Japanese Wordnet 2.0

Francis Bond
Palacký University
bond@ieee.org

Takayuki Kuribayashi
takkur@gmail.com

Abstract

This paper describes a new release of the Japanese wordnet. It uses the new global wordnet formats (McCrae et al., 2021) to incorporate a range of new information: orthographic variants (including hiragana, katakana and Latin representations) first described in Kuroda et al. (2011), classifiers, pronouns and exclamatives (Morgado da Costa and Bond, 2016) and many new senses, motivated both from corpus annotation and linking to the TUFs basic vocabulary (Bond et al., 2020). The wordnet has been moved to github and is available at https://bond-lab.github.io/wnja/.

1 Introduction

This paper describes a new release of the Japanese wordnet, v2.0. This new version of the Japanese wordnet includes orthographic variants and transliterations (Kuroda et al., 2011), classifiers, exclamatives (Morgado da Costa and Bond, 2016) and pronouns (Seah and Bond, 2014), as well as words introduced during the annotation of the NTU Multilingual Corpus (Bond et al., 2013). This is the first release in almost 10 years, and has a numerous changes.

The Japanese Wordnet was started at the National Institute of Information and Communications Technology (NICT) based on the **expand** approach of adding Japanese lemmas to existing Princeton Wordnet 3.0 (PWN: Fellbaum, 1998) synsets, with plans to follow this up by annotating a corpus and adding missing words (**extend**). This allowed us to take advantage initially of the rich information in the Princeton Wordnet.

The progress of construction is shown in Table 1. The first release (v0.9: 2009-02) contained 48,190 synsets. These were created by linking to the structure of Princeton Wordnet (Fellbaum, 1998, 3.0) through four languages: English, French, Spanish and German (Bond et al., 2008).

The second release (v0.91: 2009-08) was a bug-fix release, with slightly more synsets (50,739) but fewer senses, as we checked more of the automatically built synsets. This release included links to images in the Open Clip Art Library (OCAL Phillips, 2005) and the Suggested Upper Merged Ontology (SUMO Niles and Pease, 2001; Pease, 2011). Finally, there was one more bug-fix release (v0.92: 2009-11) this time with fewer synsets as well as senses.

The next major release (v1.0: 2010-03) saw the addition of definitions and example sentences (Kuribayashi et al., 2010). These were automatically translated from English, using a specialized corpus of example sentences, and then hand corrected. As this was part of research to produce a large parallel corpus at NICT, all definitions and examples were translated, even if they did not have any Japanese lemmas associated with them.

We next decided to do some work on producing sense-tagged corpora in order to see how well the wordnet did on describing real world Japanese text. For our first attempt, we created the Japanese SemCor (JSemCor) a (partially) sense-tagged corpus of Japanese (Bond et al., 2012). The final corpus consists of 14,169 sentences with 150,555 content words of which 58,265 are sense tagged. It allowed us to provide sense frequency data for the Japanese Wordnet.

We next annotated over 7,000 sentences in the NTU Multilingual Corpus (Tan and Bond, 2012), including news text, tourism text, short stories and an essay. This has led us to identify many missing concepts as well as many missing senses. There are 20,386 sense tagged words (including multi-word expressions) annotated in the Japanese portion of the corpus, with 6,706 distinct senses.

In 2014 a python module was developed that allowed the wnja 1.1 data to be used in NLTK (Bird et al., 2009). Goodman and Bond (2021) made a module for the new wordnet structure, which can

Year-Mon	Ver	Concepts	Words	Senses	Misc
2009-02	0.90	49,190	75,966	156,684	initial release
2009-08	0.91	50,739	88,146	151,831	SUMO, OCAL
2009-11	0.92	49,655	87,133	146,811	
2010-03	1.00	56,741	92,241	157,398	+ def, ex
2010-10	1.10	57,238	93,834	158,058	
2012-01					Japanese Semcor
2014-02					NLTK module
2023-01	2.0	58,527	90,320	148,676	262,196 forms

Table 1: Japanese wordnet milestones

be used with this release.

This release has more concepts, slightly fewer senses and words (as we delete bad entries) and many more variant forms (described in the next section).

2 Richer Information

This release of the wordnet gathers together several improvements.

2.1 Orthographic variants

The Japanese writing system is particularly complex. It consists of three separate sets of characters: hiragana, katakana and kanji. Modern Japanese also makes frequent use of Arabic numbers, Latin script and increasingly emoji.

