Some Updates on the Development of an Historical Language Wordnet

Fahad Khan[1], Daniel Prado Aranda[2], Francesca Romana Cammisa[3], Michele Cavallaro[4], Maria Francesca Carmela Giusy Germanà[4], Federica Misino[4], Chiara Tenti[4], Javier E. Díaz-Vera[2], Francisco Javier Minaya Gómez[2], Francesca Frontini[1]

[1]Istituto di Linguistica Computazionale "Antonio Zampolli", Italy,
[2]Universidad de Castilla-La Mancha, Spain,
[3]University of Bologna, Italy,
[4]University of Siena, Italy

Correspondence: fahad.khan@ilc.cnr.it

Abstract

In this article, we give an update on an ongoing initiative to build an Old English Wordnet (Old-EWN). The initial phase of this initiative was dedicated to the compilation of an emotion lexicon for Old English intended to function both as a stand-alone resource as well as a part (a sub-wordnet) of what will eventually be a wordnet covering the entire lexicon. In this phase, we worked with a pre-existing research dataset in the area of Old English emotions and re-used public domain lexicographic works as the basis of our lexicon. Another interesting aspect of this phase of our work was that it was a collaboration between researchers from the Istituto di Linguistica Computazionale, the Universidad de Castilla-La Mancha and a number of interns from the University of Siena. This initial phase is now over, and we are currently preparing to publish the wordnet in RDF format as well as to host it on a triple store. In this report we will give background on the process of creating this part of our wordnet and the challenges we faced, as well as describing the dataset itself.

1 Introduction

The term *Old English* (OE) refers to a set of related West Germanic dialects spoken in Great Britain from the 5th until the 12th centuries and which were the predecessors of modern English. OE has a corpus of surviving texts dating from the period c. 650 to c. 1150 CE, after which the language was essentially overshadowed by French for well-known historical reasons (Magennis, 2011). Although the development of computational resources for Old English has tended to lag behind that of other ancient languages, notably Ancient Greek and Latin, there has, nonetheless, been real progress on this front in the last few decades. Perhaps the most significant project in this regard has been the compilation and publication (in instalments) of a contemporary, scholarly dictionary for the language, the *Dictionary of Old English* (DOE)[1], first made available in an electronic edition as a CD-ROM and then via a web interface. In addition, the DOE project has also made a corpus available in TEI containing most surviving works in OE[2]. The DOE was intended to supersede previous legacy Old English lexicographic works, most notably the *Bosworth-Toller Anglo-Saxon Dictionary*[3] and *A Concise Anglo-Saxon Dictionary* by J. R. Clark Hall (we will refer to this latter work as CAS in what follows), both of which were first published in the 19th century and both of which have editions which are freely available in the public domain (at the time of writing). Notwithstanding the DOE's status as the single most comprehensive OE lexicon currently available – one that is informed by the latest contemporary scholarship in the field – both the Bosworth-Toller and CAS remain valuable resources and in particular for students of Old English. This is largely due to the DOE's status as a closed resource that can only be accessed from behind a paywall, but also because at the time of writing the DOE is still unfinished (although it does cover the majority of the letters in the OE alphabet). Another important electronic lexical resource for Old English is the *Thesaurus of Old English* (TOE), a lexico-semantic network whose organisational principles have been heavily inspired by Roget's Thesaurus[4] and which is also navigable via a graphical graph-based interface[5].

We have presented a very short (and incomplete) summary of the situation as regards OE language resources in order to give some background to the work which is to be described in the rest of the article in which we detail an ongoing initiative to build another kind of lexical resource for Old English,

[1] https://doe.artsci.utoronto.ca/
[2] http://hdl.handle.net/20.500.12024/2488.
[3] https://bosworthtoller.com/
[4] https://oldenglishthesaurus.arts.gla.ac.uk/
[5] http://evoke.ullet.net/app/#/view?source=toe

namely a wordnet. It is worth noting that this initiative was inspired by (and shares numerous aims in common with) previous work on the construction of Latin[6] and Ancient Greek wordnets[7]. In our particular case, however, the plan is to proceed by focusing on specific semantic fields (trying, whenever possible, to integrate the latest research on these semantic fields and the OE lexicon into our work). This will result in a number of different domain-specific wordnets, each of which is intended to be used both as a stand-alone resource and to be eventually integrated into a more comprehensive wordnet covering the whole OE lexicon (and containing those, more general concepts which weren't included in the component wordnets), this is the Old English Wordnet *(OldEWN)*. The present work is concerned with the forthcoming publication of a wordnet covering the OE emotion lexicon – that is, that part of the OE vocabulary which is dedicated to the expression and/or description of the emotions and other related concepts (an introduction to the OldEWN project in general is given in (Khan et al., 2022)). In the following section, Section 1.1, we will expand upon the specific motivations behind this work as well as the approach which we have taken while working on this part of the Old English Wordnet (OldEWN) project. In section 2 instead we give a detailed description of the resource itself.

