Cygnet: Refactoring the Open Multilingual Wordnet

Rowan Hall Maudslay
Université PSL
rowan.maudslay@psl.eu

Francis Bond
Palacký University
bond@ieee.org

Abstract

The wordnet file specification empowers the creators of different wordnets by allowing them to encode the same information in multiple different ways. A drawback of this approach is that redundancy is introduced. As a consequence, different wordnets often contain conflicting records, creating issues when one attempts to conduct multilingual research using multiple wordnets simultaneously. To address this, we present the *OMW Cygnet*, an experimental reformulation of wordnet that is designed to eliminate conflicting records and improve modularity. We convert data in 47 languages from the Open Multilingual Wordnet into this format, and release a web browser which makes it easy to navigate multilingual wordnets.

Keywords: Wordnet, Cygnet, lexical semantics, lexicography

1. Introduction

A **wordnet** (Miller, 1995) is a type of computational dictionary that has been very influential in computational research on lexical semantics. The first wordnet was developed for English (Fellbaum, 1998) and went on to spawn many spin-off projects for different languages. The proliferation of many different wordnets motivated the creation of standardised data structures, and in particular the **Wordnet LMF** (Vossen et al., 2013), a framework for encoding wordnets based on the Lexical Markup Framework (LMF; Francopoulo et al., 2006). The Wordnet LMF has evolved organically, and additional annotation layers have been added over time (e.g. Bond et al., 2020a; McCrae et al., 2021). Today it forms the backbone of the **Open Multilingual Wordnet** (OMW), which is a grid of multilingual wordnets (Bond and Foster, 2013; Vossen et al., 2016).

The Wordnet LMF was designed for exchanging wordnets: it gives individual wordnet projects flexibility, while at the same time ensuring a degree of compatibility between different resources. For multilingual research, however, this design can create problems, because different wordnets can contain conflicting records. In this paper, we review the Wordnet LMF, and relitigate several of the design decisions that were made during its development. We design an alternative wordnet format which is based on the semiotic components of signifiers, signifieds, and signs (Saussure, 1916). The result is a sign network: a **Cygnet** /sig-net/.

Internally, the format of Cygnet represents a major refactoring of the Wordnet LMF, which reduces redundancy and enables a unified semantic hierarchy across languages. Of particular note, Cygnet jettisons the notion of the synset—a concept that was key to the early development of wordnet—and instead captures the same information with fewer components. For the release of Cygnet, we convert all of the available OMW data to the Cygnet format, and also deliver an accompanying web browser which makes it easy to explore Cygnet data.

2. The History of Wordnet

The first wordnet to be produced was the **Princeton WordNet** (Fellbaum, 1998).[1] This project grew out of research in language acquisition: the initial goal of this project was to build a lexical database that would be consistent with theories of human semantic memory (Miller, 1986). It was later developed into an online human-readable dictionary (Miller et al., 1990), and into a database for language processing systems (e.g. Miller, 1995).

The Princeton WordNet has two unique design features which set it apart from other dictionaries. The first design feature is that concepts are represented by sets of synonymous lemmas. Each of these sets is called a synonym set, or **synset**. An example of a synset is {*sofa*, *couch*, *lounge*}, which corresponds to the concept of a cushioned seat for multiple people. The second design feature is that synsets are connected to each other by semantic relations, most notably hypernymy and meronymy. Hypernymy is a super–subordinate relation, commonly expressed as the IS A relation, while meronymy is a part–whole relation, commonly expressed as the HAS A relation. For example, in the Princeton WordNet, the {*oak*, *oak tree*} synset is connected by hypernymy to the {*tree*} synset, and by meronymy to the {*acorn*} synset. The inverse relations for hypernymy and meronymy are called hyponymy and holonymy respectively.

[1] In the literature, the term "wordnet" is generally used to refer to any lexical database modelled after the Princeton WordNet, while "WordNet" (with this capitalisation) is reserved for the Princeton WordNet. We follow this convention. Some wordnet projects adopt "WordNet" in their names; when referring to a specific project we use the capitalisation adopted by its authors.

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 7905–7917
11–16 May 2026.

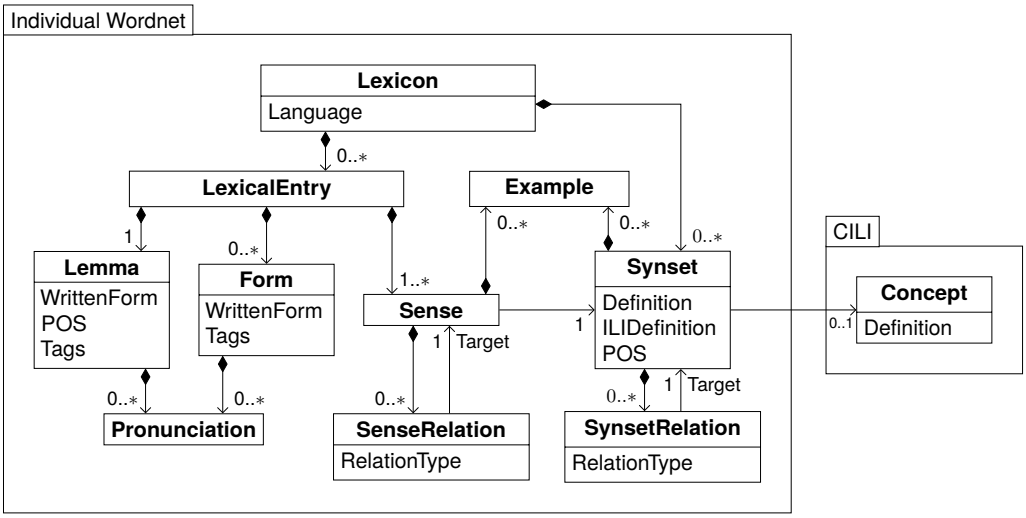

Figure 1: Structure of the Wordnet LMF

The term "synset" is a slight misnomer, since a synset is more than simply a set of synonymous wordforms: it is a complex data structure, which contains not only a set of synonyms and semantic relations, but also an ID, part-of-speech (POS) information, a definition (since Princeton WordNet version 2.0), and so on. Indeed, in the Princeton WordNet, there are many synsets which contain identical sets of words, but which correspond to distinct concepts. For example, there are two synsets comprised of {*disappear*, *go away*, *vanish*}, one that corresponds to departure without warning (e.g. "he *disappeared* without a trace"), and another that corresponds to becoming invisible. These two synsets are represented by distinct objects, which have different definitions and appear in different parts of the WordNet semantic network.

