Proceedings of ISA-18 Workshop at LREC2022, pages 86–94
Marseille, 20 June 2022

Multilingual Reference Annotation: A Case between English and Mandarin Chinese

Joanna Ut-Seong Sio, Luis Morgado da Costa
Palacký University Olomouc, The Czech Republic
Katedra asijských studií FF UP, tř. Svobody 26, 779 00 Olomouc
joannautseong.sio@upol.cz, lmorgado.dacosta@gmail.com

Abstract

This paper presents the on-going effort to annotate a cross-lingual corpus on nominal referring expressions in English and Mandarin Chinese. The annotation includes referential forms and referential (information) statuses. We adopt the RefLex annotation scheme (Baumann and Riester, 2012) for the classification of referential statuses. The data focus of this paper is restricted to [*the*-X] phrases in English (where X stands for any nominal) and their translation equivalents in Mandarin Chinese. The original English and translated Mandarin versions of 'The Adventure of the Dancing Men' and 'The Adventure of Speckled Band' from the Sherlock Holmes series were annotated. It contains 1090 instances of [*the*-X] phrases in English. Our study uncovers the following: (i) bare nouns are the most common Mandarin translation for [*the*-X] phrases in English, followed by demonstrative phrases, with the exception that when the noun phrase refers to locations/places, in such cases, demonstrative phrases are almost never used; (ii) [*the*-X] phrases in English are more likely to be translated as demonstrative phrases in Mandarin if they have the referential status of 'given' (previously mentioned) or 'given-displaced'(antecedent of an expression occurs earlier than the previous five clauses). In these Mandarin demonstrative phrases, the proximal demonstrative is more often used and it is almost exclusively used for 'given' while the distal demonstrative can be used for both 'given' and 'given-displaced'.

Keywords: referential status, referring expressions, noun phrases, definiteness, English, Mandarin Chinese

1. Introduction

1.1. Reference and referential statuses

In linguistic communication, noun phrases (e.g., *a cat*, *the song*, *that professor*) are prototypically used to refer to entities. This referring relationship is called reference. Coherence in discourse depends crucially on using an appropriate referential expression for introducing, referring, and re-introducing an entity after a long pause. This is an easy task for any competent speaker, but such knowledge is notoriously difficult to make explicit.

To pick the right referring expression is to correctly map referential forms (noun phrases) to discourse contexts. Referring expressions can be definite (e.g., *the cat*, *that cat*, *it*) or indefinite (e.g., *a cat*). Definite expressions are used to refer to identifiable entities. The former is a grammatical category and the latter a cognitive category. Identifiability can be understood as a speaker's assessment of whether a particular discourse referent is already stored in the hearer's mind or not (Lambrecht, 1996). Entities can be identifiable in different discourse contexts (Christophersen, 1939; Hawkins, 1978; Lyons, 1999). They could be identifiable from previous mentions (e.g., *I bought a shirt yesterday. The shirt was blue.*), shared general knowledge (e.g., *the sun*), association (e.g., *I bought a shirt yesterday, but the sleeves were too long.*), or being unique in the speech environments (e.g., *Please close the window.*), to name a few. Indefinite expressions are used to introduce new and non-identifiable entities (e.g., *I saw a cat yesterday*, where the particular cat cannot be identified and/or its identity is not relevant for the discourse).

These discourse contexts (e.g., previous mentions, shared knowledge, new/non-identifiable, etc.) can be understood as the relevant entity having different referential statuses and will cause the speaker to select a particular referential form to refer to it (Baumann and Riester, 2012).[1]

Reference is an essential and yet unsolved problem in Natural Language Processing (NLP) and this limits the reach of various applications (e.g., parsing, machine translation, language generation and information retrieval). A better understanding of the mapping between referential forms and referential statuses would have a tremendous impact in all NLP tasks that deal with language understanding or generation.

1.2. Expression of definiteness in English and Mandarin Chinese

Different languages can have different inventories of referring expressions. In languages like English, articles (*the*, *a/an*) provide one way of distinguishing definite and indefinite NPs. In article-less languages (e.g., Chinese and Slavic), there are more interpretative ambiguities of surface forms and word order might also play a role. It has been proposed that the Mandarin equiva-

[1] Different referential statuses are often cast in terms of accessibility, which can be understood as a property of memory representation, with some information being more privileged/accessible/salient/prominent (Arnold and Zerkle, 2019). It is usually framed as a continuum from low to high (Gundel et al., 1993).

lents of [*the*-X] phrases in English can be a bare noun, which is ambiguous between a definite and an indefinite reading (Cheng and Sybesma, 1999) or a nominal containing a demonstrative (a demonstrative phrase) (Chen, 2004). An example of 'the dog' with a bare noun as an equivalent is given in (1); an example of 'the dog' with a demonstrative phrase as an equivalent is given in (2).

