L1-L2 Parallel Treebank of Learner Chinese: Overused and Underused Syntactic Structures

Keying Li, John Lee
Department of Linguistics and Translation
City University of Hong Kong
keyingli3-c@my.cityu.edu.hk, jsylee@cityu.edu.hk

Abstract

We present a preliminary analysis on a corpus of texts written by learners of Chinese as a foreign language (CFL), annotated in the form of an L1-L2 parallel dependency treebank. The treebank consists of parse trees of sentences written by CFL learners ("L2 sentences"), parse trees of their target hypotheses ("L1 sentences"), and word alignment between the L1 sentences and L2 sentences. Currently, the treebank consists of 600 L2 sentences and 697 L1 sentences. We report the most overused and underused syntactic relations by the CFL learners, and discuss the underlying learner errors.

Keywords: learner corpus, parallel treebank, Chinese as a foreign language

1. Introduction

Learner corpora, which consist of texts written by non-native speakers, are increasingly used in quantitative studies in second language acquisition. Some of these corpora have been annotated to answer various research questions. To support analysis of grammatical mistakes made by learners, a number of them have been error-tagged (e.g., Yannakoudakis et al., 2011; Dahlmeier et al., 2013; Lee et al., 2016). To better characterize learner syntax, others have been part-of-speech (POS) tagged (e.g., Díaz-Negrillo et al., 2010; Reznicek et al., 2013), and syntactically analysed with constituent trees (e.g., Nagata and Sakaguchi, 2016) and dependency trees (e.g., Ragheb and Dickinson, 2014; Berzak et al., 2016).

Building on learner treebanks, Lee et al. (2017b) proposed to use "L1-L2 parallel treebanks" — parse trees of non-native sentences ("L2 sentences") aligned to their target hypotheses ("L1 sentences") — to facilitate analyses of learner language. Figure 1 shows an example tree pair. It includes the parse tree of the learner sentence and of its target target hypothesis, both annotated in the Universal Dependencies (UD) scheme for Chinese (Leung et al., 2016; Lee et al., 2017), as well as word alignments between the two sentences. Such a treebank has the potential to enhance Contrastive Interlanguage Analysis (CIA) (Granger, 2015) and Error Analysis (EA) by supporting a greater range of automatic, quantitative studies. For CIA, they would enable comparisons between native and interlanguages not only on the lexical level but also on the syntactic level. For EA, parallel parse trees would give more fine-grained characterization of the syntactic environment in which learner errors occur.

This paper reports on the construction of an L1-L2 parallel treebank for Chinese and presents a preliminary analysis. After summarizing previous work (Section 2), we give details on the texts in the treebank (Section 3) and on the linguistic annotations (Section 4). We then discuss the most overused and underused syntactic structures in the learner texts, as well as the underlying errors (Section 5). Finally, we conclude in Section 6.

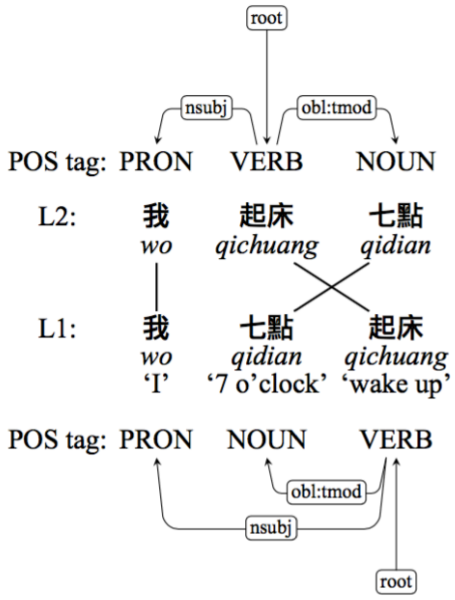

Figure 1: An example L1-L2 tree pair, including word alignments between the learner sentence ("L2") and its target hypothesis ("L1"), and the parse trees of the two sentences, annotated in the Universal Dependencies scheme for Chinese (Leung et al., 2016; Lee et al., 2017a).

