Building a Treebank for Chinese Literature for Translation Studies

Hai Hu[†] **Yanting Li**[‡] **Yina Patterson**[†] **Zuoyu Tian**[†]
Yiwen Zhang[†] **He Zhou**[†] **Sandra Kübler**[†] **Chien-Jer Charles Lin**[†]
[†]Indiana University Bloomington [‡]Northwestern University
{huhai,yinama,zuoytian,yiwezhan,hzh1,skuebler,chiclin}@indiana.edu
yanting.li@northwestern.edu

Abstract

We present a new Chinese Treebank in the literary domain, the Treebank for Chinese Literature (TCL), with an aim to foster translation studies by providing an annotated collection of Chinese texts from both translated and non-translated literature. In the current stage, our constituency treebank consists of 2 069 trees, annotated and cross-checked by six Chinese linguists, following and adapting the Chinese Penn Treebank (CTB) annotation guidelines. We discuss the issues that we encountered while annotating literary texts, and we demonstrate the usefulness of our treebank by comparing it against the news portion of CTB, and by analyzing the syntactic features of non-translated literary texts and translationese in Chinese.

1 Introduction

Despite Chinese being one of the most widely spoken languages in the world, there is still a lack of diverse treebanks in terms of genres. The largest proportion of the most widely used Penn Chinese Treebank (CTB) (Xue et al., 2005) contains texts from the news domain (plus small samples from magazines, telephone transcripts, chat messages, etc.). To the best of our knowledge, the only large-scale, freely available constituency treebank in Chinese in a different domain is the Chinese Treebank in Scientific Domain (Chu et al., 2016), with text from Chinese scientific papers.

Without the availability of high-quality, expert-annotated treebanks in domains other than the above two, it is difficult for corpus linguists to compare syntactic features of multiple domains (Xiao, 2010; Zhang, 2012; Xiao and Hu, 2015), and it is difficult to train parsers beyond the news domain. Research on domain adaptation for parsing is limited by the few available domains covered in (Chinese) treebanks.

Our overarching goal is to develop a reliable parser for Chinese for translation studies of literary texts[1]. To this end, we present our initial effort to build a Chinese treebank for literary texts. Specifically, to enable the comparison of translated and non-translated Chinese, half of our texts are originally written in Chinese and the other half translated from English to Chinese. While our intention is to create a parser for translation studies, our treebank will be a valuable resource for stylistics, translation studies (Hu et al., 2018; Lin and Hu, 2018; Rubino et al., 2016), corpus linguistics research in Chinese (Wu et al., 2010), as well as for domain adaptation for Chinese parsing (Li et al., 2019). To the the best of our knowledge, our treebank is the first sizable Chinese treebank in the literary domain[2], and also the first designed specifically for translation studies.

The paper is structured as follows: Section 2 introduces the source text and the annotation guidelines. Then section 3 presents our annotation procedure and the final annotated treebank. In section 4, we analyze the linguistic characteristics of TCL, with reference to the widely used Penn Chinese Treebank. Additionally, we compare the translation and non-translation sections within TCL.

[1]We plan to use the parser to extend prior work on translationese (Hu and Kübler, 2020; Lin, 2017; Lin and Hu, 2018) to the domain of literature.

[2]We use "literary Chinese" to mean Chinese in the domain of literature, rather than "classical Chinese", which is sometimes also referred to as "literary Chinese".

Corpus	Example sentences
$TCL_{original}$	Ex.1: 身边的小贩儿嗓门儿比他还高，低着头用小叉子拢着豆芽粗吼着：豆芽儿，绿豆的，败火，贱卖，两毛了！ 'The peddler beside him had a higher voice than him, and he lowered his head gathering the bean sprouts with a small fork and roared: bean sprouts, mung bean sprouts, relieve heatiness, low prices, only twenty cents!' Ex.2: 后辈儿孙不负浩荡皇恩，深感五坛、八庙倒可少一点儿，可那老北京的小玩艺儿：溜个马，架个鹰，斗个蛐蛐儿，玩个鸟儿的，却绝对不能少。 'The descendants live up to the mighty emperor's grace, and feel that the altars and the temples can be a little less, but the games of old Beijing: walking the horses, falconry, cricket fighting and playing with birds, definitely cannot be less.'
$TCL_{translated}$	Ex.1: 价值的确是特殊的，因为它隐而不露，所以它当然会在日后增加，尤其当这些物品被后代们视若珍宝的时候。 'The value is indeed special. Because it is hidden, it will increase in the future, especially when these objects are viewed as treasures by the descendants.' Ex.2: 新闻传媒很快就对此失去了热情，警方遮遮掩掩不知所云，联邦调查局干脆说是地方当局的事而一推了之。 'The media soon lost interest in this; the police was trying to hide something and there was nothing concrete in their statements; FBI shirked their responsibility by saying it was an issue for the local authorities.'

Table 1: Example sentences from the original and translated section of TCL.

2 Treebank Development

2.1 Data Source

Starting from our goals of creating a treebank for original and translated Chinese literature, we have selected the literary subset from two widely used corpora of Chinese. Specifically, we use the Lancaster Corpus of Mandarin Chinese (LCMC) (McEnery and Xiao, 2004) as our source for original Chinese, and the Zhejiang-University Corpus of Translated Chinese (ZCTC) (Xiao et al., 2010) for translated Chinese. LCMC has been widely used in linguistic studies of Chinese (Duanmu, 2012; Song and Tao, 2009; Zhang, 2017). Similarly, ZCTC is considered a standard resource for translation studies in Chinese (Xiao, 2010; Xiao and Hu, 2015; Hu and Kübler, 2020).

