Annotation of Pronouns in a Multilingual Corpus of Mandarin Chinese, English and Japanese

Yu Jie Seah and Francis Bond
Linguistics and Multilingual Studies
Nanyang Technological University
yjseah1@e.ntu.edu.sg,bond@ieee.org

Abstract

A qualitative and quantitative approach was used in this study to examine the distribution of pronouns in three languages, English, Mandarin Chinese and Japanese based on the parallel NTU Multilingual Corpus (NTU-MC). The pronouns are annotated with a componential analysis that allows them to be easily linked across languages. A single text (The Adventure of the Speckled Band, a short story featuring Sherlock Holmes) in three languages is tagged, annotated and linked in the corpus. The results show that although English has the highest number of pronouns, Mandarin Chinese has the highest proportion of contentful pronouns in our corpus. Also, English has more translated counterparts in Mandarin Chinese as compared to Japanese. We attributed this to the difference in usage of pronouns in the languages. Deprominalisation, surprisingly, was even for both corpora. Findings from this study can shed some light concerning translation issues on pronoun usage for learners of the languages and also contribute to improving machine translation of pronouns.
Keywords: pronoun, Chinese, English, Japanese

1. Introduction

Pronouns exist in all the world languages, although there is considerable variation in how they are used. In this paper, we offer a componential analysis of pronouns that is extended into three language (English, Mandarin Chinese and Japanese) from three totally different language families (Indo-European, Sino-Tibetan and Japonic), The way they are employed in different languages is interesting to many linguists. Furthermore, in such a globalized world like today, languages are always translated into other languages. Other than translation of content words, how pronouns are translated from language to language can allow one to learn a lot about the language and its translation. English, being the world's most globalized language, has been translated into many different languages. Comparing its translation to Mandarin Chinese and to Japanese can shed light on the usage of pronouns in each language.

There have been few corpus based studies on differences in pronoun use among languages. According to Kim (2009), there exist qualitative and quantitative differences in the usage of the second person and first person plural pronouns in texts he examined from English and Korean newspapers. In general, English uses pronouns more often, with the notable exception of the first person plural, which was more common in Korean. Our research is part of a wider study of conceptual differences between the languages (Bond et al., 2013). For this reason, we did no restrict ourselves to personal pronouns, but also considered indefinite pronouns, demonstratives and interrogative pronouns.

2. Approach

We proceeded in four steps:

1. Identify pronouns used in the corpus
2. Analyze them in terms of components
3. Tag the pronouns monolingualy in each language
4. Analyze the distribution cross lingually

2.1. Identify the pronouns

We started off by examining words tagged as pronouns in the NTU Multilingual Corpus (NTU-MC) (Tan and Bond, 2012). The NTU-MC exploits the linguistic diversity available in Singapore for the collection of a vast variety of texts from different languages. The current version is an annotated collection of around 6,000 sentences (595,000 words) in 7 languages (Arabic, English, Mandarin Chinese, Japanese, Korean, Indonesian and Vietnamese) from 7 language families (Afro-Asiatic, Indo-European, Sino-Tibetan, Japonic, Korean (language isolate), Austronesian and Austro-Asiatic). Two kinds of annotation are applied in the NTU-MC –monolingual annotation where texts are tagged for parts of speech (POS) and sense; and crosslingual annotation where texts are aligned across sentences (Bond et al., 2013; Wang and Bond, 2014).

Pronouns from the three languages (English, Mandarin Chinese and Japanese) were extracted from four data sets in the NTU-MC. They are two short stories from Sherlock Holmes –*The Adventure of the Speckled Band* and *The Adventure of the Dancing Men* (Conan Doyle, 1892; Conan Doyle, 1905), an essay named *The Cathedral and the Bazaar* (Raymond, 1999) and on-line articles about Singapore tourism (Singapore Tourist Board, 2012). In each set, English is the source language while Mandarin Chinese and Japanese translation texts are aligned to it at the sentence level. The texts have been tokenized and automatically POS tagged.

We took as pronouns anything marked as a pronoun by the part of speech tagger.[1]. This includes personal pronouns, indefinite pronouns and interrogative pronouns. Each language had slightly different part-of-speech tags, with slightly different coverage. We ended up with 60 different English pronouns, 54 Chinese and 69 Japanese. The greater number of Japanese types reflects the greater orthographic variation: the same pronoun can be written in Chinese characters (彼 *he* "kare") or using hiragana (かれ *he* "kare").