2. Previous work

Most annotation efforts on learner treebanks have focused on English. Currently, the two major dependency treebanks for learner language are the Treebank of Learner English (TLE) (Berzak et al., 2016) and the project on Syntactically Annotating Learner Language of English (SALLE) (Ragheb and Dickinson, 2014). They both contain English texts written by non-native speakers. TLE annotates a subset of sentences from the Cambridge FCE corpus (Yannakoudakis et al., 2011), while SALLE has been applied on essays written by university students. A phrase-structure treebank for learner English (Nagata and Sakaguchi, 2016) has also been constructed for the texts in the Konan-JIEM Learner Corpus (Nagata et al., 2011). None of these treebanks, however, are L1-L2 parallel treebanks: they either do not provide explicit target

hypotheses (Ragheb and Dickinson, 2014; Nagata and Sakaguchi, 2016), or have not yet provided parse trees for the target hypotheses (Berzak et al., 2016).

As interest grows in learning Chinese as a foreign language (CFL), a number of large CFL corpora have been compiled and annotated (e.g., Zhang, 2009; Wang et al., 2015; Lee et al., 2016). Lee et al. (2017a) reported the first attempts to perform dependency annotation on CFL texts.

Lee et al. (2017b) described a case study on a small-scale L1-L2 parallel treebank, to gauge the potential of using search queries on the treebank as dynamically defined error categories. Errors are typically marked in a learner corpus with error tags, each of labels a problematic text span with an error category, and sometimes provides a corrected version of the text span (Yannakoudakis et al., 2011; Dahlmeier et al., 2013). From an L1-L2 parallel treebank, a researcher can retrieve sentences that exhibit a particular kind of error with a search query that consists of the desired tree patterns with word or node alignments. The case study used such queries to identify various kinds of word-order errors in CFL texts, and argued that they can supplement error tagsets by giving researchers more flexibility and precision in error definition (Lee et al., 2017b).

3. Textual material

The treebank contains sentences written by students of Chinese as a foreign language (CFL), as well as the target hypotheses of these sentences.

3.1 Learner sentences

The texts in our corpus have been collected from CFL learners of Chinese at Xi'an Normal University, China. There are a total of 22 learners, at both intermediate and advanced levels, with 5 different native language backgrounds: Korean, Russian, Turkish, Arabic and English. To protect their privacy, we anonymized their data but retained information on their gender, number of years of CFL studies, and native language.

The students received Chinese-language classes two to three times per month at the university. These classes offered detailed instructions on writing skills, e.g., thematic exercises and trainings in aspects of diction, punctuation, sentence, together with understandings of different logic connections under various native language backgrounds. After class, students were required to submit narratives, with diverse topics such as "The most memorable trip", "Experience of cultural differences", "My family members", etc. Based on scanned copies of their essays, we manually transcribed them in digital format. So far, we have collected a total of 27 essays, which consist of 600 sentences.

3.2 Target hypotheses

It is well known that there can be multiple target hypotheses for each learner sentence (Reznicek et al., 2013). One can focus on "minimal edits", which aim to make the sentence grammatically correct with the shortest edit distance; one can also perform "fluency edits", as advocated by Sakaguchi et al. (2016), which aim to make the sentence not only grammatical correct but also native-like, often involve rewriting the sentence as a whole. Even when performing the same kind of edits, individual annotators may come up with multiple valid target hypotheses.

For our treebank, a native speaker of Mandarin Chinese performed corrections on the 600 L2 sentences to produce a "minimal edit" target hypothesis for each sentence. An L2 sentence may be split up, or several L2 sentences may be combined. There are a total of 697 L1 sentences (target hypotheses) in the treebank. In future work, we plan to include "minimal edit" target hypotheses from other annotators, as well as "fluency edit" hypotheses, and study how multiple hypotheses affects interlanguage analysis.

4. Dependency annotation

Our learner corpus is annotated in the form of an L1-L2 parallel dependency treebank. The treebank consists of sentences written by CFL learners ("L2 sentences"); their target hypothesis ("L1 sentences"); the parse trees of the L1 and L2 sentences; and word alignment between the L1 sentences and L2 sentences. Figure 1 shows an example of a tree pair in our treebank.