We select the literature genre (index "K") from both corpora, which in both cases is composed of 29 texts, each about 100 sentences. The texts are from different literary works in the 90s[3], for example, *To Live* by Yu Hua, *Memoirs of a Geisha* by Arthur Golden. We chose to annotate an equal number of sentences from each of the 29 texts since sampling from a more diverse set of texts will enhance the representativeness of the treebank.

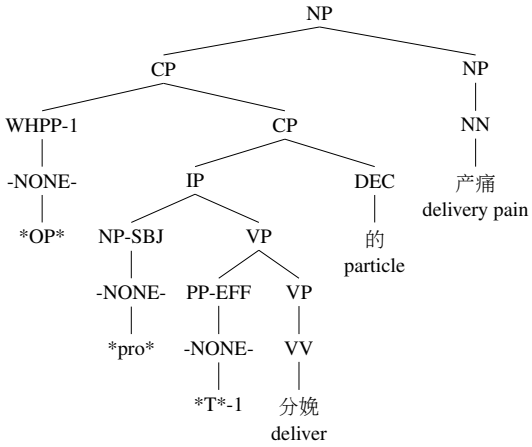
Both corpora have been segmented and part-of-speech (POS) tagged automatically using the Chinese Lexical Analysis System (Zhang et al., 2002). We did not use the segmentation and POS tags provided in the corpora because the segmentation and POS annotations are not compatible with those from the Chinese Penn Treebank, whose guidelines we follow for the syntactic annotation. In Table 1, we show example sentences from the two portions of TCL. These examples show that the language used in the literary texts is informal, and the translations show traces of English syntax.

2.2 Pre-processing

For sentence splitting, we split at the following types of punctuation signs: period (。), exclamation mark (！), question mark (？), semi-colon (；) and ellipsis (……). Then we used the default models and settings of the Stanford CoreNLP (Manning et al., 2014) to segment, POS tag, and parsed all sentences. The automatically analyzed sentences were then manually corrected by our annotators. Corrections include adjusting wrong segmentation, POS tags, and tree structures. Additionally, we add functional tags and empty categories according to our extended guidelines (see section 2.3).

In pre-processing, we encountered the following issues:

[3]The full lists can be found at https://www.lancaster.ac.uk/fass/projects/corpus/LCMC/lcmc/kat_k.htm and https://www.lancaster.ac.uk/fass/projects/corpus/ZCTC/source_K.htm.



Eng.: the pain caused by delivery (of a baby)

Figure 1: Example of an adjunct relative clause that has the new functional tag EFF.

OCR errors The text in LCMC are mostly "provided by the SSReader Digital Library in China", which has a 1-3% error rate in the OCR process (McEnery and Xiao, 2004). We corrected OCR errors if we were certain of the mis-recognized characters, based on the context and the shape of the characters: For example we corrected, 存人→ 存入 '*to deposit*', 陷阶→ 陷阱 '*trap*', and 村当于→ 相当于 '*equivalent to*'.

Normalization of punctuation signs We translated all the half-width punctuation signs to full-width ones, e.g., "." →"。", "?" →"？". We also normalized other punctuation signs, such as ellipsis, which are not consistent across LCMC and ZCTC.

2.3 Annotation Guidelines

We followed the guidelines of the Penn Chinese Treebank (Xue et al., 2005), but adopted modifications from the Chinese Treebank in Scientific Domain (SCTB) (Chu et al., 2016) where applicable for literary texts. We kept the constituent annotations in CTB as consistent with those of the Penn English Treebank (Marcus et al., 1993) as possible. The latest version of the treebank (V9) contains texts from the following genres: newswire, magazine articles, broadcast news/conversations, weblogs and discussion forums. No literary texts are included. CTB is based on the Theory of Government and Binding[4] (Chomsky, 1981), and uses empty categories and traces, which we also adopt in our annotation.

SCTB relies heavily on the annotation guidelines of CTB, but modifies them to better model scientific texts, such as creating specific POS tags for suffixes. Scientific writing is characterized by a high density of highly specialized technical terms created by suffixation. Since suffixation is very productive, as in VV + suffix (for example: 生育 '*breed*' + 期 '*period*' = 'breeding period'), SCTB treats these technical terms as two individual words and assigns separate POS labels to suffixes such as 期 '*period*'. We have incorporated those annotation rules of SCTB that are applicable for literary texts. We describe the most important extensions here[5].

Adjunct Relative Clauses Adjunct relative clauses are relative clauses where the gap in the relative clause is not clearly identifiable. For example, in Figure 1, the head noun "delivery pain" is not an argument (subject or object) in the relative clause `*pro* delivers`, but rather the *effect* or *result* that is caused by delivering a baby (see translation at bottom of Figure 1).

There has been much discussion in theoretical and psycho-linguistics on how such relative clauses are generated (Cha, 1999; Lin, 2018; Patterson, 2020; Ning, 1993). CTB treats all of these as a PP modifier inside the relative clause and provides several function tags to describe the functions of the head noun, for example, TMP (temporal) and MNR (manner). In our annotation of TCL, we found many cases of

[4] See (Xue et al., 2005, p. 4).

[5] We will release a full list of added annotation guidelines along with the treebank.

effect, where the head noun describes an effect resulting from the activity described in the relative clause. Consequently, we add EFF (effect) as a new functional tag in our guidelines.

Suffixes There are two suffixes that are frequent in our literary texts but uncommon in Chinese news texts. The first suffix is the *Erhua* (i.e., rhoticization) suffix 儿 *er* (from here on SFE), and the second is the plural suffix 们 *men* (from here on SFP).