[1] PN, WB, WRB, PRP, PRP$, WP, WP$, 名 詞-代名詞-一般

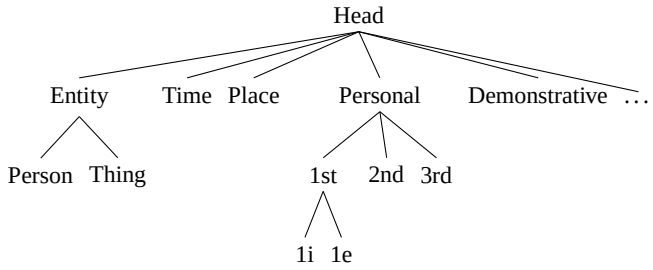

Figure 1: Head Types

2.2. Classify the Pronouns

The next stage was to analyze them componentially. The pronouns were separated into nine categories: Head, Number, Gender, Case, Type, Formality, Politeness, and Distance from Speaker. The features chosen are in line with other research and reference grammars (Backhouse, 1993; Li and Thompson, 1989; Huddleston, 1988). The purpose of this componential analysis is to code the pronouns so that we can compare and contrast them across languages. This also allows the auto-tagging programme to recognize and link the pronouns by their code. This stage took around two weeks due to the detailed componential analysis of every pronoun in the four subcorpora and analyzing ambiguous forms particularly in Japanese. The different features under each heading are shown in Table 1.

2.2.1. Head

In the first column - Head, there are altogether ten components. These are Demonstratives, Entity, Time, Manner, Person, Place, Reason, Thing, Personal and Quantifier. This feature restricts the kind of the referent, or says that the pronoun is a quantifier and thus has no restriction. We show them in Figure 1.

Every pronoun extracted will be tagged with one of these features. For example, Demonstratives refer to pronouns such as *this* and *that* while Entity are pronouns that do not have a specific category of referent, as it can refer to both person and object. Such pronouns are *all*, 俩 *lia3* "both" and いくつ *ikutsu* "some". *when*, *how*, *why* and *where* are examples of pronouns labeled under Time, Manner, Reason and Location respectively. For English pronouns, words that end with *-thing* are grouped under Thing, while for Mandarin Chinese and Japanese pronouns, they are not so clear-cut. Lastly personal pronouns and pronouns that talk about people like *everybody* and 自己*zi4ji3* "self" are categorized under Person.

Personal pronouns are further divided into 1st, 2nd and 3rd person. 1st person is then divided into exclusive and inclusive (anticipating tagging Indonesian, which makes this distinction).

Although strictly speaking not pronouns, determiners and adjectives that are closely related to pronouns (such as *both* and *many*)were also analyzed and labeled as Quantifiers. For example, we annotate *both* in both (1) and (2) of the following two sentences. They share many of the other features, so it makes sense to analyze them together. We did not attempt to cover all determiners, only those that shared some characteristics with the pronouns.

(1) *I talked to both* Entity

(2) *I talked to both authors* Quantifier

2.2.2. Number

For this feature, we identified three kinds of number –Dual, Plural and Singular. *both* is an example of Dual, *those* for Plural and 这 *zhe4* "this" for Singular. Many pronouns are not specified for number.

2.2.3. Gender

For the third column - Gender, three features were identified as well –Masculine, Feminine and Neuter. *it* in English is a neuter pronoun while 她 *ta1* "she" in Chinese is Feminine. Most pronouns are not marked for gender.

2.2.4. Case

Only English marks case. We distinguish Subjective (nominative), Objective (accusative) and Possessive pronouns: e.g. *I, me, my*. Extending to other languages may require further distinctions.

2.2.5. Type

Type differentiates the pronouns by Assertive Existential (*somebody*), Elective Existential (*anybody*), Negative (*nobody*), Reflexive (*myself*), Reciprocal (*each other*), Universal (*everybody*), Interrogative (*who*) or Other (anything else). In a decomposed semantics, we would treat all but Reflexive and Reciprocal as quantifiers: *anybody* thus becomes the equivalent of the quantifier *any* and the noun *person*.

2.2.6. Formality

The sixth column shows Formality, whether the pronouns are informal or formal. This is mainly for the Japanese pronouns, which mark for formality: 僕*boku* "I" is informal whereas 私*watashi* "I" is formal.

2.2.7. Politeness

Japanese and Chinese also encode how respected the referent is, which we call Politeness. 您 *nin3* "you" in Chinese is used to refer to high status people.

