Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 3189–3197
Marseille, 11–16 May 2020

Some Issues with Building a Multilingual Wordnet

Francis Bond[♠], **Luis Morgado da Costa**[♠], **Michael Wayne Goodman**[♠], **John P. McCrae**[♢], **Ahti Lohk**[♣]
[♠] Nanyang Technological University (NTU), Singapore
[♢] Data Science Institute, National University of Ireland Galway,
[♣] Tallinn University of Technology, Estonia
bond@ieee.org

Abstract

In this paper we discuss the experience of bringing together over 40 different wordnets. We introduce some extensions to the GWA wordnet LMF format proposed in Vossen et al. (2016) and look at how this new information can be displayed. Notable extensions include: confidence, corpus frequency, orthographic variants, lexicalized and non-lexicalized synsets and lemmas, new parts of speech, and more. Many of these extensions already exist in multiple wordnets – the challenge was to find a compatible representation. To this end, we introduce a new version of the Open Multilingual Wordnet (Bond and Foster, 2013), that integrates a new set of tools that tests the extensions introduced by this new format, while also ensuring the integrity of the Collaborative Interlingual Index (CILI: Bond et al., 2016), avoiding the same new concept to be introduced through multiple projects.

Keywords: multilingual lexicon, wordnet, collaborative development

1. Introduction

This paper provides a summary and update of some of the issues involved with coordinating multiple lexical wordnets. The Princeton WordNet (PWN) is one of the most cited lexical resources in the world with over 1.8 as many citations as the most cited paper in the ACL anthology[1] (Marcus et al. (2004) according to Joseph and Radev (2007)). This success has inspired wordnets in many languages, and many attempts to link them, such as EuroWordnet (Vossen, 1998), the Asian Language Wordnet (Charoenporn et al., 2008) and many more. In this paper, we discuss the approach of the Open Multilingual Wordnet (Bond and Foster, 2013).

In the first version of this project (OMW 1.0), wordnets were linked through the PWN. Although we knew the approach to be fundamentally incorrect (different languages lexicalize different concepts and link them in different ways (Fellbaum and Vossen, 2012)), it was chosen as a first approximation for multiple reasons. The main one was that most existing wordnets are built by translating the PWN: the **extend** model (Vossen, 1998). For example, $\text{dog}_{\text{n:1}}$ is linked to the lemmas *chien* in French, *anjing* in Malay, and so on. The overall structure of PWN serves as a useful scaffold and the fact that, for example, a $\text{dog}_{\text{n:1}}$ is an $\text{animal}_{\text{n:1}}$ is language independent. The main innovations of the OMW 1.0 were an emphasis on open licenses (so that all the data could be legally shared) and a simple shared format (so that resources could be easily converted).

The second version of the OMW (2.0) revived the idea of the InterLingual Index (ILI: Vossen et al., 1999) with the Collaborative Interlingual Index (CILI: Bond et al., 2016). In this vision, there is a shared set of concepts, which the wordnets agree to link to. In the Collaborative ILI, new wordnet projects can propose candidate ILI concepts, instantiated by synsets in the wordnet for that language. In this paper, we introduce other information added to the Open-Multilingual Wordnet and the motivation for it.

The structure of the paper is as follows: in § 2 we talk about challenges in the process of integrating the wordnets; in § 3 we look at how the wordnet format has been extended; in § 4 we look at how to ensure new concepts are not introduced multiple times; and finally, in § 5 we summarize the work in this paper.

2. Challenges for Wordnet Integration

In the first version of OMW, wordnets were mainly converted to a common format by the OMW maintainers, with a few projects giving us the data already formatted (Bond and Foster, 2013). As part of the conversion process, we ended up informally validating the wordnet structure: if we could not parse it, then we could not convert it. When we found errors, we fixed them locally and also sent bug reports upstream. However, this did not scale well, as sometimes we received notice of a new wordnet, and it took us months to find time to write a converter. Thus, we decided to try to shift this burden, as far as possible, to the individual wordnet projects. This was done in consultation with the Global Wordnet Association, with meetings at the Global Wordnet Conferences, discussions over email and through GitHub issues (Bond et al., 2016).[2]

This has largely been successful, with over 30 wordnets uploaded to the new interface. However, there were some issues along the way, which we detail below.

2.1. New OMW/CILI Interface

The Open Multilingual Wordnet has undergone a complete redesign of its system, including a new database structure and user interface. Here we discuss the reasons that contributed to this decision and the details of the reimplementation.

2.1.1. Reasons for the Redesign

The original Wordnet LMF (Vossen et al., 2013) was designed as a standardized interoperability format allowing

[1] 14,378 vs 7,839 citations according to Google Scholar (accessed 2019-12-02).

