

L1-L2 Parallel Dependency Treebank as Learner Corpus

John Lee, Keying Li, Herman Leung

Department of Linguistics and Translation

City University of Hong Kong

jsylee@cityu.edu.hk, keyingli3-c@my.cityu.edu.hk, leung.hm@gmail.com

Abstract

This opinion paper proposes the use of parallel treebank as learner corpus. We show how an L1-L2 parallel treebank — i.e., parse trees of non-native sentences, aligned to the parse trees of their target hypotheses — can facilitate retrieval of sentences with specific learner errors. We argue for its benefits, in terms of corpus reuse and interoperability, over a conventional learner corpus annotated with error tags. As a proof of concept, we conduct a case study on word-order errors made by learners of Chinese as a foreign language. We report precision and recall in retrieving a range of word-order error categories from L1-L2 tree pairs annotated in the Universal Dependency framework.

1 Introduction

A parallel treebank consists of multiple treebanks with alignments at the sentence level, and often also at the phrase and word levels. Growing interest in parallel treebanks have yielded treebanks of many language combinations (Čmejrek et al., 2004; Megyesi et al., 2010; Sulger et al., 2013; Volk et al., 2017).

So far, there has been no reported attempt to build an L1-L2 parallel treebank — i.e., parse trees of sentences written by non-native speakers (henceforth, “L2 sentences”), aligned to parse trees of their target hypotheses (henceforth, “L1 sentences”). Figure 1 shows an example parse tree pair in such a treebank. The pair consists of the parse tree of a Chinese sentence written by a learner, and the parse tree of its corrected version, or “target hypothesis”. Although a number of L2 treebanks have been built, they either do not provide explicit target hypotheses (Ragheb and Dick-

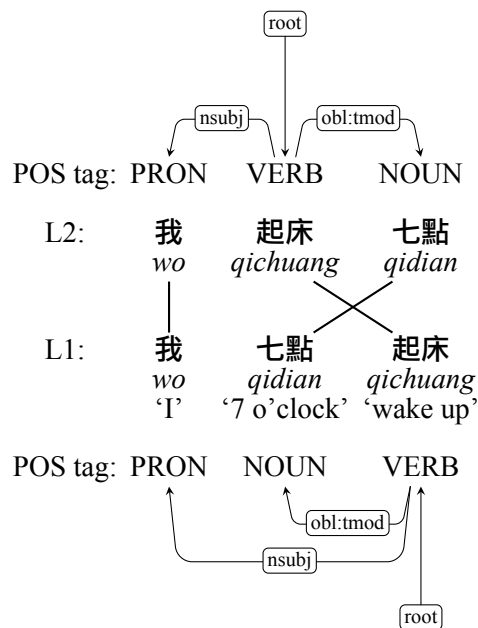


Figure 1: An example L1-L2 tree pair, including word alignments between the learner sentence (“L2”) and its target hypothesis (“L1”), and the parse trees of the two sentences, annotated in Universal Dependencies for Chinese (Leung et al., 2016; Lee et al., 2017).

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

Parallel L1-L2 treebanks can be expected to serve a number of research agendas. First, they would support quantitative studies in Contrastive Interlanguage Analysis (CIA) (Granger, 2015) and Error Analysis (EA). 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, which can inform the design of language teaching peda-

gogy. Further, just as parallel treebanks can help train machine translation (MT) systems (Čmejrek et al., 2004; Sennrich, 2015), L1-L2 treebanks can supply sentence pairs to train systems for automatic grammatical error correction (GEC). Indeed, some GEC systems obtained state-of-the-art results by casting the task as an MT problem (Rozovskaya and Roth, 2016; Junczys-Dowmunt and Grundkiewicz, 2014).

In this opinion paper, we focus on demonstrating how L1-L2 parallel treebanks can benefit learner language analysis. In the next section, we argue that these treebanks can better facilitate re-use and interoperability among learner corpora, because they provide a more precise and flexible encoding of learner errors. As a proof of concept, Section 3 presents a case study on identifying different word-order errors in Chinese L1-L2 parallel trees. Finally, Section 4 concludes.

