Linking and Disambiguating Swadesh Lists:
Expanding the Open Multilingual Wordnet Using Open Language Resources

Luís Morgado da Costa, Francis Bond, František Kratochvíl
Linguistics and Multilingual Studies,
Nanyang Technological University, Singapore
luis.passos.morgado@gmail.com, bond@ieee.com, fkratochvil@ntu.edu.sg

Abstract

In this paper we describe two main contributions in the fields of lexicography and Linked Open Data: a human corrected disambiguation, using the Princeton Wordnet's sense inventory (PWN, Fellbaum, 1998), of Swadesh lists maintained in the Internet Archive by the Rosetta Project, and the distribution of this data through an expansion of the Open Multilingual Wordnet (OMW, Bond and Foster, 2013). The task of disambiguating word lists isn't always a straightforward task. The PWN is a vast resource with many fine-grained senses, and word lists often fail to help resolve the inherent ambiguity of words. In this work we describe the corner cases of this disambiguation and, when necessary, motivate our choice over other possible senses. We take the results of this work as a great example of the benefits of sharing linguistic data under open licenses, and will continue linking other openly available data. All the data will be released in future OMW releases, and we will encourage the community to contribute in correcting and adding to the data made available.

Keywords: Swadesh lists, wordnet, lexicography, linked open data

1. Introduction

This work describes how we disambiguated and linked a large collection of over 1,200 Swadesh lists maintained by the Rosetta Project.[1] This was done as part of the Open Multilingual Wordnet (OMW, Bond and Foster, 2013), a linked collection of wordnets of multiple languages released under an open license, which includes the Princeton Wordnet's sense inventory (PWN, Fellbaum, 1998) extended with pronouns, determiners, interjections and classifiers (see Seah and Bond, 2014; Morgado da Costa and Bond, 2016).

We start by discussing the origin of Swadesh List and its multiple versions, as well as the intrinsic problems of disambiguating word lists. We discuss corner-cases of our disambiguation, and highlight the importance of working with disambiguated lists in research involving elicitation.

We introduce a new interface to OMW, designed to browse the OMW using lists and allowing users to use the data we collected in new and interesting ways. We also introduce the possibility of creating custom multilingual lists, and enjoy the benefits that come from using linked open linguistic data (i.e. senses and definitions in multiple languages).

With the exception of six prepositions and conjunctions, every word in the widely used Swadesh 207 list was mapped to a concept in the PWN, along with 72 other concepts that were spread across the many variant lists shared by the Rosetta Project. We started from an initial mapping provided by Huang et al. (2007), which was corrected and enhanced where necessary. Ultimately, this work produced a new extended version of the OMW, linking more than 270,000 new unique senses and raising the coverage of this resource from 150 to over 1,200 languages.

We commit to release the linked disambiguated Swadesh sense inventory under an open license, as part of OMW. This allows online search, downloads in a fixed format, and manipulation through the python Natural Language Toolkit (NLTK: Bird et al., 2009).[2] This inventory can continue to be used in the same way as it once was, but brings the benefits of being defined in multiple languages, and being linked to hundreds of other languages through OMW.

2. Swadesh Lists

The Swadesh List is a classic compilation of words that change at a relatively constant rate, used in comparative linguistics studies to predict language relatedness and history by tracing the retention and relative change of vocabulary among languages (Swadesh, 1952). For this reason, words in Swadesh lists are supposed to be universal, but not necessarily the most frequent. The list has seen several versions/revisions through the years of lexico-statistics work of Swadesh (1952, 1955). Each of these versions became a list in its own right, with sizes ranging from 100 to 215 words. Through the continuous revision of these words-lists, Swadesh hoped to pinpoint a list of fundamental everyday vocabulary, present in every language, as opposed to a specialized or "cultural" vocabulary. And even though Swadesh's initial intentions were to enlarge this universal vocabulary, he acknowledged that the compilation of such list should avoid problems such as potential duplication, identical roots, sound imitation and semantic shading (ambiguity). In the end, the multiple revisions of his list kept getting shorter, until reaching 100 words.

The history of the Swadesh lists can be summed as follows: in 1952 Swadesh proposes his 200 word list (see Annex A), a selected extract from a 215 word list used in his earlier work. In this version, he tries to specify the intended meaning of these words with the use of parenthetic notes. In 1955, Swadesh publishes the original 215 word list which was used to create the 200 word list, grouping them in 23 semantic groups (see Annex B). In the same publication, Swadesh proposes his final list reduction, this time to contain only 100 words: 92 words selected from the original

[1] http://rosettaproject.org

[2] http://www.nltk.org/

Min. No. Words	No. Languages	%
1	1211	1.000
50	1088	0.898
100	885	0.731
150	727	0.600
200	553	0.457
300	334	0.276
400	155	0.128
600	56	0.046
800	21	0.017
1000	9	0.007
2000	1	0.001

Table 1: Number of words per number of languages

215 list, and 8 new words (see Annex C). Finally, the widely used (non-official) 207 word Swadesh list (see Annex D) contains the 200 terms proposed in 1952, with the addition of 7 of the 8 new terms proposed (all except *claw*) in his final 100 word list (Huang et al., 2007).