[2] https://globalwordnet.github.io/schemas/

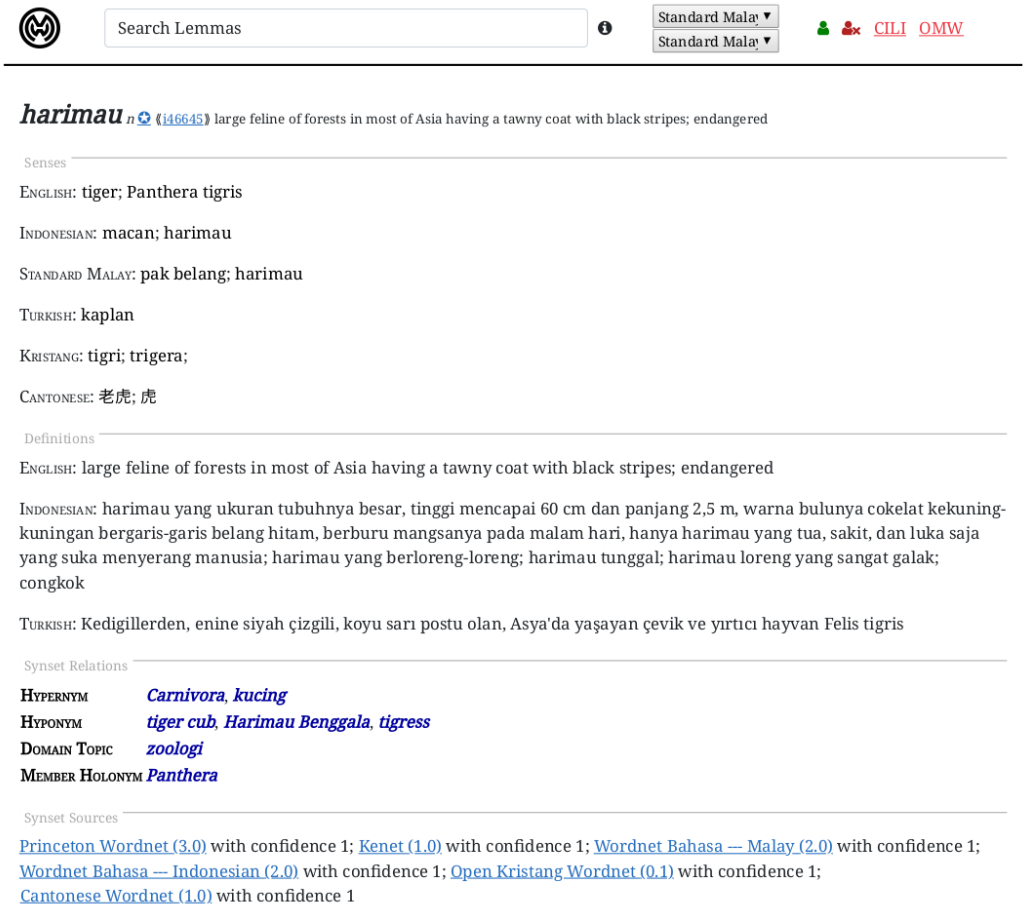

Figure 1: Screenshot of CILI concept i46645, seen through Standard Malay

the interchange of semantic information in the wordnets of the Kyoto Project. The new WN-LMF,[3] extends this format with data structures and layers that the previous OMW system was unable to deal with – some of these changes will be further discussed below (e.g. confidence, corpus frequency and orthographic variants). The main concern of the new OMW system was to be able to read this new format. The WN-LMF offers many flexible ways to encode data and while the OMW is unable to meaningfully read and represent all these new layers of information (i.e. it's an extensible format), the original form is still stored and this can be fixed in the future. The main advantage from moving to this LMF format was the ability to ensure that certain core pieces of information are in place, while giving projects the freedom to encode other layers of information that do not conflict with our system's ability to read each wordnet. In addition, there exist converters[4] from the LMF to OntoLex-Lemon (Cimiano et al., 2016), allowing these resources to be easily published as linked data.

With the release of CILI – an open, language agnostic, flat-structured index that links wordnets across languages without imposing the hierarchy of any single wordnet – the OMW needed to move away from using the Princeton Wordnet as its core, and to restructure itself around CILI. Since the original OMW system was largely committed to PWN as its core structure, the adoption of CILI required major system updates, including changes to database structure and interface design.

The new OMW system is designed to give individual wordnet projects the ability to submit their wordnets online. Then, after a series of automated validation checks, they become instantly visible in the OMW interface. This removes a big overhead cost of maintaining the OMW, since previously a lot of time was spent cleaning and converting individual wordnet projects into OMW compatible structures. This could not only be frustrating when wordnet formats changed across versions, but it also imposed a delay in their release in the OMW. In the new OMW system, wordnet projects can now upload their own wordnets, using the enhanced WN-LMF. The OMW maintainers are still helping individual projects understand and set up their own tools to output wordnets in this format. However, this is a much more manageable goal, and projects receive the immediate gratification of seeing their projects (or newer versions of their old projects) updated almost immediately.