Hiragana and **katakana** are isomorphic syllabaries made up of 46 basic characters.

The third character system is **kanji**, derived historically from Chinese characters. 2,136 kanji are in common use, based on the set of Joyo Kanji stipulated by the Japanese Ministry of Education, Culture, Sports, Science and Technology which are taught in Japanese primary and middle schools. Thousands more are used in place names, person names and historical texts.

A single kanji character generally has at least one **on**-reading which is loosely derived from its Chinese pronunciation at the time of borrowing,[1] and at least one native Japanese **kun**-reading where a Japanese word which pre-existed the orthographic borrowing was mapped onto a kanji character based on rough semantic correspondence. For example, 動 has a unique on-reading of *dō*, and a unique kun-reading of *ugo(ku/kasu)*;[2] in both cases, its basic meaning is "motion, change".

Hiragana is typically used for inflections, function words and onomatopoeic expressions. Katakana is typically used for foreign words. Words normally written in Kanji can be written in hiragana (to ease reading) or katakana (for emphasis, similar to italics in English). A single word, such as *ugoita* (動いた) "moved (intrans.)" could thus be written as うごいた or ウゴイタ. Further, some kanji have variants (typically more complicated older forms and newer simpler ones). Typically, a dictionary for human users will just list the standard form and any character variants, with possibly the pronunciation in Katakana or Hiragana (see Backhouse (1993); Bond and Baldwin (2016) for more discussion).

We have decided to list all possible forms, with one chosen as the display form. There is no universal standard for what the display form should be. However the widely used morphological analyser **juman** (Kurohashi and Nagao, 1998) lists canonical forms for all words in its dictionary ((Okabe et al., 2007)) and we use them when available.

Overall we decide as follows:

1. If there is an entry in **jumandic** we use their canonical form
2. Prefer kanji to hiragana
3. Prefer new forms to old forms
   (we compiled our own table of new and old forms)
4. If there are multiple katakana variants, prefer the longest

[1] Indeed, many kanji still have corresponding hanzi in traditional Chinese, although there are also a few kanji which were devised in Japan and are unique to Japanese, such as *hatake* (畑) "field" and *tōge* (峠) "mountain pass".

[2] The reading of 動 itself is *ugo*, and it combines with a kana-based conjugational suffix (**okurigana**) derived from *ku* or *kasu* (corresponding to intransitive and transitive verb usages, respectively), e.g. *ugoita* (動いた) "moved (intrans.)" or *ugokashiteiru* (動かしている) "is moving (trans.)".

We give an example of variants for the synset meaning "form an arch or curve" in Table 2. The first katakana entry can be used to give the pronunciation and is also used to generate a variant in Latin script, so that the dictionary can be searched by users with no Japanese input system.

We have added up to two Latin transliterations, the standard Kunrei-siki romanization (preferred by the Japanese Ministry of Education), and where it differs, the commonly used Hepburn romanization (more similar to English orthography). In Figure 1 we show the different representations of *jisho* "dictionary". Conversion is done automatically from the katakana form using the python `romkan` library.[3]

Note that due to differences in use of old and new Chinese characters and the option of omitting hiragana, a word may have many different forms: *nomikomu* "swallow" can have at least the following 飲み込む, ノミコム, 飲込む, 呑込む, 呑み込む, のみ込む, のみこむ.

Unfortunately, the display form cannot simply be the canonical form, as it can be the case that the same display form has different pronunciations for different meanings (or the same meaning), and some variants are not possible for all senses. For example *kedamono* (獣) "beast" and *shishi* (獣) "boar" are used for all mammals, but only *shishi* (獣) "boar" has the variants 猪 and 鹿. *inoshishi* (猪) "wild boar" has no variant, whereas *i* (猪) "boar (in the Chinese Zodiac)" has variants 豕 and 豬. Because of such idiosyncrasies, all entries had to be hand-checked, which was a monumental task: this is why there was such a long gap between releases. We summarize the number of forms in Table 3.

Increasing the number of variants is necessary to increase the coverage of the lexicon on corpora. It also makes the dictionary more useful to language learners, who may not be able to read the kanji, but should be able to read kana or Latin versions.

2.2 Frequencies

We include sense frequencies based on the annotation in the NTU Multilingual Corpus (Tan and Bond, 2012) and the Japanese SemCor (Bond et al., 2012).

