# A Corpus of Adpositional Supersenses for Mandarin Chinese

**Siyao Peng, Yang Liu, Yilun Zhu, Austin Blodgett, Yushi Zhao, Nathan Schneider**
Department of Linguistics, Georgetown University
Washington, DC, USA
{sp1184, yl879, yz565, ajb341, yz521, nathan.schneider}@georgetown.edu

**Abstract**

Adpositions are frequent markers of semantic relations, but they are highly ambiguous and vary significantly from language to language. Moreover, there is a dearth of annotated corpora for investigating the cross-linguistic variation of adposition semantics, or for building multilingual disambiguation systems. This paper presents a corpus in which all adpositions have been semantically annotated in Mandarin Chinese; to the best of our knowledge, this is the first Chinese corpus to be broadly annotated with adposition semantics. Our approach adapts a framework that defined a general set of *supersenses* according to ostensibly language-independent semantic criteria, though its development focused primarily on English prepositions (Schneider et al., 2018). We find that the supersense categories are well-suited to Chinese adpositions despite syntactic differences from English. On a Mandarin translation of *The Little Prince*, we achieve high inter-annotator agreement and analyze semantic correspondences of adposition tokens in bitext.

**Keywords:** adpositions, supersenses, Mandarin Chinese, corpus, annotation

## 1. Introduction

Adpositions (i.e. prepositions and postpositions) include some of the most frequent words in languages like Chinese and English, and help convey a myriad of semantic relations of space, time, causality, possession, and other domains of meaning. They are also a persistent thorn in the side of second language learners owing to their extreme idiosyncrasy (Chodorow et al., 2007; Lorincz and Gordon, 2012). For instance, the English word *in* has no exact parallel in another language; rather, for purposes of translation, its many different usages cluster differently depending on the second language. Semantically annotated corpora of adpositions in multiple languages, including parallel data, would facilitate broader empirical study of adposition variation than is possible today, and could also contribute to NLP applications such as machine translation (Li et al., 2005; Agirre et al., 2009; Shilon et al., 2012; Weller et al., 2014, 2015; Hashemi and Hwa, 2014; Popović, 2017) and grammatical error correction (Chodorow et al., 2007; Tetreault and Chodorow, 2008; De Felice and Pulman, 2008; Hermet and Alain, 2009; Huang et al., 2016; Graën and Schneider, 2017).

This paper describes the first corpus with broad-coverage annotation of adpositions in Chinese. For this corpus we have adapted Schneider et al.'s (2018) Semantic Network of Adposition and Case Supersenses annotation scheme (SNACS; see §2.2) to Chinese.[1] Though other languages were taken into consideration in designing SNACS, no serious annotation effort has been undertaken to confirm empirically that it generalizes to other languages. After developing new guidelines for syntactic phenomena in Chinese (§3), we apply the SNACS supersenses to a translation of *The Little Prince*[2] (*Xiǎo Wáng Zǐ*), finding the supersenses to be robust and achieving high inter-annotator agreement (§4). We analyze the distribution of adpositions and supersenses in the corpus, and compare to adposition behavior in a separate English corpus (see §5). We also examine the predictions of a part-of-speech tagger in relation to our criteria for annotation targets (§6). The annotated corpus and the Chinese guidelines for SNACS will be made freely available online.[3]

## 2. Related Work

To date, most wide-coverage semantic annotation of prepositions has been dictionary-based, taking a word sense disambiguation perspective (Litkowski and Hargraves, 2005, 2007; Litkowski, 2014). Schneider et al. (2015) proposed a supersense-based (unlexicalized) semantic annotation scheme which would be applied to all tokens of prepositions in English text. We adopt a revised version of the approach, known as SNACS (see §2.2). Previous SNACS annotation efforts have been mostly focused on English—particularly STREUSLE (Schneider et al., 2016, 2018), the semantically annotated corpus of reviews from the English Web Treebank (EWT; Bies et al., 2012). We present the first adaptation of SNACS for Chinese by annotating an entire Chinese translation of *The Little Prince*.

### 2.1. Chinese Adpositions and Roles

In the computational literature for Chinese, apart from some focused studies (e.g., Yang and Kuo (1998) on logical-semantic representation of temporal adpositions), there has been little work addressing adpositions specifically. Most previous semantic projects for Mandarin Chinese focused on content words and did not directly annotate the semantic relations signaled by functions words such as prepositions (Xue et al., 2014; Hao et al., 2007; You and Liu, 2005; Li et al., 2016). For example, in Chinese PropBank, Xue (2008) argued that the head word and its part of speech are clearly informative for labeling the semantic role of a phrase, but the preposition is not always the most informative element. Li et al. (2003) annotated the Tsinghua Corpus (Zhang, 1999) from *People's Daily* where the content words were

[1] Zhu et al. (2019) previewed our approach.

[2] Originally *Le Petit Prince* by Antoine de St. Exupéry, published in 1943 and subsequently translated into numerous languages.

[3] https://github.com/nert-nlp/Chinese-SNACS/

selected as the headwords, i.e., the object is the headword of the prepositional phrase. In these prepositional phrases, the nominal headwords were labeled with one of the 59 semantic relations (e.g. *Location, LocationIni, Kernel word*) whereas the prepositions and postpositions were respectively labeled with syntactic relations *Preposition* and *LocationPreposition*.[4] Similarly, in Semantic Dependency Relations (SDR, Che et al. 2012, 2016), prepositions and localizers were labeled as semantic markers *mPrep* and *mRange*, whereas semantic roles, e.g., *Location, Patient*, are assigned to the governed nominal phrases.