Swadesh lists have been used in the fields of lexical-statistics and historical-comparative linguistics, from their conception until recent times. In early days, the popularity of Swadesh's work propelled his lists into a de facto standard data-set to be collected in language description. Consequently, the large amount of collected data was eventually compiled into comparative vocabulary databases, for a number of language families. We can find examples of this in the Indo-European Lexical Cognacy Database[3] and the Austronesian Basic Vocabulary Database (Greenhill et al., 2008). These came to be the standard data-sets for computational phylogenetic language change models (see, for example, Bouckaert et al., 2012).

Until today, multiple studies continue to use and produce Swadesh-like lists as seed data to study cognates and relatedness between languages, as well as to trace language history and evolution (see Serva and Petroni, 2008; Wu et al., 2015; Pagel et al., 2013). Holman et al. (2008) introduce a fundamental study, where an automatic model was used to calculate the relative stability of the items in the 100 word Swadesh list published in 1955. And they show that the 40 most stable items on the final Swadesh list are as effective in language classification models as the full 100 word list.

3. Data

All the data collected for the work presented in this paper is readily available in the Internet Archive,[4] and was commissioned and owned by the Rosetta Project. This project is run by the Long Now Foundation, and it is a global effort, open to language specialists and native speakers, to build a publicly accessible digital library of human languages. Among other language documentation initiatives, this project maintains and openly shares a large collection of Swadesh lists in multiple languages.

A total of 1,211 lists, for 1,211 different languages, were downloaded as simple text files, along with an xml file that includes their specific meta data. This meta data includes common information, such as license, authorship, and also the language's code and full name. All lists dealt with are shared under a CC-BY 3.0 (Unported) license.[5]

Here is an excerpt from the list for Abui, an Alor-Pantar language spoken in Eastern Indonesia:

```
...
push: habi
rain: anui
rain: ʔanuy
rain: anúy
rain: anúy
rat: rui
red: arangnabake
red: kiika
red: kiika
red: ki:kɑ
red: kika
ripe: kang
ripe: ma
...
```

As can be seen above, the format of these lists includes an English word and its counterpart in the target language, separated by a colon. Multiple senses can exist for a single English word. And multiple spellings are also provided for some senses. The size of lists varied greatly. Table 1 gives an account of the distribution of list sizes (incl. duplicates). As can be seen in Table 1, all 1,211 lists contained at least one word pair, and 60% of these lists contained at least 150 word pairs. We can see that a relative large portion of these lists include a few hundred pairs, and that a few languages actually included over a thousand. Upon processing and analyzing these lists, we found that duplicates and orthographic variations were quite common, explaining why some lists have a very high number of words. This can be seen above (see, for example, the repetition of the pair *red: kiika*).

Even though the lists collected were named after Morris Swadesh, we have seen in Section 2. how this is a somewhat abstract concept. Swadesh lists often also refer to lists that include words that fall outside any of the original work of Swadesh. And this was often the case for the lists we collected. Many of the lists included English words that were not included in any of the original Swadesh lists.

A closer inspection of these extra words showed that they fell into three rough classes. Most were quite general, such as *today, son, house* and *frog*. There were also many body parts, such as *finger, arm, lip, chin, forehead*. Finally, an interesting set of words was clearly focused on Australian languages, which was made evident by a very specific lexical choice of animals such as *kangaroo, cassowary, wallaby,* and *emu*, which are only found in and around this region.

4. Disambiguation

The original design of PWN includes only contentful/referential open class words: nouns, verbs, adjectives and adverbs. In this work, however, due to the nature of the

[3] http://ielex.mpi.nl/

[4] https://archive.org/

[5] https://creativecommons.org/licenses/by/3.0/

task in question, we added two expansions of PWN that include a large set of pronouns, determiners, interjections and classifiers (see Morgado da Costa and Bond, 2016; Seah and Bond, 2014).

After pre-processing the data introduced in Section 3., removing duplicates, we decided to map every English word that had translations in at least 100 languages. While ensuring this, we found that some English words appeared, inconsistently, using multiple forms. For example, the word *fly* appeared also with the form *fly v.* and *fly (v.)*. In cases like this, all words linked to any of these forms were linked to the same concept, in this case 01940403-v – "travel through the air; be airborne".

We started with an initial mapping of the Swadesh 207 word list provided by Huang et al. (2007). We carefully rechecked this initial mapping against the cues provided in the original publications. We tried, as much as possible, to base our choices on the parenthetical notes introduced in Swadesh (1952) and the semantic grouping shown in Swadesh (1955).