Finally, after the Global Wordnet Association decided to adopt CILI, OMW became the natural place to coordinate it. This means that our system would need to provide spaces that allow other members of the Global Wordnet Association to maintain and coordinate this interlingual index. The creation of new concepts in CILI is also done through the new OMW system, which can be marked in the WN-LMF

[3] https://github.com/globalwordnet/schemas
[4] https://github.com/jmccrae/gwn-scala-api

of a wordnet, as it is uploaded. In addition to a variety of automated checks done concerning the quality of these newly proposed concepts, the OMW also offers a platform to rate, comment and deprecate CILI concepts as needed.

2.1.2. Implementation

The new OMW online system is powered by Python, Flask,[5] JavaScript and SQLite3. It is built using Bootstrap,[6] using asynchrony in some of its design, especially in the validation procedure, and to load different layers of information of displayed elements.

The new system includes a user login system, as well as a dedicated space for registered projects to directly upload their wordnets. It also includes spaces dedicated to CILI management and curation, where registered projects can vote and comment on temporary CILI concepts for a duration of time before they are officially accepted.

Visually, the OMW interface preserves the simple appeal of its original design, where the main focus is synsets. Figure 1 shows a screenshot of CILI concept i46645, using a test version of the OMW interface with only six wordnets uploaded. As mentioned above, this new system departs from the earlier PWN-centric design, in favor of CILI. The top of the concept page shows *harimau* – the label of the concept in the language used for searching (i.e. Standard Malay, in this case) – followed by the part-of speech, the CILI ID, and the synset definition. Should this concept have been found through an English search, the same information would be shown but using *tiger* instead of *harimau*. This is part of the behaviour introduced by concept labels, which will be discussed in greater detail below.

Senses are listed per language, and can be hovered over or clicked to further inspect the structure of that sense. The new OMW system is now able to store more complex layers of information for each lexical entry, which includes a system of forms and tags that will discussed in greater detail in Section 3.4, below.

The rest of the concept page lists definitions and examples in each available language (i.e. these are not strictly required to exist), and the set of semantic relations introduced by all wordnets that include a link to this CILI concept. Since the interpretation of the semantic relations is not always shared across different projects, we have started work on better online definitions for the relations, with the goal being shared documentation available for any wordnet project (or user).[7] This information is linked from the page by clicking on the semantic relation label.

Finally, the concept page ends with a list of all wordnet projects that have contributed information to this OMW/CILI concept.

2.2. Validating Wordnets

In order to upload a wordnet in the new OMW, it must be in an XML format – more specifically, in the WN-LMF format. XML has the advantage of an existing software ecosystem with a wealth of ways to validate a document's schema, i.e., the structure and the legal elements and attributes. For WN-LMF we have extended a Document Type Definition (DTD) inherited from LMF, but we are considering a change to a more modern schema definition system in future work. The DTD can be easily used to check if the XML is well-formed, which can help projects create their wordnets.

However, it turns out there were still many ways in which wordnets can be both well-formed and surprising. Some examples were: wordnets with no synsets, senses and synsets having different parts of speech, synsets with no part of speech, examples or definitions consisting of empty strings, different projects interpreting the directions of relations differently, definitions in unknown languages, projects including the PWN definitions, among others.

Therefore, the new OMW system keeps adding an increasing number of extra checks on the content of uploaded wordnets, which warn the uploader of any problems before any upload is possible. These checks are done in stages, checking a range of aspects from basic to fairly complex.

Some of the basic checks include running the wordnet through DTD validation, checking that the minimum required metadata is in place, or that confidence scores are provided (or inherited) by all entries. Higher stages of these checks include part-of-speech and language consistency across the lexicon, naming conventions for concept ids, and the quality and consistency of the graph built by semantic relations (which will be discussed in greater detail below). The last stage of this validation concerns requirements imposed by the Global Wordnet Association before being able to suggest new concepts to CILI. These include the presence of a unique English definition, and the fact that these new concepts link to the CILI hierarchy through an adequate semantic relation. Figure 2 shows an example of Stage 3 of the validation process of a test wordnet. Passed checks are shown by a tick, warnings are shown by a danger exclamation icon, and failed tests are shown by a cross mark.

This validation process usually does not take more than 15 minutes, but it is highly dependent on the size of a wordnet. Smaller wordnets take less than a couple minutes to be validated. If any of the individual checks fail, the wordnet cannot be uploaded until it is corrected. Figure 2 shows some examples of failed checks on the quality of the semantic hierarchy. On the other hand, if the validation process did not encounter any problems, then the projects will have the option to upload it onto the OMW system, which will make it immediately available on its online interface.

This automated validation effort, though sometimes challenging to set up, has been a great way to catch problems that would otherwise most certainly be overlooked in our previous system.