Sun and Jurafsky (2004) compared PropBank parsing performance on Chinese and English, and showed that four Chinese prepositions (*zài*, *yú*, *bǐ*, and *duì*) are among the top 20 lexicalized syntactic head words in Chinese PropBank, bridging the connections between verbs and their arguments. The high frequency of prepositions as head words in PropBank reflects their importance in context. However, very few annotation scheme attempted to directly label the semantics of these adposition words.

Chinese Knowledge Information Processing Group (CKIP) (1993) is the most relevant adposition annotation effort, categorizing Chinese prepositions into 66 types of senses grouped by lexical items. However, these lexicalized semantic categories are constrained to a given language and a closed set of adpositions. For semantic labeling of Chinese adpositions in a multilingual context, we turn to the SNACS framework, described below.

### 2.2. SNACS: Adposition Supersenses

Schneider et al. (2018) proposed the Semantic Network of Adposition and Case Supersenses (SNACS), a hierarchical inventory of 50 semantic labels, i.e., supersenses, that characterize the use of adpositions, as shown in Figure 1. Since the meaning of adpositions is highly affected by the context, SNACS can help distinguish different usages of adpositions. For instance, (1) presents an example of the supersense TOPIC for the adposition *about* which emphasizes the subject matter of urbanization that the speaker discussed. In (2), however, the same preposition *about* takes a measurement in the context, expressing an approximation.

(1) I gave a presentation **about:TOPIC** urbanization.[5]

(2) We have **about:APPROXIMATOR** 3 eggs left.

Though assigning a single label to each adposition can help capture its lexical contribution to the sentence meaning as well as disambiguate its uses in different scenarios, the canonical lexical semantics of adpositions are often stretched to fit the needs of the scene in actual language use.

(3) I care **about:STIMULUS⇝TOPIC** you.

For instance, (3) blends the domains of emotion (principally reflected in *care*, which licenses a STIMULUS), and cognition (principally reflected in *about*, which often marks non-emotional TOPICs). Thus, SNACS incorporates the *construal analysis* (Hwang et al., 2017) wherein the lexical semantic contribution of an adposition (its **function**) is distinguished and may diverge from the underlying relation in the surrounding context (its **scene role**). Construal is notated by SCENEROLE⇝FUNCTION, as STIMULUS⇝TOPIC in (3).[6]

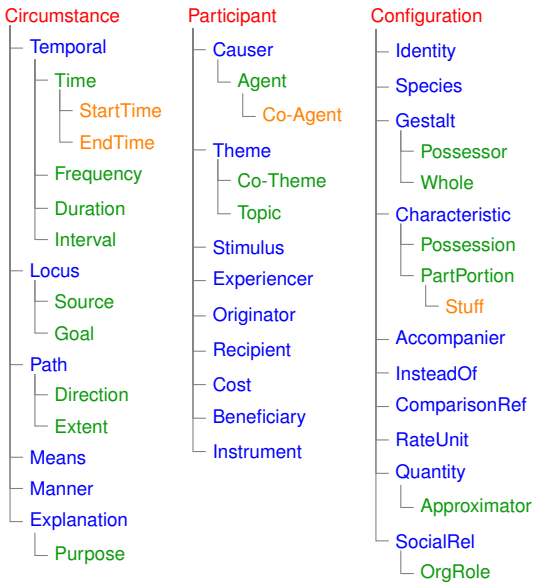


**Figure 1:** SNACS hierarchy of 50 supersenses.

Another motivation for incorporating the construal analysis, as pointed out by Hwang et al. (2017), is its capability to adapt the English-centric supersense labels to other languages, which is the main contribution of this paper. The construal analysis can give us insights into the similarities and differences of function and scene roles of adpositions across languages.

## 3. Adposition Criteria in Mandarin Chinese

Our first challenge is to determine which tokens qualify as adpositions in Mandarin Chinese and merit supersense annotations. The English SNACS guidelines (we use version 2.3) broadly define the set of SNACS annotation targets to include canonical prepositions (taking an noun phrase (NP) complement) and their subordinating (clausal complement) uses. Possessives, intransitive particles, and certain uses of the infinitive marker *to* are also included (Schneider et al., 2019).

In Chinese, the difficulty lies in two areas, which we discuss below. Firstly, prepositional words are widely attested. However, since no overt derivational morphology occurs on these prepositional tokens (previously referred to as coverbs), we need to filter non-prepositional uses of these words. Secondly, post-nominal particles, i.e., localizers, though not

[4] Though named *LocationPreposition* in Li et al. (2003), these adpositions actually occur postnominally, equivalent to localizers in this paper.

[5] Throughout this paper, adposition tokens under discussion are bolded and labeled.

[6] The supersense labels in *congruent* construals, such as TOPIC and APPROXIMATOR in (1) and (2), are both function and scene role by definition.

always considered adpositions in Chinese, deliver rich semantic information.

**Coverbs** Tokens that are considered generic prepositions can co-occur with the main predicate of the clause and introduce an NP argument to the clause (Li and Thompson, 1974) as in (4). These tokens are referred to as coverbs. In some cases, coverbs can also occur as the main predicate. For example, the coverb *zài* heads the predicate phrase in (5).

(4) tā **zài:**LOCUS xuéshù
3SG P:at academia
**shàng:**TOPIC↝LOCUS yǒusuǒzuòwéi.
LC:on-top-of successful
'He succeeded in academia.'

(5) nǐ yào de yáng jiù **zài** lǐmiàn.
2SG want DE sheep RES at inside
'The sheep you wanted is in the box.'
(zh_lpp_1943.92)

In this project, we only annotate coverbs when they do not function as the main predicate in the sentence, echoing the view that coverbs modify events introduced by the predicates, rather than establishing multiple events in a clause (Hui, 2012). Therefore, lexical items such as *zài* are annotated when functioning as a modifier as in (4), but not when as the main predicate as in (5).