Based on these, we enhanced and made a few corrections to the initial mapping provided by Huang et al. (2007). Firstly, using the expansions to PWN's concept inventory, we were now able to map 13 pronouns for which there were no previous mappings. From the remaining data, we made only 13 corrections. We provide three of these as examples:

1. the word *squeeze* had originally been mapped to 00357023-n – "the act of gripping and pressing firmly"; but since this word is presented as *to squeeze* (Swadesh, 1952), we chose instead the verbal concept 01387786-v – "squeeze or press together";

2. the word *day* had originally been mapped to 15155220-n – "time for Earth to make a complete rotation on its axis"; but since there is a parenthetical note stating "opposite of night rather than the time measure" (Swadesh, 1952), we corrected it to 15164957-n – "the time after sunrise and before sunset while it is light outside";

3. the word *louse* had originally been mapped to 02185481-n – "wingless insect with mouth parts adapted for biting, mostly parasitic on birds"; but since we thought this sense was too specific (i.e. synonym of *bird louse*), we changed the mapping to the more general concept 02183857-n – "wingless usually flattened bloodsucking insect parasitic on warm-blooded animals";

After going through the 207 mappings provided by Huang et al. (2007), we continued to map the remaining words that fell outside this list, which we collapsed into 72 other concepts. For these extra words, since little or no information was provided, we resorted to list cohesiveness and sense frequency in our disambiguation.

Through this effort, more than 270 PWN concepts received senses in at least 100 languages. The end result is an extended OMW, with more than 270,000 new unique senses and coverage for over 1,200 languages. Table 2 shows the distribution of number of concepts per number of languages. In this table we can see that all 1,211 languages received mappings to at lease one concept, and that over 66% of all languages received senses to more than 100 concepts. Only two languages received mappings for more than 250 concepts, these were Orokolo (oro) and Toaripi (tqo), both from Papua New Guinea, which received sense mappings for 251 concepts each.

Min. No. Concepts	No. Languages	%
1	1211	1.000
20	1151	0.950
40	1107	0.914
60	1011	0.835
80	962	0.794
100	806	0.666
120	666	0.550
140	595	0.491
160	501	0.414
180	345	0.285
200	145	0.120
220	63	0.052
240	21	0.017
250	2	0.002

Table 2: Number of concepts per number of languages

4.1. New and Excluded Concepts

Unfortunately, even considering an extended concept inventory from the expansion efforts mentioned above (see Section 4.), it was still insufficient to provide a complete mapping for every word. Three classes of words deserve to be mentioned here: pronouns, prepositions and conjunctions.

Pronouns were first introduced to wordnets by Seah and Bond (2014), where many pronouns were introduced and marked for a number of semantic features including, for example, number, gender and politeness.

Nevertheless, while going through the word list that extended the Swadesh lists, we found occurrences for six pronouns that had not yet been accounted for. Namely, genderless third person pronouns (listed as *he/she*), dual first person pronouns (listed as *we two*), along with their inclusive and exclusive counterparts (listed as *we two (incl.)* and *we two (excl.)*), dual second person pronouns (listed as *you two*), and dual third person pronouns (listed as *they two*).

Following the same method described in Seah and Bond (2014), we added the six missing concepts to the OMW hierarchy, and linked these missing pronouns.

Concerning prepositions and conjunctions, we find a similar situation – i.e. there are no prepositions or conjunctions in the PWN to be able to map these words. But, in this case, we know of no effort done to expand wordnet inventories in this way, and we therefore excluded these two classes of words from this work.

4.2. The Problem of Ambiguity

As it has been mentioned before, disambiguating word lists isn't a straightforward task, especially if the word lists provide little or no information that can be used to disambiguate them. Adding to this difficulty, the PWN is a vast resource

synset	lemmas	definition
00608372-v	**know**	perceive as familiar
00608502-v	**know**	be able to distinguish, recognize as being different
00595935-v	**know**	know how to do or perform something
00608670-v	**know**	know the nature or character of
00592883-v	recognize, **know**, acknowledge, recognise, today, ...	accept (someone) to be what is claimed or accept his power and authority
00594337-v	**know**	be familiar or acquainted with a person or an object
00596644-v	**know**, experience, live	have firsthand knowledge of states, situations, emotions, or sensations
00595630-v	**know**	be aware of the truth of something; have a belief or faith in something; regard as true beyond any doubt
00594621-v	**know**, cognize, cognise	be cognizant or aware of a fact or a specific piece of information; possess knowledge or information about
00596132-v	**know**	have fixed in the mind

Table 3: PWN's partial sense inventory for verbal concepts matching the lemma *know*

with many fine-grained senses. In this section we would like to highlight the difficulty of this task by describing a few corner cases.

Firstly, concerning pronouns, we would like to point out that we aware that many pronouns may not be linked correctly. The reason for this comes from the rich pronominal hierarchy that was created when adding pronouns to wordnet (Seah and Bond, 2014). This pronominal hierarchy makes use of semantic features to split pronouns in multiple concepts, depending on features like number and gender, but also politeness, formality and gender speech. We will further exemplify this problem with current situation of the first person singular pronoun in English and Japanese.

In English, the concept for the pronoun *I* is marked only for three features: *first_person, personal_pronoun, singular*. But in Japanese, the same pronoun is split in multiple concepts. We can find a concept for わたし *watashi* marked for *first_person, personal_pronoun, singular, formal, polite*; a second concept for われ *ware* marked for *first_person, personal_pronoun, singular, formal*; a third for おれ *ore* and ぼく *boku* marked for *first_person, personal_pronoun, singular, informal, men's_speech*; another one for わたくし *watakushi* marked for *first_person, personal_pronoun, singular, formal, polite, honorific*, and a few more.

