Some Considerations in the Construction of a Historical Language WordNet

Anas Fahad Khan
Istituto di Linguistica Computazionale "A. Zampolli"
Pisa, Italy
fahad.khan@ilc.cnr.it

John P. McCrae
University of Galway
Galway, Ireland
john.mccrae@insight-centre.org

Francisco Javier Minaya Gómez and **Rafael Cruz González** and **Javier E. Díaz-Vera**
University of Castilla-La Mancha
Ciudad Real, Spain
{Francisco.Minaya, Rafael.Cruz, JavierEnrique.Diaz}@uclm.es

Abstract

This article describes the manual construction of a part of the Old English WordNet (OldEWN) covering the semantic field of emotion terms. This manually constructed part of the wordnet is to be eventually integrated with the automatically generated/manually checked part covering the whole of the rest of the Old English lexicon (currently under construction). We present the workflow for the definition of these emotion synsets on the basis of a dataset produced by a specialist in this area. We also look at the enrichment of the original Global WordNet Association Lexical Markup Framework (GWA LMF) schema to include the extra information which this part of the OldEWN requires. In the final part of the article we discuss how the wordnet style of lexicon organisation can be used to share and disseminate research findings/datasets in lexical semantics.

1 Introduction

In this article, we look at the manual construction of that part of the Old English WordNet (OldEWN) dealing with the semantic field of emotion terms and which is based on previous scholarship on the emotion vocabulary for Old English (OE). This completely manual compilation process contrasts with the rest of the OldEWN which will be (primarily) the result of an initial phase of automated synset assignment followed by a subsequent post-correction phase; in this latter phase, scholars and specialists in OE will check generated synsets for correctness using a specialised platform developed for this task; more details on the full resource can be found in (Khan et al., 2022). Like the whole OldEWN, the emotion sub-wordnet is based on the second edition of Clark-Hall's *A Concise Anglo-Saxon Dictionary*[1] (Clark Hall, 1916) (CH).

We have several different aims in this article one of which is to describe some of the most recent developments in the construction of the OldEWN as a historical wordnet resource (following on from (Khan et al., 2022)). More generally, however, we wish to take a closer look into how to use legacy lexicographic resources such as the CH to create wordnets for historical languages (we present our workflow in detail in Section 2). In addition, we will present an extension of the Global WordNet Association (GWA) schema, in Section 3, that includes diachronic and etymological information and which we have developed for our emotion sub-wordnet; this may be useful for other similar wordnet projects. Finally, in Section 4, we discuss how the wordnet style of lexicon organisation can be used to share and disseminate research data in lexical semantics and how, even in cases where the coverage of a wordnet resource is low, such *sub*-wordnets can still be highly useful if they cover whole semantic fields.

2 Manually Creating an Emotion Lexicon in the Old English WordNet

Note that as the current article concentrates on the manually compiled part of the OldEWN dealing with emotions, and which we refer to as the *emotion lexicon* in what follows, we will not go into details as to the origins of the entire resource, its construction, or its scope[2]. The origin of the emotion lexicon lies in a dataset analysing emotion terms in OE which was compiled by Díaz-Vera and which

[1] We chose this edition because it has already been OCR'ed and is freely available online.

[2] These and other details of a more general nature can, however, be found in our previous article, (Khan et al., 2022).

is the result of a research program described in publications such as (Díaz-Vera, 2014). In this dataset, which is organised in a series of spreadsheets, OE words with emotion related meanings are classified on the basis of the emotion terms listed in the Geneva Emotion Wheel (GEV) (Scherer, 2005), with each word being listed in a separate spreadsheet under the appropriate GEV emotion term. Individual spreadsheets contain the following information for each of the lexical entries listed under that heading:

- The **lemma** for the entry, its **part of speech**, along with the different **orthographic** and **morphological variants** of the entry and their **distribution** in the corpus of surviving Old English texts, as well as **etymological information** on roots,
- A **gloss** of the literal sense of the entry – if the emotion term is literal or its emotion sense is primary; in cases of polysemic or derived terms where the emotion sense is secondary, both primary and secondary senses are described, as well as the **kind of figurative (metonymic/metaphoric) sense shift** (if any) which is hypothesised to have taken place between the two.