In the time since the creation of the Princeton WordNet, many other wordnets for different languages have been developed, leveraging the same design principles (see Bond and Paik, 2012). Most of these wordnets were created by adding language-specific wordforms to synsets from the Princeton WordNet. This wordnet creation method is referred to as the **expand** approach (Vossen, 1998). Other wordnets, however, were produced independently with their own synset ontologies. Some of these wordnets have been aligned with the Princeton WordNet; this wordnet creation method is referred to as **merge** (Vossen, 1998).

Data from many multilingual wordnets have been unified in the OMW. In the first version of OMW (Bond and Foster, 2013), wordnets were aligned using Princeton WordNet synsets. Language-specific synsets that did not have an equivalent in the Princeton WordNet were discarded. To address this limitation, the **Collaborative Interlingua Index** (CILI; Fellbaum and Vossen, 2012; Bond et al., 2016) was developed: it is a community-driven concept index, based on the synsets in the Princeton WordNet, but with the capacity to support new concepts. The second iteration of the OMW uses CILI to accommodate new synsets. To address difficulties in maintaining a centralised wordnet repository, this version of the OMW is decentralised: individual wordnet creators can host and update their respective wordnets based on the shared Wordnet LMF.

To interface between different languages in the OMW, a codebase has been created to merge OMW files (Bond et al., 2020a).[2] This codebase combines multiple wordnets into a single Wordnet LMF database, and rejects any wordnet which would create errors on integration. Another way to access the OMW data is through the `wn` Python module (Goodman and Bond, 2021), which allows the loading of multiple wordnets. This module allows a user to navigate from one wordnet to another (via CILI), but does not combine the wordnets.

Other efforts to construct multilingual lexica using wordnet design and data are BabelNet (Navigli and Ponzetto, 2010) and Universal Knowledge Core (UKC; Giunchiglia et al., 2017, 2018). The main differences between these resources and the OMW is that they are both based on universal inter-synset relations (whereas OMW features many language-specific ontologies), and access to them is restricted (whereas the OMW is open); for further comparison see Giunchiglia et al. (2023).

3. Reformulating the Wordnet LMF

The current version of the Wordnet LMF, as described in the Global Wordnet Association DTD[3], is illustrated in Figure 1 (some elements and attributes

[2] https://github.com/globalwordnet/OMW

[3] https://github.com/globalwordnet/schemas/blob/master/WN-LMF-1.4.dtd

are excluded for brevity). It interfaces between two components: an individual wordnet and a shared CILI, which is a flat list of concepts.

In the Wordnet LMF, each wordnet consists of one or more Lexicon objects, which contain LexicalEntry objects and Synset objects. A Synset object points to a concept in the CILI. Each LexicalEntry consists of one Lemma and one or more Senses. Each Sense points to a Synset. A LexicalEntry also optionally has one or more Form objects, which capture spelling variants or morphological inflections. Sense objects and Synset objects can both contain Example objects, which illustrate a meaning by showing a particular word usage. Finally, each Sense and Synset can have optional SenseRelation or SynsetRelation objects respectively, which capture ontological elements of the wordnet.

The design of the Wordnet LMF has been influenced by path dependency, with each new iteration maintaining compatibility with preceding versions. Moreover, to give flexibility to the creators of individual wordnet projects, the Wordnet LMF allows many pieces of information to be recorded in several different places. A downside of this design is that one often finds conflicting records when attempting to combine data from different wordnets.

In this section, we revisit the Wordnet LMF, and propose a variety of alternative formulations. These reformulations are designed to streamline multilingual research, but sacrifice backwards compatibility. Each design change is motivated by two principles: (1) to improve modularity, and (2) to enforce a single source of truth (SSOT) design. We argue that everything should be encodable in one place only, to reduce redundancy and prevent conflicting records. Changes are proposed to the following:

Ontology. In the Wordnet LMF, inter-concept relations are stored at the synset level. This design choice was made to enable language-specific ontologies. In practice, however, the vast majority of wordnets copy their synset relations directly from the Princeton WordNet, introducing redundancy. Moreover, there is nothing in the Wordnet LMF to stop the ontology becoming malformed when data from multiple wordnets is combined. For example, different wordnets combine to form loops in the hypernym hierarchy. Tools such as `wn` check for loops, but only within a single wordnet.

⇒ *Proposal.* Inter-concept relations should be in a standalone annotation layer, distinct from other wordnet content. Language-specific ontologies can be produced by creating alternative versions of this layer. The hypernym hierarchy for all concepts should be a directed acyclic graph; concept relations should not be excluded if they violate this principle.

Examples. In the Wordnet LMF, an Example object belongs to either a Sense object or a Synset object. This creates two issues. Firstly, it means that the same example can be put in multiple places. Secondly, it means that if an example contains instances of several different senses, it needs to be included multiple times (once per sense). Examples are represented by text strings: the location of the target sense in the example is not identified.

⇒ *Proposal.* Examples should exist independently, and should point to one or more senses (and not to synsets). Each example should be annotated with the location of the tokens that evoke the senses to which it is connected.

Definitions. In the Wordnet LMF, the definition of a concept is stored in three places: once in a Concept object and twice in the corresponding Synset object. Some wordnets use the definition field of the Synset to record definitions in their language, while others contain duplicate English definitions.

⇒ *Proposal.* Definitions should be in a separate annotation layer, thereby enabling a user to load a set of definitions in a language of their choosing, or to load definitions which use different defining strategies, such as full-sentence definitions (Hanks, 1987).

POS. In the current design, POS information is encoded in the Synsets of each wordnet. This creates the possibility that different wordnets could assign different POS to the same CILI concept, which should be impossible. In the Wordnet LMF, POS is also attributable to Lemma objects. In theory, the decision to include two separate POS values (in Synsets and Lemmas) is well motivated, as it can capture a distinction between grammatical categories (e.g. noun) and ontological categories (e.g. entity). Cross-linguistically, these two categories correlate: nouns typically denote entities, verbs denote events or relations, and adjectives denote properties (Croft, 1991; Dixon, 2004). However, it is possible for a concept which belongs to one ontological category to be realised using an expression which belongs to a different grammatical category. For example, in the Princeton WordNet the synset defined as "without any delay" belongs to the adverb ontology, and is expressed by *instantaneously* and *in a flash*. The former is an adverb, but the latter is a prepositional phrase. In the OMW, however, both are labelled as adverbs. In general, the distinction between ontological categories and grammatical categories is not made, meaning that the second POS encoding is usually redundant.[4]

[4]The exception is the Open English Wordnet (McCrae et al., 2019), which uses the Lemma POS attribute to add an additional distinction to adjectives.