The decision to split concepts by the set of semantic features they are marked for dictates that the English pronoun *I* and the Japanese pronoun わたし *watashi*, for example, are not senses of the same concept. This is simply an example, and other features are also used to further specialize other kinds of pronominal concepts.

Even though explaining the hierarchy and meaning of all these features is well beyond the scope of this work, it is important to note that, because we lack information about these above mentioned features, it is currently hard to pinpoint the correct mapping for pronouns collected. In cases where these features are not available (see the discussion about dual and genderless pronouns above), we decided to map pronouns to their English counterparts. While this will most certainly generate some noise in the mapping, we thought it was preferable to provide a mapping and correct it later than to exclude them.

Similar in spirit, we also felt it was difficult to choose between senses where a very fine grained distinction has been made in PWN. We exemplify this with the mapping of the verb *know*. Table 3 shows a partial sense inventory of PWN for verbal concepts matching the lemma *know*. As can be seen, this is a good example of a too fine-grained distinction of senses. In this case, even after excluding a few less likely choices, we are still invited to make a distinction between the meaning nuances of "familiar or acquainted with", "have firsthand knowledge of", "be aware of the truth of", and "be cognizant or aware of".

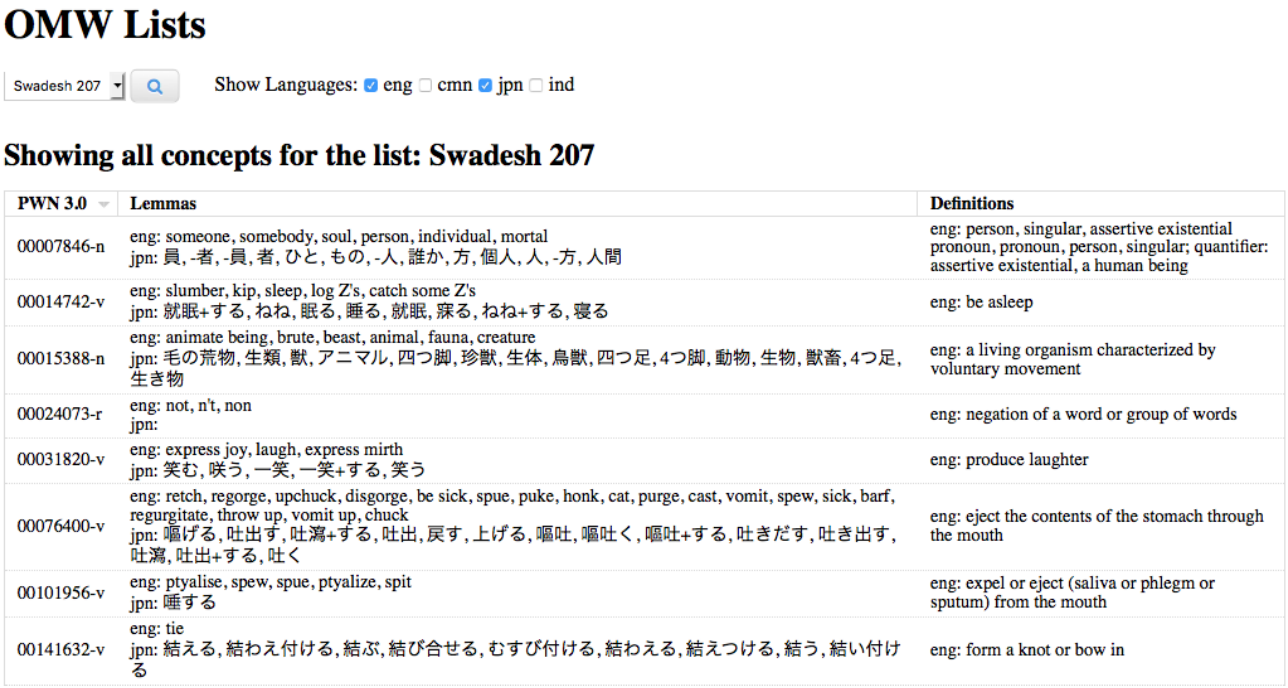
In situations like this, information on sense frequency and being consistent with the previous mapping were favored in our final choice. In this case, the concept 00594621-v – "be cognizant or aware of a fact or a specific piece of information; possess knowledge or information about" had been chosen by Huang et al. (2007), which also happened to be the most frequent sense, so we didn't change it. Out of other hard to disambiguate words, we highlight also *fat, blow, see, think* and *throw*, but many others exist.

Ultimately, what we would like to highlight is the fact that using English words as list keys is insufficient and often problematic, because it does not remove the temptation to define meanings in terms of the conceptualizations that this source language can trigger. By using language-agnostic concept-keys, the source language interference is minimized by the multilingual structure of these resources. In other words, instead of using the lemma *know* as a list key, we suggest using the equivalent, but language-agnostic concept key 00594621-v (as shown in Table 3).

5. Sharing and Visualizing the Data

Beyond the above mentioned commitment to share the processed data in subsequent OMW releases which, in turn, will also be available for manipulation through NLTK, we are also expanding the current OMW interface to allow linguists and researchers from other fields, like psychology or social sciences, to use the data described in this paper, along with the rich data already contained in OMW.

