time periods, distribution across genres, texts according to the current state of the OE scholarship as well as potentially adding etymological information including that pertaining to figurative semantic shift. Note that the CAS already includes tags indicating when an entry is only found in poetic texts or in one specific text and this information will be checked, and then incorporated and ultimately enriched with information from other sources.

Moreover, as we mentioned above, a lot of this information is already available in the OE emotions dataset which we are incorporating into the OldEWN. In the final section of this article we look at how we can extend a commonly used WN schema in order to include such information taking an example from the OE emotion dataset.

4. Enriching the OldEWN with data on Figurative Language in OE

4.1. Enriching the Global WordNet LMF Schema

Our intention in addition to publishing the finished OldEWN resource with an open licence is to make it a FAIR digital resource[20] which means, amongst other things, making it available using standardised schemas and formats. Indeed we propose to publish the resource using both the XML-based *Global WordNet Association WordNet Lexical Markup Framework* (GWA WordNet-LMF)[21] model (McCrae et al., 2021b; Bond et al., 2016) and an *OntoLex-Lemon*-based RDF encoding which is based on it. Our choice of the former was determined by its popularity as a publication and interchange format for WordNets. In the latter case, by publishing the OldEWN as linked data we make it easier to link from it to other resources as well as enabling it to be queried remotely via a SPARQL endpoint. We will focus on the GWA WordNet-LMF format in what follows.

As it currently stands the GWA WordNet-LMF schema fails to meet several of the specific expressive needs of the OldEWN or indeed of the IEWNs described in Section 2.3.3. We have therefore decided to develop an extended version of the GWA WordNet-LMF schema which we present here in part. Note that, the GWA WordNet-LMF schema is, as its name suggests, based on a prior version of the ISO Lexical Markup Framework (LMF) standard (Francopoulo, 2013). This latter has been updated in the meantime and is being republished serially as a multi-part standard (one which is however retro-compatible with the previous version) (Romary et al., 2019). One of the parts of this new version of LMF is an etymology module (Khan et al., 2020) which includes several classes and attributes which would be useful in expressing the several varieties of information which feature in the Old-

[19] Naisc, which means ‘links’ in Irish and is pronounced ‘nashk’, is available at https://github.com/insight-centre/naisc

[20] https://www.go-fair.org/fair-principles/

[21] https://github.com/globalwordnet/schemas/blob/master/WN-LMF-1.1.dtd

EWN as well as the other IEWNs[22] One such class is EtyLink representing an etymological relationship between two linguistic elements and with associated attributes @source and @target (representing the source and target of this relationship) and @type (specifying the type of the relationship). Another such class is Etymology which is defined as a container for an ordered series of EtyLink elements. We decided therefore to add these elements to a proposed update of the GWA WordNet-LMF schema. We show the use of our new enriched schema with an example taken from the OE emotion lexicon dataset mentioned above; this also allows us to illustrate what the OldEWN will ultimately look like.

4.1.1. An Example OldEWN Synset: Shame

For our example we will take a subset of the words in the OE Emotion lexicon synset for shame:

{.....,*āblysung*,..., *bismer*,...,*edwīt*, ..., ...}.

The OE word *āblysung* listed above is polysemic and has two senses 'blushing' and 'shame' with the latter the result of a metonymic shift in the former. More accurately it is an instance of *resultative metonynmy*, i.e., a relationship between two concepts in which instances of the one are generally regarded as *resulting* from the other. We encode this as follows using our new enriched schema (the definition gloss as in the other examples is taken from the CAS):

```
<LexicalEntry id = "ABLYSUNG_N">
   <Lemma writtenForm="āblysung" partOfSpeech="n"
        grammaticalGender = "f"/>
   <Sense id ="oew5_s1" synset = "example-ang-
        XXXXXX2-n">
      <Definition gloss = "blushing"/>
   </Sense>
   <Sense id ="oew5_s2" synset = "example-ang-
        XXXXXX1-n">
      <Definition gloss = "shame"/>
   </Sense>
   <etymology>
      <etyLink type = "resultative-metonymy" source=
           "oew5_s1" target="oew5_s2"/>
   </etymology>
</LexicalEntry>
```

