Old English WordNet

Existing Old English Lexical Resources

- There are a number of existing OE language resources available.
- These include retrodigitised versions of 19th century OE dictionaries such as the **Bosworth-Toller Anglo-Saxon Dictionary** and the **Clark Hall Concise Anglo-Saxon Dictionary**
- The **Dictionary of Old English** (DOE) is an ongoing project from the **University of Toronto** which aims to create a definitive lexicon of the Old English language
  - So far it covers the letters **A-K**
  - Unfortunately the resource is behind a paywall
- The **Thesaurus of Old English** (TOE) is a lexico-semantic resource that organises the OE lexicon using semantic categories derived from **Roget's Thesaurus**
  - Although accessible for research purposes, there are numerous limitations on its use
  - Can be accessed via the Evoke platform: http://evoke.ullet.net/

WordNet - an Introduction

- **WordNet** (WN) is a wide coverage, lexical database
- Initially in English, but thanks to its popularity was soon applied to other languages (currently > 200)
- WN is based on a simple organising principle of grouping **synonymous** words together
- As well as synonymy it also uses other lexico-semantic relations such as **hyponymy/hyperonymy** (as we will see)
- Has become a very popular resource for downstream NLP tasks and in particular for **word sense disambiguation** and **machine translation**
- There are also a number of corpora that have been annotated with WNs in different languages
- The basis of a number of tools and resources (such as **BabelNet**)

The History of WordNet

- The WordNet project was initiated in 1986 at **Princeton University** by the psychologist **George Miller,** a pioneering cognitive scientist
- Direction of WN was later taken over by **Christiane Felbaum**
- Resource was intended to be consistent with **contemporary theories of how the human brain stores lexical information**
- Original resource commonly referred to as **Princeton WordNet (PWN)**
- It is currently in **Version 3.1**

Organisation of WordNets 1/2

- At the basis of WN organisation is the concept of a **synset**. This is based on the linguistic concept of the **synonym**
  - Two words are traditionally defined as being synonyms if one can be substituted for the other in all/most sentences without changing the meaning of the resulting sentences
  - Words might be polysemous, have more than one sense, so can be synonymous in one sense with a given word but not in another
- A synset is a **set of synonymous word senses**
  - All words in the synset have the same **part of speech**
- E.g., The noun *car* in the sense of 'a motor vehicle with four wheels; usually propelled by an internal combustion engine' belongs to the synset:

  {*auto*, *car, automobile*, *machine*, *motorcar* }

- WN Synsets have their own ID numbers, glosses and (often) example sentences too

Organisation of WordNets 2/2

- In WN synsets can have relationships can have relationships with each other based on traditionally defined lexico-semantic relations such as **hyponymy**, **hyperonymy**, **meronymy**, as well as relationships such as **derivation**
  - A word X is a **hyponym** of Y if X is a type of Y, e.g., *car* is a hyponym of *vehicle*
  - Conversely X is a **hypernym** of Y if Y is a type of X, e.g., *vehicle* is a hypernym of *car*
  - Meronymy is a part of relation so that *wheel* is a **meronym** of *car*
- Again in WN, relations such as hyponymy/hypernymy and meronymy hold between **synsets** rather than individual words/word senses
- Crucially hyponymy/hyperonymy allow us to arrange **synsets in hierarchies**
- There are **44** such relations in PWN, some of which are specific to different parts of speech (e.g., troponymy for verbs)

WordNets Beyond Princeton

- The succes of the original WN soon led to a number of **WNs in other languages**
- It also gave birth to a number of subsequent projects and initiatives
- The **EuroWordNet project** (1996-99) was a European project that aimed to create a multilingual lexical database that linked together WNs in different European languages (**Dutch, Spanish, Italian, English, French, German, Czech, and Estonian**). These different WNs were developed independently, but were linked together by an **interlingual index (ILI)**
- The WNs in EuroWordNet were not all open; Open Multilingual WordNet worked towards the development of WNs with open licenses

The Global WordNet Association

- A **not for profit organisation** that “provides a platform for discussing, sharing and connecting wordnets for all languages in the world”
- Engages in numerous initiatives including organising the **Global WordNet conference**
- In particular it maintains a **Colloborative InterLingual Index (CILI)** which follows on from the EuroWordNet ILI index and is used to link together WN’s of different languages
- It has also defined an XML format the **Global WordNet Association format** which is based on the Lexical Markup Framework (LMF) standard
- This is used as an interchange format for WordNets