Erhua is a morpho-phonological process that adds r-coloring, in the form of the suffix *er* [ɚ], to syllables in spoken Mandarin, the written form being 儿, as in 事儿 (*thing-er*, 'thingy'), 绝活儿 (*specialty-er*, 'claim to fame'). It is usually semantically vacuous and used in informal contexts to add a diminutive sense to the stem. In Beijing Mandarin, it has been used as a marker of local identity in contrast with a cosmopolitan global identity (Zhang, 2008). In the sampled news of CTB, we found no cases of rhoticization, but in $\text{TCL}_{\text{original}}$, there are 48 such cases. This is an indication that our literary texts are more informal and colloquial than CTB news.

The plural suffix, 们 *men*, is usually attached to animate nouns, which we decided to separate from the preceding noun and label as SFP in TCL. This suffix is more frequent in TCL (145 in $\text{TCL}_{\text{original}}$ and 219 in $\text{TCL}_{\text{translated}}$, compared to 53 in the sampled CTB) and has a wider range of metaphorical usage in that it can be attached after an inanimate noun such as 眼 'eye' in the literary genre, which is rarely found in news texts.

3 The Literary Chinese Treebank

Annotation team Our tree annotation team consists of six linguists (MA/PhD students in linguistics), all native speakers of Chinese. Additionally, two experienced (computational) syntacticians are available for consultation.

Annotation procedure The annotation process consisted of four phases. In the first phase, the annotators familiarized themselves with the CTB guidelines. In the second step, each annotator annotated 10 sentences, followed by a discussion of points of uncertainty and differences in annotation. In the third phase, each annotator was assigned 230 trees to annotate. Every tree was cross-checked by a different annotator. If differences occurred, they were discussed, and the trees were corrected if necessary. Annotation issues were discussed in weekly meetings. During this process, the extended guidelines were produced, covering new cases due to the linguistic differences between news and literature, and also documenting decisions in cases of inconsistencies in the CTB. With the enhanced guidelines, each annotator annotated an additional 100 trees, after which each tree was cross-checked by a different annotator.

Size Currently, the treebank consists of 2 069 trees: 1 029 from translated literature and 1 040 from original Chinese literature, amounting to 42 054 words. These sentences are sampled from 58 works of fiction from both LCMC and ZCTC (29 each).

Inter-annotator agreement (IAA) To compute IAA, our six annotators annotated the same 47 trees, and then had a discussion to decide on the gold standard for these sentences. We compute IAA as the averaged F-measure between an annotator's trees and the agreed upon final trees. This resulted in an agreement of 92.94%, thus indicating high agreement among our annotators.

4 Analysis of TCL

It is not always clear how to evaluate a treebank, and there are many angles to investigate. In this section, our intention is to document a range of differences that give an indication of how useful the addition of this treebank will be to the existing Chinese treebanks. The investigation is mainly driven by our goal of using the treebank for translation and contrastive linguistic studies. We first look at the overall statistics of complexity across the three treebank sections. Then we investigate differences between the news and literary genres, focusing on two phenomena that are less frequent or non-existent in the CTB. Finally, we look into differences between the original and translated portions of the TCL.

In order to perform the between-genre comparison, we sampled 1 040 trees from the CTB news portion to match the number of our annotated data in $\text{TCL}_{\text{original}}$. In sampling these CTB trees, we removed the

	$TCL_{original}$	$TCL_{translated}$	CTB	Tregex pattern
# sent	1 040	1 029	1 040	
mean sent. length	19.74	17.92	27.77	
mean word length	1.36	1.41	1.73	
vocab. size	4 439	4 026	6 012	
mean tree depth	10.73	10.94	11.25	
# rules	27 250	24 960	34 042	
# rule types	1 800	1 484	2 167	
entropy of rules	6.84	6.67	7.38	
per 1 000 words				
# IP	175.85	178.15	128.46	`/^IP/`
# CP	47.59	55.37	47.82	`/^CP/`
# subordinate clause	1.17	3.31	0.52	`/^CP/ <1 (/^ADVP/<CS)`
# relative clause	17.73	20.83	23.61	`/^P/ <1 /^WH(NP|PP)/`

Table 2: Statistics of subsets of TCL, in comparison with the sampled news section in CTB. (Sentence and word lengths are computed based on the number of syllables, which is equivalent to the number of monosyllabic morphemes in Chinese. Tree depth refers to the greatest number of syntactic levels embedded in a constituent.)

header and trailing information about the name of the reporter or the dates, and only kept the content of the news.

4.1 Linguistic Characteristics of TCL

Linguistic complexity Here, we compare the linguistic complexity across the different treebank sections. We chose complexity for several reasons. First, it is an important linguistic feature, receiving attention from various branches of linguistics, e.g., typology (Juola, 2008), corpus linguistics (Covington and McFall, 2010; Kettunen, 2014), psycholinguistics (Futrell et al., 2015; Gibson, 1998; Hawkins, 2004; Lin, 2018), and language acquisition (Lu, 2010; O'Grady, 1997). Second, in translation studies, a well-known hypothesis states that translated texts are lexically and syntactically simpler than texts originally written in a language (Baker, 1993; Baker, 1996). Empirical results of this *simplification* hypothesis have been mixed (Laviosa-Braithwaite, 1996; Ilisei and Inkpen, 2011; Volansky et al., 2013; Hu and Kübler, 2020). TCL can provide a high quality data source for evaluating this hypothesis.

Table 2 presents a range of statistics on the two subsets of TCL and the sampled news section of CTB. We first notice that news texts have considerably longer sentences, longer words, slightly deeper trees, a larger vocabulary size, as well as considerably more rules and rule types. By rules we mean all non-terminal context-free rules extracted from the trees, e.g., `NP -> DP ADJP NP`. Rule type refers to the number of unique rules. All these criteria suggest that news texts are syntactically more complex than their literary counterparts.