We performed manual word segmentation, POS tagging, and dependency annotations on all sentences. Both the L1 and L2 sentences were annotated in the Universal Dependencies (UD) framework (Nivre et al., 2016). We have chosen to adopt the UD framework because of the large variety of languages for which UD treebanks exist, which can potentially facilitate contrastive analysis. Consider an investigation on the transfer hypothesis in texts written by CFL learners whose native language is X. One can examine differences in the aligned L1-L2 sentences within the portion of the treebank produced by native speakers of X. One can then further evaluate the extent to which UD treebanks of language X and of Chinese, ideally with comparable text types and topics, exhibit similar differences in linguistic properties.

For the annotation of L1 sentences, we followed the UD guidelines for Chinese (Leung et al., 2016). For the annotation of L2 sentences, we adapted the UD guidelines to take interlanguage characteristics into account (Lee et al., 2017a). Similar to current treebanks for learner English (Section 2), our guidelines adhere to the principle of "literal annotation", which asks annotators to perform syntactic analysis "as if the sentence were as syntactically well-formed as it can be, possibly ignoring meaning" (Ragheb and Dickinson, 2014). Dependency analysis on learner sentences can be challenging. According to a study on inter-annotator agreement (Lee et al., 2017a), the overall agreement can reach 94.0% for POS tags and 82.8% for labelled attachment. The agreement rate was lower within text spans that contain learner errors, dropping to 91.0% for POS tags and 75.1% for labelled attachment.

5. Overuse and underuse analysis

We would like to identify the grammatical structures with which the CFL learners in our corpus experienced the most difficulty. We applied the log-likelihood statistic (Rayson,

2008) to find the dependency relations that are most overused or underused among the L2 sentences (learner sentences), with respect to the L1 sentences (target hypotheses).

Table 1 shows the four syntactic relations, in the same Universal Dependencies (UD) scheme for Chinese (Leung et al., 2016; Lee et al., 2017a), that exhibit the most significant deviations in the learner sentences. The most underused dependency relation is `mark:adv`, which corresponds to the adverbializer. The most overused ones ones are `parataxis`, `discourse:sp`, and `compound:dir`, which correspond to parataxis, sentence-final particles and directional verb compounds. We now discuss the learner errors underlying these overuse and underuse phenomena.

Linguistic Structure	**UD Relation**	**LL Score**	**L2 Freq**	**L1 Freq**
Adverbalizer	`mark:adv`	4.38	31	50
Parataxis	`parataxis`	2.76	259	224
Sentence-final particles	`discourse:sp`	1.7	151	130
Directional verb compound	`compound:dir`	1.02	37	29

Table 1: Linguistic structures that are most overused and underused by learners, shown with their corresponding relations in Universal Dependencies (UD) and log-likelihood (LL) scores

5.1 Parataxis

Many learner sentences are written in a colloquial style, with multiple clauses placed side-by-side without any linking words between them. Although each clause is syntactically well-formed, the overall result is a run-on sentence. Consider the sentence in Table 2. Its first two clauses, headed by *kan* 'see' and *chuxian* 'appear', are linked with the relation `parataxis`(*kan*, *chuxian*).

A target hypothesis for a run-on sentence may sometimes be constructed by inserting appropriate conjunctions, but when the clauses are related only in a discursive way, the sentence may need to be split into several smaller, independent sentences. As shown in Table 2, the `parataxis` relations are thus replaced with `root`. Due to the abundant number of run-on sentences, the `parataxis` relation turned out to be the most overused one in our treebank. This phenomenon is also reflected in the higher number of L1 sentences (697 sentences) compared to L2 sentences (600 sentences).

5.2 Sentence-final particles

In Chinese, sentence-final particles, such as *le* 了, *de* 的, and *ba* 吧, can be placed at the end of clauses or sentences. They have a wide range of functions, such as modifying the modality, and expressing discourse and pragmatic information. In our UD scheme, it is annotated as a child (modifier) of the main predicate in a `discourse:sp` relation. As shown in Table 1, `discourse:sp` is among the most overused relations in the treebank, which suggests that errors related to the use of sentence-final particles is a frequent error type in learner Chinese text.