WordNets in Different Languages

- There exists a list of WNs in different languages on the GWA website http://globalwordnet.org/resources/wordnets-in-the-world/
- This list includes a wide number of modern languages (and Latin)
- A lot of these have open licenses, some are closed, many do not have clear licenses at all
- There is more than one WN for English
- Apart from the PWN there is also the Old English WordNet
- An Open Source WordNet based on PWN updated with thousands of new entries

How to create a WN for a new language

- There are at least two ways of creating a WN for a new language
- One way is to use the PWN as a **pivot** and to ‘translate’ PWN synsets into the target language (***merge***)
  - This can be used to **bootstrap** the target language WN which can subsequently be corrected
  - **Pro**: Easier than building from scratch (see below)
  - **Con**: Backbone taxonomy initially based on the Princeton one; a lot of work required for ancient languages to weed out anachronisms
- Another is to create a new WN from scratch, without reference to PWN (***expansion***)
  - Synsets in this new language can then be linked to other WNs via an Interlingual Index
  - **Pro**: The taxonomical structure is already native to the target language
  - **Con**: A lot of hard work!

An Old English WordNet?

- Idea: Build a **WN** on the basis of the work carried out in other ancient language WNs enriched with etymological and metaphoric/metonymic information
- A collaboration between various institutions/researchers including **ILC-CNR, the University of Exeter**, **National University of Ireland Galway**, **Universidad de Castilla - La Mancha**, the **University of Leiden**, **the Alpheios Project**
- The Old English WordNet will be a collaborative resource that we plan to publish with an open license
- Can be used to compare the organisation of different semantic fields/taxonomies across ancient modern languages