For example, in the synset `00174412-n` "any maneuver made as part of progress toward a goal" the Japanese senses have the following frequencies: 対策$_3$, 策$_3$, 措置$_2$, 方略, 方策, 術, 打つ手. The frequencies are used in the Open Multilingual Wordnet (OMW: Bond and Foster, 2013) to order the senses in the display, and to chose the most appropriate label for each synset. They can also be used for choosing the most frequent sense for word sense disambiguation.

2.3 Grammatical Notes

We also marked the major verb inflectional class of Sino-Japanese verbs, with a usage note (`note='sahen'`). These verbs typically appear with a support verb (such as *suru* "do" or *dekiru* "can"). On their own they look similar to nouns and typically link to a zero-derived noun. We show an example in Figure 2.

3 New Entries

We have expanded the vocabulary of the Japanese wordnet through a combination of corpus annotation and systematic expansion of lexical fields. We try to add not just individual words, but also complete semantic fields together, especially when there is a difference in conceptual structure with English. Here are some of the major additions in this release

1. Numeral classifiers (not used in English)
2. Pronouns (not in the Princeton Wordnet)
3. Exclamatives (not in the Princeton Wordnet)
4. Time/Date expressions (often split into different units than in English)
5. Japanese kinship terms (richer than English)

The semi-closed classes of pronouns, classifiers and exclamatives were added to the Chinese, English, Indonesian and Malay wordnets at the same time, as described in Seah and Bond (2014) and Morgado da Costa and Bond (2016). The numbers of new entries for the different classes are given in Table 4. We do not consider the coverage to be anywhere near complete, but we cover most common words from these classes.

Pronouns

Japanese pronouns differ on several dimensions from English — in particular there are different levels of formality for personal pronouns, and demonstrative pronouns distinguish between proximal *kono*, medial *sono* and distal *ano* as opposed

[3] https://pypi.org/project/romkan/

Display form	Pronunciation	Variants	Latin
湾曲	ワンキョク	彎曲, 弯曲, わん曲	wankyoku
反る	ソル	そる	soru
カーブ	カーブ	カーヴ	ka-bu …

Table 2: Variants of "form an arch or curve"

```
<LexicalEntry id="wnja-n-3023"> <!-- 辞書 0 n -->
  <Lemma writtenForm="辞書" partOfSpeech="n"/>
     <Form writtenForm="ジショ" script="kana"/>
     <Form writtenForm="じしょ" script="hira"/>
     <Form writtenForm="zisyo" script="latn"/>
     <Form writtenForm="jisho" script="latn-hepburn"/>
     ...
</LexicalEntry>
```

Figure 1: Different forms for *jisho*, showing scripts

Script	Number
Mixed	83,049
Katakana	89,542
Hiragana	89,605
Latin	89,542
Latin (Hepburn)	36,753
Total	388,491

Table 3: Numbers of forms by script

to English's two-way distinction: *this* proximal and *that* medial/distal.

Exclamatives

We added exclamatives (including greetings, interjections and many more), following Morgado da Costa and Bond (2016, who only added English and Chinese), which is loosely based on the classification of Jovanović (2004). Some exclamatives are similar in many languages, such as the greetings *konnichiwa* "good day" or *sayonara* "good bye". We also added some purely Japanese expressions, such as *onegai-shimasu* (1) and *otsukaresama* (2).

(1)

80002404-x (お願いします)	
lemmas:jpn	お願いします, お願い
def:jpn	よくしてくれることを求める意味合いの発話
def:eng	an expression that is uttered when you ask for a favor
exemplifies	07109847-n (utterance)
see also	00903098-v (wish)
similar to	80001988-x (please)

(2)

80002405-x (お疲れ様)	
lemmas:jpn	お疲れ様, ご苦労様
def:jpn	相手の苦労をねぎらう発話
def:eng	an expression that is uttered when you appreciate someone's work; typically used when someone leaves work
exemplifies	07109847-n (utterance)
see also	01805982-v (appreciate)
similar to	80000666-x (thank you)

Classifiers

Again we followed Morgado da Costa and Bond (2016) for the numeral classifiers. Because usage is significantly different across languages, we have no classifiers shared exactly across even such similar languages as Chinese and Japanese. We show an example of the idiosyncratic Japanese classifier for birds and rabbits in 3.