2 Learner corpora and L1-L2 parallel treebanks

A major function of a learner corpus is to facilitate retrieval of sentences with specific errors. We first discuss the limitations of the use of error tags (Section 2.1), and then propose tree search in an L1-L2 parallel treebank as an alternative approach (Section 2.2).

2.1 Error tags

Errors in a learner sentence are commonly marked with error tags. Each tag labels a problematic text span with an error category, and often also provides a corrected version of the text span (Izumi et al., 2005; Zhang, 2009; Yannakoudakis et al., 2011; Dahlmeier et al., 2013). For example, the Cambridge Learner Corpus uses XML tags to mark error categories (Nicholls, 2003), and supplements the original text with a vertical bar and the target hypothesis:

He <MV> | is </MV> happy.

The annotation above indicates that the learner sentence “He happy” lacks the verb “is”, and categorizes this error as “missing verb” (MV). Despite their widespread usage, however, error tags alone do not optimize corpus re-use and interoperability.

2.1.1 Corpus re-use

A major limitation of the error tagging approach is that learner errors must be pre-categorized. It is

difficult, or perhaps impossible, to develop a robust and general-purpose error typology that covers “all” possible types at a suitable level of granularity. Unless one can foresee research questions in the future, any tagset is by definition limited in error coverage and may not be easily reused.

As a concrete example, consider the “incomplete sentence” error in English. A typical definition of this error is a sentence without subject or finite verb, or a stand-alone subordinate clause (Bram, 1995). The Cambridge Learner Corpus does not enable automatic search for sentences with this error, however; its closest error category, MV (“missing verb”), also covers sentences that are not incomplete, for instance those that are missing modal verbs.

As another example, consider word-order errors in Chinese, which Jiang (2009) classified into a number of categories. It is impossible to directly search for sentences with these error categories in current Chinese learner corpora. The widely used HSK Dynamic Composition Corpus (Zhang, 2009) puts all word-order errors in a single category, CJX. The Test of Chinese as a Foreign Language Learner Corpus (Lee et al., 2016a), which was used in the most recent shared task on Chinese Grammatical Error Diagnosis (Lee et al., 2016b), annotates the POS involved in word-order errors but does not provide more fine-grained distinctions as in Jiang (2009).

2.1.2 Corpus interoperability

Since existing error tagsets vary widely in granularity, it is difficult to combine information from multiple learner corpora. To cite but a few examples, NUCLE (Dahlmeier et al., 2013) uses a tagset with 27 error categories; the NICT Japanese Learner English Corpus has 46 tags (Izumi et al., 2005); while different combinations in the Cambridge Learner Corpus tagset can recognize up to 80 types of different errors (Nicholls, 2003).

In general, there is no clear mapping between these error tagsets. Returning to the incomplete sentence error as example, the closest category in NUCLE is “sentence fragment” (SFrag). However, it applies not only to the kinds of incomplete sentences described by Bram (1995), but also more broadly to complete sentences that suffer from stylistic issues, or those that should be merged with their neighbors. As such, SFrag only partially overlaps with the MV category in the Cambridge Learner Corpus.

2.2 Tree query for learner error retrieval

In view of the limitations of error tags described above, we propose the use of L1-L2 parallel treebank for learner error retrieval. A search query on such a treebank, consisting of a pair of parse tree patterns with alignments (Table 1), can be viewed as a dynamically defined error category.

The idea of leveraging linguistic annotations to search for learner errors is not new. As noted by Reznicek et al. (2013), when both learner sentences and target hypotheses are POS-tagged and word-aligned, a search query with constraints on POS and word positions can effectively express an error category. This approach is becoming more widely applicable, as more learner corpora are enriched with POS annotation (Lüdeling et al., 2008; Díaz-Negrillo et al., 2010) and enhanced alignments (Felice et al., 2016).