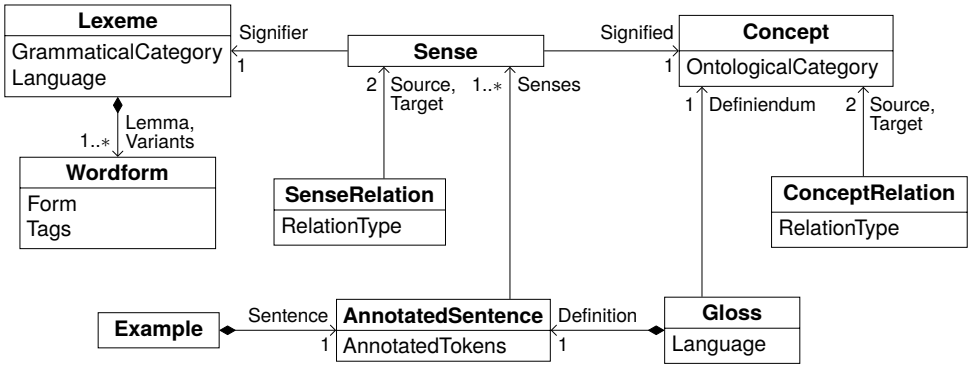

Figure 2: Structure of Cygnet

⇒ *Proposal.* POS should be split into two distinct classes: a grammatical category and an ontological category. The grammatical category should be a property of lexical entries, and the ontological category should be a property of concepts. A synset should not contain POS information, because the ontological category of a synset is conditioned on the category of the concept to which it is connected.

Language. In the Wordnet LMF, language is a property of the Lexicon class. This means that a wordnet lexicon is inherently monolingual.

⇒ *Proposal.* Language should be a property of lexical entries, not a property of a dataset.

Relations. In the Wordnet LMF, some types of relation are the inverse of other types. For example, both hypernymy and hyponymy are supported in the Wordnet LMF: every time there is a hypernymy relation in one direction, there *should* be hyponymy in the other direction. The wordnets differ as to whether they store the inverse in the XML: most do, but OdeNet (Siegel and Bond, 2021) and OpenWordnet-PT (de Paiva and Rademaker, 2012) do not. In the data used in this paper, we found that 2,132 relations were missing their inverse.

⇒ *Proposal.* Symmetric and asymmetric relations should be represented as distinct relation classes; each should only be encoded once.

4. OMW Cygnet

In this section, we describe the design of **OMW Cygnet**. Cygnet is an experimental reformulation of the OMW 2, which implements the design proposals outlined in §3. An overview of the structure of Cygnet is shown in Figure 2 (metadata and indices are not shown). Cygnet consists of seven main classes:

1. A **Concept** consists of an ontological category and an index (not depicted in Figure 2). The meaning of a concept comes from its gloss and from inter-concept relations (see below).

2. A **ConceptRelation** captures a relation between two concepts, such as hypernymy. Inverse relations are not included; information about which relations are symmetric and asymmetric is stored separately.

3. A **Lexeme** is a collection of related wordforms in a particular language. It consists of a lemma, a language attribute, a grammatical category, and optional variants. The lemma and variants are both represented using an auxiliary Wordform class, which contains a string and language-specific tags which can denote properties of the form and contain pronunciation information.

4. A **Sense** is a tuple which points to one lexeme and one concept: it is a bridge between a language-specific signifier and a language-agnostic signified, as shown in the top row of Figure 2.

5. A **SenseRelation** captures relationships between senses; its structure is the same as a ConceptRelation.

6. An **Example** consists of a sentence which illustrates one or more senses. The sentence is represented as an AnnotatedSentence object, which contains markup that points to senses evoked within it.

7. A **Gloss** provides a concept with a definition written in natural language. The definition is represented as an AnnotatedSentence, which can optionally identify the senses it contains.

In the Wordnet LMF, most objects are in composition relationships, which means that they are owned by another object and do not exist independently (see Figure 1). As a result, these objects are tightly coupled, reducing modularity. In Cygnet, the main classes all exist independently.

Layer	Dependencies
ConceptLayer	—
LexemeLayer	—
SenseLayer	ConceptLayer & LexemeLayer
GlossLayer	ConceptLayer
ExampleLayer	SenseLayer
ConceptRelationLayer	ConceptLayer
SenseRelationLayer	SenseLayer

Table 1: Cygnet annotation layers

Cygnet is structured in seven distinct annotation layers, each corresponding to one of the main classes above. These layers, and their dependencies, are shown in Table 1.[5] This design means that Cygnet is heavily modularised, allowing users to plug in data from different sources with ease. If a user wanted to change definitions to a different language, they could do that by changing the definition layer; if they wanted to add data from another language, they could do that by adding additional lexeme and sense layers; if they wanted to change to a different ontology, they could do that by changing the concept relation layer; if they wanted to add examples from a particular corpus, they could do that by loading another example layer; and so on.

5. Release

For the first release of Cygnet we produce a database and an accompanying web browser. The database is archived at Zenodo, DOI: 10.5281/zenodo.19050701.

5.1. Dataset

We design an XML Schema File to describe the Cygnet structure, and write a script to convert OMW 2 data into this form and into an SQL database for more efficient access. This script merges data from multiple Wordnet LMF files into a single file. The result is a dataset covering 47 languages from 59 wordnets. For all included wordnets we use the most up-to-date version available at the time of writing.[6] The languages and wordnets in Cygnet are shown in Table 2, and details about the data processing procedure are given below. Additional detail about the included wordnets can be found in Appendix B.

[5]The GlossLayer may include pointers to the SenseLayer, but this information is optional and not required for compatibility.

[6]The only exception is DanNet, where we use the OMW 2.0 release even though a newer version is available; we use this version because definitions in the newer version are truncated due to licensing restrictions.

Merging Ontologies. We include all concepts from CILI 1.0 and add a new concept for each synset that exists in a wordnet but is not yet in CILI. In the concept relation layer, we combine synset relations from every wordnet. We remove duplicate relations and transitive relations: if concept A is connected to B, and B is connected to C, then if an edge exists directly from A to C we remove it. We also exclude 431 relations which introduce loops. For every relation, if its inverse is missing, we recover it; we then only store one direction for each relation type. One problem we encounter is that sometimes a new concept's POS is mislabelled, which means that it is included in the ontology of another class (e.g. a verb is labelled as a hypernym of a noun). To identify these conflicts, we start at the root nodes of each POS's ontology, and relabel the POS of new concepts in each graph that are mislabelled.

Combining Duplicate Languages. Some languages are covered by multiple wordnets. In Cygnet we combine this data. Repeat entries (e.g. the same sense appearing in multiple wordnets of the same language) are only included once.

English Definitions. Some of the merged wordnets contain new concepts, and only provide a definition in the target language. In order to ensure that every concept in Cygnet has an English definition, we machine translate the extra definitions. Machine translation is performed using ArgosTranslate, a library based on OpenNMT (Klein et al., 2017).