(3)

76100129-x (羽)	
lemmas:jpn	羽
def:jpn	ツバメやタカやペンギンなどの鳥、またウサギに対しても用いられる分類辞
exe:jpn	日本では、月で一羽のウサギが餅を搗いていると考えられています; 彼は４羽のオウムを飼っています
def:eng	a sortal classifier used for birds such as a swallow, a hawk or a penguin, and also specifically for rabbits
exe:eng	in Japan, people think a rabbit is making rice cake on the moon; he has 4 parrots
exemplifies	06308436-n (classifier)
classifies	01503061-n (bird)
classifies	02324045-n (rabbit)

```
<LexicalEntry id="wnja-v-74345" note="sahen"> <!-- 読書 0 v -->
      <Lemma writtenForm="読書" partOfSpeech="v"/>
        <Form writtenForm="ドクショ" script="kana"/>
        <Form writtenForm="どくしょ" script="hira"/>
        <Form writtenForm="dokusyo" script="latn"/>
        <Form writtenForm="dokusho" script="latn-hepburn"/>
       <Sense id="wnja-00625119-v-74345" synset="wnja-00625119-v" confidenceScore="1.0"/>
 </LexicalEntry>
```

Figure 2: Entry for *dokusho*, showing the usage note sahen

Class	Synsets	Lemmas	Examples
Classifier	47	47	人, 匹, 機
Exclamation	24	37	ああ, なるほど, さよなら
Pronoun	21	70	あちら, こちら
Personal Pronoun	19	29	私, あなた, 彼, 彼女
Reflexive Pronoun	2	6	自分, 己れ
Demonstrative Pronoun	22	25	これ, それ, あれ
Interrogative pronoun	10	13	どれ

Table 4: New Classes of Words

Time Expressions

Many time expressions which are phrases in English are single words in Japanese (such as 今週 *konshuu* "this week", or 今朝 *kesa* "this morning"). Historically, these were compounds in Chinese, but have been borrowed as single words. We added some 280 time senses, looking simultaneously at Japanese, Chinese and English. These included days of the month, compound dates and holidays. English was added for two reasons. The first was that it is useful for those that use the wordnets as bilingual lexicons. The second is that there is some lexicalization: we say *last year, this year, next year* but *yesterday morning, this morning, tomorrow morning* and *last night, tonight, tomorrow night*.[4] Chinese equivalents are arguably also lexicalized (and were typically segmented as two character expressions by the Penn Chinese Treebank (Xue et al., 2005)), adding them also made crosslingual linking easier. We give an example of an entry (including English and Chinese) in (4).

(4)

90000501-n (last year)	
lemmas:jpn	昨年, 去年
lemmas:eng	last year
lemmas:cmn	去年
def:jpn	現在の属する年の直前の年
exe:jpn	去年は盛りだくさんな年だった
def:eng	the year before this year
exe:eng	last year was an eventful one
def:cmn	今年的前一年
hypernym	15203791-n (year)

Kinship Terms

As well as distinguishing older and younger brothers and sisters, Japanese distinguishes aunts and uncles older and younger than the parent they are related to. For example, *oba* (伯母) "an aunt who is older than one's parent" vs *oba* (叔母) "an aunt who is younger than one's parent". Most kin terms have formal and informal variants, for the moment they are added to the same synset, in future work we wish to distinguish them using sense-based usage links.

Other new vocabulary

One other interesting difference between Japanese and English is in describing temperature. English uses the same words for temperature experienced by touching or as a general feeling (5). Japanese on the other hand distinguishes a general feeling (6) used for example when feeling cold, or

[4] Ross (1995) argues that English temporal nouns are **defective**: they are typically pronominalized by *then* and have idiosyncratic determiner use.

cold weather; and experiencing by touch (7) used for example for a cold soup or cold hands.

(5) ⟨*cold, cool, warm, hot*⟩

(6) feel: ⟨寒い, 涼しい, 暖かい, 暑い⟩

(7) touch: ⟨ 冷たい, 温かい, 熱い⟩

In fact, the words for warm and hot are pronounced the same whether for feeling or to-touch: *atatakai* and *atsui*, the difference is only written. These words were identified due to their presence in the TUFS basic vocabulary for teaching (Bond et al., 2020). We show their structure in 3.

Finally, we have added many new synsets that came up in the corpora being annotated: altogether 770 new synsets have been added. We give some examples below, some are from Japanese culture (8,9), some from Singapore (10: as we annotated Singapore tourist documents) and some from news and essays (11). Many of these should also be added to the Open English Wordnet (McCrae et al., 2020).