2.3. Graph Checks

When we attempted to use the OMW for sense disambiguation, it turned out that it had cycles. This led us to check the individual wordnets, and we found that some wordnets (including PWN 3.0!) had ill-formed graph structures such as loops and cycles (Lohk et al., 2019). These are not caught by the XML structure, but make the wordnet graph impossible to manage, so we added checks for these, and pro-

[5] https://flask.palletsprojects.com/en/1.1.x/

[6] `https://getbootstrap.com/`

[7] `https://globalwordnet.github.io/gwadoc/`

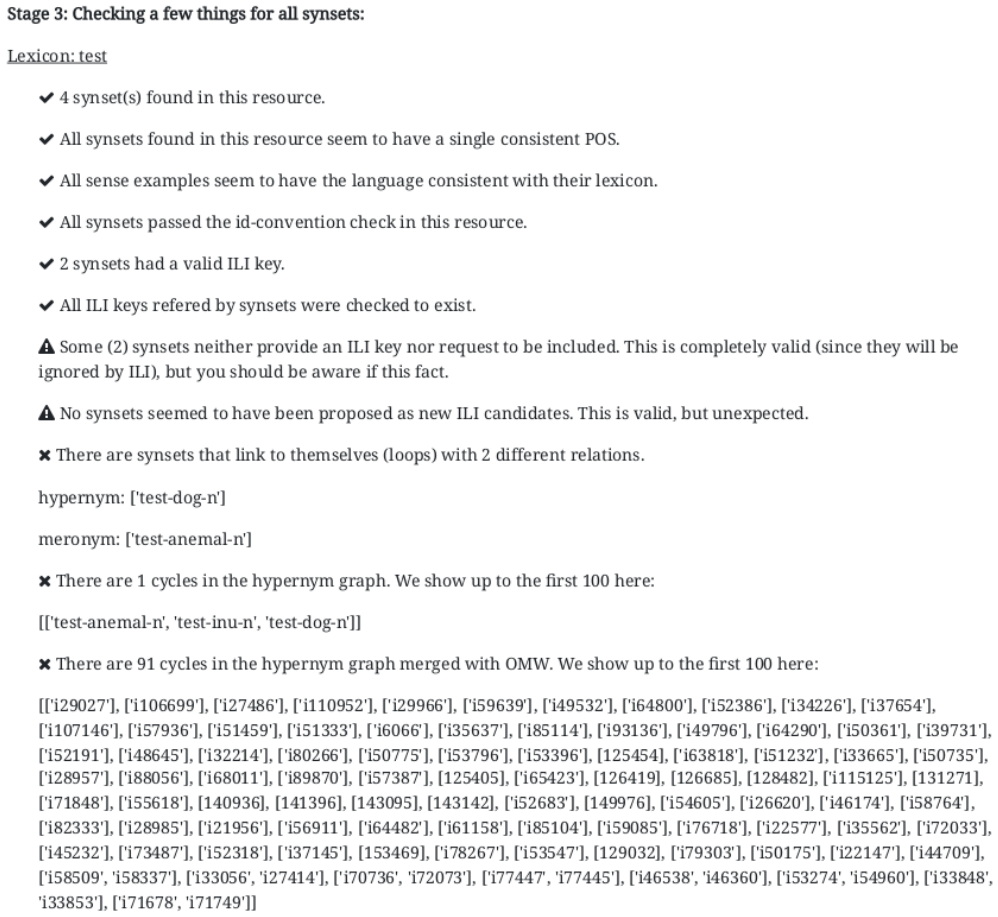

Figure 2: Screenshot of Stage 3 of the Validation Report

vided feedback to all wordnets for which we found errors. These checks are now done after upload as part of the validation, and must be passed for the wordnet to be accepted into OMW 2.0.

We check for three things. The first is loops: does a synset in the new wordnet link to itself (using any semantic relation). The second is cycles in the hypernym graph of the new wordnet: if we claim A is-a B, B is-a C and C is-a A, then this is problematic, both conceptually and when we try to use the graph. These checks follow Lohk et al. (2019). The last check combines the hypernym graph of the new wordnet with the combined hypernym graph of OMW, with nodes linked to the same CILI ID treated as the same. We then check again that no cycles occur. So if one wordnet introduced A is-a B, a second B is-a C and a third C is-a A, we should still catch it. This check cannot be done just with the wordnet on its own, it requires the wordnets to be merged to find the errors. The goal of these checks is to ensure that the final hypernym graph produced by the OMW is a directed acyclic graph.

Finally, we also check for possible duplicate concepts (§ 4), but these are not currently treated as blocking for validation.

3. Extensions for OMW 2.0

In this section we discuss some of the extensions to the representation of the wordnets. Confidence, corpus frequency, orthographic variants and lexicalization are stored in the wordnets themselves. Labels for synsets are computed by the system, as we need labels for synsets that may not appear in any particualr wordnet.