Note that Formality and Politeness are somewhat different from the T-V distinctions made in European languages which typically only mark second person, and show the relation between speaker and hearer (historically a power difference, now more often a difference in familiarity between the speakers). Japanese pronouns encode a more absolute level of respect for their referent.

2.2.8. Proximity

The final feature is meant for pronouns that mark for Proximal, Medial or Distal distance from the speaker. These pronouns are used for demonstratives (*this* "distal entity") and by extension location pronouns such as あそこ *soko* "there: distal place" and time pronouns (*then* "distal time"). Chinese and English only have a two way distinction (proximal and distal: *this* and *that*). Japanese has a three way system: これ *this* "kore", それ *sore* "that" and あれ *are* "that over there".

Head	Number	Gender	Case	Type	Formality	Politeness	Proximity
Demonstratives	Dual	Feminine	Objective	Assertive	Formal	Polite	Distal
Entity	Plural	Masculine	Possessive	Elective	Informal		Medial
Time	Singular	Neuter	Subjective	Negative			Proximal
Manner				Other			
Person				Reciprocal			
Place				Universal			
Reason				Interrogative			
Thing				Reflexive			
Personal (1e, 1i, 2, 3)							
Quantifier							

Table 1: The 8 types of pronoun features

2.2.9. Summary

The features are used to define a concept, which we treat as a wordnet synset (Fellbaum, 1998). A single synset may have multiple lemmas associated with it: for example, the synset with features (Person, Assertive) has two English lemmas *someone* and *somebody*. We also linked the types to appropriate wordnet senses (for example Person is `person`$_{n:1}$, Place is `location`$_{n:1}$). The other components were kept as a separate table, linked using the wordnet synset IDs. We ended up with 107 different synsets for the 60 English, 54 Chinese and 69 Japanese pronouns.

2.3. Monolingual Tagging

After analyzing the pronouns by their different components we added them to our local wordnets' sense inventories (14 were already there, mainly interrogatives and indefinite pronouns). For English we use the Princeton Wordnet (Fellbaum, 1998), for Chinese the Chinese Open Wordnet (Wang and Bond, 2013) and for Japanese the Japanese Wordnet (Isahara et al., 2008). Treating the pronouns as synsets enabled us to use our existing wordnet tagging tools.

We carried out tagging on a single subcorpus: *The Adventure of the Speckled Band* and its Chinese and Japanese translations. We chose it as it had more (reported) speech than the other genres, and was thus had a greater variety of pronouns. We show the numbers of pronouns found in each language in Table 2. This includes all types, including quantifiers.

The main issue in the monolingual tagging was distinguishing what we shall call contentful pronouns (such as those described above) from purely structural pronouns such as dummy *it*, existential *there*, relative pronouns (*the dog who barked*) and pronouns in idiomatic expressions (*Oh My God!*). We expect contentful pronouns would introduce a quantifier into a formal semantic representation, while the structural ones would not.

In addition, there were some tokenization errors, mainly in the Chinese and Japanese corpora. These we fixed as we carried out the annotation.

2.4. Cross-lingual tagging

In the initial annotation, each pronoun was linked to the pronoun in the corresponding translation with the best feature match. If there was a tie, the leftmost pronoun pair was linked first, then the next and so on. The annotator then went through the bilingual corpus and checked each pair.

Language	English	Chinese	Japanese
Contentful	1,370	1,177	463
Other	75	19	51
Total	1,445	1,196	514
Sentences	599	620	702
Words	11,628	12,433	13,902

Table 2: Number of pronouns found in the corpora

At this stage they checked both whether they are tagged as pronouns correctly by the auto-tagging programme and whether the concept links between the source language and target language are accurate. This was done several times to ensure accuracy. This stage took around four weeks to complete both English-Chinese and English-Japanese corpora, with a longer time needed for the English-Chinese one due to the greater number of pronouns present there. On average, three to four sentences can be done every hour. An example of matching pronouns is given in (3) where the English is followed by Chinese. The first two English pronouns match the Chinese, the third has no equivalent.

(3) a. You see that we have been as good as our word

b. 你 瞧， 我们 是 说到做到
ni3 qiao2, wo3men shi4 shuo1dao4zuo4dao4
的
de4

'You see, we do what (we) say'

3. Results

Having linked and tagged the relationships between words, we proceeded to count the number of pronouns in each language and their links. The number of contentful and non-contentful (structural or segmentation errors) are shown in Table 2. Differences in word and sentence tokenization give different numbers of words and sentences for the three languages, even though the content is basically the same. Even allowing for these light differences, English has more pronouns than Chinese which has far more than Japanese. The non-contentful pronouns are mainly structural for English, while they are mainly tokenization errors for Chinese and Japanese.