(8)

80001626-n (soba_noodle)	
lemmas:jpn	蕎麦
lemmas:eng	soba
def:jpn	そば粉で作られた細い麺
def:eng	narrow noodle made from buckwheat
hypernym	(noodle)

(9)

80000338-n (Shunto)	
lemmas:jpn	春闘
lemmas:eng	spring wage negotiation
def:jpn	毎年労働組合が、賃金引き上げなどの要求を掲げて行う全国的な闘争
def:eng	annual event by Japanese workers union when wages are renegotiated
hypernym	(protest)

(10)

80002377-n (castle construction)	
lemmas:jpn	築城
def:jpn	城の建設
def:eng	the construction of castles
hypernym	(construction)

(11)

90000315-n (hajjah)	
lemmas:jpn	ハジャ
lemmas:eng	hajjah
def:jpn	メッカへの巡礼を行った女性
def:eng	a woman who has made the pilgrimage to Mecca
hypernym	(haji)
category	(muslim)

(12)

80001731-n (exchange student)	
lemmas:jpn	留学生
lemmas:eng	exchange student
def:jpn	海外で勉強する学生
def:eng	a student who studies abroad
hypernym	(student)

4 More Accessible

Earlier versions of the Japanese wordnet were available at a university web site, with the data stored in sourceforge. For this release, data and documentation are stored in github, to make them more permenant. The wordnet is available online, both as plain xml, and as a released tarball with the license and canonical citation. This can be loaded directly from the Python WN module (Goodman and Bond, 2021), or the OMW interface. The Japanese wordnet can be found here: `https://bond-lab.github.io/wnja/`.

5 Conclusions

This paper presents the current state of the Japanese Wordnet: **wnja**. We hope that **wnja** will continue to be a useful resource not only for natural language processing, but also for language education/learning and linguistic research.

In future work, we want to look more at the description of formality and politeness, as well as to increase the coverage.

Acknowledgements

This research was supported in part by the JSPS/NUS Grant *Automatically determining meaning by comparing a text to its translation*, MOE Tier 2 grant *That's what you meant: a Rich Representation for Manipulation of Meaning* (MOE ARC41/13), the MOE Tier 1 grant on *Shifted in Translation: An Empirical Study of Meaning Change Across Languages (RG51/12)*, the Creative Commons grant on *Assessing the effect of license choice on the use of lexical resources* and joint research with Fuji-Xerox on *Multilingual Semantic Analysis*. Especial thanks to the other Japanese wordnet developers, Hitoshi Isahara, Kyoko Kanzaki, Kow Kuroda, Kiyotaka Uchimoto, Masao Utiyama, Darren Cook, Asuka Sumida and Kentaro Torisawa, as well as the many contributors who gave feedback. Some of this work was done while visiting the Humanities Center at Tokyo University, thanks to Tsuneko Nakazawa and Tsuneaki Kato.

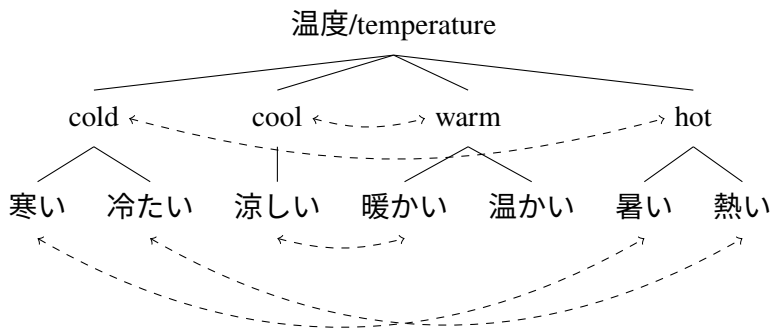

Figure 3: Structure for temperature words

Some nodes are not lexicalized in Japanese, but are still useful for the structure
temperature is linked by ATTRIBUTE (属性); tree is HYPONYM; dashed arrows are ANTONYM

References

Anthony E. Backhouse. 1993. *The Japanese Language: An Introduction*. Oxford University Press, Oxford.

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

Francis Bond and Timothy Baldwin. 2016. Introduction to Japanese computational linguistics. In Francis Bond, Timothy Baldwin, Kentaro Inui, Shun Ishizaki, Hiroshi Nakagawa, and Akira Shimazu, editors, *Readings in Japanese Natural Language Processing*, chapter 1, pages 1–28. CSLI Publications.