**Localizers** Localizers are words that follow a noun phrase to refine its semantic relation. For example, *shàng* in (4) denotes a contextual meaning, 'in a particular area,' whereas the co-occurring coverb *zài* only conveys a generic location. It is unclear whether localizers are syntactically postpositions, but we annotate all localizers because of their semantic significance. Though coverbs frequently co-occur with localizers and the combination of coverbs and localizers is very productive, there is no strong evidence to suggest that they are circumpositions. As a result, we treat them as separate targets for SNACS annotation: for example, *zài* and *shàng* receive LOCUS and TOPIC↝LOCUS respectively in (4).

Setting aside the syntactic controversies of coverbs and localizers in Mandarin Chinese, we regard both of them as adpositions that merit supersense annotations. As in (4), both the coverb *zài* and the localizer *shàng* surround an NP argument *xuéshù* ('academia') and they as a whole modify the main predicate *yǒusuǒzuòwéi* ('successful'). In this paper, we take the stance that coverbs co-occur with the main predicate and precede an NP, whereas localizers follow a noun phrase and add semantic information to the clause.

## 4. Corpus Annotation

We chose to annotate the novella *The Little Prince* because it has been translated into hundreds of languages and dialects, which enables comparisons of linguistic phenomena across languages on bitexts. This is the first Chinese corpus to undergo SNACS annotation. Ongoing adpositional supersense projects on *The Little Prince* include English, German, French, and Korean. In addition, *The Little Prince* has received large attention from other semantic frameworks and corpora, including the English (Banarescu et al., 2013) and Chinese (Li et al., 2016) AMR corpora.

### 4.1. Preprocessing

We use the same Chinese translation of *The Little Prince* as the Chinese AMR corpus (Li et al., 2016), which is also sentence-aligned with the English AMR corpus (Banarescu et al., 2013). These bitext annotations in multiple languages and annotation semantic frameworks can facilitate cross-framework comparisons.

Prior to supersense annotation, we conducted the following preprocessing steps in order to identify the adposition targets that merit supersense annotation.

**Tokenization** After automatic tokenization using Jieba,[7] we conducted manual corrections to ensure that all potential adpositions occur as separate tokens, closely following the Chinese Penn Treebank segmentation guidelines (Xia, 2000). The final corpus includes all 27 chapters of *The Little Prince*, with a total of 20k tokens.

**Adposition Targets** All annotators jointly identified adposition targets according to the criteria discussed in §3. Manual identification of adpositions was necessary as an automatic POS tagger was found unsuitable for our criteria (§6).

**Data Format** Though parsing is not essential to this annotation project, we ran the StanfordNLP (Qi et al., 2018) dependency parser to obtain POS tags and dependency trees. These are stored alongside supersense annotations in the *CoNLL-U-Lex* format (modeled after the STREUSLE corpus; Schneider and Smith, 2015; Schneider et al., 2018). CoNLL-U-Lex extends the CoNLL-U format used by the Universal Dependencies (UD; Nivre et al., 2016) project to add additional columns for lexical semantic annotations.[8]

### 4.2. Reliability of Annotation

The corpus is jointly annotated by three native Mandarin Chinese speakers, all of whom have received advanced training in theoretical and computational linguistics. Supersense labeling was performed cooperatively by 3 annotators for 25% (235/933) of the adposition targets, and for the remainder, independently by the 3 annotators, followed by cooperative adjudication. Annotation was conducted in two phases, and therefore we present two inter-annotator agreement studies to demonstrate the reproducibility of SNACS and the reliability of the adapted scheme for Chinese.

Table 1 shows raw agreement and Cohen's kappa across three annotators computed by averaging three pairwise comparisons. Agreement levels on scene role, function, and full construal are high for both phases, attesting to the validity of the annotation framework in Chinese. However, there is a slight decrease from Phase 1 to Phase 2, possibly due to the seven newly attested adpositions in Phase 2 and the 1-year interval between the two annotation phases.

IAA SAMPLES

| Phase | Time | Chapters | # Adpositions |
|---|---|---|---|
| Phase 1 | July 2018 | 15–20 | 111 |
| Phase 2 | Sept 2019 | 26–27 | 124 |

RAW AGREEMENT

| Phase | Scene | Function | Construal |
|---|---|---|---|
| Phase 1 | .92 | .95 | .90 |
| Phase 2 | .93 | .90 | .89 |

KAPPA

| Phase | Scene | Function | Construal |
|---|---|---|---|
| Phase 1 | .90 | .93 | .88 |
| Phase 2 | .92 | .88 | .88 |

**Table 1:** Inter-annotator agreement (IAA) results on two samples from different phases of the project.

| | Toks. | Types |
|---|---|---|
| Chapters | 27 | NA |
| Sentences | 1,597 | NA |
| Tokens | 20,287 | NA |
| Adpositions | 933 | 70 |
| Prepositions | 667 | 42 |
| Postpositions | 266 | 28 |
| Supersenses | 933 | 29 |
| Scene roles | 933 | 28 |
| Functions | 933 | 26 |
| Construals | 933 | 41 |
| Congruent (scene=fxn) | 803 | 25 |
| Divergent (scene≠fxn) | 130 | 16 |

**Table 2:** Statistics of the final Mandarin *The Little Prince* Corpus (the Chinese SNACS Corpus). Tokenization, identification of adposition targets, and supersense labeling were performed manually.