In the second part of Table 2, which focuses on grammatical rules, we calculated the entropy of the distribution of grammar rules. The numbers show that the news domain has a higher entropy, indicating more uncertainty and complexity of its grammar rule distribution. The numbers in the third part of the table, however, are more diverse: While both parts of TCL has a higher number of IPs[6] (indicating more main clauses) and a higher number of subordinate clauses, CTB has more relative clauses than both $TCL_{original}$ and $TCL_{translated}$. In terms of CPs (small clauses), the translated text $TCL_{translated}$ outnumbers both $TCL_{original}$ and CTB.

Focusing on $TCL_{original}$ and $TCL_{translated}$, we observe that the original literature domain is more complex in terms of mean sentence length, vocabulary size, as well as the number of rules and rule types. This lends some support for the simplification hypothesis at both the lexical and sentence levels. However, for the other measures in Table 2, the differences are either too small or even reversed. We will look at the simplification hypothesis more closely in section 4.4.

[6]These structures were extracted using `Tregex` patterns (Levy and Andrew, 2006).

	$TCL_{original}$		$TCL_{translated}$		CTB (news, sampled)	
No.	POS tag	Percentage	POS tag	Percentage	POS tag	Percentage
1	VV	17.55%	VV	16.34%	NN	28.30%
2	NN	16.29%	NN	15.23%	PU	12.46%
3	PU	13.06%	PU	12.43%	VV	11.73%
4	AD	10.09%	AD	9.85%	-NONE-	7.02%
5	-NONE-	7.37%	PN	7.84%	NR	6.37%
6	PN	4.89%	-NONE-	7.30%	AD	4.90%
7	M	2.80%	P	3.05%	P	3.68%
8	AS	2.75%	DEG	2.89%	CD	3.17%
9	NR	2.69%	VA	2.56%	JJ	3.00%
10	CD	2.63%	M	2.41%	M	2.87%

Table 3: The 10 most frequent POS tags in $TCL_{original}$ and CTB news (sampled).

POS distribution We also had a closer look at the distribution of POS tags in $TCL_{original}$ and CTB news, to check for differences on the morpho-sytactic level. Table 3 presents the 10 most frequent POS tags and their proportions per corpus. A comparison shows interesting differences:

One clear difference concerns the proportion of nouns (NN) in the two corpora. In $TCL_{original}$, 16.29% of the words are nouns, in CTB, the proportion is almost twice as high, 28.30%. The prominence of NN in news texts is in line with previous empirical results (e.g., Zhang (2012)). A more detailed analysis shows that 经济 '*economy*', 企业 '*enterprise*', 公司 '*company*', 发展 '*development*' and 国 '*country*' are the five most frequent nouns in CTB, compared to 人 '*human*', 事 '*thing*', 话 '*speech*', 家 '*home*' and 父亲 '*father*' in $TCL_{original}$. They also show the trend that monosyllabic nouns are generally preferred in spoken and less formal genres, as previously observed by Zhang (2012). The lower proportion of nouns in $TCL_{original}$ corresponds to a higher frequency of verbs (VV), which indicates the "verbi-ness" of Chinese literature texts (Zhang, 2012). Directly related is the high frequency of adverbs (AD) since literary texts tend to use more adverbs for detailed and vivid description of actions.

Previous corpus studies (e.g., Zhang (2017)) have shown that personal pronouns, especially in third person, are associated with narrative discourse while first and second persons are linked to interactive discourse. Our analysis provides supporting evidence: We see a much higher frequency of pronouns (PN) in literary texts overall: 4.89% in $TCL_{original}$ vs. 0.87% in news texts (ranked 18th in CTB, not shown in Table 3). This is due to the fact that literature uses both narrative and interactive discourse while news mainly uses narrative discourse. While 他 '*he*' is the most frequent pronoun in both texts, the other frequent pronouns have different distributions: In the literary texts, we have first and second person pronouns (我 '*I*', 你 '*you*') along with the reflexive (自己 '*self*'). In contrast, for news, we find the neutral third person pronoun, two demonstratives, and finally the first person pronoun: 其 '*it*', 此 '*this*', 这 '*this*' and 我 '*I*'.

We also observe a wider range of POS tags used in $TCL_{original}$. Apart from the two new tags we created for suffixes (SFE and SFP), there are two tags that occur in $TCL_{original}$ but not in CTB: IJ (interjection) and ON (onomatopoeia), both typical for colloquial expressions. From the POS distribution of $TCL_{translated}$, in contrast, we see that translated Chinese overuses pronouns (PN), prepositions (P) and the marker 的(DEG), confirming the results from previous translation studies in Chinese (Xiao and Hu, 2015; Hu et al., 2018; Hu and Kübler, 2020).

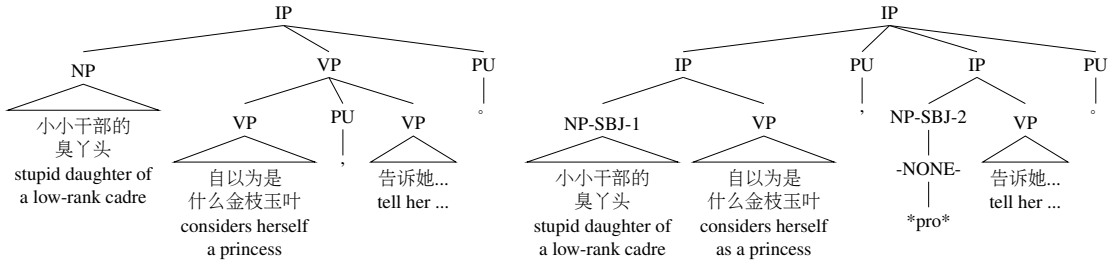
4.2 Comparing the News and Literary Genres

In this section, we provide a comparison of $TCL_{original}$ and CTB. We focus on two syntactic phenomena that are either less frequent in CTB or completely absent, (a) the pro-drop phenomena and (b) fragments and incomplete sentences. Both phenomena would cause lower parser performance in a domain adaptation scenario where the parser needs to parse literary texts but has been trained on CTB.