The words *bismer* and *edwīt* are also polysemic. Both of them have the sense of 'shame' as well as in the case of *bismer* having the additional meaning 'filthiness, defilement' and in the case of *edwīt* of 'reproach, scorn, abuse'. In both cases we are dealing with *causative metonymy* i.e., of a relationship between two concepts in which instances of the one are generally regarded as *causing* the other. This can be represented as follows:

```
<LexicalEntry id = "BISMER_N">
<Lemma writtenForm="bismer" partOfSpeech="n"/
     grammaticalGender = "nmf">
   <Sense id ="oew10_s1" synset = "example-ang-
        XXXXXX1-n">
     <Definition gloss = "disgrace, scandal, shame,
          mockery, insult, reproach, scorn,"/>
   </Sense>
   <Sense id ="oew10_s2" synset = "example-ang-
        XXXXXX4-n">
     <Definition gloss = "filthiness, defilement"/>
   <etymology>
     <etyLink type = "causative-metonymy" source="
          oew10_s2" target="oew10_s1"/>
   </etymology>
   </Sense>
 </LexicalEntry>


<LexicalEntry id = "ED-WIT_N">
<Lemma writtenForm="edwīt" partOfSpeech="n"
     grammaticalGender = "n"/>
   <Sense id ="oew13_s1" synset = "example-ang-
        XXXXXX6-n">
      <Definition gloss = "reproach, scorn, abuse"/>
   </Sense>
   <Sense id ="oew13_s2" synset = "example-ang-
        XXXXXX1-n">
      <Definition gloss = " shame, disgrace,"/>
   <etymology>
      <etyLink type = "causative-metonymy" source="
           oew13_s1" target="oew13_s2"/>
   </etymology>
   </Sense>
</LexicalEntry>
```

We can potentially express more complex kinds of etymological information as well as adding different kinds of morpho-syntactic information using the GWA WordNet-LMF schema by adding other additional classes and properties from the new version of LMF. Our plan is to publish a version of the schema which will capture the expressivity needed to publish the OldEWN as well as the other WNs in the IEWN family. In addition we are also looking to add classes and properties describing periods, dialects, and genre.

5. Conclusion

In this article we have described our plans for the creation of a WordNet for the Old English language. Being largely a statement of intent most of the work still remains to be done. However the experience of planning and conceiving the OldEWN along with the experiences which the authors have accumulated working on other WNs as well as similar lexico-semantic resources has led us to formulate a number of conclusions. The first concerns the importance of articulating the target use cases when planning any WordNet, and for combining automated processes with manual curation. For some purposes, having dated textual citations is crucial. For others, frequency data derived from specific corpora. When it comes to the specific exigencies of working with historic languages a number of additional issues arise. For instance is the chief interest going from modern categories to ancient ones or the reverse? Do we want to know how many species of flowering plants were distinguished in the ancient vocabulary or do we want to know how many different kinds of plants we recognize today are referred to by specific Greek words? Many of the problems inherent in using a pivot language to construct new WordNets among modern languages (eg. mismatched metaphors and polysemy) are much more intractable when comparing a modern with an ancient language where not only the

[22] At the moment our enriched schema has an unofficial status but we are in contact with the relevant ISO group with a view to regularising it.

categories but even the relationships among the categories may differ. It seems inevitable that some subject domains will be much more successfully mapped than others, but those difficulties can of course be among the most interesting discoveries if the purpose is to compare conceptual spaces. Our hope is that in our efforts towards the construction of an Old English WN we will be able to make a contribution to these and other related discussions, in addition, of course to making a useful scholarly lexical resource available to scholars, students and indeed anyone interested in Old English.

6. Acknowledgements

Many thanks to Thijs Porck for his helpful suggestions. This work has also been informed by discussions held during the ongoing COST Action NexusLinguarum – European network for Web-centered linguistic data science (CA18209), supported by COST (European Cooperation in Science and Technology) and the H2020 program ELEXIS (grant agreeement no. 731015).

7. Bibliographical References

Bakay, Ö., Ergelen, Ö., Sarmış, E., Yıldırım, S., Arıcan, B. N., Kocabalcıoğlu, A., Özçelik, M., Sanıyar, E., Kuyrukçu, O., Avar, B., et al. (2021). Turkish wordnet kenet. In *Proceedings of the 11th global wordnet conference*, pages 166–174.

Biagetti, E., Zanchi, C., and Short, W. M. (2021). Toward the creation of Wordnets for ancient Indo-european languages. In *Proceedings of the 11th Global Wordnet Conference*, pages 258–266.

Bizzoni, Y., Boschetti, F., Diakoff, H., Del Gratta, R., Monachini, M., and Crane, G. (2014). The making of ancient greek wordnet. In *Proceedings of the ninth international conference on language resources and evaluation (LREC'14)*, pages 1140–1147.