## 5. Corpus Analysis

Our corpus contains 933 manually identified adpositions. Of these, 70 distinct adpositions, 28 distinct scene roles, 26 distinct functions, and 41 distinct full construals are attested in annotation. Full statistics of token and type frequencies are shown in Table 2. This section presents the most frequent adpositions in Mandarin Chinese, as well as quantitative and qualitative comparisons of scene roles, functions, and construals between Chinese and English annotations.

### 5.1. Adpositions in Chinese

We analyze semantic and distributional properties of adpositions in Mandarin Chinese. The top 5 most frequent prepositions and postpositions are shown in Table 3. Prepositions include canonical adpositions such as *yīnwèi* and coverbs such as *zài*. Postpositions are localizers such as *shàng* and *zhōng*. We observe that prepositions *zài* and *duì* are dominant in the corpus (greater than 10%). Other top adpositions are distributed quite evenly between prepositions and postpositions. On the low end, 27 out of the 70 attested adposition types occur only once in the corpus.

[7] https://github.com/fxsjy/jieba

[8] https://github.com/nert-nlp/streusle/blob/master/CONLLULEX.md

| Prep. | Trans. | % | Count |
|---|---|---|---|
| zài | on | 18.4 | 172 |
| duì | to | 11.0 | 103 |
| bǎ | *theme marker* | 7.2 | 67 |
| yīnwèi | due to | 4.7 | 44 |
| gěi | to | 3.5 | 33 |
| **Total** | | 44.9 | 419 |
| **Postp.** | **Trans.** | **%** | Count |
| shàng | on top of | 9.5 | 89 |
| zhōng | in the middle of | 4.9 | 46 |
| lǐ | inside of | 3.9 | 36 |
| láishuō | to one's regard | 2.7 | 25 |
| shí | at the time of | 2.1 | 20 |
| **Total** | | 23.2 | 216 |

**Table 3:** Percentages and counts of the top 5 prepositions and postpositions in Chinese *Little Prince*. The percentages are out of all adpositions.

### 5.2. Supersense & Construal Distributions in Chinese versus English

The distribution of scene role and function types in Chinese and English reflects the differences and similarities of adposition semantics in both languages. In table 4 we compare this corpus with the largest English adposition supersense corpus, STREUSLE version 4.1 (Schneider et al., 2018), which consists of web reviews. We note that the Chinese corpus is proportionally smaller than the English one in terms of token and adposition counts.[9] Moreover, there are fewer scene role, function and construal types attested in Chinese. The proportion of construals in which the scene role differs from the function (scene≠fxn) is also halved in Chinese. In this section, we delve into comparisons regarding scene roles, functions, and full construals between the two corpora both quantitatively and qualitatively.

**Overall Distribution of Supersenses** Figures 2 and 3 present the top 10 scene roles and functions in Mandarin Chinese and their distributions in English. It is worth noting that since more scene role and function types are attested in the larger STREUSLE dataset, the percentages of these supersenses in English are in general lower than the ones in Chinese.

There are a few observations in these distributions that are of particular interest. For some of the examples, we use an annotated subset of the English *Little Prince* corpus for qualitative comparisons, whereas all quantitative results in English refer to the larger STREUSLE corpus of English Web Treebank reviews (Schneider et al., 2018).

**Fewer Adpositions in Chinese** As shown in Table 4, the percentage of adposition targets over tokens in Chinese is only half of that in English. This is due to the fact that Chinese has a stronger preference to convey semantic information via verbal or nominal forms. Examples (6, 7) show that the prepositions used in English, *of* and *in*, are translated as copula verbs (*shì*) and progressives (*zhèngzài*) in Chinese. Corresponding to Figures 2 and 3, the proportion

[9] We exclude possessives and multi-word expressions that are annotated in the English corpus since possessives are not formed by adpositional phrases in Mandarin Chinese.

| | toks | % adps | uniq adps | uniq scene | uniq fxn | uniq cons | scene≠fxn | % scene≠fxn |
|---|---|---|---|---|---|---|---|---|
| **Chinese:** *Little Prince* | 20k | 4.6% | 70 | 28 | 26 | 41 | 16 | 14% |
| **English:** *EWT Reviews* | 55k | 7.4% | 111 | 47 | 40 | 170 | 130 | 27% |

**Table 4:** Statistics of Adpositional Supersenses in Chinese versus English. *% adps* presents the proportion of adposition targets over all token counts; *uniq adps/scene/fxn/cons* demonstrates the type frequency of adposition tokens, scene role and function supersense and construals; *scene≠fxn* and *% scene≠fxn* shows the type frequency and proportion of divergent construals.