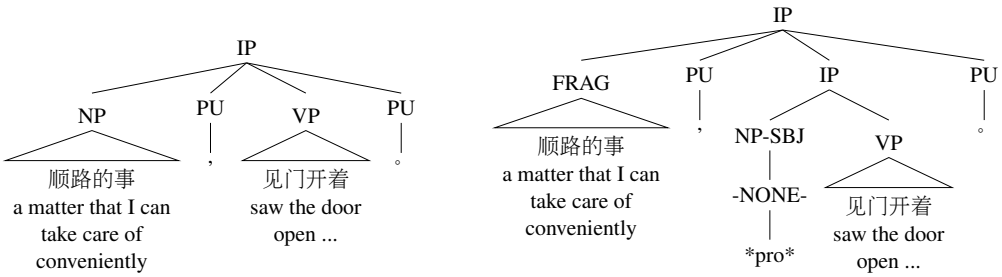
Pro-drop phenomena Chinese is known for its extensive use of pro-drop, especially in informal language. Xiao and Hu (2015) suggest that pro-drop is a significant indicator for specific genres. However, in order to test this hypothesis, they need syntactically annotated texts, or a parser that can produce

Structure	$TCL_{original}$	CTB (news, sampled)
pro-drop	614	343
pro-drop (per 1000 words)	30.0	11.9
subject pro-drop	602	334
object pro-drop	12	9

Table 4: Statistics of pro-drop phenomena in $TCL_{original}$ and CTB.



Eng.: The stupid daughter of a low-rank cadre considers herself a princess, [you] tell her ...



Eng.: (This is) A matter that I can take care of conveniently, [I] saw the door open, ...

Figure 2: Parsing errors involving pro-drop phenomena. Left: incorrect parser analysis. Right: gold tree in TCL. The dropped pronouns are in square brackets in the English translations.

empty categories. Since neither option was available, their hypothesis could not be tested empirically. However, the annotated $TCL_{original}$ and CTB do include empty categories, thus allowing us to investigate this hypothesis. We present the statistics of pro-drop in the two treebanks in Table 4. Since the treebanks contain a similar number of sentences, but CTB's sentences are considerably longer, we do not only report the absolute counts but also the counts normalized per 1 000 words. Pro-drop is much more common in literary texts: 614 occurrences in $TCL_{original}$ vs. 343 in CTB, or 30.0 normalized occurrences vs. 11.9.

Table 4 also shows that subject pro-drop is much more prevalent in both genres. Object pro-drop is rarely used and only occurs around 10 times in either treebank. However, the high percentage of subject pro-drop (602 cases in TCL0 vs. 334 cases in CTB) can provide challenges for the automatic parser and may cause systematic errors in the sentence structure. We show some parsing errors related to pro-drop in Figure 2.

In the first example, the gold tree is composed of two independent clauses: [NP_1 + VP_1] + [NP_2 (pro-drop) + VP_2], where the second clause has a dropped subject pronoun. However, since the parser cannot generate empty categories and would have to create an untypical IP with a single VP daughter, it failed to recognize the two clauses and instead grouped VP_1 and VP_2 into a coordinated VP with NP_1 acting as the shared subject. For the second example, we see that a dangling NP (a fragment) was incorrectly parsed as the subject whereas the correct analysis should insert a dropped pronoun in the subject position.

Fragments and incomplete phrases There are 30 fragments (`FRAG`) and incomplete phrases (`INC`) in $TCL_{original}$, which are often dangling PPs or NPs. In CTB, in contrast, the only fragments and incomplete phrases are found in the headers of the news articles, which we excluded from our sample. This means

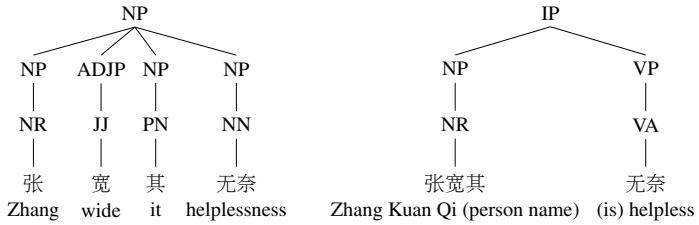


Figure 3: Parser error involving a person name. Left: parser mistake. Right: gold tree in TCL.

that in a formal genre such as news, all sentences are complete. Thus, if we train a parser on news texts, and then use it to parse literary texts, the parser may not be able to parse the incomplete structures in literary texts (see the second example in Figure 2).

This comparison only scratches the surface of the differences between the two genres. Considering the unique features of literary texts, our treebank will not only be a valuable resource for linguists interested in specific syntactic phenomena (such as pro-drop), but also be useful for building more reliable parsers for the literary domain.

4.3 Analysis of Parser Errors

Following the analysis above, we also looked at the actual parser errors. Since the trees in TCL are first automatically parsed using the default Chinese parser in Stanford CoreNLP (Manning et al., 2014) trained on news texts, we can analyze the errors in an out-of-domain parsing setting, after having manually corrected the trees. Here, we show two of the most common types of errors that the parser has made.