The results for the linkage of the pronouns are separated into two parts for better understanding — the first part being the results for the English-Chinese corpus (Table 3) and the

	Linked Pronouns						Non-linked Pronouns	
	# Matching Features					Pronoun	English	Chinese
	5	6	7	8	9	to Noun		
# Pronouns	5	19	54	789	58	134	369	215

Table 3: English-Chinese pronoun translation

	Linked Pronouns						Non-linked Pronouns	
	# Matching Features					Pronoun	English	Japanese
	5	6	7	8	9	to Noun		
# Pronouns	15	120	114	37	32	139	943	109

Table 4: English-Japanese pronoun translation

second part for the results found from the English-Japanese corpus (Table 4).

There are in total 925 English to Chinese pronouns linked to each other, with 0.5% of them having only 5 pronoun features match, 2.1% having 6 pronoun features match, 5.8% having 7 pronoun features match, 85.3% having 8 features match and 6.3% having 9 pronoun features match where 9 is the maximum match. Most pronouns match everything except Case. Those that matched exactly were mainly indefinite pronouns, which don't show case.

There are also 134 pronouns that are linked to non-pronouns. 76 of them are English pronouns while 58 of them are Chinese pronouns. These typically linked to common nouns.

Out of the 1,370 contentful English pronouns, 26.9% of them are not linked. For the Chinese contentful pronouns, only 18.2% were not linked to anything.

For English and Japanese, far fewer pronouns were linked. There are in total 318 linked English to Japanese pronouns. Out of these, 4.7% have 5 matched features, 37.7% have 6 matched features, 35.8% have 7 matched features, 11.6% have 8 matched features and 10% have 9 matched features. The majority of the linked English-Japanese pronouns, unlike the English-Chinese corpus, have around 6 to 7 matched features. This is because they typically mismatch on both Case in English and Politeness or Proximity in Japanese.

Similar to the English-Chinese corpus, there are 139 pronouns in the English-Japanese corpus that are linked to non-pronouns. 109 of the pronouns are English pronouns and the other 30 are Japanese pronouns. In contrast to the English-Chinese corpus, most (68.8%) of the English contentful pronouns do not link to anything at all. Surprisingly, for the Japanese pronouns, 23.5% of them are not linked to any English words in the English source text.

4. Discussion

English has the most pronouns, followed by Mandarin Chinese and lastly Japanese. If we include non-contentful pronouns (such as dummy *it*, existential *there* and also complementizers like *that* and *which*), this becomes even more pronounced. Also, in English, many pronouns can also double up as determiners (Collins COBUILD, 2005). Determiners share many common words with pronouns such as *this*, *that* and indefinite ones such as *all* and *some*. In contrast, Chinese almost only uses contentful pronouns, and Japanese tends to drop pronouns altogether.

English personal pronouns have more different forms: Subjective, Accusative and Possessive. English also has other categories of pronouns that both Mandarin Chinese and Japanese do not have. For example, for the component Negative, English has *none* and *nothing* which do not have identical correspondents in Mandarin Chinese and Japanese: which do not negate inside noun phrases. This is because both languages tend to use verbs to express negativity instead of marking it in the pronoun like in (4) where the English is followed by Chinese and Japanese.

(4) a. . . . but none commonplace

b. 但是 却 没有 一 例 是
Dan4shi4 que4 mei2you3 yi1 li4 shi4
平淡无奇 的
ping2dan4wu2qi2 de
'But, there is not one case that is featureless.'

c. どれ も 尋常では ない事件 である
Dore mo jinjode wa nai jiken dearu
'There is not any unusual case.'

In addition, Mandarin Chinese and Japanese are topic-prominent languages (Li and Thompson, 1989; Obana, 2000). Once the topic is established, sentences following it omit any pronouns, as there is no need for them to refer back as the readers can infer from contextual knowledge the subject of the sentence.

Furthermore, out of the three languages, only Japanese marks politeness and some evidentiality on the verb (Backhouse, 1993), making the use of pronouns rather unnecessary and this seems to play an important role in reducing the numbers of pronouns found in the corpus as compared to the English source text and Chinese translation text, resulting in the low rate of links to the English pronouns in the original text. One example can be seen below in (5), with English and Japanese:

(5) a. I have heard of you, Mr. Holmes

b. あなたの ことは、以前からお聞きして
Anata no koto wa, izen kara o kiki shite
います。
imasu .
'About you, (I) humbly heard previously.'