(1)	狗	要	過	馬路。
	gǒu	yào	guò	mǎlù
	dog	needs	cross	road

'The dog needs to cross the road' (Cheng and Sybesma, 1999)

(2)	有	一	個	獵人	養著	一	隻	狗，
	yǒu	yī	gè	lièrén	yǎng-zhe	yī	zhī	gǒu,
	have	one	CL	hunter	keep-ASP	one	CL	dog.

這	隻	狗	很	懂事。
zhè	zhī	gǒu	hěn	dǒngshì
this	CL	dog	very	intelligent

'There was a hunter who had a dog, the dog was very intelligent.' (Chen, 2004)

Demonstratives in Mandarin have been claimed to share some of the functions of the definite article in English (Chen, 2004). A comparison of the original English version of *Alice in Wonderland* and its Mandarin translations shows that demonstratives used in the Mandarin versions outnumber those in the English text, up to 3 times more in one translated version (Lu et al., 2018). This suggests the usage of the Mandarin demonstratives are less restricted than the English ones.

1.3. Goal of the paper

The annotation effort reported in this paper serves several goals: (i) to map [*the*-X] phrases in English to its equivalents in Mandarin; (ii) to map referential information statuses to referential forms within each language; (iii) to test and revise the RefLex annotation scheme (Baumann and Riester, 2012) using both English and Mandarin data.

As discussed earlier on, Mandarin bare nouns can be definite (Cheng and Sybesma, 1999) and demonstrative-containing phrases in Mandarin are always definite. We are particularly interested in finding out whether these two kinds of referring expressions are indeed the Mandarin translation equivalents of [*the*-X] phrases in English, as suggested in the literature, and if so, what the distribution is like with respect to the different referential information statuses.

The rest of this paper is structured as follows. In Section 2 we will describe the methodology of our analysis. Section 3 discusses the key findings of the annotation effort. Section 4 concludes and points to new directions for future studies. Sections 5 and 6 include some notes on the release of the tagged data and all necessary acknowledgments respectively.

2. Methodology

2.1. The Data

This project is currently using texts from the NTU Multilingual Corpus (Tan and Bond, 2014) – an open corpus with parallel data in multiple languages containing, among other genres, the full canon of Sherlock Holmes, by Sir Arthur Conan Doyle. The work presented in this paper is based on the annotation of two full short stories: *The Adventure of the Specked Band* (Conan Doyle, 1892) and *The Dancing Men* (Conan Doyle, 1905).

The English version of the *The Adventure of the Specked Band* short story contains 599 sentences and 11,741 words. Its Mandarin translation contains 620 sentences and 12,444 words. The English version of the *The Dancing Men* short story contains 666 sentences and 12,602 words. Its Mandarin translation contains 606 sentences and 11,339 words.

Short stories made an excellent data source for this project since its nature required the use of text suitable to analyze discourse structure. This means that we could not use use corpora comprised only of single sentences or text snippets since many discourse features are only present in longer narratives. At the same time, in order to future-proof our project and plan for future analyses that may require us to trace entities across their entire discourse life, we needed to find narratives that were not too long.

In addition, we were happy to choose the NTU Multilingual Corpus because it was readily available under an open license, allowing us to openly share all new layers of annotation we produce. Furthermore, both short stories chosen for this project have had a substantial amount of annotation and analysis from previous projects including sense-tagging using the Princeton Wordnet for English (Fellbaum, 1998) and the Chinese Open Wordnet for Mandarin (Wang and Bond, 2013). Even though we are not currently making full use of this layer of annotation, we believe that the information made available through sense tagging (e.g., relations of hyponymy and hypernymy across different discourse entities) will be of great valuable for the future directions of this project. Finally, the fact that the NTU Multilingual Corpus contains not only more short stories, but also parallel translations of some of these stories in other languages (including Japanese, Dutch, German, Indonesian and Italian) made this corpus extremely interesting to support the study of multilingual and cross-lingual referential analysis.

2.2. Expanding IMI: a Multilingual Semantic Annotation Environment

An important step of the methodology of this project was to decide on an annotation system to support our current and future goals within this project. While there were no shortage of options – e.g., Slate (Kummerfeld, 2019) or SALTO (Burchardt et al., 2006) have both been used as annotation benches for the RefLex Scheme –, we ended up choosing to expand IMI – a multilingual semantic annotation environment (Bond et al., 2015)[2]. Despite each annotation system having their strengths, our decision to use IMI was based on the fact that it was an online, open-source project specifically designed for multilingual semantic annotation – able to enrich a corpus with multiple layers of morphosyntactic and semantic information, as well as interfaces to manage cross-lingual links between sentences, concepts and words.

IMI was originally designed for sense tagging, using Open Multilingual Wordnet (Bond and Foster, 2013). It provides multiple layers of annotation that include lemmatization, POS tagging, sense tagging, sentiment annotation and interlingual-mapping. It is developed in Python and SQLite, and supports both concurrent annotation (i.e., multiple taggers tagging the same data at the same time), as well as parallel tagging (i.e., multiple taggers tagging the same set of data in parallel, using multiple databases). This annotation tool has been tested for a wide selection of languages, including English, Mandarin, Japanese and Indonesian. Finally, its flexibility and ease of customization had been proven by the development of multiple project-specific layers of annotation including: sentiment analysis (Bond et al., 2016), grammatical error analysis (Winder et al., 2017), and semantic role labeling (Choi, 2019).