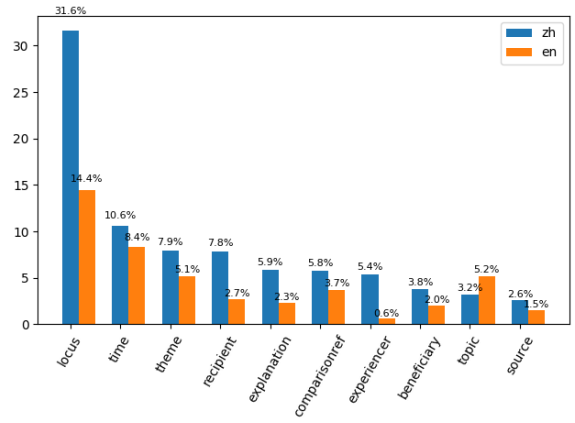


**Figure 2:** Top 10 most frequent scene roles in Chinese versus English.

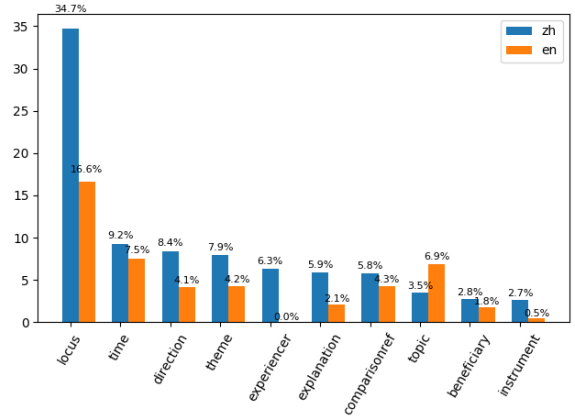


**Figure 3:** Top 10 most frequent functions in Chinese versus English.

of the supersense label TOPIC in English is higher than that in Chinese; and similarly, the supersense label IDENTITY is not attested in Chinese for either scene role or function.

(6) It was a picture **of:TOPIC** a boa constrictor **in:MANNER** the act **of:IDENTITY** swallowing an animal . (en_lpp_1943.3)

(7) [huà de] **shì** [[yì tiáo mǎngshé] **zhèngzài** draw DE COP one CL boa PROG tūnshí [yì zhī dà yěshòu]] swallow one CL big animal
'The drawing is a boa swallowing a big animal'. (en_lpp_1943.3)

**Larger Proportion of LOCUS in Chinese** In both Figure 2 and Figure 3, the percentages of LOCUS as scene role and function are twice that of the English corpus respectively. This corresponds to the fact that fewer supersense types occur in Mandarin Chinese than in English. As a result, generic locative and temporal adpositions, as well as adpositions tied to thematic roles, have larger proportions in Chinese than in English.

**EXPERIENCER as Function in Chinese** Despite the fact that there are fewer supersense types attested in Chinese, EXPERIENCER as a function is specific to Chinese as it does not have any prototypical adpositions in English (Schneider et al., 2019). In (8), the scene role EXPERIENCER is expressed through the preposition *to* and construed as GOAL, which highlights the abstract destination of the 'air of truth'. This reflects the basic meaning of *to*, which denotes a path towards a goal (Bowerman and Choi, 2001). In contrast, the lexicalized combination of the preposition *duì* and the localizer *láishuō* in (9) are a characteristic way to introduce the mental state of the experiencer, denoting the meaning 'to someone's regard'. The high frequency of *láishuō* and the semantic role of EXPERIENCER (6.3%) underscore its status as a prototypical adposition usage in Chinese.

(8) **To:EXPERIENCER⇝GOAL** those who understand life, that would have given a much greater air of truth to my story. (en_lpp_1943.185)

(9) [**duì:EXPERIENCER** [dǒngdé shēnghuó de P:to know-about life DE rén] **láishuō:EXPERIENCER**], zhèyàng shuō people LC:one's-regard this-way tell jiù xiǎndé zhēnshí RES seems real
'It looks real to those who know about life.' (zh_lpp_1943.185)

**Divergence of Functions across Languages** Among all possible types of construals between scene role and function, here we are only concerned with construals where the scene role differs from the function (scene≠fxn). The basis of Hwang et al.'s (2017) construal analysis is that a scene role is construed as a function to express the contexual meaning of the adposition that is different from its lexical one. Figure 4 presents the top 10 divergent (scene≠fxn) construals in Chinese and their corresponding proportions in English. Strikingly fewer types of construals are formed in Chinese. Nevertheless, Chinese is replete with RECIPIENT⇝DIRECTION adpositions, which constitute nearly half of the construals.

The 2 adpositions annotated with RECIPIENT⇝DIRECTION are *duì* and *xiàng*, both meaning 'towards' in Chinese. In (10, 11), both English *to* and Chinese *duì* have RECIPIENT as the scene role. In (10), GOAL is labelled as the function of *to* because it indicates the completion of the "saying"

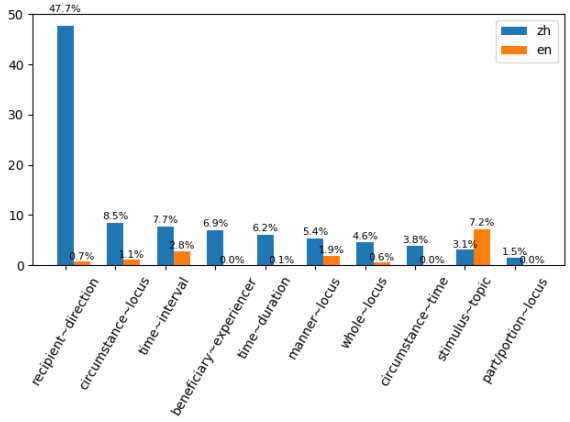


**Figure 4:** Top 10 Construals where scene≠function in Chinese versus English.

event.[10] In Chinese, *duì* has the function label DIRECTION provided that *duì* highlights the orientation of the message uttered by the speaker as in (11). Even though they express the same scene role in the parallel corpus, their lexical semantics still requires them to have different functions in English versus Chinese.

(10) You would have to say **to:RECIPIENT⇝GOAL** them: "I saw a house that costs $20,000." (en_lpp_1943.172).

(11) (nǐ) bìxū [**duì:RECIPIENT⇝DIRECTION**
2SG must P:to
tāmen] shuō: "wǒ kànjiàn le yí dòng shíwàn
3PL say 1SG see ASP one CL 10,000
fǎláng de fángzi."
franc DE house
'You must tell them: "I see a house that costs 10,000 francs." ' (zh_lpp_1943.172).