Between the English-Chinese corpus and the English-Japanese corpus, another major difference is the number of corresponding features that majority of the linked pronouns have. For the English-Chinese corpus, majority of the linked pronouns have 8 matching pronoun features while for the English-Japanese corpus, majority of the linked pronouns have around 6 to 7 matching pronoun features. This is most likely due to Japanese language having different speech levels (Obana, 2000). The different speech levels cause a differentiation between the pronouns, resulting in Japanese having a few different words for the same pronoun. For example, for the first person pronoun, in Japanese there are variations such as わし*washi* which also marks for masculine speaker and informal and 私*watashi* which marks for formal and politeness. These features do not exist in English but from the perspective of semantics, they should be linked to the first person pronouns in English. This problem does not exist in Mandarin Chinese, as there is no such differentiation in speech levels in Mandarin Chinese. Therefore, more features can be matched.

Also, from the linking of the pronouns, there were many cases where English pronouns were linked to Mandarin Chinese and Japanese pronouns that are different in meaning such as the third person pronoun *it* in the English text to the demonstrative pronoun それ*sore* "that" or even to そこ*there* "there" in Japanese. Although this happens in the English-Chinese corpus as well, they are less frequent, thus resulting in more of the pronouns linked have more matched features as compared to those in the English-Japanese corpus. We give an example of this in (6), with English and Chinese, where *it* is linked to 这 *zhe4* "this".

(6) a. It is a swamp adder!
b. 这 是 一 条 沼地 蝰蛇!
Zhe4 shi4 yi1 tiao2 zhao3di4 kui2she2
'This is a swamp adder!'

Deprominalisation (a pronoun linking to a noun) occurs almost evenly in both the English-Chinese and English-Japanese corpora As seen in the results, the number of pronouns matched to non-pronouns in the English-Chinese corpus is around the same. This result is not expected as deprominalisation was predicted to occur much more frequently in the English-Japanese corpus than in the English-Chinese corpus. It could be a case of the source language effecting the translation: although native speakers said the translations were good, they almost certainly have more pronouns than texts written originally in Chinese or Japanese.

From the tagging of the pronouns and their concept links, there were a few interesting cases that were found. In the English source text, we realized that pronouns often exist in idiomatic phrases. However, these pronouns do not actually have any particular antecedent to refer to as they are almost always used in the same way regardless of its environment and this means that they cannot be linked.

(7) a. My God!
b. 天哪 !
Tian1na
'Heaven!'
なんてこったい！
Nan te kottai
'What the heck'

We see in (7) that *my* is used here as a pronoun in an idiomatic phrase and after translation, no pronouns were seen. In both the Mandarin Chinese and Japanese text, the idiomatic translation has no pronoun in it. We give another example in (8).

(8) a. It is very kind of you.
b. 非常 感谢!
Fei1chang2 gan3xie4
'Very grateful'
c. 感謝 しているよ
Kansha shite iru yo
'(I) am grateful.'

We give one final example in (9). The Chinese translation here again choses to take the figurative meaning of *I am in your hands* and translated it to "I will obey all your instructions". However, in the Japanese text, this is literally translated, possibly because the phrase is commonly used in translating prayers and is thus somewhat established.

(9) a. I assure you that I am in your hands.
b. 我 向 你 保证, 我 一切
Wo3 xiang4 ni3 bao3zheng4, wo3 yi2qie4
听从 你 的吩咐
ting1cong2 ni3 de fen1fu4.
'I promise you, I will obey all your instructions'
c. あなたの 手にすべてをおゆだねします
Anata no te ni subete o o yudane shimasu
わ
wa
'I will leave everything in your hands'

Another interesting note was that other than pronouns, both Mandarin Chinese and Japanese tend to use classifier phrases anaphorically. Numeral classifiers (like the *head* in *two head of cattle*) are used for most nouns in Japanese. The classifier can combine with numerals, interrogatives and in Chinese determiners. The resulting phrase can be used anaphorically: for example 那*na4jian1* "that room (CLASSIFIER)" which can mean '"hat house/room". Without the need of the proper noun in Mandarin Chinese, the determiner+classifier word can be used to refer to a certain room, thus acting like a pronoun. Although classifiers are not as widely used in English as in Mandarin Chinese and Japanese, numerals in English can sometimes take on anaphoric roles as well.

The annotation scheme we use here has two parts: the lexicon, which in this case is richly structured with components, and the corpus, which allows annotation of concepts and links between them. The two have to be kept synchronized.