Another strong motivation to use IMI was the fact that this system was designed to develop the NTU Multilingual Corpus (Tan and Bond, 2014) – the open corpus used for this project. This means that the short stories tagged in the context of this project were already in the required format to be used with this annotation system.

Within the context of this project we developed a new annotation interface within IMI which we named 'The RefLex Corpus Tagger' (see Figure 1). This interface allows the use of any custom tagset, which can be organized in different classes/types of tags. Tagsets can be language specific or shared across all languages. The system allows the tagset to grow incrementally, which is ideal for projects of exploratory nature such as this one. In future iterations, new tags can be easily added to the interface without jeopardizing the integrity of the data.

The annotation process is done sentence by sentence (but the tool also provides annotators with access to the full text, for reference). To add a new annotation, annotators can select a single word or any number of words (contiguous or non-contiguous) – referred as *chunks* within the system. Multiple tags can be provided for the same *chunk*, allowing the adoption of flexible tagset of varied classes/types. Finally, total and partial overlap of *chunks* are also allowed within the tagging system – which was an essential feature to allow the independent annotation of embedded phrases within larger phrases.

2.3. Annotation Schema

Currently, the RefLex Corpus Tagger has two separate tagsets: one for English and one for Mandarin. Each tagset is divided in different layers of annotation. The English tagset has three such layers: NP Structure (referential forms), Referential Status and Modification. The Mandarin tagset has an additional layer – i.e., Semantic Class –, totaling four layers. For each *chunk* created within a sentence, annotators had to select one value for each layer, which is selected from a predefined dropdown box (see Figure 2 for the interface of the Mandarin Chinese Tagger and Figure 3 for a full list of available tags, English and Mandarin Chinese combined).

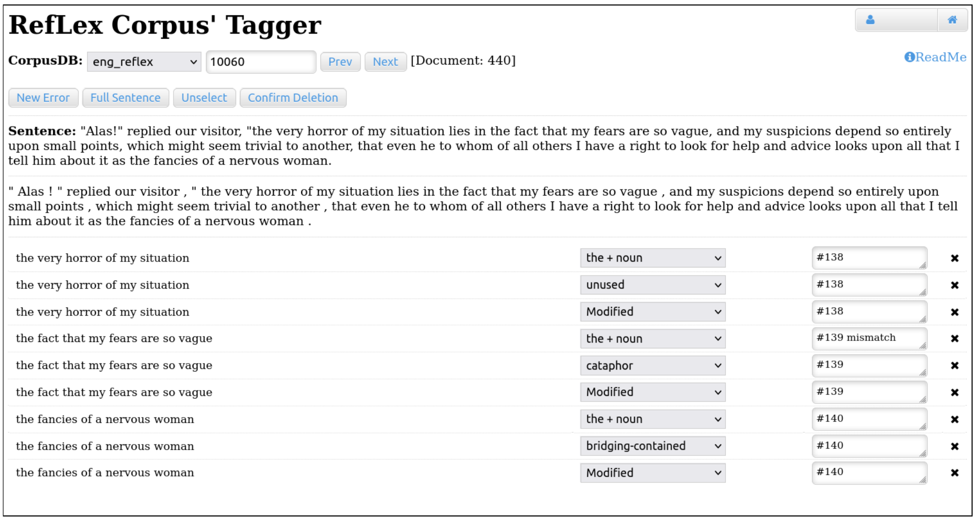
The inventories of referential forms are different in English and in Mandarin. In this study, we only focus on [*the*-X] phrases in English (e.g., *the boys*), this also includes [*the*-numeral-X] phrases (e.g., *the three boys*) and [*the*-numeral] (e.g., *the three*) phrases, though the last two types only constitute 3% of the total, see Table 1. For Mandarin, we include a full range of possible referential forms (see Figure 3). It should be noted that the NP structure includes only the functional elements within a noun phrase. Elements such as modifiers do not affect the classification. For example, a modified bare noun in Mandarin is still considered a bare noun.

Regarding referential statuses, we annotate using the RefLex Scheme (Baumann and Riester, 2012). RefLex provides a fine-grained classification scheme to annotate texts on two levels, referential and lexical. On the referential level, the scheme provides a list of contexts for the use of referring and non-referring expressions. These contexts are distinguished based on referential statuses (e.g., different kinds of anaphoric contexts and discourse-new contexts). The lexical level provides explicit evaluation of the degree of relatedness of lexical expressions (e.g., hypernym). In this paper, we only focus on the referential level. We plan to deal with the lexical level at a later stage of the project, by exploiting the existing sense annotations available through the NTU Multilingual Corpus and consider a fuller range of lexical relations provided by wordnets.

