Sinitic Wordnet: Laying the Groundwork with Chinese Varieties Written in Traditional Characters

Chih-Yao Lee
National Taiwan University
Taipei, Taiwan.
chihyaolee@gmail.com

Shu-Kai Hsieh
National Taiwan University
Taipei, Taiwan.
shukai@gmail.com

Abstract

The present work seeks to make the logographic nature of Chinese script a relevant research ground in wordnet studies. While wordnets are not so much about words as about the concepts represented in words, synset formation inevitably involves the use of orthographic and/or phonetic representations to serve as headword for a given concept. For wordnets of Chinese languages, if their synsets are mapped with each other, the connection from logographic forms to lexicalized concepts can be explored backwards to, for instance, help trace the development of cognates in different varieties of Chinese. The Sinitic Wordnet project is an attempt to construct such an integrated wordnet that aggregates three Chinese varieties that are widely spoken in Taiwan and all written in traditional Chinese characters.

1 Introduction

As with Romance languages descending from Classical Latin that stand on their own in present days, Sinitic languages[1], or major descendants of Archaic Chinese, have developed into fully-fledged languages without or with very limited mutual intelligibility (Tang and van Heuven, 2007). However, thanks to a shared logographic writing system that has not seen drastic changes in modern times[2], speakers of distinct Chinese languages can use a common set of logographic characters to communicate.

As language is not merely a vehicle for the expression of thought, but the way to thought itself, writing systems not only represent a language, but reflect and record its ever-changing nature. As both linguists and wordnet builders, we see a great potential for wordnets to assist in lexical-semantic studies across Chinese languages, synchronic and diachronic alike, and serve as a handy repository where logograph-based searches are enabled.

In this paper, we present the initial version of a new resource named "Sinitic Wordnet", which not only includes the lexicons of Mandarin, Southern-Min and Hakka, but makes use of Collaborative Interlingual Index to link them to other wordnet projects.

2 Methodology

In this section, we explain how synsets were organized based on the dictionaries and how they were interlinked afterwards.

2.1 Conversion of Individual Lexicons into Wordnets

We retrieved from the website of gov-zero[3] machine-readable versions of Mandarin-to-Mandarin, Southern-Min-to-Mandarin, and Hakka-to-Mandarin dictionaries compiled by the Ministry of Education, Taiwan. The statistics of the three dictionaries are given in Table 1.

Dictionary Type	Entry Count
Mandarin-to-Mandarin	166,119
Southern-Min-to-Mandarin	20,377
Hakka-to-Mandarin	15,487

Table 1: Entry counts of the three dictionaries.

Assuming that (nearly) synonymous word senses were glossed largely the same (Sinha et al., 2006), we started by using sense glosses as the

[1]The term "Sinitic" was chosen to suggest that the varieties of Chinese are distinct languages rather than different dialects of a same language.

[2]At least not until after the 1950s, when the Chinese Character Simplification Scheme was introduced in China.

[3]More commonly referred to as g0v, gov-zero is a civic tech community that promotes the ideas of open government, open data, civic participation, and new media in Taiwan.

unique identifier for a synset entry. By means of comparing the similarities of sense definition between every two pairs of synsets, we were able to merge entries of synsets that are similar, if not identical, in meaning. After automated matching and merging, the resulting synsets were manually checked. Table 2 gives the numbers of synsets derived from each of the three dictionaries.

Dictionary Type	Synset Count
Mandarin-to-Mandarin	25,761
Southern-Min-to-Mandarin	3,158
Hakka-to-Mandarin	2,400

Table 2: Synset counts of the three lexicons.

2.2 Alignment of Individual Wordnets

To align synsets from the individual wordnets, we resorted to pattern-matching in three pieces of information that may (or may not) be included in an entry:

1. **Sense definition**: a gloss, (near-)synonym or translation equivalent (in the case of bilingual dictionaries). As we were able to compare the degree to which two sense definitions are alike to organize synsets within an individual wordnet, by the same token, we could map between synsets of different wordnets by computing their similarities based on synset glosses. Also, in Southern-Min-to-Mandarin and Hakka-to-Mandarin dictionaries, some of the definitions are not really glosses, but simply translation equivalents in Mandarin. We used those Mandarin equivalents as links to Southern-Min and Hakka. While this link is from sense to lemma rather than between senses, we selected the first and usually the most salient sense of the lemma to be the represented concept.

2. **Example words**: when the lemma of an entry has usages as bound-morpheme, there can be compound words to illustrate how the lemma combines with others. Along with such example words in the two bilingual dictionaries, there are usually Mandarin equivalents to Southern-Min and Hakka, respectively. Again, by choosing the first sense of the translation in Mandarin, we were able to establish links between the three languages.

3. **Multilingual translation**: in a separate section of the dictionaries, there are translations to European languages (including French, German and Spanish) as well as between the three Chinese varieties. Once again, the first sense of the translation words was used to connect the three lexicons.

2.3 Mapping with Princeton WordNet

In order to facilitate integration with other resources as well as enable queries in English, Sinitic Wordnet has its synsets mapped with those of Princeton WordNet (Fellbaum, 2010). The mapping was done by bilinguals of English and one of the Chinese varieties.