POS Conversion. Cygnet supports a separate encoding for the grammatical categories and ontological categories. Present wordnet data does not make this distinction. POS information is recorded separately for both lemmas and synsets, and while there are cases of divergence (e.g. 77 times in the most recent version of DanNet, Pedersen et al., 2009), these cases appear to be the result of erroneous annotation. For this reason, we discard the POS of lemmas, and set the grammatical category of lexemes and the ontological category of concepts both to match the synset POS. To simplify cross-resource mapping, we convert these POS values to values that more closely resemble the POS labels in Universal Dependencies (Nivre et al., 2017, 2020). The mapping from wordnet POS categories to their new labels is shown in Table 3. The satellite adjective POS class is lost: this POS class is a legacy of the Princeton WordNet, and is not commonly found in other POS specifications.

Example Annotation. We annotate each example sentence with the position of the sense which it evokes. For examples belonging to Sense objects,

Language	Wordnet	Version	Citation
Abui	Abui Wordnet	abwn:0.1	Kratochvil and Morgado da Costa, 2022
Albanian	Albanet	omw-sq:2.0	Ruci, 2008
Arabic	Arabic WordNet	omw-arb:2.0	Elkateb et al., 2006
	TUFS	tufs-ar:2.0	Bond et al., 2020b
Assamese	TUFS	tufs-as:2.0	Bond et al., 2020b
Basque	Multilingual Central Repo.	omw-eu:2.0	Gonzalez-Agirre et al., 2012
Bulgarian	BulTreeBank Wordnet	omw-bg:2.0	Simov and Osenova, 2023
Burmese	TUFS	tufs-my:2.0	Bond et al., 2020b
Cantonese	Cantonese Wordnet	cantown-yue:1.0	Sio and Costa, 2019
Catalan	Multilingual Central Repo.	omw-ca:2.0	Gonzalez-Agirre et al., 2012
Croatian	Croatian Wordnet	omw-hr:2.0	Oliver et al., 2015
Danish	DanNet	omw-da:2.0	Pedersen et al., 2009
Dutch	Open Dutch WordNet	omw-nl:2.0	Postma et al., 2016
English	Open English Wordnet	oewn:2025	McCrae et al., 2019
	TUFS	tufs-en:2.0	Bond et al., 2020b
Filipino	TUFS	tufs-tl:2.0	Bond et al., 2020b
Finnish	FinnWordNet	omw-fi:2.0	Lindén and Carlson., 2010
French	Wordnet Libre du Français	omw-fr:2.0	Sagot and Fišer, 2008
	TUFS	tufs-fr:2.0	Bond et al., 2020b
Galician	Multilingual Central Repo.	omw-gl:2.0	Gonzalez-Agirre et al., 2012
German	Offenes Deutsches WordNet	odenet:1.4	Siegel and Bond, 2021
	TUFS	tufs-de:2.0	Bond et al., 2020b
Greek	Greek Wordnet	omw-el:2.0	N/A
Hebrew	Hebrew Wordnet	omw-he:2.0	Ordan and Wintner, 2007
Icelandic	IceWordNet	omw-is:2.0	N/A
Indonesian	Wordnet Bahasa	omw-id-2.0	Noor et al., 2011
	TUFS	tufs-id:2.0	Bond et al., 2020b
Italian	MultiWordNet	omw-it:2.0	Pianta et al., 2002
	ItalWordNet	omw-iwn:2.0	Roventini et al., 2000
Japanese	Japanese Wordnet	omw-ja:2.0	Bond and Kuribayashi, 2023
	TUFS	tufs-ja:2.0	Bond et al., 2020b
Khmer	TUFS	tufs-km:2.0	Bond et al., 2020b
Korean	TUFS	tufs-ko:2.0	Bond et al., 2020b
Kurdish	KurdNet	kurdnet:1.0	Aliabadi et al., 2014
Lao	TUFS	tufs-lo:2.0	Bond et al., 2020b
Latvian	Latvian WordNet	wordnet_lv:1.0	Paikens et al., 2023
Lithuanian	Lithuanian WordNet	omw-lt:2.0	Garabík and Pileckytė, 2013
Standard Malay	Wordnet Bahasa	omw-zsm-2.0	Noor et al., 2011
	TUFS	tufs-ms:2.0	Bond et al., 2020b
Mandarin Chinese	Chinese Open Wordnet	omw-cmn:2.0	Huang et al., 2010
	TUFS	tufs-zh:2.0	Bond et al., 2020b
Mongolian	TUFS	tufs-mn:2.0	Bond et al., 2020b
Norwegian Bokmål	Norwegian Wordnet	omw-nb:2.0	N/A
Norwegian Nynorsk	Norwegian Wordnet	omw-nn:2.0	N/A
Polish	plWordNet	omw-pl:2.0	Piasecki et al., 2009
Portuguese	OpenWordnet-PT	own-pt:1.0.0	de Paiva and Rademaker, 2012
	TUFS	tufs-pt:2.0	Bond et al., 2020b
Romanian	Romanian Wordnet	omw-ro:2.0	Tufiș and Barbu Mititelu, 2014
Russian	TUFS	tufs-ru:2.0	Bond et al., 2020b
Slovak	Slovak WordNet	omw-sk:2.0	Ondrej Dzurjuv and Garabík, 2011
Slovenian	sloWNet	omw-sl:2.0	Fišer and Erjavec, 2010
Spanish	Multilingual Central Repo.	omw-es:2.0	Gonzalez-Agirre et al., 2012
	TUFS	tufs-es:2.0	Bond et al., 2020b
Swedish	WordNet-SALDO	omw-sv:2.0	Borin et al., 2013
Thai	Thai Wordnet	omw-th:2.0	Thoongsup et al., 2009
	TUFS	tufs-th:2.0	Bond et al., 2020b
Turkish	TUFS	tufs-tr:2.0	Bond et al., 2020b
Urdu	TUFS	tufs-ur:2.0	Bond et al., 2020b
Vietnamese	TUFS	tufs-vi:2.0	Bond et al., 2020b

Table 2: Languages and wordnets in Cygnet

Wordnet	Description	Cygnet
n	noun	NOUN
v	verb	VERB
a	adjective	ADJ
s	satellite	
r	adverb	ADV
p	adposition	ADP
u	unknown	UNK
c	conjunction	CONJ
x	non-referential	NREF

Table 3: POS conversion

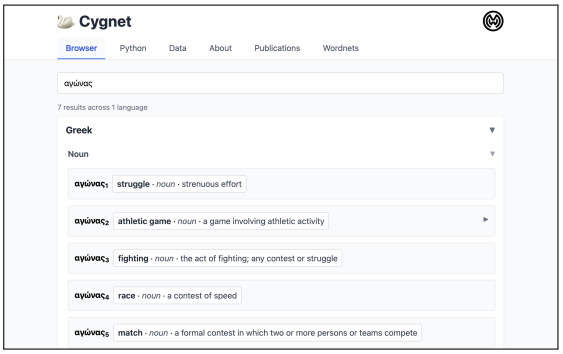

Figure 3: Screen capture of Cygnet website

we perform a string search for the wordform that belongs to the sense, as well as all morphological and inflectional variants. For examples belonging to Synset objects, we search for all of the synset's senses, and match with whichever fits. Any matching substring is extended to token boundaries. The Python module `spacy` (Honnibal et al., 2020) is used for tokenisation and lemmatisation; we initialise `spacy` using the small model of whichever language we are treating, or the multilingual model if no tailored model is present. If we are unable to identify the wordform of a sense in an example then we discard it and log the discrepancy. Details about the number of examples kept and discarded are given in Appendix A.