Among the various sentence-final particles, *le* is most overused. Table 3 shows an example sentence with an unnecessary *le*, and hence a superfluous relation `discourse:sp`(*gandong* 感动 'moved', *le* 了). In general, *le* should not be used in a simple assertion of a past event that did not involve a change of state (Li and Thompson, 1989).

The overuse of sentence-final particles appears to correlate with the native language. Native speakers of Korean, for example, overuse *le* more often than native speakers of English. This may be explained by the fact that Korean, similar to Chinese, uses sentence-final particles to mark clause types (Pak, 2006), while English does not. This finding also corresponds to a study of two CFL learners' acquisition of *le* (Sun, 1993).

Overuse of the `discourse:sp` relation is also in part due to the confusion between *le* as a sentence-final particle and *le* as an aspect marker. Though they share the same form, they function differently and appear at different positions. A verb typically occurs with the aspect marker *le* if the direct object is definite (Li and Thompson, 1989). Learners sometimes place it by mistake at the end of the sentence, such as in the sentence **yisheng gei wo kai yiping shenjinganyao le* 医生给我开一瓶神经安药了 'The doctor gave me some drugs for the nervous system'. It is annotated as a sentence-final particle, producing a superfluous `discourse:sp` relation.

Despite its overall trend of overuse, learner usage of *le* exhibits much variation. There are also many cases of omission, for example when the predicate is an Accomplishment or an Achievement verb. In such cases, *le* is required to present the resultative state after the attainment of the goal, such as in the sentence *women likai de shijian dao le* (我们离开的时间到了) 'Our time to leave has come'. Failure to use *le* in this context is also a rather common error in our data.

5.3 Directional verb compound

A directional verb compound is frequently used to express motion events, and is a typological feature of Chinese as a serial-verb language (Peyraube, 2006). The compound consists of at least two verbs, where the second verb is a the directional or deictic motion verb. Consider a sentence such as *ta pa shang le dingfeng* 他爬上了顶峰 'he climbed up to the peak'. In the two-verb series *pa* 'climb' and *shang* 'up', the second verb *shang* serves as a deictic motion verb. The UD scheme marks this type of compound with the relation `compound:dir`(*pa*, *shang*). As shown in Table 1, `compound:dir` is one of the more overused relations, indicating that unnecessary use of directional verbs is a significant learner error.

L2: … *偶然看到微信有一个功能是附近的人，我按了以后附近的人中出现了他的微信号，我申请加了他 …   … **ouran kan dao weixin you yi ge gongneng shi fujin de ren, wo an le yihou fujin de ren zhong chuxian le ta de weixin hao, wo shenqing jia le ta, …*   'By chance I saw a feature in WeChat called "People Nearby", after I clicked the button his number appeared among the people nearby, I sent him a friend invite, …'   `parataxis`(*kan* 看 'see', *chuxian* 出现 'appear')
L1: … 偶然看到微信有一个功能是附近的人。我按了以后，附近的人中出现了他的微信号。我申请加了他 …   *… ouran kan dao weixin you yi ge gongneng shi fujin de ren. wo an le yihou fujin de ren zhong chuxian le ta de weixin hao. wo shenqing jia le ta, …*   'By chance I saw a feature in WeChat called "People Nearby". After I clicked the button, his number appeared among the people nearby. I sent him a friend invite, …'   `root`(*kan* 看 'see')   `root`(*chuxian* 出现 'appear')

Table 2. Example of L1 and L2 sentences illustrating overuse of the `parataxis` relation (see Section 5.1).

L2: *但我去"麦加"朝觐时我很感动了   **dan wo qu "maijia" chaojin shi wo hen gandong le*.   'But when I went on pilgrimage to Mecca, I was very moved.'   `discourse:sp`(*gandong* 感动 'moved', *le* 了)
L1: 但我去"麦加"朝觐时我很感动   'But when I went on pilgrimage to Mecca, I was very moved.'

Table 3. Example of L1 and L2 sentences illustrating overuse of the `discourse:sp` relation (see Section 5.2).