3 Results

In this section, we show how it is possible to track in Sinitic Wordnet concepts that are encoded in different logographs, and vice versa. Also, an entry is given the Turtle format for the sake of illustration.

3.1 Sinitic Wordnet as Bridge between Concepts, Synsets and Logographs

When converting dictionaries into wordnets, we focused on concepts as expressed by sense glosses written in Mandarin, grouped them into synsets according to how similar their describing texts were, and mapped them with counterparts in Princeton WordNet. Now that the foregoing steps have been completed, the entire procedure can be examined in the opposite direction to help discover lexicalization patterns as well as to observe the way word senses distribute from variety to variety and determine whether new ones have developed in one language, or whether old ones have ceased to exist in another.

Take for example the concept POT. If one is curious about how this idea is encoded in traditional Chinese characters across different varieties, they can run a query using English words (e.g. *pot*) or phrases (e.g. *cooking vessel*) that may express the concept. As shown in Figure 1, a query of "pot" would lead to the English synset {pot}, which is in turn linked with equivalents from each of the Chinese lexicons. To encode the concept POT, Mandarin chooses '鍋' (guō) over '鼎' (tiánn) and '鑊' (vók), which are respectively adopted in Southern-Min and Hakka.

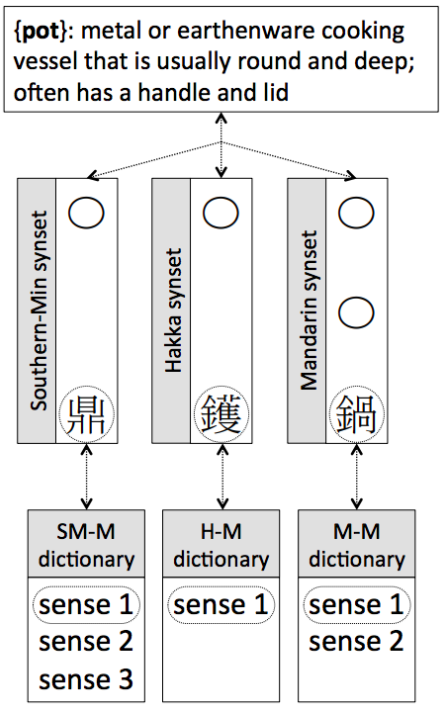

Figure 1: From a concept to synsets, from a synset to lemmas as represented by different holographs.

Reversely, it is equally possible to look at what distinct meanings a single logograph carries in different varieties, as illustrated in Figure 2, where the logograph '鼎' (tiánn) is taken as query to look for synsets whose consisting members are represented by the same character.

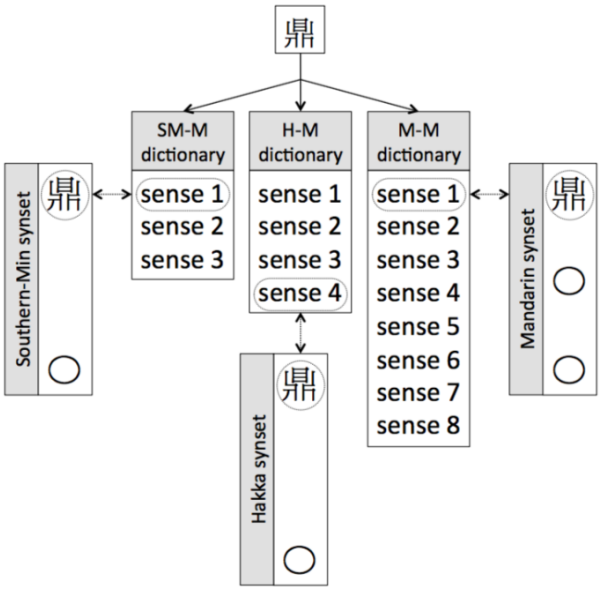

Figure 2: From a holograph to synsets in different lexicons.

3.2 Sinitic Wordnet as Linked Data

To improve its interoperability with other lexical resources, Sinitic Wordnet has been converted in RDF format using the *lemon* model (McCrae et al., 2011; McCrae et al., 2012). Figure 3 shows what a *lemon*ized sense looks like in Turtle format[4].

```
@prefix owl: <http://www.w3.org/2002/07/
    ↪ owl#> .
@prefix rdf: <http://www.w3.org
    ↪ /1999/02/22-rdf-syntax-ns#> .
@prefix lemon: <http://www.lemon-model.
    ↪ net/lemon#> .
@prefix wordnet-ontology: <http://
    ↪ wordnet-rdf.princeton.edu/
    ↪ ontology#> .
<http://lope.linguistics.ntu.edu.tw/swn/
    ↪ mandarin/dong4wu4/052268> a lemon
    ↪ :LexicalEntry ;
    lemon:canonicalForm <#CanonicalForm>
        ↪ ;
    lemon:sense <#1> ;
    wordnet-ontology:part_of_speech
        ↪ wordnet-ontology:noun .
<#CanonicalForm> a lemon:Form ;
    lemon:writtenRep "@cmn" .
<#1> a lemon:LexicalSense ;
    lemon:reference <http://lope.
        ↪ linguistics.ntu.edu.tw/swn/
        ↪ mandarin/2068> ;
    wordnet-ontology:gloss
        ↪
        ↪ "@cmn" ;
    owl:sameAs <http://wordnet-rdf.
        ↪ princeton.edu/wn31/100015568-
        ↪ n> .
```

Figure 3: The first sense of *dong4wu4* in Turtle.