3.1. Confidence

Many wordnets are built using automatically constructed entries as a base, and they are then hand corrected and extended (Bond et al., 2008; Fišer et al., 2012). In this case the wordnet maybe be a mixture of automatically built and hand-checked entries. For systems which use the data from OMW as a basis for learning further multilingual networks (such as BabelNet: Navigli and Ponzetto, 2012), automatically built entries are not suitable training data. We were therefore asked to add **confidence** as a feature. Each major element can have a `ConfidenceScore`. We show an example in Figure 3, where the confidence score for the sense is less than 1.0, suggesting that the entry is not hand checked, but may be crowd-sourced or automatically constructed. de Melo and Weikum (2008) have shown that automatically made wordnets are useful for some tasks, so we do not want to exclude such data altogether, but it is best to distinguish it from fully hand-built data.

In the online interface, the default is to only show high confidence entries (the default is 0.85) and the confidence is used to control the opacity: low confidence entries appear grayed out.

```
<LexicalEntry id="w1">
    <Lemma writtenForm=" 头发" partOfSpeech="n">
        <Tag category="hanzi">Hans</Tag>
    </Lemma>
    <Form  writtenForm=" 頭髮">
        <Tag category="hanzi">Hant</Tag>
    </Form>
    <Form  writtenForm="tóufa">
        <Tag category="romanization">pīnyīn</Tag>
    </Form>
    <Form  writtenForm="tou2fa5">
        <Tag category="romanization">pin1yin1</Tag>
    </Form>
    <Form  writtenForm="toufa">
        <Tag category="romanization">pinyin</Tag>
    </Form>
    <Sense id="example-en-1-n-1" synset="example-en-1-n"
           ConfidenceScore='0.9'>
      <SenseRelation relType="derivation" target="example-en-10161911-n-1"/>
      <Count dc:source='corpus1'>12</Count>
      <Count dc:source='corpus2'>5</Count>
    </Sense>
</LexicalEntry>
```

Figure 3: Example of extended LMF (XML)

3.2. Corpus frequency

A sense can have zero or more counts (see Figure 3), these are stored in OMW as meta information about the sense. When a sense is queried, either to display on its own or as part of a synset, the frequencies are used in various ways.
When displaying lemmas for a synset, they are (i) ordered by frequency and (ii) more frequent synsets are made relatively larger, as can be seen in Figure 4.
(iii) When displaying senses, the frequency is shown for the sense (this can also be seen in the mouse over for a sense). The frequency is also used to choose the label (see below).

3.3. Labels

In a monolingual wordnet, typically each synset has one or more lemmas in the language. In this case, wordnets typcially use the collection of lemmas of a given synset as its label, or possible one lemma chosen at random. However, in the OMW, a synset may be linked to a synset that has no lemmas in the language being investigated. For example, in the example in Figure 1 the synset for ***tiger cub*** has no Malay lemmas, so the English label is used.
We chose a single label for each synset in each language. First, we choose among the lemmas using the following characteristics in order. If there is a tie, we go down to the next level.

1. Choose the most frequent lemma
2. Choose the most unique lemma (across all senses)
3. Choose the shortest lemma
4. Choose the first listed lemma (in the database)

If it has no lemmas then we back off to the English label; if there are no lemmas in English or in the language that was searched then we pick any language that has lemmas.
This gives each synset a label that is the best representative we can manage automatically. Of course, if wordnets rank their lemmas, or come with labels, then we should use the first ranked lemma, but currently this is not represented in the input LMF.

3.4. Orthographic variants

The new WN-LMF used by OMW allows a more detailed organization of a word's orthographic variants. Figure 3 shows a snippet of the WN-LMF for a single lexical entry (`<LexicalEntry id="w1">`). In this example we will use Mandarin Chinese to illustrate how we can make full usage of forms and tags to encode orthographic variants alongside romanizations.
Each `LexicalEntry` element in the WN-LMF expects a single `Lemma`. The writtenForm of the `Lemma` is taken as the canonical orthographical form of this lexical entry – and it will be used throughout OMW as such (i.e. to create labels, etc.). In our case, the canonical form of our lexical entry is 头发, which means "hair". In addition to the required `Lemma`, each lexical entry may contain zero or more elements of the type `Form`. Forms are the way of encapsulating orthographic variation within the new WN-LMF. In our example, forms are being used to provide both different spellings, and romanizations of that lexical entry.
OMW search is done through all `writtenForm` elements within a `LexicalEntry` – which include the `Lemma` and all `Form` elements. The lexical entry shown in Figure 3 contains one true orthographic variant (頭髮, in traditional