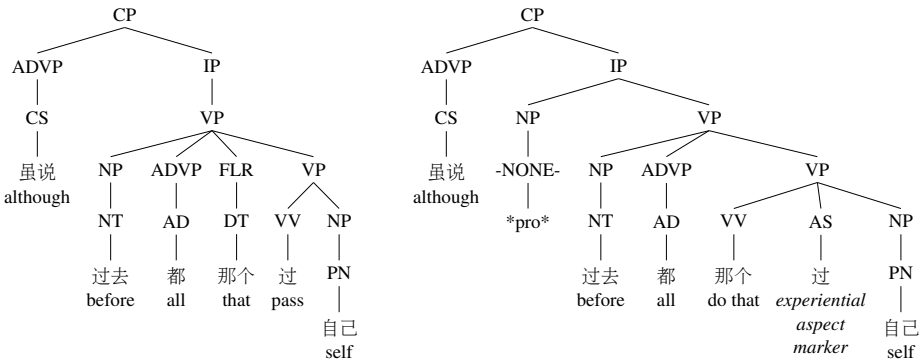
Named entities It is difficult for the parser to detect the named entities in literary texts, especially person names. We manually checked 20 named entities in 58 trees. Out of these 20 named entities, the parser only correctly recognized 8. All errors were due to over-segmentation. For example, the person name 苦根 (literally '*bitter root*') was segmented into two words and tagged `JJ NN`, rather than `NR` as a whole. The person name 张宽其 (literally '*Zhang wide he/it*') was segmented into three words and tagged `NR JJ PN`. I.e., only the surname 'Zhang' was recognized correctly (see Figure 3). There are also cases where a surname was labeled `VV` (e.g., 许, which can be a verb meaning '*allow*'). In general, the names in literary texts are more atypical and thus present a challenge to the parser.

Creative use of words In the literary treebank, there are cases where a word is used atypically, often as a part of speech different from its typical use. For instance, the word 臭 is an adjective meaning '*smelly/stinky*'. However, in one sentence, it is used as a verb meaning 'to trash (sth.)': 臭广告 'to trash the commercials'. The parser analyzed the phrase as an NP 'stinky commercials': (NP (JJ 臭) (NN 广告)). Another example is given in Figure 4. Here the demonstrative 那个 '*that*' is used as a verb to mean 'do so', which is a euphemism in spoken Chinese where the unspoken action it refers to needs to be reconstructed from the context. This type of flexibility and creative use in terms of parts of speech almost exclusively happens in literary texts. Such cases tend to lead to parse trees with very low accuracy since these wrong analyses require major changes to the rest of the tree.

4.4 Comparing Original and Translated Chinese Literary Texts

In this section, we have a closer look at the linguistic complexity of translated and original Chinese in literary texts.

As described above, one prominent hypothesis from translation studies states that translated texts are lexically and syntactically simpler than the texts originally written in the same language (Baker, 1993). This is often referred to as the *simplification hypothesis*, and is often assumed to be a universal feature of all translations. With our human-annotated, high-quality treebank, we can provide empirical evidence for/against the hypothesis in a language vastly different from Indo-European languages, for which the hypothesis has mostly been investigated (Ilisei and Inkpen, 2011; Volansky et al., 2013).



Eng.: ... although (they) all did that to (my)self before ...

Figure 4: Parser error involving creative use of words. Left: wrong parse from the parser. Right: gold tree in TCL.

	count		mean XP length in words			mean XP depth		
XP	orig	trans	orig	trans	p value	orig	trans	p value
CP	1277	1368	4.51	4.77	0.0623	5.60	5.80	0.0124
DNP	460	580	2.68	2.63	0.5738	3.45	3.44	0.8977
PP	676	704	3.49	**4.09**	0.0011	4.28	**4.70**	0.0011
NP	8830	8449	1.60	**1.72**	0.0001	2.64	**2.72**	0.0006
VP	9314	8090	4.14	4.31	0.0333	3.98	**4.25**	0.0
IP	3610	3285	10.83	10.95	0.6553	6.80	**7.27**	0.0
DP	344	350	1.60	1.51	0.1481	2.59	2.50	0.1214
ADVP	2216	2012	1.01	1.01	0.7233	2.01	2.01	0.2044
LCP	297	345	3.39	**4.10**	0.0014	4.11	**4.58**	0.0019
ADJP	384	279	1.03	**1.09**	0.0069	2.02	**2.06**	0.0025

Table 5: Statistics for XP structures in $\text{TCL}_{\text{original}}$ and $\text{TCL}_{\text{translated}}$. Greater values are in bold if $p < 0.01$, indicating more complexity, i.e., longer or deeper XP.

There are many ways to determine the complexity of sentences. Here we focus on two measures for linguistic complexity: the *length* and the *tree depth* of a linguistic unit. Specifically, we extract the treelets of the major phrases such as NPs, and VPs, and compare their complexity in literary texts of translated Chinese and those written in Chinese originally.

The comparisons of mean XP lengths and mean XP depths are shown in Table 5, along with the p values of the t-tests. For all the phrase types that show a significant difference between $\text{TCL}_{\text{original}}$ and $\text{TCL}_{\text{translated}}$, it is the *translated* texts that are more complex: PP, NP, LCP, and ADJP have longer mean lengths while PP, NP, VP, IP, LCP and ADJP have greater depths. This means that translated literary texts tend to have more complex (i.e., longer and deeper) linguistic units. These results contradict the simplification hypothesis and show that for many important phrases in Chinese, translations exhibit greater complexity.

While it is difficult to determine the exact reasons, for Chinese, these phrases are more complex in translations, there have been attempts. For example, Lin (2011) argues that the relative position of the modifier and the head inside a phrase has critical influence on human sentence processing. That is, for complex NPs with relative clauses, "the later the head noun is encountered, the greater temporary uncertainty exists in (human) parsing, and therefore the more difficult for (human) parsing" (Lin, 2011). Since Chinese is head-final in NPs and VPs (see the left two trees in Figure 5), long pre-head modifiers are generally dispreferred because they put too much processing pressure on the human processor. In contrast, English does not have such problems of "uncertainty" because the head precedes the modifier (see trees on the right in Figure 5), allowing the human processor to be able to comprehend and produce long RC and PP modifiers inside NPs and VPs respectively[7].