Bond, F. and Foster, R. (2013). Linking and extending an open multilingual wordnet. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1352–1362.

Bond, F., Vossen, P., McCrae, J. P., and Fellbaum, C. (2016). CILI: the Collaborative Interlingual Index. In *Proceedings of the Global WordNet Conference 2016*.

Bosworth, J. (1882). *An Anglo-Saxon Dictionary: Based on the Manuscript Collections of the Late Joseph Bosworth...*, volume 1. Clarendon Press.

Cameron, A. and diPaolo Healey, A. (1979). The dictionary of old english. *Dictionaries: Journal of the Dictionary Society of North America*, 1(1):87–96.

Clark Hall, J. R. (1916). *A concise Anglo-Saxon dictionary: for the use of students*. Swan Sonnenschein & Company, second edition.

Díaz-Vera, J. E. and Manrique-Antón, T. (2014). 'better shamed before one than shamed before all': Shaping shame in Old English and Old Norse texts. In *Metaphor and Metonymy across Time and Cultures*, pages 225–264. De Gruyter Mouton.

Fellbaum, C. (2010). Wordnet. In *Theory and applications of ontology: computer applications*, pages 231–243. Springer.

Francopoulo, G. (2013). *LMF lexical markup framework*. Wiley Online Library.

Franzini, G., Peverelli, A., Ruffolo, P., Passarotti, M., Sanna, H., Signoroni, E., Ventura, V., and Zampedri, F. (2019). Nunc Est Aestimandum: Towards an evaluation of the Latin Wordnet. In *Proceedings of CLiC-it 2019*.

Geeraerts, D. and Gevaert, C. (2008). Hearts and (angry) minds in Old English. In *Culture, Body, and Language*, pages 319–348. De Gruyter Mouton.

Khan, F., Romary, L., Salgado, A., Bowers, J., Khemakhen, M., and Tasovac, T. (2020). Modelling Etymology in LMF/TEI. In *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*. European Language Resources Association (ELRA).

Lockett, L. (2018). *Anglo-Saxon psychologies in the vernacular and Latin traditions*. University of Toronto Press.

Magennis, H. (2011). *The Cambridge Introduction to Anglo-Saxon Literature*. Cambridge University Press.

McCrae, J. P. and Buitelaar, P. (2018). Linking Datasets Using Semantic Textual Similarity. *Cybernetics and Information Technologies*, 18(1):109–123.

McCrae, J. P., Rademaker, A., Bond, F., Rudnicka, E., and Fellbaum, C. (2019). English WordNet 2019 – An Open-Source WordNet for English. In *Proceedings of the 10th Global WordNet Conference – GWC 2019*.

McCrae, J. P., Rademaker, A., Rudnicka, E., and Bond, F. (2020). English WordNet 2020: Improving and Extending a WordNet for English using an Open-Source Methodology. In *Proceedings of the Multimodal Wordnets Workshop at LREC 2020*, pages 14–19.

McCrae, J. P., Ahmadi, S., bin Yim, S., and Bajčetić, L. (2021a). The ELEXIS system for monolingual sense linking in dictionaries. In *Proceedings of The Seventh Biennial Conference on Electronic Lexicography, eLex 2021*, pages 542–559.

McCrae, J. P., Goodman, M. W., Bond, F., Rademaker, A., Rudnicka, E., and Costa, L. M. D. (2021b). The GlobalWordNet Formats: Updates for 2020. In *Proceedings of the 11th Global Wordnet Conference*, pages 91–99.

Miller, G. A. (1998). *WordNet: An electronic lexical database*. MIT press.

Minozzi, S. (2017). Latin Wordnet, una rete di conoscenza semantica per il latino e alcune ipotesi di utilizzo nel campo dell'information retrieval.

Strumenti digitali e collaborativi per le Scienze dell'Antichità, (14):123–134.

Pianta, E., Bentivogli, L., and Girardi, C. (2002). Multiwordnet: developing an aligned multilingual database. In *First international conference on global WordNet*, pages 293–302.

Romary, L., Khemakhem, M., George, M., Bowers, J., Khan, F., Pet, M., Lewis, S., Calzolari, N., and Banski, P. (2019). LMF reloaded. In *Asialex 2019*.

Stolk, S. (2021). Evoke: Exploring and extending a thesaurus of old english using a linked data approach. *Amsterdamer Beiträge zur älteren Germanistik*, 81(3-4):318–358.