Annotated Definitions. The Princeton WordNet Gloss Corpus contains sense annotation for all definitions in Princeton WordNet 3.0, based on the senses in that same version. Many of these definitions and senses exist unchanged in the present releases of CILI and the Open English Wordnet. We therefore extract the sense markup from the Princeton WordNet Gloss Corpus for all concepts that persist in CILI, and include this information in the Gloss layer.

Additional Attributes. Every class in Cygnet contains a Provenance object (not shown in Figure 2) that contains metadata about its origin. The Wordnet LMF also contains several other pieces of information not discussed thusfar (and not depicted in Figure 1). We discard confidence scores, lexicalisation labels, and syntactic patterns. In the future, this information could easily be added to Cygnet with additional properties or in new annotation layers.

5.2. Website

To explore Cygnet data we design a web browser. A screenshot of the website is shown in Figure 3. Users are able to search for words, and filter by language and by POS. Selecting a concept shifts the view to reveal additional concept information, including every wordform associated with the concept, inter-concept relations, multilingual definitions, and language-specific examples. Other tabs provides information about the Cygnet project and the ability to download Cygnet data.

6. Contextualising Cygnet

A comparison between Cygnet and other multilingual resources based on wordnet is shown in §6. In some respects, Cygnet is similar to BabelNet and UKC, as these resources also enable a universal concept ontology. However, Cygnet additionally supports language-specific ontologies, and unlike these resources, it is open. A notable omission from Cygnet, which makes it distinct from all other wordnet-based multilingual lexica, is that it contains no formal notion of a synset. The reason for this is that, having implemented the proposals outlined in §3, the Synset class is itself redundant: it contains no additional information that cannot be recovered from a combination of senses and concepts.

To evaluate the extent of the database conflicts and redundancy removed in Cygnet we compare the Cygnet SQL database to the SQL database generated by the `wn` module for the same wordnets. This comparison is shown in Table 5. In megabytes, Cygnet is 72% of the size of the `wn` version. The size of Cygnet reduces to 303.8 MB (38%) when Provenance metadata is removed. The most pronounced difference between the two databases is the number of concepts included in both versions. The reason for this difference is that Cygnet only includes each concept once, whereas in the `wn` version a concept is included once for each synset that represents it. Redundancy is further reduced in Cygnet by merging equivalent synsets and senses together and by removing inverse relations.

Cygnet is built from community-contributed wordnets, and we are committed to give back to those communities. During processing, Cygnet records all data-quality events in structured JSON logs, which include information about duplicate identifiers, unresolvable sense links, hypernym conflicts,

Resource	Centralised Database?	Universal Ontology?	Tagged Examples?	Open?
OMW	✓	✗	✗	✓
OMW 2	✗	✗	✗	✓
OMW Cygnet	✓	✓	✓	✓
BabelNet	✓	✓	✗	✗
UKC	✓	✓	✗	✗

Table 4: Comparison of lexical-semantic resources

Metric	Cygnet	OMW 2 [wn]
Size (MB)	569.8	796.5
Concepts	127,606	1,259,680
Lexemes	1,415,546	1,496,227
Senses	2,165,289	2,245,911
Concept relations	282,960	374,218
Sense relations	123,048	123,441

Table 5: Size comparison of Cygnet and OMW 2

morphological match failures, and so on. A companion script distils these logs into human-readable reports, grouped by severity (CRITICAL, WARNING, INFO). These reports are targeted at upstream wordnet maintainers who may not be familiar with Cygnet's internals. This closes a feedback loop that benefits the wider lexical-resource ecosystem: errors revealed during integration are reported back to the source, with the goal of improving data quality not only in Cygnet but in the original resources themselves. We view this as a form of community stewardship and ecosystem maintenance, whereby downstream consumers actively contribute to the health of the shared resources they depend on.

7. Conclusion

In this paper, we presented the OMW Cygnet. Externally, the OMW Cygnet looks very similar to OMW 1 or 2. Internally, however, it is based on a new data format which reduces redundancy and improves modularity. This new data format means that a range of database conflicts have been removed, resolving many issues that have until now frustrated multilingual research. We converted data from OMW 2 into the Cygnet format, and produced a website to interface with this data. The website and data are made openly available.[7]

In ongoing work, we are developing a Python module that interfaces with Cygnet, with endpoints that resemble the endpoints of existing Python modules for processing wordnet data, namely `nltk` (Bird et al., 2009) and `wn` (Goodman and Bond, 2021). We plan to describe this module in greater depth in a separate technical report. In future work, we want to add further refinements to the sense and concept relation types. We also want to convert more wordnets into the OMW family of formats, so that these can then be integrated into Cygnet.

8. Acknowledgements

We thank the developers and maintainers of all the wordnets we are using here. This work has received support under the Major Research Program of PSL Research University "CultureLab" launched by PSL Research University and implemented by ANR with the references ANR-10-IDEX-0001.

[7] https://cygnet.maudslay.eu

9. Bibliographical References

Purya Aliabadi, Mohammad Sina Ahmadi, Shahin Salavati, and Kyumars Sheykh Esmaili. 2014. Towards building KurdNet, the Kurdish WordNet. In *Proceedings of the Seventh Global Wordnet Conference (GWC 2014)*, pages 1–6, Tartu, Estonia. Global Wordnet Association (GWA).

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. O'Reilly Media, Inc.

Francis Bond and Ryan Foster. 2013. Linking and extending an Open Multilingual Wordnet. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL 2013)*, volume 1, pages 1352–1362, Sofia, Bulgaria. Association for Computational Linguistics (ACL).

Francis Bond and Takayuki Kuribayashi. 2023. The Japanese Wordnet 2.0. In *Proceedings of the 12th Global Wordnet Conference (GWC 2023)*, pages 179–186, San Sebastian, Basque Country. Global Wordnet Association (GWA).

Francis Bond, Luis Morgado da Costa, Michael Wayne Goodman, John Philip McCrae, and Ahti Lohk. 2020a. Some issues with building a multilingual Wordnet. In *Proceedings of the Twelfth Language Resources and Evaluation Conference (LREC 2020)*, pages 3189–3197, Marseille, France. European Language Resources Association (ELRA).