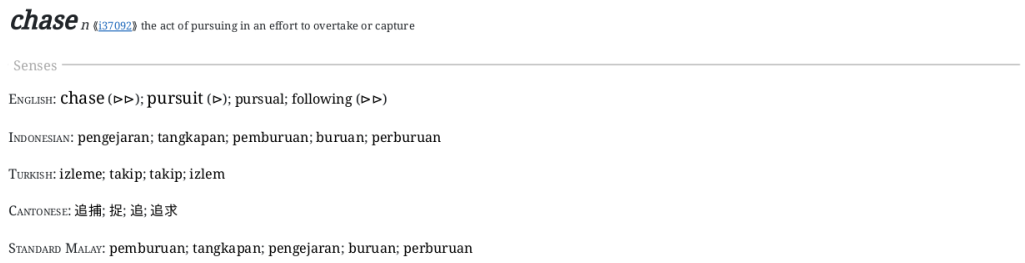

Figure 4: Screen shot of *chase*

Chinese script) and three romanizations – pīnyīn and two simplifications of this romanization system. This means that users are able to find senses linked to this lexical entry through any of the five written forms associated with it. Both lemmas and forms allow the use of the element `Tag`, which can be used to further describe that particular writtenForm. Tags are able to qualify variants using a category and a value. In our example, the canonical form, 头发, has a tag with the category "hanzi", denoting that is encoded in Chinese script, and the value "Hans", which is an official handle to indicate that it is written in simplified Mandarin Chinese. The orthographic variant 頭髮 has a `Tag` element with the same category "hanzi" but value "Hant", which is the official handle indicating that it is written in traditional Mandarin Chinese. Similarly, the three romanizations of this lexical entry all have category defined as "romanization", but slightly different values: pīnyīn, pin1yin1 and pinyin. This is done because the official romanization for Mandarin Chinese, known as pīnyīn, encodes tones using diacritics that are not normally found on standard English keyboards. In order to facilitate search, many dictionaries allow the use a numeric pīnyīn system (e.g. pin1yin1) or a stripped down version where tones are not encoded at all (e.g. pinyin).

With the new WN-LMF, individual projects are invited to use the full structure available in LexicalEntry elements as best as they can. This can include both regional variation – including classic cases such as *color* (US) and *colour* (UK) – and diachronic variation – such as orthographic practices in social media. In addition, this same system can be used to provide romanization systems that can improve the discoverability of each entry.

The flexible nature of this data structure means that we cannot always interpret the full range of information provided in individual wordnet projects. Nevertheless, this information can be safely stored in the WN-LMF. OMW will keep monitoring the type of information individual projects provide, and work on the best ways to display them as they become relevant to our project as a whole.

3.5. Non-lexicalized Synsets

In a multilingual wordnet, if a synset has no lemmas in one language, it is not clear if that is because it cannot be expressed in that language, or just that no lexicographer has worked on it. To deal with this we have an optional attribute `lexicalized` on the synset and sense types, with a boolean value, defaulting to `true`. This was inspired by the MultiWordNet and Basque projects (Pianta et al., 2002; Pociello et al., 2011).

If a synset is marked as `lexicalized=false` then it means it has no lemmas, and this is a deliberate decision on the part of the wordnet builders (but the synset may be included to keep the hierarchy in sync with other projects). For example the synset 冷い ***tsumetai*** "cold to touch" in Japanese would be `lexicalized=false` in English.

If a sense has `lexicalized=true` then it has been validated in some standard lexicon for the language. If it has `lexicalized=false`, then it is believed to be compositional and only added as an aid to multilingual users (similar to **phrase** in multiwordnet). For example *harimau anak* "young tiger" in the Indonesian synset for ***tiger cub*** is `lexicalized=false`, or *dedos pedas* "foot finger-and-toe" in the synset for ***toe*** in Spanish. These allow the lexicographers to put in useful translation equivalents while acknowledging that they are not necessarily part of the language.

Vincze and Almázi (2014) discuss other synsets that may not be lexicalized in Hungarian, such as place names or culturally specific concepts.

4. Duplicate Sense Detection

One difficult challenge with integrating many wordnets of different languages is that they may define identical concepts and as such this would lead to duplicates in the CILI index. As such, we have introduced a system for duplicate sense detection based on the Naisc system introduced by McCrae and Buitelaar (2018). This system is intended for dataset linking and is being specialized for sense linking in the ELEXIS project (Krek et al., 2018). In this case we use it to find duplicate senses based on the definitions given in English in the CILI and provide suggestions to uploaders as to cases where they may have defined a concept that has already been introduced into the CILI by another user.

The Naisc system is based on a multi-stage processing of datasets in order to find the matching links between the elements, this consists of the following steps.

1. **Blocking:** The first step is to analyze which elements of the datasets may potentially match. In this case we are looking for synsets that are similar to the new definition. This is achieved by means of the VGram algorithm (Li et al., 2007). The goal here is to avoid comparing the incoming definition in great detail with every definition already in the CILI as this is computationally expensive and can lead to poor results.