The tags used for the referential level are as follow: **given-sit**: an expression whose referent is immediately present in the text-external context; **given**: an expression whose referent mentioned in previous discourse context; **given-displaced**: an expression whose referent is mentioned in the previous discourse context earlier than 5 clauses before; **cataphor**: an expression whose referent is established only in the subsequent text; **bridging**: a non-coreferential anaphoric expression which is dependent on a unique referent established in a previously introduced scenario; **bridging-contained**: a non-coreferential anaphoric expression that is anchored to an embedded phrase; **unused**: a discourse-new expression which is unique; **new**: an expression denoting a discourse-new and non-uniquely identifiable referent;

In addition to our main interests, referential forms (NP structure) and referential statuses, we have also included the layers of annotation indicating whether the

[2] https://github.com/bond-lab/IMI



RefLex Corpus' Tagger

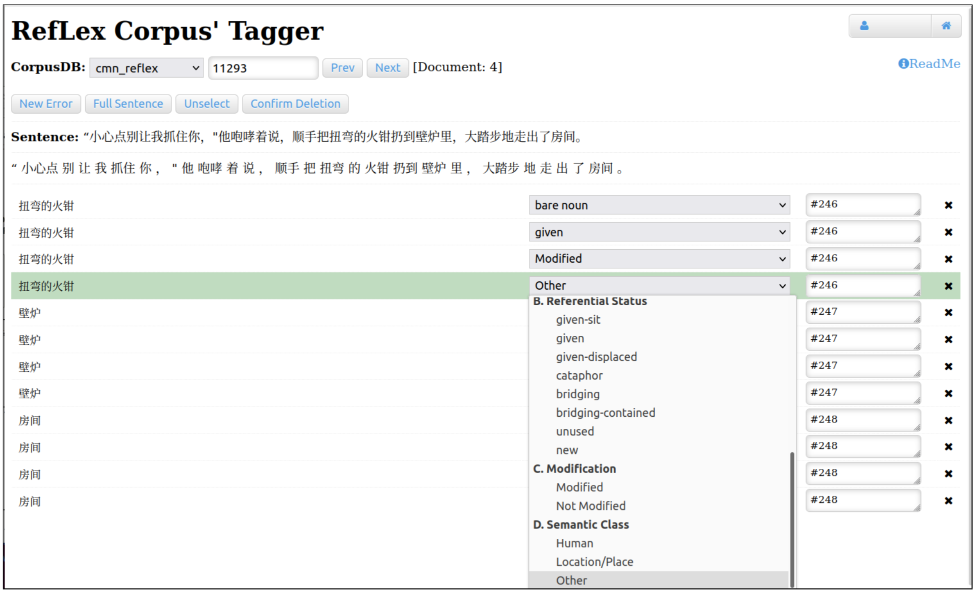
CorpusDB: eng_reflex 10060 [Document: 440]

Sentence: "Alas!" replied our visitor, "the very horror of my situation lies in the fact that my fears are so vague, and my suspicions depend so entirely upon small points, which might seem trivial to another, that even he to whom of all others I have a right to look for help and advice looks upon all that I tell him about it as the fancies of a nervous woman.

" Alas ! " replied our visitor , " the very horror of my situation lies in the fact that my fears are so vague , and my suspicions depend so entirely upon small points , which might seem trivial to another , that even he to whom of all others I have a right to look for help and advice looks upon all that I tell him about it as the fancies of a nervous woman .

the very horror of my situation	the + noun	#138	✖
the very horror of my situation	unused	#138	✖
the very horror of my situation	Modified	#138	✖
the fact that my fears are so vague	the + noun	#139 mismatch	✖
the fact that my fears are so vague	cataphor	#139	✖
the fact that my fears are so vague	Modified	#139	✖
the fancies of a nervous woman	the + noun	#140	✖
the fancies of a nervous woman	bridging-contained	#140	✖
the fancies of a nervous woman	Modified	#140	✖

Figure 1: New IMI annotation layer developed: the IMI RefLex Tagger (English)



RefLex Corpus' Tagger

CorpusDB: cmn_reflex 11293 [Document: 4]

Sentence: “小心点别让我抓住你，"他咆哮着说，顺手把扭弯的火钳扔到壁炉里，大踏步地走出了房间。

“ 小心点 别 让 我 抓住 你 ， " 他 咆哮 着 说 ， 顺手 把 扭弯 的 火钳 扔到 壁炉 里 ， 大踏步 地 走 出 了 房间 。

扭弯的火钳	bare noun	#246	✖
扭弯的火钳	given	#246	✖
扭弯的火钳	Modified	#246	✖
扭弯的火钳	Other	#246	✖
壁炉		#247	✖
壁炉		#247	✖
壁炉		#247	✖
壁炉		#247	✖
房间		#248	✖
房间		#248	✖
房间		#248	✖
房间		#248	✖

Figure 2: IMI RefLex Tagger (Mandarin Chinese): view of dropdown box with flexible tagging schema

noun phrase is modified and some information regarding the semantic class of the noun, whether it refers to human, location, or others. For a simple illustration, an English example is given below:

(3) I met a man yesterday. [The man] told me a story.