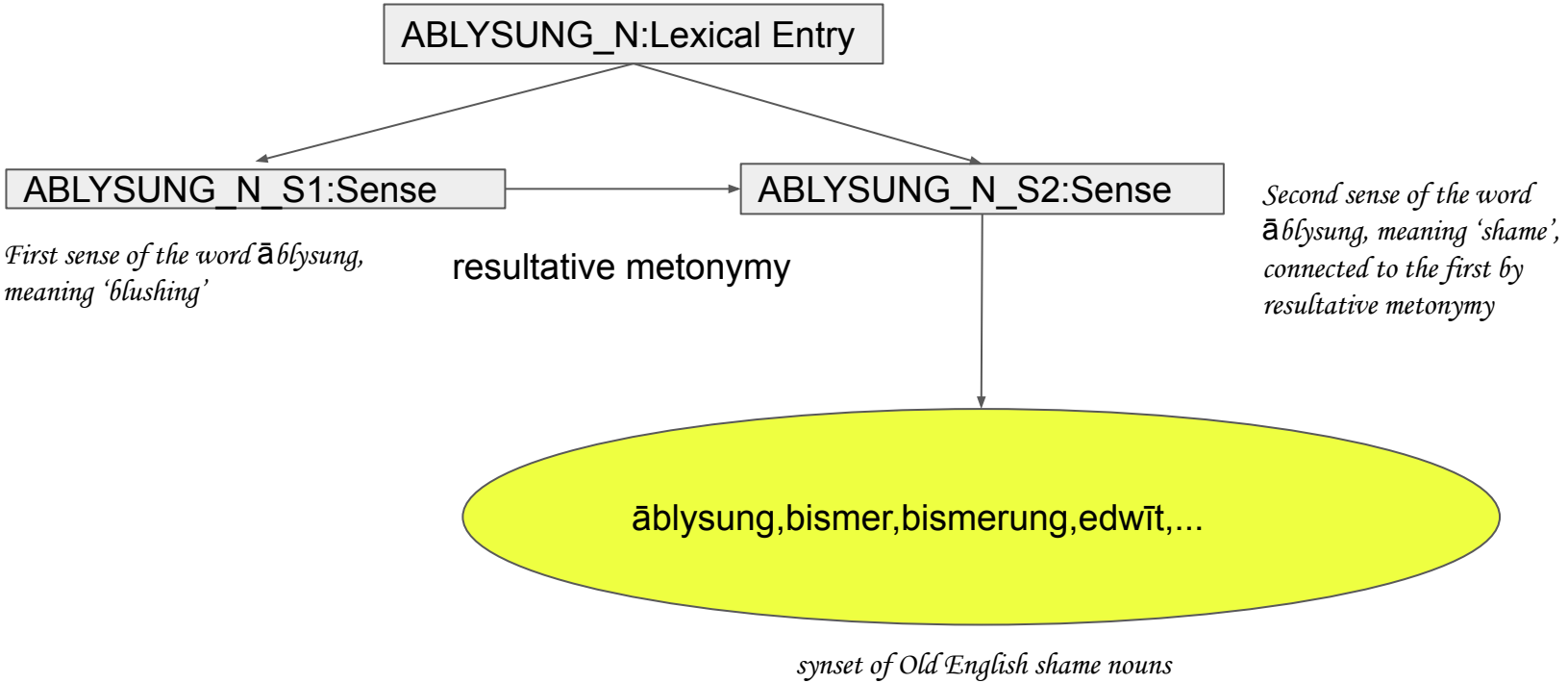
Building an Old English WordNet

- Our approach is to combine **automated methods** to derive synsets and link them to the Open English WordNet with **post-correction** via a validation platform **PLUS** a curated part that is based on previous research on the **Old English lexicon of emotions**
- The research in question was conducted by **Javier Diaz Vera** and looked at polysemy in **OE emotion terms** and in particular the metaphorical/metonymic processes underlying sense shift
- This produced an organisation of OE emotion terms in lists of synonyms along with etymology and this will form the basis of **the emotion component** of the Old English WordNet
- This curated part will hopefully expand to encompass other semantic domains

Old English WordNet - SHAME

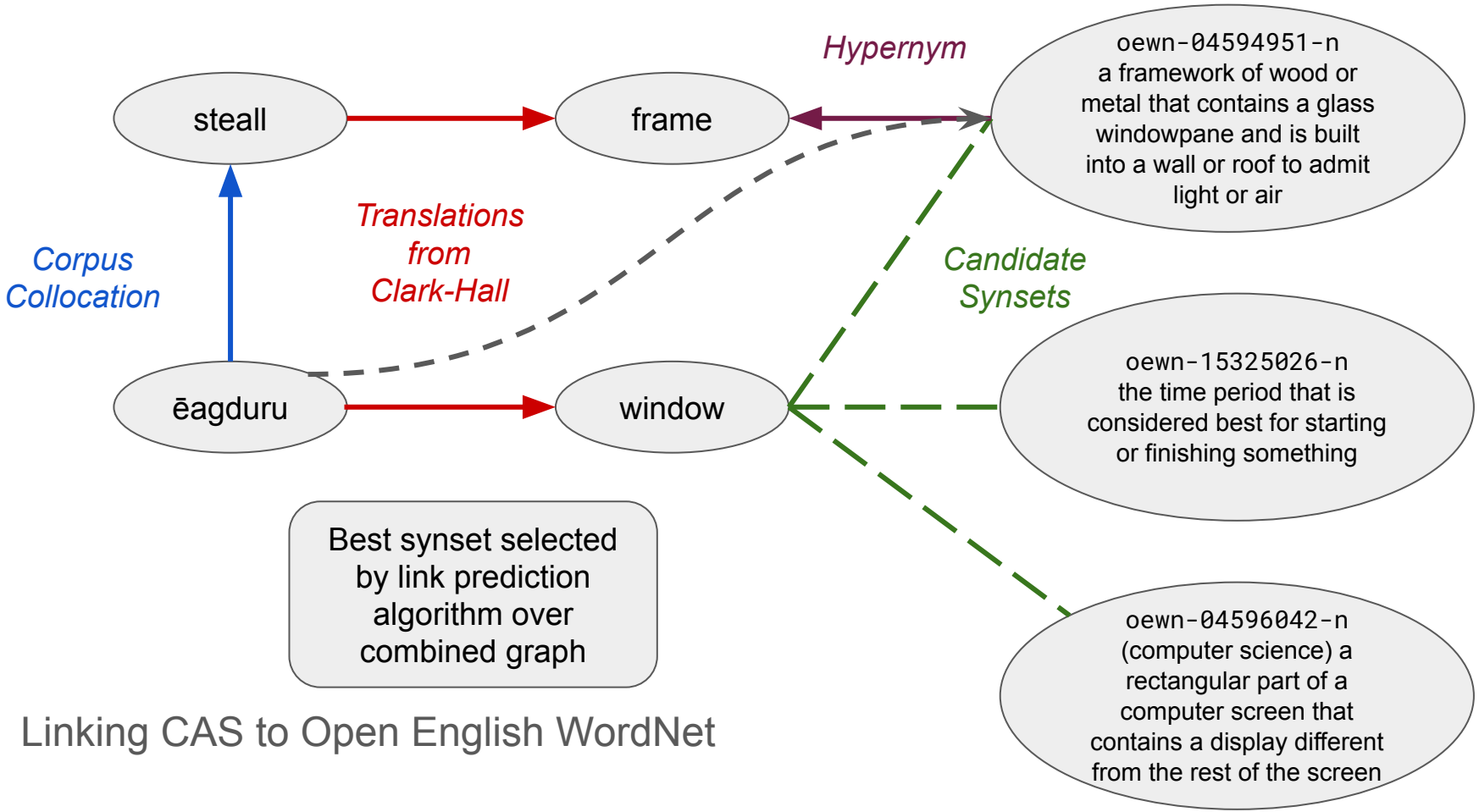
- We are currently working with an OE expert on organising SHAME expressions into synsets as part of the Old English WN emotion lexicon
- Taking advantage of the Diaz Vera dataset we can enrich this WN with data on polysemic shame terms and the processes of metaphor and metonymy involved in sense shifts
  - Want to align this with work being carried out in metaphor in Latin WN
- This first manual work on constructing synsets will inform the automated part of the construction of the WN