For instance, in the spreadsheet listing *shame* related terms in OE, we currently have 77 entries. These include the noun *scand* which literally means 'shame', but also include the polysemic verb *ablysian* which means both 'to blush' and 'to be ashamed'. The lexical information in these spreadsheets is derived from several different sources but crucially, lemma and sense information is based on that given in the Dictionary of Old English (DOE)[3]. Having become aware of this dataset our feeling was that it would lend itself very well to being incorporated within the OldEWN, especially since the lexical entries in the spreadsheets were already grouped together (provisionally) into synsets based on emotion terms. On the other hand, we were also eager to begin integrating the kind of information on figurative sense shifts included in the original Díaz-Vera dataset into OldEWN and extending the basic wordnet framework in order to do so; indeed data on figurative sense shifts is already being added to the Latin WordNet[4]. Once we made the decision to build the emotion lexicon part of the OldEWN on the basis of the Díaz-Vera dataset, we had to reconcile this with our previous choice to use the CH as the basis of the whole OldEWN; this is further discussed in Section 4. In what follows we give a description of our workflow for constructing the emotion lexicon[5]. In what follows we give a description of our workflow for the creation of the emotion lexicon.

For each of the emotion words in a spreadsheet, we look for the corresponding entry in the CH; we then use the information contained in the latter as the basis of the OldEWN lexical entry in the emotion lexicon[6]. In case either the entry or one or more of the senses does not exist in the CH we use another OE dictionary, the Bosworth-Toller *An Anglo-Saxon dictionary* (Bosworth, 1882) (BT) as the basis of a new lexical entry and/or senses. As regards the creation of OE synsets in the emotion lexicon, we use the synset which is the closest modern day English equivalent to the word sense in question in the Open English WordNet[7] (OEWN)[8] as a reference. For instance, in the case of OE words in the *shame* spreadsheet we look for synsets in OEWN containing the verb *to shame*, the noun *shame*, the adjective *ashamed*, etc. This gives us a set of relevant (modern) English synsets which we use as pivots to define new Old English synsets: using the definitions in the CH (or the BT in case of missing definitions) to decide which synset to link to (this is a purely manual process for now). We then map our new Old English synsets to their corresponding Open English synsets using the latter's Collaborative Interlingual Index ID (described below in Section 3). Finally, we add information on figurative sense shifts between the entries at the level of the sense (rather than at the synset level) using a modified version of the GWA LMF format; see the next section for more details.

3 Extending the Global WordNet LMF format

The Global WordNet Association (GWA) formats were introduced by Bond et al. (2016) and Mc-

[3]The electronic version of the latest draft can be found here https://doe.artsci.utoronto.ca/

[4]See https://latinwordnet.exeter.ac.uk/lexicon

[5]Note that although the emotion lexicon takes the Díaz-Vera dataset as a starting point, we do not necessarily keep to the synset assignation proposed therein.

[6]In particular we take the lemma and the sense definitions from CH. Although, these defintions may also sometimes be modified in case they do not accord with latter scholarship.

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

[8]Although their acronyms are similar, the OEWN is not to be confused with OldEWN.

Crae et al. (2021) to serve as a common set of schemata for the representation of wordnets and to enable their integration in the Open Multilingual Wordnet[9] through the Collaborative Interlingual Index (CILI). The formats describe three fully convertible serializations: an XML format based on Kyoto-LMF (Soria et al., 2009), a JSON serialization, and a RDF serialization that is a subset of the OntoLex-Lemon (McCrae et al., 2012) model. The three formats have been adopted by a number of projects and initiatives in the wordnet community including the OEWN mentioned above. Since all of the formats are fully interoperable and have the same underlying conceptual model, we focus on the XML based LMF format (GWA LMF) in what follows. These formats, which are closely based on the original Princeton WordNet (Miller, 1995) data model, model wordnets as containing **lexical entries** which have a number of **senses** that are linked to **synsets**[10]. As the formats are designed for the interchange of wordnets, they were developed with the goal of providing only a minimal number of common features. As such, the intention was for users to extend the set of elements in these schemas to represent their own data. And in fact, this is the strategy we pursued in order to be able to encode the OldEWN, and in particular the emotion lexicon, as we describe next.