2. **Lens** The next step is to analyse the incoming data and detect the properties that can match. In particular, we extract the lemmas, definitions and examples of the entries. This includes not only the English lemma, definition and example but also any other languages if they are in the language of the wordnet and already in another language in the Open Multilingual WordNet.

3. **Text Similarity** We then apply to each of the candidates a text similarity function as defined in McCrae and Buitelaar (2018). These functions are interchangable and will evolve as we discover better similarity.

4. **Graph Similarity** In addition to analyzing the labels, we also apply link prediction in order to predict the similarity between these new candidate synsets in the graph and the existing synset in the CILI. In order to do this we construct a graph as follows.

    - Collect all links that are proposed between the new synset and other synsets in the submitted wordnet
    - Collect all links between any CILI concepts in any wordnet already in the Open Multilingual WordNet, i.e., the union of all wordnets.
    - For concepts in the submitted wordnet with an ILI already specified create a link from the submitted concept to the corresponding CILI concept

    In this way we create a graph that connects the submitted concept to all concepts in the CILI. We then use this to predict links by calculating the Personalized PageRank (Page et al., 1999) using the FastPPR method (Lofgren et al., 2014).

5. **Scoring** As we have three scores coming from the similarity of lemmas, definitions and examples and one score from the PPR similarity in the graph we need a way to combine them. While a supervised process would be best, we do not yet have many examples of duplicate senses and so have opted for an unsupervised approach. In this case we simply apply a min-max scaling to each of the input features to bring it into the range $[0, 1]$ and then apply an average to output a score.

6. **Matching** For many cases of application of Naisc, matching is a hard task where we try to avoid multiple matches between concepts. However, in this case we only want to suggest potential matches and so we simply return the top percentage of most similar matches detected by the system, by default we return the top 30%.

In practice, this is implemented by means of a subprocedure which is called when a new wordnet is submitted to the CILI. The Naisc system analyzes the uploaded dataset and uses the SQLite database that is used as a backend for Open Multilingual WordNet, it then outputs a JSON document listing the top 30% of similar senses. It is planned that as more training data becomes available we will transfer to a supervised system whereby we can predict those senses that are likely to be duplicates based on past results.

5. Conclusion and Future Work

Collaborative construction of linguistic resources involves cooperation at many levels. In this paper, we have discussed some of the issues involved in building the Open Multilingual Wordnet, and outlined some solutions. One problem with the OMW 1.0 was that the source code was not made open: for OMW 2.0 we have put the code online from the start.[8]

By restructuring the OMW 2.0 around the Collaborative Interlingual Index (CILI) instead of the English-only Princeton WordNet, and by providing a means for individual wordnet projects to propose new concepts for CILI, we have enabled the organic development of truly multilingual resources that can exploit the rich structure of existing wordnets to accelerate their own development without being restricted by the linguistic realities of unrelated languages.

Our adoption of the flexible WN-LMF format and the new web-based submission system whereby wordnet maintainers can upload and immediately validate their own data removes us as the sole gatekeepers and bottleneck in the process, thus handing power back to the maintainers to resolve detected issues in their data, while still allowing us to assist when necessary. The WN-LMF format requires a core layer of information but allows for additional layers of extensions so wordnet maintainers can experiment with new kinds of annotations while still being compatible with the upload system. We recognize a number of extensions already: confidence scores, corpus frequencies, robust synset labels, orthographic variants of lemmas, and non-lexical synsets. Finally, in order to handle the union of many wordnets for many different languages, we have developed a validation "gauntlet" that checks each uploaded wordnet for basic schema violations, annotation inconsistencies, common practical errors, graph structure errors, and duplicated senses.

The adaptation of existing wordnets to the new WN-LMF format, the resolution of their errors detected in validation, and their subsequent inclusion in OMW 2.0 is ongoing work. In future work, we will compile statistics from the uploaded wordnets in OMW 2.0 in order to further document and publicize them. We would also like to offer more flexible uploading methods; rather than only uploading entire wordnets at a time, it would be convenient to upload partial information and updates. For example, corpus frequencies are generated when a new corpus is annotated and not necessarily when a wordnet's structure changes, and corpus annotation may even be performed by a different team from the wordnet's maintainers, so the ability to upload the frequencies without changing the wordnet would simplify the necessary validation and reduce the burden on the maintainer.

We would like to move from using a DTD to validate the XML formatted wordnets to using a more powerful XSD schema. This would allow us to constrain data types more fully.

6. Acknowledgements

We would like to thank the members of the Global Wordnet Association and our three anonymous reviewers for their

[8] `https://github.com/globalwordnet/omw`

many helpful comments. John McCrae's work is funded in part under the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015 (ELEXIS) and in part by an Irish Research Council Laureate Consolidator Award ("Cardamom", IRCLA/2017/129). Ahti Lohk's work is supported by the Estonian Research Council grant PSG227. Michael Wayne Goodman is a research fellow with NTU's Digital Humanities Cluster.