Relevant for this work, we have produced a list browser (see Figure 1), where well known vocabulary lists will be made readily available for browse and download. Currently, we include the four Swadesh lists: commonly referred to



OMW Lists

Swadesh 207

Show Languages: ☑ eng □ cmn ☑ jpn □ ind

Showing all concepts for the list: Swadesh 207

PWN 3.0	Lemmas	Definitions
00007846-n	eng: someone, somebody, soul, person, individual, mortal jpn: 員, -者, -員, 者, ひと, もの, -人, 誰か, 方, 個人, 人, -方, 人間	eng: person, singular, assertive existential pronoun, pronoun, person, singular; quantifier: assertive existential, a human being
00014742-v	eng: slumber, kip, sleep, log Z's, catch some Z's jpn: 就眠+する, ねね, 眠る, 睡る, 就眠, 寐る, ねね+する, 寝る	eng: be asleep
00015388-n	eng: animate being, brute, beast, animal, fauna, creature jpn: 毛の荒物, 生類, 獣, アニマル, 四つ脚, 珍獣, 生体, 鳥獣, 四つ足, 4つ脚, 動物, 生物, 獣畜, 4つ足, 生き物	eng: a living organism characterized by voluntary movement
00024073-r	eng: not, n't, non jpn:	eng: negation of a word or group of words
00031820-v	eng: express joy, laugh, express mirth jpn: 笑む, 咲う, 一笑, 一笑+する, 笑う	eng: produce laughter
00076400-v	eng: retch, regorge, upchuck, disgorge, be sick, spue, puke, honk, cat, purge, cast, vomit, spew, sick, barf, regurgitate, throw up, vomit up, chuck jpn: 嘔げる, 吐出す, 吐瀉+する, 吐出, 戻す, 上げる, 嘔吐, 嘔吐く, 嘔吐+する, 吐きだす, 吐き出す, 吐瀉, 吐出+する, 吐く	eng: eject the contents of the stomach through the mouth
00101956-v	eng: ptyalise, spew, spue, ptyalize, spit jpn: 唾する	eng: expel or eject (saliva or phlegm or sputum) from the mouth
00141632-v	eng: tie jpn: 結える, 結わえ付ける, 結ぶ, 結び合せる, むすび付ける, 結わえる, 結えつける, 結う, 結い付ける	eng: form a knot or bow in

Figure 1: Excerpt from the Swadesh 207 list as show in the OMW Lists interface

as "Swadesh 200", "Swadesh 215", the final reduced list known as "Swadesh 100", and the ubiquitous unofficial "Swadesh 207".

This new interface allows the user to select any number of languages and a predefined list, for which a table-like array of data is produced, allowing to compare data across languages. We hope that this interface may help field linguists and other types of works involving elicitation, since lists can be tailor-made with a specific language selection in mind. Using lists produced in this way will guarantee that the data is pre-disambiguated, and can later be merged back and compared against other linked data.

The array of data produced can not only provide lists of lemmas in multiple languages, but also definitions where available. And since an English definition is a requirement to be a part of the OMW, lemmas in any language can always be accompanied with a definition to help disambiguate the respective sense.

This interface also has an option to produce a custom list of concepts (i.e. a list that has not been predefined). And we hope to further enrich this interface with other known lists such as the Sign Language Swadesh List[6] (Woodward, 1993), or the Holman et al. (2008) most stable 40 word list.

6. Conclusion and Future Work

Using open data provided by the Rosetta Stone project, we have been able to link over 270,000 new senses to the Open Multilingual Wordnet. As a consequence of this, we have also greatly expanded the language knowledge this resource, which previously had data for 150 languages, but now contains data for over 1200 languages.

To accomplish this, we have carefully disambiguated and linked over 1200 lists of words based on the work of Morris Swadesh. This disambiguation redefines the English words previously being used as Swadesh keys to a language-agnostic concept key in the Open Multilingual Wordnet.

This work has obvious practical benefits for lexicographic elicitation, setting an example on how sense disambiguated lexicons can, by linking to a language-agnostic concept key, enrich the knowledge we have of world languages. We believe that, to be able to do comparative work in the field of lexical semantics, it is important to control elicitation through an agreed upon sense inventory, as provided here. This kind of linked data, can provide enough resolution to study semantic typology (i.e. word similarity, language families, word loaning, etc.).

We have also shown that wordnets can be expanded through the use of open data. And following this trend, we want to continue linking known lexicographic lists and resources, especially when these can be sense disambiguated. Unfortunately, data in enough quantity is necessary to justify this time-consuming work.

Our next target will be to link the World Loanword Database (WoLD) (Haspelmath and Tadmor, 2009), which provides linked mini-dictionaries (1000-2000 words) for 41 languages, with comprehensive information about the loanword status of each word. This is a well organized project, with well curated data, but disambiguating a much larger list will also have higher costs associated. To link a project of this size, since WoLD also provides definitions, we will most likely look into methods of automatic word sense disambiguation.