An Extension of the GWA LMF Format for Diachronic Lexical Data

To start with, our resource is closely aligned to a pre-existing dictionary but with various new additions to the original content, including new lemmas and senses (and therefore sense definitions). We therefore felt it would be desirable to add definitions for individual senses along with metadata for specifying when entries/senses/definitions have been added or modified[11] to our wordnet. None of these features is available in the current GWA formats, and neither are a number of others that are important for historical languages such as OE (although these features can also be important for contemporary languages). For instance, we would like to include markers of rarity/uniqueness such as are found in the CH, as well as, more generally, information regarding dating, variations in forms along with information about word etymologies and specifically sense shifts. Finally, the GWA formats do not permit for the inclusion of salient (to OE) morpho-syntactic features like grammatical gender which we would also like to include in our resource[12]. Consequently, we made the following modifications to the GWA LMF format:

- The introduction of an **Etymology** element to be associated with both **LexicalEntry** and **Sense** elements from the original schema; this element consists of a series of one or more **EtyLink**s, where the latter represent an etymological link between two elements.
- This new **EtyLink** element carries attributes for specifying the source and target of an etymological link as well as for type of link; this allows us to indicate the kind of figurative conceptual shift which has taken place between two senses.
- The addition of a @grammaticalGender attribute to the **LexicalEntry** element.
- The addition of a **SenseDefinition** element related to the Sense element (with relevant Dublin Core metadata attributes for provenance information).

Our intention is for this extended schema to be reusable across a more general family of diachronic wordnet use cases. Indeed, in order to enhance this re-usability, we based the etymological part our expanded schema on a pre-existing ISO standard, namely, the latest multi-part version of LMF (Romary et al., 2019). We have made our new extended version of the GWA LMF format with these new features available as a DTD[13]. We have also defined an XSLT transformation from our extended version of the GWA LMF format to the original GWA LMF format[14].

In the listing below, we use our new extended schema to represent the OE noun *āblysung* which means both 'blushing' and 'shame' and where there is a resultative metonymy relation between the two senses of the word which we have listed:

[9] https://omwn.org/

[10] Further documentation can be found at https://globalwordnet.github.io/schemas/.

[11] Adding definitions for individual senses would help users to see what we based our decisions on when assigning synsets to individual senses.

[12] One way of circumventing these restrictions would be to include this information in another resource to be linked to the OldEWN, perhaps a digital edition of the CH dictionary in a format like TEI-XML. However our intention is to make OldEWN as self contained a resource as possible.

[13] https://github.com/anasfkhan81/OldEnglish/blob/main/WN-IELMF-0.DTD

[14] https://github.com/anasfkhan81/OldEnglish/blob/main/IELMF2GWALMF.xsl

```
<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>
```

In our resource, shifts don't directly apply to synsets themselves but to individual senses; the networks of synsets and their relations, then, help us to 'locate' such changes in meaning within the wider lexicon. In the next section, we look in more detail at some of the issues behind the use of pre-existing, legacy resources in the creation of the OldEWN and the use of the wordnet format for disseminating and sharing research data.

4 Discussion: the use of Pre-Existing Dictionaries and focusing on semantic fields in creating a wordnet

As previously reported in (Khan et al., 2022), we made the decision to use a dictionary as the basis of our wordnet for OE quite early on in its development, in part as an experiment in how to create such a resource for a historical language using freely available, legacy lexicographic resources. The idea being to use dictionary definitions, along with collocation information from the corpus of existing Old English texts, to help bootstrap a first provisional round of synsets. For reasons of convenience, the dictionary we chose was the CH[15] since its definitions are shorter and generally simpler than the BT's (e.g., without the latter's nested sense structure) and the CH generally follows a consistent and straight-forward separation of terms into different senses, all of which make entries easier to process. On the other hand, the BT includes far more semantic information and indeed more senses than the CH and is generally much more comprehensive than the latter (which was targeted specifically towards students). This became abundantly clear during the process of putting together our emotion lexicon: indeed, we very quickly came up against cases where Díaz-Vera's original dataset – which takes the even more comprehensive DOE as its reference – described senses which were present neither in the CH or the BT. In many cases, these senses occurred just once in the corpus of Old English texts and in several cases only as translation glosses, i.e., these were senses which wouldn't necessarily be seen as good candidates for inclusion in a general purpose wordnet.