4 Publishing the Resource

Once the wordnets and their mappings derived from this project are made more tidy, we will release the data under an open license in order to ensure that it can be put into use as widely as possibly. Before that, we have made the resource available by integrating it with two best practices in the WordNet community, namely with the Linguistic Linked Open Data Cloud and the Collaborative Interlingual Index.[5]

4.1 Publishing the Resource as Linked Data

By way of synset mapping, Sinitic Wordnet not only has its consisting lexicons interlinked, but also links directly to Princeton Wordnet. As shown in Figure 3, there is an outward link to Princeton WordNet because the synset referenced

[4] http://www.w3.org/TR/turtle/

[5] http://lope.linguistics.ntu.edu.tw/swn

to by the lexical sense has an equivalent in English. Meanwhile, the links to WordNet serve as key to the Linguistic Linked Open Data cloud (Chiarcos et al., 2013) and interface with other linguistic resources. Moreover, Sinitic Wordnet can be integrated into the Global WordNet Grid when organized by the ontology consisting of 71 Base Types proposed by the Global WordNet Association.[6] An initial mapping has identified 169 synsets comparable to the Base Types.[7]

4.2 Integrating the Resource with Collaborative Interlingual Index

The Collaborative Interlingual Index (Bond et al., 2016) has been proposed as a method to enable cross-lingual development of wordnets. Chief among the primary objectives of the project is to establish a standard operating procedure by which new synsets can be defined and added to a common repository, resolving compatibility issues that may occur when wordnets for languages other than English introduce concept not lexicalized in English. In order to facilitate the integration of Sinitic Wordnet with the Collaborative Interlingual Index, we are making the full version of the resource available in the Global WordNet Association's recommended formats and release under an open license.

5 Conclusion

Based on monolingual (Mandarin to Mandarin) and bilingual (Southern-Min/Hakka to Mandarin) dictionaries by the Ministry of Education, Taiwan, we have presented a method for developing an integrated wordnet that includes and interlinks the lexicons of Mandarin, Southern-Min and Hakka. The resource was generated semi-automatically, relying on bilinguals of the Chinese varieties to map synsets with Princeton WordNet. To align the synsets, a mixture of methods was employed, including looking for synonyms in sense definitions, and translation equivalents in example words as well as in a section of the dictionaries that gives translation words in European and Chinese languages.

In addition to more thorough check-ups upon the integrity and quality of existing lexicons, our plans for future development include the addition of Cantonese as spoken in Hong Kong, another Chinese variety that is also written in traditional Chinese characters, and the construction of a web-based graphical user interface for public access of Sinitic Wordnet.

[6] http://w.globalwordnet.org/gwa/ewn_to_bc/BaseTypes.htm

[7] http://lope.linguistics.ntu.edu.tw/swn/gwn/

References

Francis Bond, Piek Vossen, John P McCrae, and Christiane Fellbaum. 2016. Cili: the collaborative interlingual index.

Christian Chiarcos, John McCrae, Philipp Cimiano, and Christiane Fellbaum. 2013. Towards open data for linguistics: Linguistic linked data. In *New Trends of Research in Ontologies and Lexical Resources*, pages 7–25. Springer.

Christiane Fellbaum. 2010. Wordnet. In *Theory and applications of ontology: computer applications*, pages 231–243. Springer.

John McCrae, Dennis Spohr, and Philipp Cimiano. 2011. Linking lexical resources and ontologies on the semantic web with lemon. In *Proceedings of the 8th Extended Semantic Web Conference on The Semantic Web: Research and Applications - Volume Part I*, ESWC'11, pages 245–259, Berlin, Heidelberg. Springer-Verlag.

John McCrae, Guadalupe Aguado-de Cea, Paul Buitelaar, Philipp Cimiano, Thierry Declerck, Asunción Gómez-Pérez, Jorge Gracia, Laura Hollink, Elena Montiel-Ponsoda, Dennis Spohr, and Tobias Wunner. 2012. Interchanging lexical resources on the semantic web. *Language Resources and Evaluation*, 46(4):701–719.

Manish Sinha, Mahesh Reddy, and Pushpak Bhattacharyya. 2006. An approach towards construction and application of multilingual indo-wordnet. In *3rd Global Wordnet Conference (GWC 06), Jeju Island, Korea*.

Chaoju Tang and Vincent J van Heuven. 2007. Mutual intelligibility and similarity of chinese dialects: Predicting judgments from objective measures. *Linguistics in the Netherlands*, 24(1):223–234.