Nevertheless, even though the data released by the Rosetta Project is much simpler, and arguably even ill formatted (i.e. spurious repetition, spelling inconsistencies in the English

[6] This list modifies the Swasdesh list in order to study sign languages. In particular, the proportion of indexical signs (body parts and pronouns) was reduced, as they are more likely to be similar.

keys, etc.), we have shown with this work that only a certain amount of coherence and consistency are necessary to make data useful. The Rosetta Project is an excellent example of the benefits that come from crowd-sourced open data, which can be achieved with minimal supervision.

Concerning future work, and following the discussion introduced in Section 4.2., it would be important to do some error analysis on the mapped senses. We hope to do this in two ways. Firstly, we would like to use the multilingual structure of the OMW to automatically check the overlap between existing and newly mapped senses in languages for which we already have data. This will give us a rough estimate of the quality of the mapping, as we expect to have most of the Swadesh senses in human curated projects. We will use the results of this method to provide a confidence score to new senses added to the OMW. A second way to account for the quality of the data will be to encourage lexicographers and native speakers around the world to check, correct and enrich these lists once the data has been published. Following a crowd-sourced schema similar to the one used by Rosetta Project to produce these lists, we hope to ask subscribers of well-known listservers, such as the Linguistlist,[7] for help correcting an enriching this data-set.

A second line of future work will focus on the further development of tools to disseminate and make this data useful for as many people as possible. As it was mentioned in Section 5., enriching the newly created interface with other lists used in research, e.g. linguistics or psychology, can help create a positive feedback of open data. Also, by providing the ability to create and save custom lists, we hope that people can be creative in the way they use these open resources.

Finally, concerning the words that were excluded during this round of linking, we hope to continue the expansion trend of wordnets, and soon include prepositions and conjunctions as two new classes of concepts. And since prepositions are an specially interesting class of word to study crosslingually, our first target will be prepositions. English prepositions are often translated as nouns in Chinese and Japanese: for example *between* is translated as 間 *aida* "space or region between" in Japanese. Towards this end, we hope to build on existing semantic taxonomies for prepositions such as Schneider et al. (2015).

7. Acknowledgements

This research was supported in part by the MOE Tier 2 grant *That's what you meant: a Rich Representation for Manipulation of Meaning* (MOE ARC41/13). We would also like to We would also like to thank Shu-Kai Hsieh, Huang Chu-Ren and Laurent Prevot for providing the initial mapping that was used in this work.

References

Stephen Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python*. O'Reilly. (`www.nltk.org/book`).

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*, pages 1352–1362. ACL, Sofia, Bulgaria.

Remco Bouckaert, Philippe Lemey, Michael Dunn, Simon J Greenhill, Alexander V Alekseyenko, Alexei J Drummond, Russell D Gray, Marc A Suchard, and Quentin D Atkinson. 2012. Mapping the origins and expansion of the indo-european language family. *Science*, 337(6097):957–960.

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

Simon J Greenhill, Robert Blust, and Russell D Gray. 2008. The austronesian basic vocabulary database: from bioinformatics to lexomics. *Evolutionary Bioinformatics*, 4:271.

Martin Haspelmath and Uri Tadmor. 2009. *Loanwords in the world's languages: a comparative handbook*. Walter de Gruyter.

Eric W Holman, Søren Wichmann, Cecil H Brown, Viveka Velupillai, André Müller, and Dik Bakker. 2008. Explorations in automated language classification. *Folia Linguistica*, 42(3-4):331–354.

Chu-Ren Huang, Laurent Prevot, I-Li Su, and Jia-Fei Hong. 2007. Towards a conceptual core for multicultural processing: A multilingual ontology based on the swadesh list. In *Intercultural Collaboration*, pages 17–30. Springer.

Luis Morgado da Costa and Francis Bond. 2016. Wow! What a useful extension! Introducing non-referential concepts to wordnet. In *Proceedings of the International Conference on Language Resources and Evaluation*. Slovenia.

Mark Pagel, Quentin D Atkinson, Andreea S Calude, and Andrew Meade. 2013. Ultraconserved words point to deep language ancestry across eurasia. *Proceedings of the National Academy of Sciences*, 110(21):8471–8476.

Nathan Schneider, Vivek Srikumar, Jena D. Hwang, and Martha Palmer. 2015. A hierarchy with, of, and for preposition supersenses. In *Proceedings of The 9th Linguistic Annotation Workshop*, pages 112–123.

Yu Jie Seah and Francis Bond. 2014. Annotation of pronouns in a multilingual corpus of Mandarin Chinese, English and Japanese. In *10th Joint ACL - ISO Workshop on Interoperable Semantic Annotation*. Reykjavik.

Maurizio Serva and Filippo Petroni. 2008. Indo-european languages tree by levenshtein distance. *EPL (Europhysics Letters)*, 81(6):68005.

Morris Swadesh. 1952. Lexico-statistic dating of prehistoric ethnic contacts: with special reference to north american indians and eskimos. *Proceedings of the American philosophical society*, 96(4):452–463.

[7] `http://linguistlist.org/`

Morris Swadesh. 1955. Towards greater accuracy in lexicostatistic dating. *International journal of American linguistics*, 21(2):121–137.

James Woodward. 1993. Intuitive judgments of Hong Kong signers about the relationship of sign language varieties in Hong Kong and Shanghai. *CUHK Papers in Linguistics*, 4:88–96.