The annotated nominal is ‘the man’. The NP structure will be ‘the + noun’; the referential status tag is ‘given’. The referring expression is ‘not modified’.

These layers of annotation allow for a fine-grained analysis of the use and distribution of referential statuses. In this analysis, we will provide a summary of how these layers of annotation relate with each other. However, in the interest of scope, it will be impossible to cover all available relations between layers. Concerning semantic classes, in particular, we will only be able to discuss key relations that concern ‘location/place’. A detailed analysis of other semantic classes will be left as future work.

2.4. Annotation Process

Annotation was conducted in two stages, by two trained taggers fluent in both English and Mandarin Chinese. The data was first tagged by a single annotator who worked with an early version of the tagging schema. Once the tagging schema was revised and expanded, the data was re-tagged by a second annotator.

The annotation process also included multiple discussions with both authors of this paper, which guided the annotation process, clarified any doubts and made final decisions in difficult or ambiguous cases. The annotation process was also aided by a series of automatic checks (written in Python) that flagged inconsistencies in the data annotation.

A. NP Structure (English)
- the + noun
- the + numeral + noun
- the + numeral

A. NP Structure (Mandarin)
- Ø
- pronoun
- proper names (bare)
- bare noun
- classifier + noun
- one + classifier + noun
- any other numeral + classifier + noun
- dem
- dem + noun
- dem + (one) + classifier
- dem + any other numeral + classifier
- dem + (one) + classifier + noun
- dem + any other numeral + classifier + noun

B. Referential Status (English and Mandarin)
- given-sit
- given
- given-displaced
- cataphor
- bridging
- bridging-contained
- unused
- new

C. Modification (English and Mandarin)
- Modified
- Not Modified

D. Semantic Class (Mandarin)
- Human
- Location/Place
- Other

Figure 3: IMI RefLex Tagger tagset

In total, the annotation of the two short stories took a total of 200 hours to complete (divided among the two annotators).

3. Discussion of Results

The tagging process resulted in 1090 [*the*-X] nominal phrases tagged for English and their Mandarin counterparts. As can be seen in Table 1, the overwhelming majority of [*the*-X] nominal phrases in these two short stories were of the form 'the + noun' (i.e., without the use of numerals). This means that we do not currently have enough data to fully discuss if the use of numerals has a measurable impact in the use of referential expressions.

Out of all English nominal phrases, 46% were modified in some way. This includes mostly adjectival modification and relative clauses. For comparison, only 38% of the Mandarin Chinese translations were modified.

NP Structure	Freq.
the + noun	**0.97**
the + numeral + noun	0.02
the + numeral	0.01
Total (n=1090)	1.00

Table 1: English by NP Structure

The most common Mandarin translation for [*the*-X] phrases in English are bare nouns (64%), followed by demonstrative phrases (including different kinds in the list, and adding up to 17%) – see Table 2. It is expected that bare nouns and demonstrative phrases are the most common equivalents for the English simple definite [*the*-X], as has been suggested in the literature. The new and interesting finding is that the use of bare nouns is much more common (almost 4 times more) than the use of demonstrative phrases.

Even though Chen (2004) proposes that demonstrative phrases in Mandarin can be semantic equivalents of [*the*-X] phrases in English, he also mentions some differences. Consider the following scenario. A and B enter A's house. B is aware that A has a baby but the baby is not in sight. A can say:

(4)	安靜	點兒，	別	把	孩子	吵醒	了。
	ānjìng	diǎnr,	bié	bǎ	háizi	chǎo-xǐng	le.
	quiet	bit,	don't	BA	baby	wake-up	SFP.

'Be quiet. Don't wake up the baby.'(Chen, 2004)

The translation of 'the baby' in 4 is *háizi*, a bare noun, The use of a demonstrative phrase will be inappropriate in this case as the baby is not in sight. Visibility seems to be a condition that governs the use of the demonstratives in some situations. It is also not possible to use demonstratives in Mandarin to translate phrases like 'the sun', as it is globally unique. These semantic restrictions on the use of the Mandarin demonstratives might explain why the percentage of demonstrative phrases as the translation equivalents of [*the*-X] phrases is much lower than that of bare nouns.

Another interesting find worth mentioning from Table 2 is the fact that about 10% of [*the*-X] phrases in English are completely missing from the Mandarin translation (shown as '—missing—'). We are currently not equipped to discuss exactly when this can happen, but our findings suggest that dropping a reference in the translation may be related to the referential status of the original expressions. Table 3 shows that English expressions missing Mandarin counterparts were mostly tagged as 'unused', 'given' or 'cataphor'. A fuller study including these expressions' permanence in the discourse could shed further light into this topic.