In this example we look at one of the members of the Old English shame synset for nouns as it is structured in our provisional Old English WN Emotion lexicon dataset



Building an Old English WordNet

- We will use the **Clark Hall** *Concise Anglo Saxion Dictionary* to extract lemmas and definitions (initially)
- The NUIG linking tool **Naisc** will be used to build a collocation graph based on Old English corpora (thanks to John for this!)
- This graph will be compared to the Open English WordNet, using link prediction techniques to rank candidate links between the collocation graph and the Open English WordNet graph
- This will give us a first set of **candidate synsets**
- These will then be validated/checked by experts in Old English using **a similar platform** to that used for other ancient language WNs



Linking CAS to Open English WordNet

Aims

- By publishing the Old English WordNet as linguistic linked open data we can take advantage of different ontologies and vocabularies and link to other linked data datasets in order to represent different historical and linguistic aspects of our dataset
- We are also looking into extending the Global WordNet Association LMF model to include enriched information for representing semantic shift/diachronic information
- We are considering creating a shared pool of concepts for the ancient language IE WNs, similar to what has been done for other culturally related languages
- This will become part of the GWA Collaborative Interlingual Index

Building a WN for OE emotions

Approach based on JDV's spreadsheets of emotions in OE:

- For each spreadsheet lemma in the emotion list find the corresponding lemma in the Clark Hall (CH)
- Create an entry in the LMF format with senses and definitions from the CH
- Create emotion synsets and link them to OEWN using ILI's as ID's
- Create synsets for other senses too
- If emotion sense isn't in the CH look at BT and (as a last resort) the DOE
  - Create a new sense in the entry.
  - Add all this info to the metadata (just comments for now).
  - Some of the senses mentioned by JDV are found only once in the corpus

Future Work

- Convert CH to TEI/OntoLex - only add definitions in case these aren't present in the CH other reference sense ID
- Add diachronic/morphological root/sense shift information when available
  - Promote WN's as a way of publishing and making accessible research on lexical semantic shift (idea for GWN workshop article)
  - Can be used to annotate an OE corpus (e.g., the DOE corpus)
- Short to Medium Term: Finish (as far as possible) converting JDV's list to LMF
  - Check if our taxonomies of emotions are appropriate for OE (need OE experts here)
- Skeleton vocabulary for OE converted to lexical entries in the OldEWN (use e.g., Skeat's list, or Barney's Word-Hoard?)