[7]We note that the issue of headedness has been extensively investigated by Liu (2010). Unfortunately, Liu (2010) only offers

Figure 5: NP and VP structures in Chinese (left) and English (right)

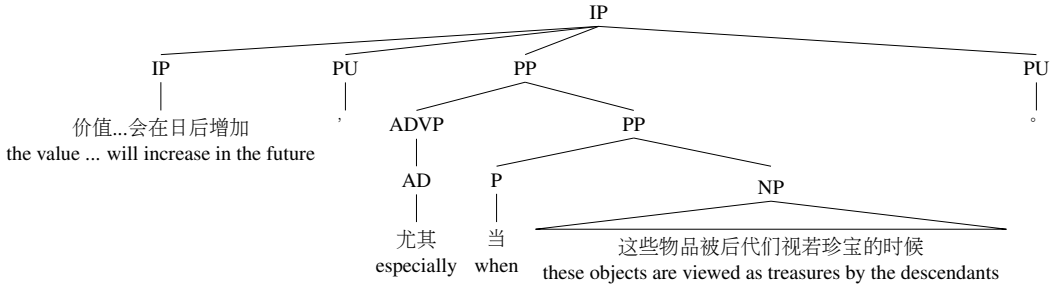


Figure 6: Example of a long dangling PP at the end of a sentence in $\text{TCL}_{\text{translated}}$, a feature of translated Chinese.

If the assumption that English has more complicated RCs and PPs is true (for which Lin (2011) provides preliminary corpus evidence), then the trend of longer PPs in English-to-Chinese translations that we find can be attributed to the *interference* effect; i.e., the syntax of the source language interferes with the production of the same structures in the translations (Toury, 1995).

We can further investigate this hypothesis in TCL. As an example, we find interference of word order from English PP structures in sentences in $\text{TCL}_{\text{translated}}$, as is illustrated by Figure 6. The sentence has a sentence-final PP, which is not the typical position for PPs in original Chinese. As shown in Figure 5, PPs usually precede the verbal head inside the VP in Chinese. The structure presented in Figure 6 is common in English as in *IP, especially when ...*. Furthermore, PPs of the structure "当..." (*when ...*) have been identified as a characteristic of Europeanized Chinese (Wang, 1944; He, 2008). Here we see an example, which gives an indication of the reason for this phenomenon: Chinese texts translated from English inherit the linear ordering of constituents.

In sum, our preliminary analysis provides counter-evidence for the simplification hypothesis but some evidence for the interference hypothesis. Putting together the findings in Table 5 and the results from Table 2, which showed that translations have shorter sentences but longer words and slightly deeper trees, we conclude that the simplification hypothesis may be an over-simplification of the complex correlations between translations and originals, and we may need a combination of the simplification and interference hypotheses to explain the syntactic differences between translations and originals.

5 Conclusion and Future Work

In this paper, we have presented the Treebank for Chinese Literature (TCL), a novel Chinese treebank in the literary domain. The treebank contains texts from both translated and original Chinese and is thus suitable for translation and contrastive linguistic studies. We have compared our treebank with the news section of the Penn Chinese Treebank, and we have carried out a comparison of the translated and original portions of the new treebank. We have shown significant differences between the treebanks, from which we conclude that having such a treebank will be invaluable not only for linguistic analyses of literary texts but also for training parsers.

statistics for subject-verb or adjective-noun orders, but not for PPs and RCs. Thus we leave it for future work to follow this line of research and use dependency treebanks to look into the order and complexity of PPs and RCs in Chinese.

Acknowledgements

We thank our anonymous reviewers for helpful suggestions. This work is supported by the Indiana University – Renmin University Strategic Seed Fund and a summer incubator grant from the Institute for Digital Arts and Humanities at Indiana University. He Zhou is funded by China Scholarship Council.

References

Mona Baker. 1993. Corpus linguistics and translation studies: Implications and applications. In Mona Baker, Gill Francis, and Elena Tognini-Bonelli, editors, *Text and Technology: In Honour of John Sinclair*, pages 233–250. Amsterdam: John Benjamins.

Mona Baker. 1996. Corpus-based translation studies: The challenges that lie ahead. In Harold Somers, editor, *Terminology, LSP and Translation. Studies in Language Engineering in Honour of Juan C. Sager*, volume 18, pages 175–186. Amsterdam and Philadelphia: Benjamins.

Jong-Yul Cha. 1999. Semantics of Korean gapless relative clause constructions. *Studies in the Linguistic Sciences*.

Noam Chomsky. 1981. *Lectures on Government and Binding: The Pisa Lectures*. Walter de Gruyter.

Chenhui Chu, Toshiaki Nakazawa, Daisuke Kawahara, and Sadao Kurohashi. 2016. SCTB: A Chinese treebank in scientific domain. In *Proceedings of the 12th Workshop on Asian Language Resources (ALR12)*, pages 59–67, Osaka, Japan.

Michael A Covington and Joe D McFall. 2010. Cutting the Gordian knot: The moving-average type-token ratio (MATTR). *Journal of Quantitative Linguistics*, 17(2):94–100.

San Duanmu. 2012. Word-length preferences in Chinese: A corpus study. *Journal of East Asian Linguistics*, 21(1):89–114.

Richard Futrell, Kyle Mahowald, and Edward Gibson. 2015. Large-scale evidence of dependency length minimization in 37 languages. *Proceedings of the National Academy of Sciences*, 112(33):10336–10341.

Edward Gibson. 1998. Linguistic complexity: Locality of syntactic dependencies. *Cognition*, 68(1):1–76.