Table 4 and Table 5 show the distribution of referential statuses across the tagged English and Mandarin nominal expressions. The information provided in these tables is not unexpected. While it is possible to see a broad parallelism in the overall referential statuses of

NP Structure	Freq.
bare noun	0.64
dem + (one) + classifier + noun	**0.14**
—missing—	0.10
one + classifier + noun	0.06
dem + noun	**0.02**
any other numeral + classifier + noun	0.01
pronoun	0.01
classifier + noun	0.01
dem + any other numeral + classifier + noun	**0.01**
Ø	0.01
proper names (bare)	0.00
dem + any other numeral + classifier	**0.00**
Total (n=1090)	1.00

Table 2: Mandarin by NP Structure

Referential Status	Freq.
unused	**0.36**
given	**0.19**
cataphor	**0.17**
given-sit	0.12
given-displaced	0.06
bridging	0.06
bridging-contained	0.04
Total (n=108)	1.00

Table 3: English referential statuses for missing Mandarin translations

the tagged nominal expressions, we can also observe small discrepancies across these classes (e.g., slightly higher values for 'given' expressions in Mandarin, or slightly higher 'unused' expressions in English). These discrepancies can be explained in part by the fact that some expressions were missing in the Mandarin translations, as already discussed, but also by how normal translation practices of literary texts do not always follow a strict literal translation method, and often change slightly the focus and even the flow of information.

Referential Status	Freq.
given	0.30
given-sit	0.20
unused	0.19
given-displaced	0.10
bridging	0.09
bridging-contained	0.07
cataphor	0.04
Total (n=1090)	1.00

Table 4: English referential statuses across all nominal expressions

One of the main goals of this paper was to map referential information statuses to referential forms. Figure 4 shows a summary of this mapping by providing the top two NP structures for each referential status in the Mandarin text. The categories 'given' and 'given-displaced' have the highest percentage of demonstrative phrases (29% and 31% respectively) while the overall average is 17% only. The use of the demonstratives in 'given-displaced' contexts can be understood as a way to re-introduced/activate a referent that has been mentioned not too recently (the threshold is set at 5 clauses before in RefLex). If the use of the demonstratives in Mandarin is used as a way to reactivate a certain referent, it can then be understood why in 'given-sit' and 'bridging-contained' the percentages are low. In these contexts, the referent is in the immediate text-external context in the former, and anchored to an element in the embedded phrase (linguistically very proximal) in the latter.

Referential Status	Freq.
given	0.36
given-sit	0.20
unused	0.15
—missing—	0.10
given-displaced	0.10
bridging	0.07
bridging-contained	0.03
cataphor	0.01
Total (n=1090)	1.00

Table 5: Mandarin referential statuses across all nominal expressions

A deeper analysis of demonstrative phrases in Mandarin shows an interesting distribution between proximal and distal usages. Out of the 185 tokens of demonstrative phrases in the Mandarin tagset (which are translations of [*the*-X]), proximal demonstratives are more frequent: 59% (n=109) phrases use the proximal demonstrative 這 *zhè* and 41% (n=76) phrases use the distal demonstrative 那 *nà*.

When we look at their distribution with respect to referential statuses, a clearer difference emerges. Out of the 109 tokens of the proximal demonstrative *zhè*, 89% are 'given' (an expression whose referent is mentioned in previous discourse) – see Table 6. On the other hand, out of the 76 tokens of the distal demonstrative *nà*, 41% are 'given' while 42% are 'given-displaced' (an expression whose referent is mentioned in the previous discourse context earlier than 5 clauses before) – see Table 7. Both 'given' and 'given-displaced' are anaphoric in nature. They differ in that the former is used for recent mentions (fewer than 5 clauses) and the latter is for distant mentions (more than 5 clauses). The proximal demonstrative *zhè* is overwhelmingly used for recently mentioned antecedents. The distal demonstrative *nà* can be used for both, recent mentions or earlier mentions.[3]

Our data does not have noun phrases under the cate-

[3] Chen (2004) suggests that the proximity of *zhè* makes it a better anaphoric device than the distal *nà* in referring to an antecedent recently introduced into discourse.

Referential Status	Freq.
given	0.89
given-sit	0.06
given-displaced	0.04
unused	0.02
Total (n=109)	1.00

Table 6: Mandarin referential statuses across proximal (这, zhè) demonstrative nominal expressions

Referential Status	Freq.
given-displaced	0.42
given	0.41
unused	0.07
bridging	0.05
given-sit	0.04
bridging-contained	0.01
Total (n=76)	1.00

Table 7: Mandarin referential statuses across distal (那, nà) demonstrative nominal expressions

gory 'new'. This is expected as generally only indefinite noun phrases are used to refer to new entities, and for this study, we are only looking at [*the*-X] phrases in English and its Mandarin equivalents.