Francis Bond, Hiroki Nomoto, Luís Morgado da Costa, and Arthur Bond. 2020b. Linking the TUFS basic vocabulary to the Open Multilingual Wordnet. In *Proceedings of the Twelfth Language Resources and Evaluation Conference (LREC 2020)*, pages 3181–3188, Marseille, France. European Language Resources Association (ERLA).

Francis Bond and Kyonghee Paik. 2012. A survey of WordNets and their licenses. In *Proceedings of the 6th Global WordNet Conference (GWC 2012)*, pages 64–71, Matsue, Japan. Global Wordnet Association (GWA).

Francis Bond, Piek Vossen, John McCrae, and Christiane Fellbaum. 2016. CILI: The Collaborative Interlingual Index. In *Proceedings of the 8th Global WordNet Conference (GWC 2016)*, pages 50–57, Bucharest, Romania. Global Wordnet Association (GWA).

Lars Borin, Markus Forsberg, and Lennart Lönngren. 2013. SALDO: a touch of yin to WordNet's yang. *Language Resources and Evaluation*, 47(4):1191–1211.

William Croft. 1991. *Syntactic Categories and Grammatical Relations: The Cognitive Organization of Information*. University of Chicago Press.

Valeria de Paiva and Alexandre Rademaker. 2012. Revisiting a Brazilian WordNet. In *Proceedings of the 6th Global WordNet Conference (GWC 2012)*, Matsue, Japan. Global Wordnet Association (GWA).

Robert M. W. Dixon. 2004. Adjective classes in typological perspective. In Robert M. W. Dixon and Alexandra Aikhenvald, editors, *Adjective Classes: A Cross-Linguistic Typoogy*, pages 1–49. Oxford University Press.

Sabri Elkateb, William Black, Horacio Rodríguez, Musa Alkhalifa, Piek Vossen, Adam Pease, and Christiane Fellbaum. 2006. Building a WordNet for Arabic. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC 2006)*, pages 29–34, Genoa, Italy. European Language Resources Association (ELRA).

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. The MIT Press.

Christiane Fellbaum and Piek Vossen. 2012. Challenges for a multilingual wordnet. *Language Resources and Evaluation*, 46(2):313–326.

Darja Fišer and Tomaz Erjavec. 2010. sloWNet: construction and corpus annotation. In *Proceedings of Fifth International Conference of the Global WordNet Association (GWC 2010)*.

Gil Francopoulo, Monte George, Nicoletta Calzolari, Monica Monachini, Nuria Bel, Mandy Pet, and Claudia Soria. 2006. Lexical Markup Framework (LMF). In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC 2006)*, Genoa, Italy. European Language Resources Association (ELRA).

Radovan Garabík and Indrė Pileckytė. 2013. From multilingual dictionary to Lithuanian Wordnet. In *Proceedings of the 7th Internation Conference for Natural Language Processing, Corpus Linguistics, and E-learning (SLOVKO 2013)*, pages 74–80.

Fausto Giunchiglia, Khuyagbaatar Batsuren, and Abed Alhakim Freihat. 2018. One world — seven thousand languages. In *Proceedings of the 7th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing 2018)*, pages 220–235, Hanoi, Vietnam. Springer.

Fausto Giunchiglia, Khuyagbaatar Batsuren, and Gabor Bella. 2017. Understanding and exploiting language diversity. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI 2017)*, pages 4009–4017, Melbourne, Australia. International Joint Conferences on Artificial Intelligence (IJCAI) Organization.

Fausto Giunchiglia, Gábor Bella, Nandu C. Nair, Yang Chi, and Hao Xu. 2023. Representing interlingual meaning in lexical databases. *Artificial Intelligence Review*, 56(10):11053–11069.

Aitor Gonzalez-Agirre, Egoitz Laparra, and German Rigau. 2012. Multilingual Central Repository version 3.0. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012)*, pages 2525–2529, Istanbul, Turkey. European Language Resources Association (ELRA).

Michael Wayne Goodman and Francis Bond. 2021. Intrinsically interlingual: The Wn Python library for Wordnets. In *Proceedings of the 11th Global Wordnet Conference (GWC 2021)*, pages 100–107, Pretoria, South Africa. Global Wordnet Association (GWA).

Patrick Hanks. 1987. Definitions and explanations. In John Sinclair, editor, *Looking up: An account of the COBUILD project in lexical computing*. Collins.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Chu-Ren Huang, Shu-Kai Hsieh, Jia-Fei Hong, Yun-Zhu Chen, I-Li Su, Yong-Xiang Chen, and Sheng-Wei Huang. 2010. Chinese WordNet: Design and implementation of a cross-lingual knowledge processing infrastructure. *Journal of Chinese Information Processing*, 24(2):14–23.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. 2017. OpenNMT: Open-source toolkit for neural machine translation. In *Proceedings of ACL 2017, System Demonstrations*, pages 67–72, Vancouver, Canada. Association for Computational Linguistics (ACL).

Frantisek Kratochvil and Luís Morgado da Costa. 2022. Abui Wordnet: Using a toolbox dictionary to develop a wordnet for a low-resource language. In *Proceedings of the First Workshop on NLP Applications to Field Linguistics*, pages 54–63, Gyeongju, Republic of Korea. International Conference on Computational Linguistics.

Krister Lindén and Lauri Carlson. 2010. FinnWordNet — WordNet påfinska via översättning. *LexicoNordica — Nordic Journal of Lexicography*, 17:119–140.

John P. McCrae, Michael Wayne Goodman, Francis Bond, Alexandre Rademaker, Ewa Rudnicka, and Luis Morgado Da Costa. 2021. The Global Wordnet formats: Updates for 2020. In *Proceedings of the 11th Global Wordnet Conference (GWC 2021)*, pages 91–99, Pretoria, South Africa. Global Wordnet Association (GWA).

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)*, pages 245–252, Wroclaw, Poland. Global Wordnet Association (GWA).

George A. Miller. 1986. Dictionaries in the mind. *Language and Cognitive Processes*, 1(3):171–185.

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

George A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J. Miller. 1990. Introduction to WordNet: An on-line lexical database*. *International Journal of Lexicography*, 3(4):235–244.

Roberto Navigli and Simone Paolo Ponzetto. 2010. BabelNet: Building a very large multilingual semantic network. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL 2010)*, pages 216–225, Uppsala, Sweden. Association for Computational Linguistics (ACL).

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal Dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the Twelfth International Conference on Language Resources and Evaluation (LREC 2020)*, pages 4034–4043, Marseille, France. European Language Resources Association (ELRA).

Joakim Nivre, Daniel Zeman, Filip Ginter, and Francis Tyers. 2017. Universal Dependencies. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2017)*, Valencia, Spain. Association for Computational Linguistics (ACL).