A qualitative analysis suggested that many of these errors involve Chinese directional verbs that can play multiple roles: used independently (e.g., *xia* 下 'descend', *jin* 进 'enter') or modified by a directional complement (e.g., *xialai* 下来 'come down', *jinlai* 进来 'come in'); or serving as a directional complement itself (*fangxia* 放下 'put down', *paojin* 跑进 'run into'). Table 4 shows an example where *jin* 'enter' unnecessarily takes the directional complement *lai* 'come', yielding the superfluous relation `compound:dir`(*jin*, *lai*).

L2: *她进来了教室，坐在我左边。   **ta jin lai le jiaoshi, zuo zai wo zuobian.*   'She came into the classroom and sat on my left.'   `compound:dir`(*jin* 进 'came', *lai* 来 'into')
L1:她进了教室，坐在我左边。   *ta jin le jiaoshi, zuo zai wo zuobian.*   'She entered the classroom and sat on my left.'

Table 4. Example of L1 and L2 sentences illustrating overuse of the `compound:dir` relation (see Section 5.3).

5.4 Adverbializer

As shown in Table 1, the most underused structure in the learner texts is the adverbializer. In Chinese, the manner adverbializer *de* 地 turns an adjective that follows it into an adverb. Consider the sentence *nabian de ren qiguai de kan zhe wo* 那边的人奇怪地看着我 'The person over there looked at me strangely'. In this sentence, the adverbializer *de* turns the preceding adjective *qiguai* 'strange' into the adverb 'strangely'. In our UD scheme, this grammatical function is annotated as the relation `mark:adv`(*qiguai*, *de*).

The adverbializer *de* is easily confusable with a homonym, the particle *de* 的, which may follow an adjective that modifies a noun. For example, in the noun phrase *qiguai de wenti* 奇怪的问题 'strange problem', the particle *de* is inserted between the adjective *qiguai* 'strange' and the noun *wenti* 'problem'. The UD scheme uses the relation `mark:rel`(*qiguai*, *de*) to mark this structure.

The adverbializer *de* and the particle *de* share a similar linguistic environment in that they both modify adjectives. In our data, learners tend to overuse the particle and underuse the adverbializer. As shown in the example in Table 5, this phenomenon results in the relation `mark:rel`(*tebie* 特别 'special', *de* 的) in place of the expected relation `mark:adv`(*tebie* 特别 'special', *de* 地). This kind of error was a major contributor to the underuse of `mark:adv`.

6. Conclusions

We have described an on-going effort to build a large-scale L1-L2 parallel dependency treebank — i.e., parse trees of non-native sentences ("L2 sentences"), aligned to the parse trees of their target hypotheses ("L1 sentence") — for Chinese. The treebank is annotated in the Universal Dependencies (UD) framework.

We presented a preliminary analysis on the treebank, identifying the most overused and underused syntactic

relations in the learner text, with respect to the log-likelihood score. The adverbializer is the most underused, while parataxis, sentence-final particles, directional verb compounds are the most underused.

We are currently expanding the size of the treebank, including the number and kinds of target hypotheses. In future work, we plan to perform a most exhaustive analysis of overused and underused grammatical structures, and apply the treebank data to evaluate the transfer hypothesis in conjunction with other UD treebanks.

L2: *当时我特别的同感他们 **dangshi wo tebie de tonggan tamen* 'At that time, I felt special sympathetic to them.' `mark:rel`(*tebie* 特别 'special', *de* 的)
L1: 当时我特别地同感他们 *dangshi wo tebie de tonggan tamen* 'At that time, I felt especially sympathetic to them.' `mark:adv`(*tebie* 特别 'special', *de* 地)

Table 5. Example of L1 and L2 sentences illustrating underuse of the `mark:adv` relation (see Section 5.4).

7. Acknowledgements

The work reported in this paper was partially supported by a Strategic Research Grant (Project no. 7004494) from City University of Hong Kong.

8. Bibliographical References

Berzak, Y., Kenney, J., Spadine, C., Wang, J. X., Lam, L., Mori, K. S., Garza, S., and Katz, B. (2016). Universal Dependencies for Learner English. In *Proc. ACL*.

Dahlmeier, D., Ng, H. T., and Wu, S. M. (2013). Building a Large Annotated Corpus of Learner English: The NUS Corpus of Learner English. In *Proc. 8th Workshop on Innovative Use of NLP for Building Educational Applications*.