unused
- bare noun — 81%
- one + classifier + noun — 8%

given
- bare noun — 56%
- dem + (one) + classifier + noun — 29%

given-sit
- bare noun — 83%
- one + classifier + noun — 10%

bridging
- bare noun — 92%
- dem + (one) + classifier + noun — 5%

given-displaced
- bare noun — 63%
- dem + (one) + classifier + noun — 31%

bridging-contained
- bare noun — 97%
- dem + any other numeral + classifier + noun — 3%

cataphor
- bare noun — 75%
- dem + any other numeral + classifier + noun — 25%

Figure 4: Top two Mandarin NP structures per referential status

When the semantic class of the noun is 'location/place', the percentage of Mandarin demonstrative phrases as the translation equivalents of English [*the*-X] phrases is very low. Among the 140 tokens of 'location/place' noun phrases, only 4% of the tokens are translated with a demonstrative (see Table 8). This could be related to the issue of discourse persistent/prominence. Entities that are 'props' rather than regular participants in discourse are often marked differently in discourse. Quoting Recasens et al. (2013): 'not all discourse entities are created equal. Some lead long lives and appear in a variety of discourse contexts (coreferents), whereas others never escape their birthplaces, dying out after just one mention (singletons).' Hopper (1986) observes that in Malay, the absence of the classifier in an NP correlates with the entity having a short 'discourse persistence' (or 'thematic importance' in Givón (1984), referring to the importance of a referent in discourse). This could be the reason why in Mandarin, the demonstrative is almost never used when translating 'location/places' from English because locations/places very often have low discourse persistent/prominence. As discussed earlier on, the categories 'given' and 'given-displaced' have a high percentage of demonstrative phrases, around 30% in average. Among NPs tagged as 'location/place' in the two categories, the percentage of NP forms using demonstratives were much lower than the average – only 8%.

NP Structure	Freq.
bare noun	0.89
dem + (one) + classifier + noun	0.04
one + classifier + noun	0.04
proper names (bare)	0.01
pronoun	0.01
Total (n=140)	1.00

Table 8: Mandarin NP type for NPs tagged as Location/Place

Referential Status	Freq.
given-sit	0.49
given	0.18
given-displaced	0.17
unused	0.14
bridging	0.01
bridging-contained	0.01
Total (n=140)	1.00

Table 9: Mandarin NP Ref-Status Mapping for NPs tagged as Location/Place

4. Conclusion and Future Work

In this study, even though we have only tagged [*the*-X] phrases in English and their Mandarin equivalents, we have been able to detect patterns that are interesting for the study of reference. In tracking the Mandarin translation equivalents for English [*the*-X] phrases, we observe that Mandarin bare nouns are the most common Mandarin translation, followed by demonstrative phrases, with the exception that when the noun phrase refers to locations/places. In fact, when the noun phrase refers to locations/places demonstrative phrases are almost never used. We show that [*the*-X] phrases in English are more likely to be translated as demonstrative phrases in Mandarin if they have the referential status of 'given' (previously mentioned) or 'given-displaced'(antecedent of an expression occurs earlier than the previous five clauses). Finally, we also show evidence for a clear functional difference between the Mandarin proximal demonstrative and the distal demonstrative: the Mandarin proximal demonstrative appears more frequently and is almost exclusively used to refer to referents with recent antecedents (fewer than 5 clauses before) while the Mandarin distal demonstrative can be used for both recent and distant referents.

This study is very limited in scope, we only look at [*the*-X] phrases in English and their Mandarin equivalents. However, even maintaining its scope, there are still ways to expand our analysis, for example, adding more tagging categories. We would like to add 'shell nouns' as one of the semantic classes. Shell nouns are nouns that conceptually encapsulate complex pieces of information (Schmid, 2018), such as *fact*, *reason*, *problem*, *position*, *fact*, etc. Similar to semantically empty nouns for 'fellow' or 'person' in Mandarin, when unmodified, we expect such nouns to be less likely to appear bare due to the lack of semantic content. Instead, a demonstrative will be expected.[4] We would also like to add the tags for proximal and distal demonstratives since our discussion of results has shown this to be a dimension worthy of further exploration.

To expand the scope in the future, we want to include other English phrases and their Mandarin equivalents using more parallel texts, ideally also including other genres. Furthermore, we would like to track the referential forms referring to specific referents throughout the whole discourse. This would allow us to study the relationship between fluctuation in salience/accessibility and referential forms in a referent's discourse life.

In addition to expanding our project's depth of analysis through new layers of annotation, we would also like to better exploit the multilingual nature of the dataset we used. The NTU Multilingual Corpus includes sense-tagged translations of shorts stories and of texts in other genres for Japanese, Italian, and Indonesian. Adopting a widely multilingual research agenda looking into mapping referential statuses to structural forms, abandoning English-centric analyses, could help gain new insights on the distinction between general trends and language specific features of referential analysis.

Finally, another important area we believe worth pursuing is the further development of the IMI RefLex Tagger. While it serves its current purpose, the annotation interface could still be improved further, especially in the cross-lingual link of expressions in two languages (which is currently done manually), and also in the ability to tag both languages side by side (which currently has to be emulated by opening two browser windows).

5. Release Notes

The raw text and all annotations produced as part of this corpus will be released under a Creative Commons Attribution 4.0 International License (CC BY 4.0)[5]. This data will be deposited in the following GitHub repository: https://github.com/lmorgadodacosta/multilingual-referential-annotation-corpus.