Nurril Hirfana Bte Mohamed Noor, Suerya Sapuan, and Francis Bond. 2011. Creating the Open Wordnet Bahasa. In *Proceedings of the 25th Pacific Asia Conference on Language, Information and Computation (PACLIC 25)*, pages 255–264, Singapore. Institute of Digital Enhancement of Cognitive Processing, Waseda University.

Antoni Oliver, Krešimir Šojat, and Matea Srebačić. 2015. Enlarging the Croatian WordNet with WN-Toolkit and Cro-Deriv. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, pages 480–487, Hissar, Bulgaria. INCOMA Ltd.

Ján Genči Ondrej Dzurjuv and Radovan Garabík. 2011. Generating sets of synonyms between languages. In *Natural Language Processing, Multilinguality. Proceedings of the SLOVKO 2011 Conference*.

Noam Ordan and Shuly Wintner. 2007. Hebrew WordNet: A test case of aligning lexical databases across languages. *International Journal of Translation*, 19(1):39–58.

Peteris Paikens, Agute Klints, Ilze Lokmane, Lauma Pretkalniņa, Laura Rituma, Madara Stāde, and Laine Strankale. 2023. Latvian WordNet. In *Proceedings of the 12th Global Wordnet Conference (GWC 2023)*, pages 187–196, San Sebastian, Basque Country. Global Wordnet Association (GWA).

Bolette Sandford Pedersen, Sanni Nimb, Jørg Asmussen, Nicolai Hartvig Sørensen, Lars Trap-Jensen, and Henrik Lorentzen. 2009. DanNet — the challenge of compiling a WordNet for Danish by reusing a monolingual dictionary. *Language Resources and Evaluation*, 43(3):269–299.

Emanuele Pianta, Luisa Bentivogli, and Christian Girardi. 2002. MultiWordNet: Developing an aligned multilingual database. In *Proceedings of the First International WordNet Conference*, pages 293–302, Mysore, India. Global Wordnet Association (GWA).

Maciej Piasecki, Stan Szpakowicz, and Bartosz Broda. 2009. *A Wordnet from the Ground Up*. Wroclaw University of Technology Press.

Marten Postma, Emiel van Miltenburg, Roxane Segers, Anneleen Schoen, and Piek Vossen. 2016. Open Dutch WordNet. In *Proceedings of the 8th Global WordNet Conference (GWC 2016)*, Bucharest, Romania. Global Wordnet Association (GWA).

Adriana Roventini, Antonietta Alonge, Nicoletta Calzolari, Bernardo Magnini, and Francesca Bertagna. 2000. ItalWordNet: a large semantic database for Italian. In *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC 2000)*, Athens, Greece. European Language Resources Association (ELRA).

Ervin Ruci. 2008. On the current state of Albanet and related applications. Technical report, University of Vlora.

Benoît Sagot and Darja Fišer. 2008. Building a free French WordNet from multilingual resources. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC 2008)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Ferdinand de Saussure. 1916. *Cours de Linguistique Générale*. Payot, Paris.

Melanie Siegel and Francis Bond. 2021. OdeNet: Compiling a GermanWordNet from other resources. In *Proceedings of the 11th Global Wordnet Conference (GWC 2021)*, pages 192–198, Pretoria, South Africa. Global Wordnet Association (GWA).

Kiril Simov and Petya Osenova. 2023. Recent developments in BTB-WordNet. In *Proceedings of the 12th Global Wordnet Conference (GWC 2023)*, pages 220–227, San Sebastian, Basque Country. Global Wordnet Association (GWA).

Joanna Ut-Seong Sio and Luis Morgado Da Costa. 2019. Building the Cantonese Wordnet. In *Proceedings of the 10th Global Wordnet Conference (GWC 2019)*, pages 206–215, Wroclaw, Poland. Global Wordnet Association (GWA).

Sareewan Thoongsup, Thatsanee Charoenporn, Kergrit Robkop, Tan Sinthurahat, Chumpol Mokarat, Virach Sornlertlamvanich, and Hitoshi Isahara. 2009. Thai WordNet construction. In *Proceedings of the 7th Workshop on Asian Language Resources*, pages 139–144, Singapore. Association for Computational Linguistics (ACL).

Dan Tufiș and Verginica Barbu Mititelu. 2014. *The Lexical Ontology for Romanian*, volume 48 of *Text, Speech and Language Technology*, pages 491–504. Springer.

Piek Vossen, editor. 1998. *EuroWordNet: A multilingual database with lexical semantic networks*. Kluwer Academic Publishers.

Piek Vossen, Francis Bond, and John McCrae. 2016. Toward a truly multilingual GlobalWordnet grid. In *Proceedings of the 8th Global WordNet Conference (GWC 2016)*, pages 424–431, Bucharest, Romania. Global Wordnet Association (GWA).

Piek Vossen, Claudia Soria, and Monica Monachini. 2013. Wordnet-LMF: A standard representation for multilingual Wordnets. In Gil Francopoulo, editor, *LMF: Lexical Markup Framework*, chapter 4, pages 51–66. John Wiley & Sons, Ltd.

10. Language Resource References

The language resources below are the resources from which all Cygnet data is sourced.

Sina Ahmadi and others. 2014. *KurdNet*. PID https://github.com/sinaahmadi/kurdnet.

Francis Bond and others. 2013. *Open Multilingual Wordnet*. Global Wordnet Association. PID https://omwn.org/.

Francis Bond and others. 2020. *TUFS Basic Vocab to OMW*. PID https://github.com/omwn/tufs.

Valeria de Paiva and others. 2012. *OpenWordnet-PT*. PID https://github.com/own-pt/openWordnet-PT/.

Frantisek Kratochvil and others. 2022. *Abui Wordnet*. PID https://github.com/fanacek/abuiwn.

John P. McCrae and others. 2019. *Open English Wordnet*. Global Wordnet Association. PID https://en-word.net/.

Peteris Paikens and others. 2023. *Latvian WordNet*. PID https://wordnet.ailab.lv/.

Melanie Siegel and others. 2021. *OdeNet*. PID https://github.com/hdaSprachtechnologie/odenet.

Joanna Ut-Seong Sio and Luis Morgado da Costa. 2019. *Cantonese Wordnet*. PID https://github.com/lmorgadodacosta/CantoneseWN.

A. Example Processing Details

Table 6 shows the number of example sentences kept or discarded for each wordnets that feature examples. For the majority, only a small number were discarded. The anomaly is the Japanese Wordnet, for which many examples were discarded. This is because its examples are direct translations of the English examples from Princeton WordNet, and do not necessarily include the relevant Japanese senses. Of the examples discarded for other languages, some were due to limitations with the parsing approach we use. Other cases were sentences that included a word to evoke a concept which did not appear in the synset, but arguably could be added to the synset in future versions. For example, the synset $\{proper, right\}$ in the Open English Wordnet, defined as "appropriate for a condition or purpose or occasion or a person's character", is connected to the example sentence "she is not suitable for the position". The word *suitable* expresses the target concept. Although it is not included in the synset, *suitable* could reasonably be added. Logs with detail of the examples that were excluded from Cygnet will be passed to the developers of the upstream wordnets.