References

Francis Bond and Ryan Foster. 2013. Linking and extending an open multilingual wordnet. In *51st Annual Meeting of the Association for Computational Linguistics: ACL-2013*, pages 1352–1362. Sofia. URL http://aclweb.org/anthology/P13-1133.

Francis Bond, Hitoshi Isahara, Kyoko Kanzaki, and Kiyotaka Uchimoto. 2008. Using multilingual WordNets to compile a Japanese WordNet. In *14th Annual Meeting of the Association for Natural Language Processing*, pages 853–856. Tokyo.

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.

Thatsanee Charoenporn, Virach Sornlerlamvanich, Chumpol Mokarat, and Hitoshi Isahara. 2008. Semi-automatic compilation of Asian WordNet. In *14th Annual Meeting of the Association for Natural Language Processing*, pages 1041–1044. Tokyo.

Philipp Cimiano, John P. McCrae, and Paul Buitelaar. 2016. Lexicon Model for Ontologies: Community Report.

Gerard de Melo and Gerhard Weikum. 2008. On the utility of automatically generated wordnets. In Attila Tanács, Dóra Csendes, Veronika Vincze, Christiane Fellbaum, and Piek Vossen, editors, *4th Global Wordnet Conference: GWC-2008*, pages 146–161. Szeged, Hungary.

Christiane Fellbaum and Piek Vossen. 2012. Challenges for a multilingual wordnet. *Language Resources and Evaluation*, 46(2):313–326. Doi=10.1007/s10579-012-9186-z.

Darja Fišer, Jernej Novak, and Tomaž Erjavec. 2012. sloWNet 3.0: development, extension and cleaning. In *Proceedings of 6th International Global Wordnet Conference (GWC 2012)*, pages 113–117. The Global WordNet Association.

Mark T Joseph and Dragomir R Radev. 2007. Citation analysis, centrality, and the ACL anthology. Technical report, Citeseer.

Simon Krek, John McCrae, Iztok Kosem, Tanja Wissek, Carole Tiberius, Roberto Navigli, and Bolette Sandford Pedersen. 2018. European Lexicographic Infrastructure (ELEXIS). In *Proceedings of the XVIII EURALEX International Congress on Lexicography in Global Contexts*, pages 881–892. URL http://euralex.org/wp-content/themes/euralex/proceedings/Euralex%202018/118-4-2986-1-10-20180820.pdf.

Chen Li, Bin Wang, and Xiaochun Yang. 2007. Vgram: Improving performance of approximate queries on string collections using variable-length grams. In *Proceedings of the 33rd international conference on Very large data bases*, pages 303–314. VLDB Endowment.

Peter A Lofgren, Siddhartha Banerjee, Ashish Goel, and C Seshadhri. 2014. FAST-PPR: scaling personalized pagerank estimation for large graphs. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1436–1445. ACM.

Ahti Lohk, Heili Orav, Kadri Vare, Francis Bond, and Rasmus Vaik. 2019. New polysemy structures in wordnets induced by vertical polysemy. In *Proceedings of the Tenth Global Wordnet Conference*, pages 394–403. CLARIN-PL digital repository. URL http://hdl.handle.net/11321/718.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 2004. Building a large annotated corpus of English: The Penn treebank. In Geoffrey Sampson and Diana McCarthy, editors, *Corpus Linguistics: Readings in a Widening Discipline*, chapter 21, pages 242–257. Continuum.

John P. McCrae and Paul Buitelaar. 2018. Linking Datasets Using Semantic Textual Similarity. *Cybernetics and Information Technologies*, 18(1):109–123. URL http://www.cit.iit.bas.bg/CIT_2018/v-18-1/10_paper.pdf.

Roberto Navigli and Simone Paolo Ponzetto. 2012. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250.

Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. 1999. The PageRank citation ranking: Bringing order to the web. Technical report, Stanford InfoLab.

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

Elisabete Pociello, Eneko Agirre, and Izaskun Aldezabal. 2011. Methodology and construction of the Basque wordnet. *Language Resources and Evaluation*, 45(2):121–142.

Veronika Vincze and Attila Almázi. 2014. Non-lexicalized concepts in wordnets: A case study of English and Hungarian. In *Proceedings of the 7th Global WordNet Conference (GWC 2014)*, pages 118–126. Tartu.

Piek Vossen, editor. 1998. *Euro WordNet*. Kluwer.

Piek Vossen, Francis Bond, and John McCrae. 2016. Toward a truly multilingual global wordnet grid. In *Proceedings of the 8th Global Wordnet Conference (GWC 2016)*. 419–426.

Piek Vossen, Wim Peters, and Julio Gonzalo. 1999. Towards a universal index of meaning. In *Proceedings of ACL-99 Workshop, Siglex-99, Standardizing Lexical Resources*, pages 81–90. Maryland.

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. ISTE Ltd + John Wiley & sons, Inc.