1.1 Constructing An Old English Wordnet for Emotions

One strong impetus behind the current work was the desire to see how wordnets could be used to compare conceptualisations in a given semantic field across languages, or across different stages of the same language. In the case of emotion terms, we were originally interested in comparing Old English with Latin in order to track the influence of cultural exchanges and language contact on the conceptualisation of emotions, as well as making comparisons between Old English and Modern English, to see how stable certain distinctions were over time. Note that, although the work presented in the current paper remains within the framework of the standard wordnet schema, we are planning on enriching and extending this schema to allow for more sophisticated descriptions of linguistic and historical relationships between senses and between languages; see (Khan et al., 2023).

When it comes to actually putting together the lexicon our approach throughout the whole OldEWN project has been based on the use of pre-existing lexicographic resources as a means of bootstraping our resource (an approach which isn't uncommon in the creation of wordnets for historical and under-resourced languages). Primary among these pre-existing resources is Clark-Hall's Concise Anglo-Saxon dictionary (CAS) which is the foundation of our attempts to construct OldEWN thanks to its fairly authoritative lemma list[8], the relative brevity of its definitions (making them easier to process and work with), and of course the fact that it is now in the public domain. In the case of the emotion lexicon, we also based our efforts on a research dataset previously compiled by one of the co-authors of the current work, Javier Díaz-Vera, and which gives a comprehensive description of the emotion lexicon of OE from a cognitive linguistics point of view. Although our initial choice of OE words was based on the list in this dataset, the CAS supplied us with the exact form of each lemma, as well as an initial list of senses for each entry (including those senses that weren't in Díaz-Vera's original research dataset). We used the definitions in the CAS to assign synsets to senses by matching them to glosses from the Open English Wordnet [9] (McCrae et al., 2019, 2020) and then deriving synsets on the basis of shared ILI (Interlingual Index) identifiers. This process wasn't always straightforward, and often the CAS defintions were only a starting point, to be modified according to our original research dataset and augmented with other lexicographic material, including definitions from the Bosworth Toller. In view of these difficulties we decided to carry out the whole process manually. However, the experiences gained during this stage will help us in future experiments on automating this process.

In order to add an extra layer of conceptual organisation to our wordnet, and to make it easier to navigate, we adopted the classification of the emotions proposed by the Geneva Wheel of Emotions (GWE) (Scherer, 2005). This classification can be found in Table 1. Here we summarise the procedure we used to generate the OldEWN emotion lexicon:

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

[7] https://greekwordnet.chs.harvard.edu

[8] The orthography used in the CAS is followed by other resources such as the DOE and the TOE, something which can be vital in a language such as OE with such a varied orthography.

[9] https://en-word.net

- *Generating a list of lexical entries*: the list of entries was taken from the Díaz-Vera dataset classifying emotion words in Old English (see (Khan et al., 2023) for more details on this dataset and its relationship with the OldEWN), this list was subsequently added to on the basis of the CAS and the BT,
- *Establishing a list of lemmas and senses*: lemmas (i.e., the canonical forms of the words chosen in the last step) were taken from a public domain OE lexicon (CAS); where possible the senses for these terms were based on the definitions for the entries in CAS (even if our senses don't always correspond to those listed as individual, separate senses in the dictionary); sometimes we modified and/or added senses when they were in our original dataset but not in the CAS,
- *Mapping synsets to senses*: next, we searched for the Open English Wordnet synset gloss that best matched the meaning of each of the senses derived in the previous step (we compared synset glosses with the senses and their definitions as determined in last stage) and assigned senses Interlingual Index ILI identifiers (Bond et al., 2016) accordingly; in many cases there wasn't a synset gloss that matched exactly, so we we looked for potential hyponyms; in future work we intend to propose new interlingual index concepts using the pre-existing Global Wordnet Association workflow on the basis of our observations in this stage,
- Finally we generated new Old English synsets based on those senses that are mapped to the same ILI.

It is worth noting that the work described here was the fruit of a collaboration between three different institutions, and in particular of researchers from the the Istituto di Linguistic Computazionale «A. Zampolli» and the University of Castilla-La Mancha who worked along side student interns from the *Lingue e comunicazione interculturale e d'impresa* BA of the University of Siena. A large part of the success of the work being described here can be attributed to the success of this collaboration.