Francis Bond, Timothy Baldwin, Richard Fothergill, and Kiyotaka Uchimoto. 2012. Japanese SemCor: A sense-tagged corpus of Japanese. In *Proceedings of the 6th Global WordNet Conference (GWC 2012)*, pages 56–63, Matsue.

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.

Francis Bond, Hitoshi Isahara, Kyoko Kanzaki, and Kiyotaka Uchimoto. 2008. Boot-strapping a WordNet using multiple existing WordNets. In *Sixth International Conference on Language Resources and Evaluation (LREC 2008)*, Marrakech.

Francis Bond, Hiroki Nomoto, Luís Morgado da Costa, and Arthur Bond. 2020. Linking the TUFS basic vocabulary to the open multilingual wordnet. In *Twelth International Conference on Language Resources and Evaluation (LREC 2020)*, Marseilles.

Francis Bond, Shan Wang, Eshley Huini Gao, Hazel Shuwen Mok, and Jeanette Yiwen Tan. 2013. Developing parallel sense-tagged corpora with wordnets. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse (LAW 2013)*, pages 149–158, Sofia.

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

Michael Wayne Goodman and Francis Bond. 2021. Intrinsically interlingual: The Wn Python library for wordnets. In *11th International Global Wordnet Conference (GWC2021)*.

Vladimir Ž Jovanović. 2004. The form, position and meaning of interjections in English. *FACTA UNIVERSITATIS-Linguistics and Literature*, (Vol. 3/11):17–28.

Takayuki Kuribayashi, Francis Bond, Kow Kuroda, Kiyotaka Uchimoto, Hitoshi Isahara, Takayuki Kuribayashi, and Kyoko Kanzaki. 2010. Japanese WordNet 1.0. In *16th Annual Meeting of the Association for Natural Language Processing*, pages A5–3, Tokyo.

Kow Kuroda, Takayuki Kuribayashi, Francis Bond, Kyoko Kanzaki, and Hitoshi Isahara. 2011. Orthographic variants and multilingual sense tagging with the Japanese WordNet. In *17th Annual Meeting of the Association for Natural Language Processing*, pages A4–1, Toyohashi.

Sasao Kurohashi and Makoto Nagao. 1998. *Japanese morphological analysis system JUMAN version 3.6 manual*. Kyoto University.

John P. McCrae, Michael Wayne Goodman, Francis Bond, Alexandre Rademaker, Ewa Rudnicka, and Luís Morgado da Costa. 2021. The global wordnet formats: Updates for 2020. In *11th International Global Wordnet Conference (GWC2021)*.

John P. McCrae, Alexandre Rademaker, Ewa Rudnicka, and Francis Bond. 2020. English wordnet 2020: Improving and extending a wordnet for English usingan open-source methodology. In *Workshop on Multimodal wordnets at LREC 2020*.

Luís Morgado da Costa and Francis Bond. 2016. Wow! what a useful extension to wordnet! In *10th International Conference on Language Resources and Evaluation (LREC 2016)*, Portorož.

Ian Niles and Adam Pease. 2001. Towards a standard upper ontology. In *Proceedings of the 2nd International Conference on Formal Ontology in Information Systems (FOIS-2001)*, Maine.

Kouji Okabe, Daisuke Kawahara, and Sadao Kurohasi. 2007. Improving nlp resources using canonical forms. In *13th Annual Meeting of The Association for Natural Language Processing*, Kyoto.

Adam Pease. 2011. *Ontology: A Practical Guide*. Articulate Software Press, Angwin, CA.

Jonathan Phillips. 2005. Introduction to the open clip art library. http://rejon.org/media/writings/ocalintro/ocal_intro_phillips.html. (accessed 2007-11-01).

John Robert Ross. 1995. Defective noun phrases. In *Papers from the Regional Meeting of the Chicago Linguistic Society*, volume 31, pages 398–440. University of Chicago.

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.

Liling Tan and Francis Bond. 2012. Building and annotating the linguistically diverse NTU-MC (NTU-multilingual corpus). *International Journal of Asian Language Processing*, 22(4):161–174.

Naiwen Xue, Fei Xia, Fu-Dong Chiou, and Martha Palmer. 2005. The Penn Chinese TreeBank: Phrase structure annotation of a large corpus. *Natural Language Engineering*, 11(2):207–238.