**New Construals in Chinese** Similar to the distinction between RECIPIENT⇝GOAL and RECIPIENT⇝DIRECTION in English versus Chinese, language-specific lexical semantics contribute to unique construals in Chinese, i.e. semantic uses of adpositions that are unattested in the STREUSLE corpus. Six construals are newly attested in the Chinese corpus:

- BENEFICIARY⇝EXPERIENCER
- CIRCUMSTANCE⇝TIME
- PARTPORTION⇝LOCUS
- TOPIC⇝LOCUS
- CIRCUMSTANCE⇝ACCOMPANIER
- DURATION⇝INSTRUMENT

Of these new construals, BENEFICIARY⇝EXPERIENCER has the highest frequency in the corpus. The novelty of this construal lies in the possibility of EXPERIENCER as function in Chinese, shown by the parallel examples in (12, 13), where *duì* receives the construal annotation BENEFICIARY⇝EXPERIENCER.

(12) One must not hold it **against:BENEFICIARY** them . (en_lpp_1943.180)

[10]The prototypical function of *to* indicates telic motion events. Telicity, however, is not required for DIRECTION.

(13) xiǎohǎizimen
children
**duì:BENEFICIARY⇝EXPERIENCER** dàrénmen
P:to adults
yìnggāi kuānhòu xie
should lenient COMP
'Children should not hold it against adults.' (zh_lpp_1943.180)

Similarly, other new construals in Chinese resulted from the lexical meaning of the adpositions that are not equivalent to those in English. For instance, the combination of *dāng ... shí* (during the time of) denotes the circumstance of an event that is grounded by the time (*shí*) of the event. Different lexical semantics of adpositions necessarily creates new construals when adapting the same supersense scheme into a new language, inducing newly found associations between scene and function roles of these adpositions. Fortunately, though combinations of scene and function require innovation when adapting SNACS into Chinese, the 50 supersense labels are sufficient to account for the semantic diversity of adpositions in the corpus.

## 6. POS Tagging of Adposition Targets

We conduct a post-annotation comparison between manually identified adposition targets and automatically POS-tagged adpositions in the Chinese SNACS corpus. Among the 933 manually identified adposition targets that merit supersense annotation, only 385 (41.3%) are tagged as ADP (adposition) by StanfordNLP (Qi et al., 2018). Figure 5 shows that gold targets are more frequently tagged as VERB than ADP in automatic parses, as well as a small portion that are tagged as NOUN. The inclusion of targets with POS=VERB reflects our discussion in §3 that coverbs co-occurring with a main predicate are included in our annotation. The automatic POS tagger also wrongly predicts some non-coverb adpositions, such as *yīnwéi*, to be verbs.

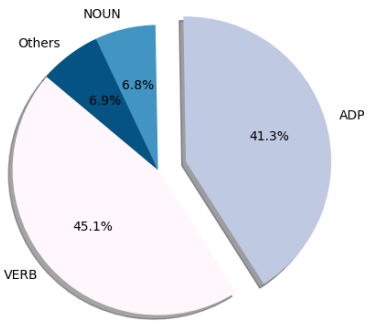


**Figure 5:** POS Distribution of Gold Adposition Tokens.

The StanfordNLP POS tagger also suffers from low precision (72.6%). Most false positives resulted from the discrepancies in adposition criteria between theoretical studies on Chinese adpositions and the tagset used in Universal Dependencies (UD) corpora such as the Chinese-GSD corpus. For instance, the Chinese-GSD corpus considers subordinating

conjunctions (such as *rúguǒ*, *yídàn*, *jìrán*, *zhǐyào*) adpositions; however, theoretical research on Chinese adpositions such as Li and Thompson (1989) differentiates them from adpositions, since they can never syntactically precede a noun phrase.

Hence, further SNACS annotation and disambiguation efforts on Chinese adpositions cannot rely on the StanfordNLP ADP category to identify annotation targets. Since adpositions mostly belong to a closed set of tokens, we apply a simple rule to identify all attested adpositions which are not functioning as the main predicate of a sentence, i.e., not the *root* of the dependency tree. As shown in Table 5, our heuristic results in an $F_1$ of 82.4%, outperforming the strategy of using the StanfordNLP POS tagger.

| | $P$ | $R$ | $F_1$ |
|---|---|---|---|
| StanfordNLP ADP | 72.6 | 41.3 | 52.6 |
| attested `dep≠root` adpositions | 75.1 | 91.3 | 82.4 |

**Table 5:** Adposition identification performance on Chinese SNACS corpus.

## 7. Conclusion

In this paper, we presented the first corpus annotated with adposition supersenses in Mandarin Chinese. The corpus is a valuable resource for examining similarities and differences between adpositions in different languages with parallel corpora and can further support automatic disambiguation of adpositions in Chinese. We intend to annotate additional genres—including native (non-translated) Chinese and learner corpora—in order to more fully capture the semantic behavior of adpositions in Chinese as compared to other languages.

## Acknowledgements

We thank anonymous reviewers for their feedback. This research was supported in part by NSF award IIS-1812778 and grant 2016375 from the United States–Israel Binational Science Foundation (BSF), Jerusalem, Israel.

## References

Agirre, Eneko, Atutxa, Aitziber, Labaka, Gorka, Lersundi, Mikel, Mayor, Aingeru, and Sarasola, Kepa (2009). Use of rich linguistic information to translate prepositions and grammatical cases to Basque. In Màrquez, Lluís and Somers, Harold, editors, *Proc. of EAMT*, pages 58–65. Barcelona, Catalonia, Spain.