However, as we mentioned above, one of our central aims in this project is to show the usefulness of publishing specialised datasets using the wordnet model: even if we subsequently end up with a wordnet or subwordnet where the coverage of various different parts of the lexicon of the language in question is very uneven or perhaps non-existent[16]. Such resources be valuable for what they tell us about single semantic fields or thematic parts of the lexicon. Therefore, in our opinion, the wordnet format should be promoted as a shared semantic framework for disseminating and sharing research in lexical semantics and related fields, with a view to making such research data as interoperable as possible.

It is worth pointing out here that the original inspiration behind the creation of the Old English Wordnet was to enable the comparison of concepts (and their interrelationships) across different ancient Indo-European language lexicons. Our work is based on previous efforts on the creation of Latin, Ancient Greek and Sanksrit WordNets and the effort to harmonise their structure using a shared schema (Biagetti et al., 2021); with the inclusion of semantic shift information, we facilitate even richer kinds of comparison between languages.

5 Conclusion

In this article we have reported on some recent experiences of the authors' in the development of an emotion lexicon as an (enriched) part of a wordnet for Old English. We are currently only part way through the encoding of the original Díaz-Vera dataset. When completed it will be made available in all three GWA formats as a separate wordnet

[15]The are three main dictionaries for Old English, two of which (CH and BT) date from the late 19th century and are both in the public domain. The third, the *Dictionary of Old English* (DOE), is still very much under copyright – indeed, users require a paid subscription in order to access it – and we could not therefore use it as the basis of our resource, which we intend to be published with a Creative Commons licence. The DOE is the most authoritative of the three and includes an extensive if not exhaustive list of citations for each entry. It is however currently unfinished and covers the letters A to I.

[16]This entails, however, that the kind of metadata which we referred to above, dealing with e.g., provenance, distribution, etc in Section 3 becomes especially important for the usability of the OldEWN.

based motion lexicon as well as being integrated into the main OldEWN resource.

References

Erica Biagetti, Chiara Zanchi, and William Michael Short. 2021. Toward the creation of Wordnets for ancient Indo-european languages. In *Proceedings of the 11th Global Wordnet Conference*, pages 258–266.

Francis Bond, Piek Vossen, John P. McCrae, and Christiane Fellbaum. 2016. CILI: the Collaborative Interlingual Index. In *Proceedings of the Global WordNet Conference 2016*.

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

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

Javier E Díaz-Vera. 2014. From cognitive linguistics to historical sociolinguistics: The evolution of old english expressions of shame and guilt. *Cognitive Linguistic Studies*, 1(1):55–83.

Fahad Khan, Francisco J Minaya Gómez, Rafael Cruz González, Harry Diakoff, Javier E Diaz Vera, John Philip 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.

John McCrae, Guadalupe Aguado de Cea, Paul Buitelaar, Philipp Cimiano, Thierry Declerck, Asunción Gómez-Pérez, Jorge Gracia, Laura Hollink, Elena Montiel-Ponsoda, Dennis Spohr, and Tobias Wunner. 2012. Interchanging lexical resources on the Semantic Web. *Language Resources and Evaluation*, 46(6):701–709.

John P. McCrae, Michael Wayne Goodman, Francis Bond, Alexandre Rademaker, Ewa Rudnicka, and Luis Morgado Da Costa. 2021. The GlobalWordNet Formats: Updates for 2020. In *Proceedings of the 11th Global Wordnet Conference*, pages 91–99.

George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Laurent Romary, Mohamed Khemakhem, Fahad Khan, Jack Bowers, Nicoletta Calzolari, Monte George, Mandy Pet, and Piotr Bański. 2019. Lmf reloaded. *arXiv preprint arXiv:1906.02136*.

Klaus R Scherer. 2005. What are emotions? and how can they be measured? *Social science information*, 44(4):695–729.

Claudia Soria, Monica Monachini, and Piek Vossen. 2009. Wordnet-LMF: fleshing out a standardized format for wordnet interoperability. In *Proceedings of the 2009 International Workshop on Intercultural Collaboration*, pages 139–146.