Díaz-Negrillo, A., Meurers, D., Valera, S., and Wunsch, H. (2010). Towards Interlanguage POS Annotation for Effective Learner Corpora in SLA and FLT. *Language Forum*, 36(1-2):139–154.

Granger, S. (2015). Contrastive Interlanguage Analysis: A Reappraisal. *International Journal of Learner Corpus Research* 1(1):7–24.

Lee, L. H., Gaoqi, R. A. O., Yu, L. C., Endong, X. U. N., Zhang, B., & Chang, L. P. (2016). Overview of NLP-TEA 2016 Shared Task for Chinese Grammatical Error Diagnosis. In *Proc. 3rd Workshop on Natural Language Processing Techniques for Educational Applications (NLPTEA2016)*.

Lee, J., Leung, H., & Li, K. (2017a). Towards Universal Dependencies for Learner Chinese. In *Proc. NoDaLiDa Workshop on Universal Dependencies*.

Lee, J., Li, K., & Leung, H. (2017b). L1-L2 Parallel Dependency Treebank as Learner Corpus. In *Proc. 15th International Conference on Parsing Technologies*.

Leung, H., Poiret, R., Wong, T. S., Chen, X., Gerdes, K. and Lee, J. (2016). Developing Universal Dependencies for Mandarin Chinese. In *Proc. 12th Workshop on Asian Language Resources*.

Li, C. N. and Thompson, S. A. (1989). *Mandarin Chinese: A Functional Reference Grammar*. University of California Press.

Nagata, R., Whittaker, E., and Sheinman, V. (2011). Creating a Manually Error-tagged and Shallow-parsed Learner Corpus. In *Proc ACL*.

Nagata, R. and Sakaguchi, K. (2016). Phrase Structure Annotation and Parsing for Learner English. In *Proc. ACL*.

Nivre, J., de Marneffe, M.-C., Ginter, F., Goldberg, Y., Hajic, J., Manning, C. D., McDonald, R., Petrov, S., Pyysalo, S., Silveira, N., Tsarfaty, R., and Zeman, D. (2016). Universal Dependencies v1: A Multilingual Treebank Collection. In *Proc. LREC*.

Pak, M. (2006). Jussive clauses and agreement of sentence final particles in Korean. *Japanese/Korean Linguistics*, *14*, 295-306.

Peyraube, A. (2006). Motion events in Chinese. *Space in languages: Linguistic systems and cognitive categories*, 66, 121-135.

Ragheb, M. and Dickinson, M. (2014). Developing a Corpus of Syntactically-Annotated Learner Language for English. In *Proc. 13th International Workshop on Treebanks and Linguistic Theories (TLT)*.

Rayson, P. (2008). From Key Words to Key Semantic Domains. *International Journal of Corpus Linguistics*, *13*(4), 519-549.

Reznicek, M., Lüdeling, A., and Hirschmann, H. (2013). Competing Target Hypotheses in the Falko Corpus: A Flexible Multi-Layer Corpus Architecture. In Ana Díaz-Negrillo, editor, *Automatic Treatment and Analysis of Learner Corpus Data*, pages 101–123, Amsterdam. John Benjamins.

Sakaguchi, K., Napoles, C., Post, M., & Tetreault, J. (2016). Reassessing the goals of grammatical error correction: Fluency instead of grammaticality. *Transactions of the Association for Computational Linguistics*, *4*, 169-182.

Sun, D. (1993). Preliminary analysis of the acquisition of le 了 by foreign learners. *Language Teaching and Linguistic Studies*, 2, pages 65-75.

Wang, M., Malmasi, S., and Huang, M. (2015). The Jinan Chinese Learner Corpus. In *Proc. 10th Workshop on Innovative Use of NLP for Building Educational Applications*.

Yannakoudakis, H., Briscoe, T., and Medlock, B. (2011). A New Dataset and Method for Automatically Grading ESOL Texts. In *Proc. ACL*.

Zhang, B. (2009). The Characteristics and Functions of the HSK Dynamic Composition Corpus. *International Chinese Language Education*, 4(11).