Banarescu, Laura, Bonial, Claire, Cai, Shu, Georgescu, Madalina, Griffitt, Kira, Hermjakob, Ulf, Knight, Kevin, Koehn, Philipp, Palmer, Martha, and Schneider, Nathan (2013). Abstract Meaning Representation for sembanking. In *Proc. of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*.

Bies, Ann, Mott, Justin, Warner, Colin, and Kulick, Seth (2012). English Web Treebank. Technical Report LDC2012T13, Linguistic Data Consortium, Philadelphia, PA.

Bowerman, Melissa and Choi, Soonja (2001). Shaping meanings for language: universal and language-specific in the acquisition of spatial semantic categories. In Bowerman, Melissa and Levinson, Stephen, editors, *Language Acquisition and Conceptual Development*, pages 475–511. Cambridge University Press.

Che, Wanxiang, Shao, Yanqiu, Liu, Ting, and Ding, Yu (2016). SemEval-2016 Task 9: Chinese Semantic Dependency Parsing. In *Proc. of SemEval*, pages 1074–1080. San Diego, California.

Che, Wanxiang, Zhang, Meishan, Shao, Yanqiu, and Liu, Ting (2012). SemEval-2012 Task 5: Chinese Semantic Dependency Parsing. In *Proc. of *SEM/SemEval*, pages 378–384. Montréal, Canada.

Chinese Knowledge Information Processing Group (CKIP) (1993). Chinese part-of-speech analysis. Technical Report 93-05, Taipei.

Chodorow, Martin, Tetreault, Joel R., and Han, Na-Rae (2007). Detection of grammatical errors involving prepositions. In *Proc. of the Fourth ACL-SIGSEM Workshop on Prepositions*, pages 25–30. Prague, Czech Republic.

De Felice, Rachele and Pulman, Stephen G. (2008). A classifier-based approach to preposition and determiner error correction in L2 English. In *Proc. of Coling*, pages 169–176. Manchester, UK.

Graën, Johannes and Schneider, Gerold (2017). Crossing the border twice: reimporting prepositions to alleviate L1-specific transfer errors. In *Proc. of the Joint 6th Workshop on NLP for Computer Assisted Language Learning and 2nd Workshop on NLP for Research on Language Acquisition at NoDaLiDa, Gothenburg, 22nd May 2017*, pages 18–26. Gothenburg, Sweden.

Hao, Xiao-yan, Liu, Wei, Li, Ru, and Liu, Kai-ying (2007). Description systems of the Chinese FrameNet database and software tools. *Journal of Chinese Information Processing*, 5.

Hashemi, Homa B. and Hwa, Rebecca (2014). A comparison of MT errors and ESL errors. In Calzolari, Nicoletta, Choukri, Khalid, Declerck, Thierry, Loftsson, Hrafn, Maegaard, Bente, Mariani, Joseph, Moreno, Asuncion, Odijk, Jan, and Piperidis, Stelios, editors, *Proc. of LREC*, pages 2696–2700. Reykjavík, Iceland.

Hermet, Matthieu and Alain, Désilets (2009). Using first and second language models to correct preposition errors in second language authoring. In *Proc. of the Fourth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 64–72. Boulder, Colorado.

Huang, Hen-Hsen, Shao, Yen-Chi, and Chen, Hsin-Hsi (2016). Chinese preposition selection for grammatical error diagnosis. In *Proc. of COLING*, pages 888–899. Osaka, Japan.

Hui, Audrey Li Yen (2012). *Order and constituency in Mandarin Chinese*, volume 19. Springer.

Hwang, Jena D., Bhatia, Archna, Han, Na-Rae, O'Gorman, Tim, Srikumar, Vivek, and Schneider, Nathan (2017). Double trouble: the problem of construal in semantic annotation of adpositions. In *Proc. of *SEM*, pages 178–188. Vancouver, Canada.

Li, Bin, Wen, Yuan, Bu, Lijun, Qu, Weiguang, and Xue, Nianwen (2016). Annotating The Little Prince with Chinese AMRs. In *Proc. of LAW X – the 10th Linguistic Annotation Workshop*, pages 7–15. Berlin, Germany.

Li, Charles N. and Thompson, Sandra A. (1974). Co-verbs in Mandarin Chinese: verbs or prepositions? *Journal of Chinese Linguistics*, 2(3):257–278.

Li, Charles N. and Thompson, Sandra A. (1989). *Mandarin Chinese: A functional reference grammar*. Univ of California Press.

Li, Hui, Japkowicz, Nathalie, and Barrière, Caroline (2005). English to Chinese Translation of Prepositions. In Kégl, Balázs and Lapalme, Guy, editors, *Advances in Artificial Intelligence*, number 3501 in Lecture Notes in Computer Science, pages 412–416. Springer, Berlin.

Li, Mingqin, Li, Juanzi, Dong, Zhendong, Wang, Zuoying, and Lu, Dajin (2003). Building a large Chinese corpus annotated with semantic dependency. In *Proc. of the Second SIGHAN Workshop on Chinese Language Processing*, pages 84–91. Sapporo, Japan.

Litkowski, Ken (2014). Pattern Dictionary of English Prepositions. In *Proc. of ACL*, pages 1274–1283. Baltimore, Maryland, USA.

Litkowski, Ken and Hargraves, Orin (2005). The Preposition Project. In *Proc. of the Second ACL-SIGSEM Workshop on the Linguistic Dimensions of Prepositions and their Use in Computational Linguistics Formalisms and Applications*, pages 171–179. Colchester, Essex, UK.

Litkowski, Ken and Hargraves, Orin (2007). SemEval-2007 Task 06: Word-Sense Disambiguation of Prepositions. In *Proc. of SemEval*, pages 24–29. Prague, Czech Republic.