2 Resource Description

The output of the work described in the last section is a lexical resource describing the whole emotion vocabulary of Old English (as well as related words) with the senses of words organised in synsets according to the standard wordnet schema. We intend to publish this resource as an RDF dataset with an open licence (CC-BY) both in a triple store, making it available via a public SPARQL endpoint, as well as depositing it in a CLARIN repository. We also plan to continue working on other parts of the OE lexicon in the near future. In addition to this, we plan to enrich the emotion lexicon by integrating information on semantic shifts and adding etymological links to modern English words. The current version of the resource can be found here: http://lari-datasets.ilc.cnr.it/OldEWN_Emotions#

2.1 The Old English Lexicon of Emotions: A First Version

We decided to generate our lexicon directly as an RDF dataset in the turtle format[10]; this is one of the formatting/publication choices recommended by the Global Wordnet Association[11]. One of the main motivations behind this choice was that it made it easier to make our data publically available for querying using the powerful SPARQL query language, and for eventually linking to other datasets. For instance, we can easily extract the list of lemmas in the lexicon with the following simple query:

```
SELECT ?f
    WHERE
    {
        ?l ontolex:writtenForm ?f .
    }
    ORDER BY ?f
}
```

Listing 1: Simple SPARQL Query.

In effect then, with this dataset we have created a linguistic knowledge graph of Old English emotion words in which lexical semantic information is organised using the wordnet schema. The initial work of compiling the data was carried out using Google sheets, we subsequently downloaded each sheet in the TSV format and which we then converted to RDF using an adhoc Python script that made use of the Python library RDFLib. In particular we modelled our data using the OntoLex-Lemon vocabulary[12] as well as the specialised wordnet RDF vocabulary[13] which has been made available by the

[10] https://www.w3.org/TR/turtle/
[11] https://globalwordnet.github.io/schemas/
[12] https://www.w3.org/2016/05/ontolex/
[13] https://globalwordnet.github.io/schemas/wn

GWA for the purpose of publishing and exchanging wordnets as linked data (we will refer to this latter vocabulary as wn in what follows). After carrying out the conversion and following an initial data cleaning process we ended up with a lexicon consisting of 1522 Old English lexical entries, along with 2358 lexical senses and 1021 synsets. Our senses are tagged for the emotional semantic fields featured in Table 1 and taken from the GWE.

Emotion (GWE)	**Senses**	**Examples**
General Emotions	136	*brēostwylm*
Involvement-Interest	51	*georn, ellen*
Amusement-Laughter	24	*gamen, ārēran*
Pride-Elation	86	*þrútian*
Joy-Happiness	167	*hyht, drēam*
Enjoyment-Pleasure	70	*bliss*
Tenderness-Feeling Love	71	*lufu, lufian*
Wonderment-Feeling Awe	78	*āblycgan, ege*
Feeling Disburdened-Relief	12	*līhtan*
Astonishment-Surprise	15	*styltan, ofwundrian*
Longing-Nostalgia	164	*gītsung*
Pity-Compassion	18	*efensārgung, frōfornes*
Sadness-Despair	228	*biter*
Worry-Fear	804	*sēoðan*
Embarrassment-Shame	215	*scamu*
Guilt-Remorse	68	*gylt, scyld*
Disappointment-Regret	64	*hrēow, bētan*
Envy-Jealousy	24	*æfeste, anda*
Disgust-Repulsion	151	*fēogan, hatian*
Contempt-Scorn	30	*forsēon*
Irritation-Anger	126	*irisian, irre*

Table 1: Emotion, Number of Entries, and Example Words

We use the wn property note to associate this information with individual senses. One can easily write a SPARQL query to count the number of senses belonging to each category:

```
SELECT (COUNT(?s) AS ?triples)
WHERE {
?s a ontolex:LexicalSense ;
    wn:note "Involvement-Interest"@en .
  }
```

Listing 2: Ontolex-Lemon example in turtle.