6. Acknowledgements

We would like to thank Belinda Juan and Jaroslav Tecl for helping us to annotate data. The research described here is supported by the European Regional Development Fund - Project 'Sinophone Borderlands - Interaction at the Edges' CZ.02.1.01/0.0/0.0/16_019/0000791 and by the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement H2020-MSCA-IF-2020 CHILL – No.101028782.

7. Bibliographical References

Arnold, J. E. and Zerkle, S. A. (2019). Why do people produce pronouns? pragmatic selection vs. rational models. *Language, Cognition and Neuroscience*, 34(9):1152–1175.

Baumann, S. and Riester, A. (2012). Referential and lexical givenness: Semantic, prosodic and cognitive aspects. *Prosody and meaning*, 25:119–162.

Bond, F. and Foster, R. (2013). Linking and extending an Open Multilingual Wordnet. In *51st Annual Meeting of the Association for Computational Linguistics: ACL-2013, Sofia*, pages 1352–1362.

Bond, F., Morgado da Costa, L., and Le, T. A. (2015). IMI – A multilingual semantic annotation environment. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, System Demonstrations (ACL 2015)*, pages 7–12, Beijing, China.

Bond, F., Ohkuma, T., Da Costa, L. M., Miura, Y., Chen, R., Kuribayashi, T., and Wang, W. (2016). A multilingual sentiment corpus for chinese, english and japanese. In *6th Emotion and Sentiment Analysis Workshop (at LREC 2016)*.

Burchardt, A., Erk, K., Frank, A., Kowalski, A., Pado, S., and Pinkal, M. (2006). Salto-a versatile multi-level annotation tool. In *LREC*, pages 517–520. European Language Resource Association.

Chen, P. (2004). Identifiability and definiteness in chinese.

[4] This is motivated by some preliminary work done with manual annotation between English and Mandarin text. We found that there is a higher chance for relatively semantically empty nouns, e.g., *fellow*, *fact*, etc. to be translated in Mandarin with a demonstrative (Sio and Juan, 2019).

[5] https://creativecommons.org/licenses/by/4.0/

Cheng, L. L.-S. and Sybesma, R. (1999). Bare and not-so-bare nouns and the structure of np. *Linguistic Inquiry*, 30(4):509–542.

Choi, H. Y. J. (2019). A corpus based analysis of-kan and-i in Indonesian. Master's thesis, Nanyang Technological University, Singapore.

Christophersen, P. (1939). The articles: A study of their theory and use in english.

Conan Doyle, A. (1892). *The Adventures of Sherlock Holmes*. George Newnes, London.

Conan Doyle, A. (1905). *The Return of Sherlock Homes*. George Newnes, London.

Christiane Fellbaum, editor. (1998). *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA.

Givón, T. (1984). The pragmatics of referentiality. *Georgetown University round table on language and linguistics*, pages 120–138.

Gundel, J. K., Hedberg, N., and Zacharski, R. (1993). Cognitive status and the form of referring expressions in discourse. *Language*, pages 274–307.

Hawkins, J. A. (1978). Definiteness and indefiniteness. atlantic highlands.

Hopper, P. J. (1986). Some discourse functions of classifiers in malay. *Noun classes and categorization*, 7:309–325.

Kummerfeld, J. K. (2019). SLATE: A super-lightweight annotation tool for experts. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 7–12, Florence, Italy, July. Association for Computational Linguistics.

Lambrecht, K. (1996). *Information structure and sentence form: Topic, focus, and the mental representations of discourse referents*, volume 71. Cambridge university press.

Lu, W., Verhagen, A., and Su, I. (2018). A multiple-parallel-text approach for viewpoint research across languages. *Expressive minds and artistic creations: Studies in cognitive poetics*.

Lyons, C. (1999). *Definiteness*. Cambridge University Press.

Recasens, M., Danescu-Niculescu-Mizil, C., and Jurafsky, D. (2013). Linguistic models for analyzing and detecting biased language. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1650–1659.

Schmid, H.-J. (2018). Shell nouns in english-a personal roundup. *Caplletra. Revista Internacional de Filologia*, (64):109–128.

Sio, J. U.-S. and Juan, L.-T. B. (2019). Investigate the differences between chinese bare nouns and demonstrative phrases using parallel texts. *Presented at the Annual Conference on Asian Studies (ACAS), Olomouc, the Czech Republic*.

Tan, L. and Bond, F. (2014). NTU-MC toolkit: Annotating a linguistically diverse corpus. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: System Demonstrations*, page 86–89, Dublin.

Wang, S. and Bond, F. (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.

Winder, R. V. P., MacKinnon, J., Li, S. Y., Lin, B., Heah, C., Morgado da Costa, L., Kuribayashi, T., and Bond, F. (2017). NTUCLE: Developing a corpus of learner English to provide writing support for engineering students. In *Proceedings of the 4th Workshop on NLP Techniques for Educational Applications (NLPTEA 2017)*, Taipei, Taiwan.