Lorincz, Kristen and Gordon, Rebekah (2012). Difficulties in learning prepositions and possible solutions. *Linguistic Portfolios*, 1(1):14.

Nivre, Joakim, Marneffe, Marie-Catherine de, Ginter, Filip, Goldberg, Yoav, Hajič, Jan, Manning, Christopher D., McDonald, Ryan, Petrov, Slav, Pyysalo, Sampo, Silveira, Natalia, Tsarfaty, Reut, and Zeman, Daniel (2016). Universal Dependencies v1: a multilingual treebank collection. In Calzolari, Nicoletta, Choukri, Khalid, Declerck, Thierry, Grobelnik, Marko, Maegaard, Bente, Mariani, Joseph, Moreno, Asuncion, Odijk, Jan, and Piperidis, Stelios, editors, *Proc. of LREC*, pages 1659–1666. Portorož, Slovenia.

Popović, Maja (2017). Comparing language related issues for NMT and PBMT between German and English. *The Prague Bulletin of Mathematical Linguistics*, 108(1):209–220.

Qi, Peng, Dozat, Timothy, Zhang, Yuhao, and Manning, Christopher D. (2018). Universal Dependency parsing from scratch. In *Proc. of CoNLL*, pages 160–170. Brussels, Belgium.

Schneider, Nathan, Hwang, Jena D., Bhatia, Archna, Srikumar, Vivek, Han, Na-Rae, O'Gorman, Tim, Moeller, Sarah R., Abend, Omri, Shalev, Adi, Blodgett, Austin, and Prange, Jakob (2019). Adposition and Case Supersenses v2.3: Guidelines for English. *arXiv:1704.02134v4 [cs]*. August 18 version: `https://arxiv.org/abs/1704.02134v4`.

Schneider, Nathan, Hwang, Jena D., Srikumar, Vivek, Green, Meredith, Suresh, Abhijit, Conger, Kathryn, O'Gorman, Tim, and Palmer, Martha (2016). A corpus of preposition supersenses. In *Proc. of LAW X – the 10th Linguistic Annotation Workshop*, pages 99–109. Berlin, Germany.

Schneider, Nathan, Hwang, Jena D., Srikumar, Vivek, Prange, Jakob, Blodgett, Austin, Moeller, Sarah R., Stern, Aviram, Bitan, Adi, and Abend, Omri (2018). Comprehensive supersense disambiguation of English prepositions and possessives. In *Proc. of ACL*, pages 185–196. Melbourne, Australia.

Schneider, Nathan and Smith, Noah A. (2015). A corpus and model integrating multiword expressions and supersenses. In *Proc. of NAACL-HLT*, pages 1537–1547. Denver, Colorado.

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

Shilon, Reshef, Fadida, Hanna, and Wintner, Shuly (2012). Incorporating linguistic knowledge in statistical machine translation: translating prepositions. In *Proc. of the Workshop on Innovative Hybrid Approaches to the Processing of Textual Data*, pages 106–114. Avignon, France.

Sun, Honglin and Jurafsky, Daniel (2004). Shallow semantic parsing of Chinese. In *Proc. of HLT-NAACL*, pages 249–256. Boston, Massachusetts, USA.

Tetreault, Joel R. and Chodorow, Martin (2008). The ups and downs of preposition error detection in ESL writing. In *Proc. of Coling*, pages 865–872. Manchester, UK.

Weller, Marion, Fraser, Alexander, and Schulte im Walde, Sabine (2015). Target-side generation of prepositions for SMT. In *Proc. of EAMT*, pages 177–184. Antalya, Turkey.

Weller, Marion, Schulte im Walde, Sabine, and Fraser, Alexander (2014). Using noun class information to model selectional preferences for translating prepositions in SMT. In *Proc. of the 11th Conference of the Association for Machine Translation in the Americas*, pages 275–287. Vancouver, Canada.

Xia, Fei (2000). The segmentation guidelines for the Penn Chinese Treebank (3.0). Technical Report IRCS-00-06, University of Pennsylvania, Philadelphia, PA.

Xue, Nianwen (2008). Labeling Chinese predicates with semantic roles. *Computational Linguistics*, 34(2):225–255.

Xue, Nianwen, Bojar, Ondřej, Hajič, Jan, Palmer, Martha, Urešová, Zdeňka, and Zhang, Xiuhong (2014). Not an interlingua, but close: comparison of English AMRs to Chinese and Czech. In Calzolari, Nicoletta, Choukri, Khalid, Declerck, Thierry, Loftsson, Hrafn, Maegaard, Bente, Mariani, Joseph, Moreno, Asuncion, Odijk, Jan, and Piperidis, Stelios, editors, *Proc. of LREC*, pages 1765–1772. Reykjavík, Iceland.

Yang, York Chung-Ho and Kuo, June-Jei (1998). The Chinese temporal coverbs, postpositions, coverb-postposition pairs, and their temporal logic. In *Proc. of PACLIC*, pages 20–32. Singapore.

You, Liping and Liu, Kaiying (2005). Building Chinese FrameNet database. In *2005 International Conference on Natural Language Processing and Knowledge Engineering*, pages 301–306. IEEE.

Zhang, Jianping (1999). *A Study of Language Model and Understanding Algorithm for Large Vocabulary Spontaneous Speech Recognition*. Ph.D. thesis, Tsinghua University.

Zhu, Yilun, Liu, Yang, Peng, Siyao, Blodgett, Austin, Zhao, Yushi, and Schneider, Nathan (2019). Adpositional Supersenses for Mandarin Chinese. *Proceedings of the Society for Computation in Linguistics*, 2(1):334–337.