In what follows, we will look at the word *bliss* from the Enjoyment-Pleasure semantic field to show how the OldEWN is structured. The lexical entry with its three senses is as follows:

```
:BLISS_N a ontolex:LexicalEntry ;
    lexinfo:gender lexinfo:masculine ;
    ontolex:canonicalForm :BLISS_N_lemma ;
    ontolex:sense [
                a ontolex:LexicalSense ;
                skos:definition "'bliss,'␣
                     merriment,␣happiness"@en ;
                ontolex:isLexicalizedSenseOf :
                     olde_924 ;
                wn:note "Enjoyment-Pleasure"@en,
                "Joy-Happiness"@en
                ],
                [
                a ontolex:LexicalSense ;
                skos:definition "kindness,␣
                     friendship,␣grace,␣favour"
                     @en ;
                ontolex:isLexicalizedSenseOf :
                     olde_468 ;
                wn:note "Enjoyment-Pleasure"@en,
                "Joy-Happiness"@en
                ],
                [
                a ontolex:LexicalSense ;
                skos:definition "cause␣of␣
                     happiness"@en ;
                wn:note "Enjoyment-Pleasure"@en,
                "Joy-Happiness"@en
                ];
    wn:partOfSpeech wn:noun .
```

Listing 3: Example Entry.

The first of these senses is associated with an Old English synset with the identifier old_924. This latter synset is defined as follows. Note the link to the ILI concept with identifier i110442.

```
:olde_924 a ontolex:LexicalConcept ;
    skos:inScheme <http://example.org/olde/> ;
    wn:definition [ rdf:value "a state of extreme
         happiness"@en ] ;
    wn:ili ili:i110442 ;
    wn:partOfSpeech wn:noun .
```

Listing 4: Synset

Again one of the most important advantages of making our resource available in RDF is that we can use the SPARQL language to create queries such as the following which lists all the senses belonging to the same synset.

```
SELECT DISTINCT ?s
WHERE {
?l ontolex:sense ?s .
?s ontolex:isLexicalizedSenseOf :olde_294 .
?s  skos:definition ?d .
  }
```

Listing 5: SPARQL Query

2.2 Conclusions and Future Work

In this submission we have given an update on the development of a wordnet for Old English, focusing on the publication of that part of the resource which describes emotion terms in Old English and which is now complete. Aside from beginning to work on other semantic fields we have listed a number of future aims throughout the paper (in particular enhancing our resource with information on semantic shifts). To these we add the following:

- Adding links to other relevant resources. This includes linking to Old English entries for the words in our wordnet (where they exist) in the linked data version of Wiktionary, DBnary[14] (Sérasset, 2015); we would also like to add

[14] https://kaiko.getalp.org/about-dbnary/

etymological links from our wordnet to Open English Wordnet[15].

- Another goal is to add new concepts derived from the vocabulary of Old English to the Collaborative Interlingual Index (CILI). This latter serves as a bridge between different wordnets, facilitating cross-linguistic comparisons. We aim to enhance the representation of ancient and historical languages in global wordnet resources in collaboration with other researchers working on wordnets for e.g., Latin and Ancient Greek, making it easier for researchers to draw connections between Old English and other languages represented in the CILI. (Vossen et al., 1999; Bond et al., 2016)

References

Francis Bond, Piek Vossen, John Philip McCrae, and Christiane Fellbaum. 2016. Towards a universal index of meaning. In *Proceedings of the 8th Global WordNet Conference (GWC)*, pages 50–57.

Fahad Khan, John P. McCrae, Francisco Javier Minaya Gómez, Rafael Cruz González, and Javier E. Díaz-Vera. 2023. Some considerations in the construction of a historical language WordNet. In *Proceedings of the 12th Global Wordnet Conference*, pages 101–105, University of the Basque Country, Donostia - San Sebastian, Basque Country. Global Wordnet Association.

Fahad Khan, Francisco J. Minaya Gómez, Rafael Cruz González, Harry Diakoff, Javier E. Diaz Vera, John P. McCrae, Ciara O'Loughlin, William Michael Short, and Sander Stolk. 2022. Towards the construction of a WordNet for Old English. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3934–3941, Marseille, France. European Language Resources Association.

Hugh Magennis. 2011. *The Cambridge Introduction to Anglo-Saxon Literature*. Cambridge University Press.

John P. McCrae, Alexandre Rademaker, Francis Bond, Ewa Rudnicka, and Christiane Fellbaum. 2019. English WordNet 2019 – An Open-Source WordNet for English. In *Proceedings of the 10th Global WordNet Conference – GWC 2019*.

John Philip McCrae, Alexandre Rademaker, Ewa Rudnicka, and Francis Bond. 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.

[15] https://en-word.net

Klaus R. Scherer. 2005. What are emotions? and how can they be measured? *Social Science Information*, 44(4):693–727.

Gilles Sérasset. 2015. Dbnary: Wiktionary as a lemon-based multilingual lexical resource in rdf. *Semantic Web*, 6(4):355–361.

Piek Vossen, Wim Peters, and Julio Gonzalo. 1999. Towards a universal index of meaning. In *Proceedings of ACL-99 workshop, Siglex-99, standarizing lexical resources*, pages 81–90.