Ren Wu, Yuya Matsuura, and Hiroshi Matsuno. 2015. On generating language family-trees based on basic vocabulary. In *ITC-CSCC: International Technical Conference on Circuits Systems, Computers and Communications*, pages 272–275.

Annex A: The 200 word list proposed by Morris Swadesh in 1952, including parenthetical explanations

I, thou, he, we, ye, they, this, that, here, there, who?, what?, where?, when?, how, not, all, many, some, few, other, one, two, three, four, five, big, long, wide, thick, heavy, small, short, narrow, thin, woman, man (male human), person, child (young person rather than as relationship term), wife, husband, mother, father, animal, fish, bird, dog, louse, snake, worm, tree, woods, stick (of wood), berry (of fruit), seed, leaf, root, bark (of tree), flower, grass, rope, skin (person's), meat (flesh), blood, bone, fat (organic substance), egg, tail, feather (larger feathers rather than down), hair, head, ear, eye, nose, mouth, tooth (front rather than molar), tongue, foot, leg, hand, wing, belly, guts, neck, back (person's), heart, liver, to drink, to eat, to bite, to suck, to spit, to vomit, to blow (of wind), breathe, to laugh, to see, to hear, to know (facts), to think, to smell (perceive odor), to fear, to sleep, to live, to die, to kill, to fight, to hunt (game), to hit, to cut, to split, to stab (or stick), to scratch (as with fingernails to relieve itch), to dig, to swim, to fly, to walk, to come, to lie (on side), to sit, to stand, to turn (change one's direction), to fall (drop rather than topple), to give, to hold (in hand), to squeeze, to rub, to wash, to wipe, to pull, to push, to throw, to tie, to sew, to count, to say, to sing, to play, to float, to flow, to freeze, to swell, sun, star, water, to rain, river, lake, sea (ocean), salt, stone, sand, dust, earth (soil), cloud, fog, sky, wind, snow, ice, smoke (of fire), fire, ashes, to burn (intrans.), road (or trail), mountain, red, green, yellow, white, black, night, day (opposite of night rather than the time measure), year, warm (of weather), cold (of weather), new, old, good, bad (deleterious or unsuitable), rotten (especially log), dirty, straight, sharp (as knife), dull (knife), smooth, wet, dry (substance), right (correct), near, far, right (hand), left (hand), at, in, with (accompanying), and, if, because, name

Annex B: The 215 word list organized by semantic groups, published by Morris Swadesh in 1955

(a) **personal pronouns:** *I, thou, we, he, ye, they*

(b) **interrogatives:** *who, where, what, when, how*

(c) **correlatives** *and, if, because*

(d) **locatives:** *at, in, with*

(e) **location:** *there, far, near, right (side), here, that, this, left(side)*

(f) **position and movement:** *come, sit, give, fly, stand, hold, fall, swim, turn, walk, throw, pull, float, flow, lie, push*

(g) **manipulations:** *wash, split, tie, hit, wipe, cut, rub, dig, scratch, squeeze*

(h) **time periods:** *year, day, night*

(i) **numerals:** *one, two, three, four, five, six, seven, eight, nine, ten, twenty, hundred*

(j) **quantitatives:** *all, few, many, some*

(k) **size:** *wide, thick, long, thin, narrow, big, small, short*

(l) **natural objects and phenomena:** *ice, salt, star, sun, wind, sky, cloud, rain, water, sea, smoke, snow, sand, stone, mountain, ashes, earth, dust, lake, fog, river, fire*

(m) **plants and plant parts:** *bark, leaf, grass, tree, root, flower, woods, seed, berry (fruit), stick*

(n) **animals:** *worm, snake, louse, fish, dog, animal, bird*

(o) **persons:** *person (human being), woman, child, man*

(p) **body parts and substances:** *blood, ear, hand, tongue, tooth, foot, egg, back, tail, meat (flesh) eye, feather, skin, bone, head, mouth, nose, wing, heart, fat, guts, belly, neck, hair, liver, leg*

(q) **body sensations and activities:** *drink, die, hear, see, sleep, live, eat, know, bite, fear, think, breathe, vomit, smell*

(r) **oral activities:** *laugh, sing, suck, cry, spit, speak*

(s) **colors:** *black, green, red, white, yellow*

(t) **descriptives:** *old, dry, good, new, warm, rotten, cold, sharp, right (correct), straight, smooth, bad, wet, dull, dirty*

(u) **kinship:** *brother, sister, father, mother, husband, wife*

(v) **cultural objects and activities:** *sew, rope, shoot, hunt, cook, count, play, clothing, work, dance, spear, stab, fight*

(w) **miscellaneous:** *name, other, not, burn, blow, freeze, swell, road, kill*

Annex C: The 100 word list proposed by Morris Swadesh in 1955

all, ashes, bark, belly, big, bird, bite, black, blood, bone, burn, cloud, cold, come, die, dog, drink, die, ear, earth, eat, egg, eye, fat (grease), feather, fire, fish, fly, foot, give, good, green, hair, hand, head, hear, heart, I, kill, know, leaf, lie, live, long, louse, man, many, meat (flesh), mountain, mouth, name, neck, new, night, nose, not, one, person (human being), rain, red, road (path), root, sand, see, seed, sit, skin, sleep, small, smoke, stand, star, stone, sun, swim, tail, that, this, thou, tongue, tooth, tree, two, walk, warm (hot), water, we, what, white, who, woman, yellow, say, moon, round, full, knee, claw, horn, breast

Annex D: The widely used 207 word list with Princeton Wordnet 3.0 Mappings

Synsets starting with the numeral 7 are part of the expanded wordnet and are not in yet included in PWN.