5. Future Work

We would like to extend the annotation in a few ways. One is to tag more texts in the NTU-MC. The pronoun distributions in this paper are solely extracted from one story and thus we cannot generalize the results across genres.[2] The second is to add more languages to the pronoun analysis: our next language will be Indonesian, again from a different language family. We also want to extend the componential analysis to related words such as terms of address and numeral classifiers. Chinese, Japanese and Indonesian all use kinship terms to refer to non-kin: you may address a stranger as *uncle* or *older sister*.

We would also like to examine further the cases of pronouns linking to different pronouns and non-pronouns: Are the synsets always compatible? and what cues drive the choice of pronoun or demonstrative or common noun phrase? We hope that the crosslingual analysis will give some insights into the different strategies employed in the different languages.

Our distinction between contentful and structural pronouns is still only informally described. We would like to sharpen this distinction.

Finally, our analysis is compatible with (and partly inspired by) the decompositional analysis of pronouns in the English Resource Grammar (ERG), an HPSG implementation of English (Flickinger, 2000). We would like to check that all our pronouns are in the ERG and add them to the corresponding grammars of Chinese, Indonesian and Japanese. The HPSG grammars distinguish clearly between contentful and structural pronouns, and could be used to help in the monolingual annotation.

The annotated corpora and extended wordnets will be made available from the NTU-MC website: `http://compling.ntu.edu.sg/ntumc`. The corpus is licensed with the Creative Commons Attribution Only License (CC BY)[3], and the wordnets under their respective (open) licenses.

6. Conclusions

In this paper we introduced an annotation scheme for pronouns based on a componential analysis. It was tested on three languages, and used to tag a Chinese, English and Japanese tritext. The results show that pronouns, though universal, are used differently across languages, resulting in a difference in distribution among the three languages and a difference in the concept links between the English-Chinese corpus and English-Japanese corpus. We have began to account for these differences and presented examples of some interesting cases.

With this study, we hope that translation issues regarding pronoun usage would be useful and clearer to those who are learning the language and that the material from this study can contribute to pronoun translation across languages.

Acknowledgments:

This work was supported in part by the Singapore Ministry of Education (Academic Research Fund Tier 2: MOE2013-T2-1-084). Thanks to Shan Wang, David Moeljadi and an anonymous reviewer for their helpful comments.

[2] Ideally, we would like to also annotate text with Chinese and Japanese as their source to control for translationese.

[3] `creativecommons.org/licenses/by/3.0/`

7. References

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

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.

Collins COBUILD. 2005. *English grammar*. Harper Collins, 2 edition.

Arthur Conan Doyle. 1892. *The Adventures of Sherlock Homes*. George Newnes, London.

Arthur Conan Doyle. 1905. *The Return of Sherlock Homes*. George Newnes, London. Project Gutenberg `www.gutenberg.org/files/108/108-h/108-h.htm`.

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

Dan Flickinger. 2000. On building a more efficient grammar by exploiting types. *Natural Language Engineering*, 6 (1):15–28.

Rodney Huddleston. 1988. *English grammar: an outline*. Cambridge University Press, Cambridge.

Hitoshi Isahara, Francis Bond, Kiyotaka Uchimoto, Masao Utiyama, and Kyoko Kanzaki. 2008. Development of the Japanese WordNet. In *Sixth International conference on Language Resources and Evaluation (LREC 2008)*, Marrakech.

Chul-Kyu Kim. 2009. Personal pronouns in English and Korean texts: A corpus-based study in terms of textual interaction. *Journal of Pragmatics*, 41:2086–2099.

Charles N. Li and Sandra A. Thompson. 1989. *Mandarin Chinese: A Functional Reference Grammar*. University of California Press.

Yasuko Obana. 2000. *Understanding Japanese: A handbook for learners and teachers.* Kurusio Publishers.

Eric S. Raymond. 1999. *The Cathedral & the Bazaar*. O'Reilly.

Singapore Tourist Board. 2012. Your Singapore. Online: `http://www.yoursingapore.com`. [Accessed 2012].

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.

Shan Wang and Francis Bond. 2013. Building the chinese open wordnet (cow): Starting from core synsets. In *Proceedings of the 11th Workshop on Asian Language Resources, a Workshop at IJCNLP-2013*, pages 10–18, Nagoya.

Shan Wang and Francis Bond. 2014. Building sense-tagged multilingual corpora. In *Ninth International Conference on Language Resources and Evaluation (LREC 2014)*, Reykjavik.