Wordnet	# Included	# Excluded
Albanet	3,320	4
BulTreeBank Wordnet	14,473	0
Japanese Wordnet	15,967	32,309
Latvian WordNet	72,598	0
Multilingual Central Repo. (Basque)	2,374	1
Multilingual Central Repo. (Catalan)	2,880	111
Multilingual Central Repo. (Galician)	4,496	0
Multilingual Central Repo. (Spanish)	1,145	51
MultiWordNet	1,839	116
Open English Wordnet	49,308	288
Offenes Deutsches WordNet	683	39
OpenWordnet-PT	2,234	587
sloWNet	3,970	192
TUFS (Arabic)	7,479	0
TUFS (Assamese)	2,013	0
TUFS (Burmese)	1,206	0
TUFS (English)	736	2
TUFS (Filipino)	778	0
TUFS (French)	1,637	70
TUFS (German)	1,137	47
TUFS (Indonesian)	1,014	0
TUFS (Japanese)	1,170	36
TUFS (Khmer)	1,015	0
TUFS (Korean)	1,291	407
TUFA (Lao)	1,219	2
TUFS (Malaysian)	1,206	0
TUFS (Mandarin Chinese)	848	0
TUFS (Mongolian)	795	0
TUFS (Portuguese)	651	92
TUFS (Russian)	757	38
TUFS (Spanish)	1,437	113
TUFS (Thai)	745	0
TUFS (Turkish)	1,717	0
TUFS (Urdu)	718	0
TUFS (Vietnamese)	1,236	0

Table 6: Statistics of example extraction

B. Additional Details of Included Wordnets

Table 7 contains information about the licenses of the different wordnets included in Cygnet, as well as the statistics of data included in Cygnet from each wordnet.

Language	Wordnet	# Concepts	# Senses	# Lexemes	Licence
Abui	Abui Wordnet	1,474	3,606	1,476	CC BY
Albanian	Albanet	4,675	9,599	6,489	CC BY 3.0
Arabic	Arabic WordNet	9,916	37,335	18,000	CC BY-SA 3.0
	TUFS	603	868	589	CC BY 4.0
Assamese	TUFS	501	636	403	CC BY 4.0
Basque	Multilingual Central Repo.	29,414	48,933	26,388	CC BY 3.0
Bulgarian	BulTreeBank Wordnet	4,959	8,936	6,737	CC BY 3.0
Burmese	TUFS	578	1,043	684	CC BY 4.0
Cantonese	Cantonese Wordnet	5,110	16,295	11,672	CC BY 4.0
Catalan	Multilingual Central Repo.	60,462	100,120	69,301	CC BY 3.0
Croatian	Croatian Wordnet	23,115	47,890	29,081	CC BY 3.0
Danish	DanNet	4,476	5,859	4,521	WordNet
Dutch	Open Dutch WordNet	30,177	60,259	43,667	CC BY-SA 4.0
English	Open English Wordnet	107,519	185,129	135,911	CC BY 4.0
	TUFS	259	311	471	CC BY 4.0
Filipino	TUFS	642	717	475	CC BY 4.0
Finnish	FinnWordNet	116,763	189,226	130,741	CC BY 3.0
French	Wordnet Libre du Français	59,091	102,647	59,612	CeCILL-C
	TUFS	288	485	587	CC BY 4.0
Galician	Multilingual Central Repo.	34,769	53,120	40,873	CC BY 3.0
German	Offenes Deutsches WordNet	19,717	144,440	118,530	CC BY-SA 4.0
	TUFS	480	547	467	CC BY 4.0
Greek	Greek Wordnet	18,049	24,106	18,263	Apache 2.0
Hebrew	Hebrew Wordnet	5,123	6,543	5,374	WordNet
Icelandic	IceWordNet	4,951	15,897	11,573	CC BY 3.0
Indonesian	Wordnet Bahasa	38,085	106,688	41,478	MIT
	TUFS	276	385	542	CC BY 4.0
Italian	MultiWordNet	34,756	62,125	43,004	CC BY 3.0
	ItalWordNet	7,875	12,163	19,680	ODC-BY
Japanese	Japanese Wordnet	57,184	158,069	94,002	WordNet
	TUFS	194	243	440	CC BY 4.0
Khmer	TUFS	642	718	464	CC BY 4.0
Korean	TUFS	642	739	494	CC BY 4.0
Kurdish	KurdNet	2,144	6,240	3,885	CC BY-SA 4.0
Lao	TUFS	636	712	469	CC BY 4.0
Latvian	Latvian WordNet	8,623	16,980	12,085	CC BY 4.0
Lithuanian	Lithuanian WordNet	9,462	16,032	11,428	CC BY-SA 3.0
Standard Malay	Wordnet Bahasa	36,911	105,028	38,755	MIT
	TUFS	269	347	438	CC BY 4.0
Mandarin Chinese	Chinese Open Wordnet	42,300	79,797	63,339	WordNet
	TUFS	615	704	480	CC BY 4.0
Mongolian	TUFS	620	703	448	CC BY 4.0
Norwegian Bokmål	Norwegian Wordnet	4,455	5,586	4,244	WordNet
Norwegian Nynorsk	Norwegian Wordnet	3,671	4,762	3,436	WordNet
Polish	plWordNet	33,826	52,378	45,458	WordNet
Portuguese	OpenWordnet-PT	52,670	83,762	59,211	CC BY 4.0
	TUFS	252	317	478	CC BY 4.0
Romanian	Romanian Wordnet	56,026	84,638	52,600	CC BY-SA
Russian	TUFS	637	711	465	CC BY 4.0
Slovak	Slovak WordNet	18,507	44,029	29,228	CC BY-SA 3.0
Slovenian	sloWNet	42,583	70,945	40,340	CC BY-SA 3.0
Spanish	Multilingual Central Repo.	78,417	145,641	93,834	CC BY 3.0
	TUFS	217	276	466	CC BY 4.0
Swedish	WordNet-SALDO	6,796	6,904	5,872	CC BY 3.0
Thai	Thai Wordnet	73,350	95,517	83,481	WordNet
	TUFS	327	393	477	CC BY 4.0
Turkish	TUFS	629	720	471	CC BY 4.0
Urdu	TUFS	642	710	470	CC BY 4.0
Vietnamese	TUFS	641	768	510	CC BY 4.0

Table 7: Wordnet resources in Cygnet