English	PWN3.0	English	PWN3.0	English	PWN3.0	English	PWN3.0
I	77000015-n	stick/ club	04317420-n	smell	02124748-v	sand	15019030-n
thou	77000021-n	fruit	13134947-n	fear	01780202-v	dust	14839846-n
he	77000046-n	seed	11683989-n	sleep	00014742-v	earth	14842992-n
we	77000002-n	leaf	13152742-n	live	02614181-v	cloud	09247410-n
you	77000019-n	root	13125117-n	die	00358431-v	fog	11458314-n
they	77000031-n	bark	13162297-n	kill	01323958-v	sky	09436708-n
this	77000061-n	flower	11669335-n	fight	01090335-v	wind	11525955-n
that	77000079-n	grass	12102133-n	hunt	01143838-v	snow	11508382-n
here	08489497-n	rope	04108268-n	hit	01400044-v	ice	14915184-n
there	08489627-n	skin	01895735-n	cut	01552519-v	smoke	13556893-n
who	77000095-n	meat	07649854-n	split	02030158-v	fire	13480848-n
what	77000091-n	blood	05399847-n	stab	01230350-v	ashes	14769160-n
where	77000084-n	bone	05269901-n	scratch	01250908-v	burn	00377002-v
when	77000104-n	fat	05268965-n	dig	01309701-v	road	04096066-n
how	77000090-n	egg	01460457-n	swim	01960911-v	mountain	09359803-n
not	00024073-r	horn	01325417-n	fly	01940403-v	red	04962784-n
all	02269286-a	tail	02157557-n	walk	01904930-v	green	04967191-n
many	01551633-a	feather	01896031-n	come	01849221-v	yellow	04965661-n
some	01552634-a	hair	05254393-n	lie	01547001-v	white	04960729-n
few	01552885-a	head	05538625-n	sit	01543123-v	black	04960277-n
other	02069355-a	ear	05320899-n	stand	01546768-v	night	15167027-n
one	13742573-n	eye	05311054-n	turn	01907258-v	day	15164957-n
two	13743269-n	nose	05598147-n	fall	01972298-v	year	15201505-n
three	13744044-n	mouth	05301908-n	give	02199590-v	warm	02529264-a
four	13744304-n	tooth	05282746-n	hold	01216670-v	cold	01251128-a
five	13744521-n	tongue	05301072-n	squeeze	01387786-v	full	01211531-a
big	01382086-a	fingernail	05584265-n	rub	01249724-v	new	01640850-a
long	01433493-a	foot	05563266-n	wash	00557686-v	old	01638438-a
wide	02560548-a	leg	05560787-n	wipe	01392237-v	good	01123148-a
thick	02410393-a	knee	05573602-n	pull	01609287-v	bad	01125429-a
heavy	01184932-a	hand	02440250-n	push	01871979-v	rotten	01070538-a
small	01415219-a	wing	02151625-n	throw	01508368-v	dirty	00419289-a
short	01436003-a	belly	05556943-n	tie	00141632-v	straight	02314584-a
narrow	02561888-a	guts	05534333-n	sew	01329239-v	round	02040652-a
thin	02412164-a	neck	05546540-n	count	00948071-v	sharp	00800826-a
woman	10787470-n	back	05558717-n	say	00979870-v	dull	00800248-a
man	10287213-n	breast	05554405-n	sing	01729431-v	smooth	02236842-a
person	00007846-n	heart	05388805-n	play	01072949-v	wet	02547317-a
child	09918248-n	liver	05385534-n	float	01904293-v	dry	02551380-a
wife	10780632-n	drink	01170052-v	flow	02066939-v	correct	00631391-a
husband	10193967-n	eat	01168468-v	freeze	00445711-v	near	00444519-a
mother	10332385-n	bite	01445932-v	swell	00256507-v	far	00442361-a
father	10080869-n	suck	01169704-v	sun	09450163-n	right	02031986-a
animal	00015388-n	spit	00101956-v	moon	09358358-n	left	02032953-a
fish	02512053-n	vomit	00076400-v	star	09444783-n	at	*excluded*
bird	01503061-n	blow	02100632-v	water	14845743-n	in	*excluded*
dog	02084071-n	breath	00001740-v	rain	15008607-n	with	*excluded*
louse	02183857-n	laugh	00031820-v	river	09411430-n	and	*excluded*
snake	01726692-n	see	02150948-v	lake	09328904-n	if	*excluded*
worm	01922303-n	hear	02169702-v	sea	09426788-n	because	*excluded*
tree	13104059-n	know	00594621-v	salt	07813107-n	name	06333653-n
forest	09284015-n	think	00629738-v	stone	09416076-n		