John Hawkins. 2004. *Efficiency and Complexity in Grammars*. Oxford University Press.

Yang He. 2008. *A Study of Grammatical Features in Europeanized Chinese*. Commercial Press. In Chinese.

Hai Hu and Sandra Kübler. 2020. Investigating translated Chinese and its variants using machine learning. *Natural Language Engineering (Special Issue on NLP for Similar Languages, Varieties and Dialects)*.

Hai Hu, Wen Li, and Sandra Kübler. 2018. Detecting syntactic features of translated Chinese. In *Proceedings of the Second Workshop on Stylistic Variation*, pages 20–28.

Iustina Ilisei and Diana Inkpen. 2011. Translationese traits in Romanian newspapers: A machine learning approach. *International Journal of Computational Linguistics and Applications*, 2(1-2):319–32.

Patrick Juola. 2008. Assessing linguistic complexity. *Language complexity: Typology, Contact, Change*, 89:107.

Kimmo Kettunen. 2014. Can type-token ratio be used to show morphological complexity of languages? *Journal of Quantitative Linguistics*, 21(3):223–245.

Sara Laviosa-Braithwaite. 1996. *The English Comparable Corpus (ECC): A Resource and a Methodology for the Empirical Study of Translation*. Ph.D. thesis, University of Manchester.

Roger Levy and Galen Andrew. 2006. Tregex and Tsurgeon: Tools for querying and manipulating tree data structures. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation*, pages 2231–2234, Genoa, Italy.

Zhenghua Li, Xue Peng, Min Zhang, Rui Wang, and Luo Si. 2019. Semi-supervised domain adaptation for dependency parsing. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2386–2395.

Chien-Jer Charles Lin and Hai Hu. 2018. Syntactic complexity as a measure of linguistic authenticity in modern Chinese. In *26th Annual Conference of International Association of Chinese Linguistics and the 20th International Conference on Chinese Language and Culture*, Madison, WI.

Chien-Jer Charles Lin. 2011. Chinese and English relative clauses: Processing constraints and typological consequences. In *Proceedings of the 23rd North American Conference on Chinese Linguistics (NACCL-23)*, Eugene, OR.

Chien-Jer Charles Lin. 2017. Head-modifier relations in Europeanized Chinese: Linguistic authenticity and sentence processing. In *29th North American Conference on Chinese Linguistics (NACCL)*.

Chien-Jer Charles Lin. 2018. Subject prominence and processing filler-gap dependencies in prenominal relative clauses: The comprehension of possessive relative clauses and adjunct relative clauses in Mandarin Chinese. *Language*.

Haitao Liu. 2010. Dependency direction as a means of word-order typology: A method based on dependency treebanks. *Lingua*, 120(6):1567–1578.

Xiaofei Lu. 2010. Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics*, 15(4):474–496.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics*, pages 55–60, Baltimore, MD.

Mitchell Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Anthony McEnery and Zhonghua Xiao. 2004. The Lancaster Corpus of Mandarin Chinese: A corpus for monolingual and contrastive language study. In *LREC*, pages 1175–1178.

Chunyan Ning. 1993. *The Overt Syntax of Topicalization and Relativization in Chinese*. Ph.D. thesis, University of California, Irvine, CA.

William O'Grady. 1997. *Syntactic development*. University of Chicago Press.

Yina Patterson. 2020. *A study of nominal-clausal relations in Mandarin Chinese*. Ph.D. thesis, Indiana University Bloomington, IN.

Raphael Rubino, Ekaterina Lapshinova-Koltunski, and Josef Van Genabith. 2016. Information density and quality estimation features as translationese indicators for human translation classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 960–970.

Zuoyan Song and Hongyin Tao. 2009. A unified account of causal clause sequences in Mandarin Chinese and its implications. *Studies in Language*, 33(1):69–102.

Gideon Toury. 1995. *Descriptive Translation Studies and Beyond*. John Benjamins.

Vered Volansky, Noam Ordan, and Shuly Wintner. 2013. On the features of translationese. *Digital Scholarship in the Humanities*, 30(1):98–118.

Li Wang. 1944. *Theory of Chinese Grammar*. Commercial Press. In Chinese.

Fuyun Wu, Elsi Kaiser, and Elaine Andersen. 2010. Subject preference, head animacy and lexical cues: A corpus study of relative clauses in Chinese. In *Processing and Producing Head-Final Structures*, pages 173–193. Springer.

Richard Xiao and Xianyao Hu. 2015. *Corpus-Based Studies of Translational Chinese in English-Chinese Translation*. Springer.

Richard Xiao, Lianzhen He, and Ming Yue. 2010. In pursuit of the third code: Using the ZJU corpus of translational Chinese in translation studies. In *Using Corpora in Contrastive and Translation Studies*, pages 182–214. Cambridge Scholars Newcastle.

Richard Xiao. 2010. How different is translated Chinese from native Chinese?: A corpus-based study of translation universals. *International Journal of Corpus Linguistics*, 15(1):5–35.

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

Kevin Zhang, Qun Liu, Hao Zhang, and Xue-Qi Cheng. 2002. Automatic recognition of Chinese unknown words based on roles tagging. In *Proceedings of the first SIGHAN workshop on Chinese language processing-Volume 18*, pages 1–7. Association for Computational Linguistics.

Qing Zhang. 2008. Rhotacization and the “Beijing Smooth Operator": The social meaning of a linguistic variable. *Journal of Sociolinguistics*, 12(2):201–222.

Zheng-Sheng Zhang. 2012. A corpus study of variation in written Chinese. *Corpus Linguistics and Linguistic Theory*, 8(1):209–240.

Zheng-Sheng Zhang. 2017. *Dimensions of Variation in Written Chinese*. Routledge.