Many learner errors, however, cannot be adequately specified with POS alone. Take subject-verb agreement as an example. It does not suffice to search for two aligned verbs with different tags (e.g., VB and VBZ), since the change in conjugation may be a result of other errors (e.g., noun number). The tree query in Table 1(a) provides a more precise and transparent definition of the error. It requires the aligned verbs in both the L1 and L2 sentences to have a singular noun (NN) as subject. Hence, it specifically targets the subject-verb agreement error where the learner mistakes the root form of the verb for the third-person singular present tense. Similarly, to search for Chinese word-order errors involving time expressions, the tree query in Table 1(b) requires a specific dependency relation between the aligned verbs and nouns. This requirement helps exclude other errors that exhibit similar POS patterns, for example violations of the SVO word order.

This proposed approach promotes both corpus re-use and interoperability. Free from a fixed error typology, the user may interrogate the corpus with any suitable tree query, at an arbitrary level of granularity; the learner corpus is thus re-usable to the extent that the desired error type can be defined with POS tags and dependency relations. In terms of interoperability, mappings between error tagsets are no longer necessary; instead, this approach requires mappings between POS tagsets and dependency relations. This is admittedly still a considerable problem, but one that is arguably easier to solve, especially with the emergence of

(a) Subject-verb agreement error	(b) Time expression word-order error

Table 1: Tree queries for (a) subject-verb agreement in English (in Stanford Dependencies); and (b) time expression word-order errors in Chinese (in Universal Dependencies).

Error type	Freq	P	R
Time Expressions	21.1%	0.92	0.92
Modifiers + V	15.8%	0.50	0.50
Action Series	11.4%	0.65	0.85
Locative Expressions	11.4%	0.91	0.77
Subsidiary Relations	8.8%	1.00	0.80
Beneficiary	7.9%	1.00	0.56
Modifiers + N	7.0%	0.89	1.0
DE position	7.0%	1.00	0.38
Topic-comment	6.1%	0.83	0.71
Question	3.5%	1.00	0.50

Table 2: Precision (P) and recall (R) of the manually crafted tree queries in retrieving various error types in the test set. See Jiang (2009) for a description of the error types.

international standards such as Universal Dependencies (Nivre et al., 2016).

3 Case study

As a proof of concept, we conducted a case study on word-order errors, a frequent error in learner Chinese (Lee et al., 2016a), and measured the extent to which the proposed approach succeeded in retrieving sentences with specific error categories.

3.1 Set-up

At least three taxonomies have been proposed for Chinese word-order errors, by Yu (1986), Ko (1997), and Jiang (2009), respectively. We selected the one by Jiang, the most fine-grained of the three, with 27 categories grouped under 9 principles. This taxonomy has been applied on a dataset of 408 sentences, written by students at various proficiency levels, labelled as levels 1 (least proficient) through 3 (most proficient). We

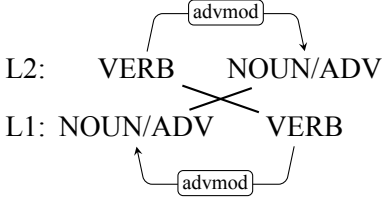
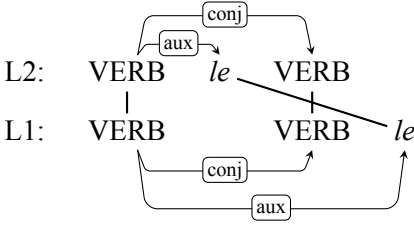
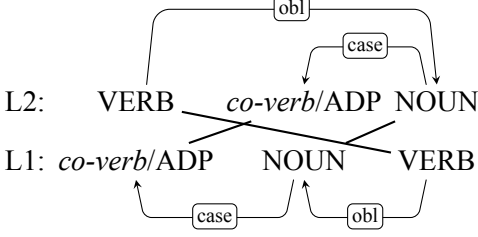
<p>(a) Modifiers + V (Adverb + V) L2: 我去第一次中國... <i>wo qu/VERB diyici/NOUN zhongguo</i> ‘I’ ‘go’ ‘first time’ ‘China’ L1: 我第一次去中國... <i>wo diyici/NOUN qu/VERB zhongguo</i> ‘I’ ‘first time’ ‘go’ ‘China’ “I go for the first time to China ...”</p>	
<p>(b) Action Series (LE position) L2: 我們去了參觀故宮 <i>women qu/VERB le cangan/VERB gugong</i> ‘we’ ‘go’ LE ‘visit’ ‘Forbidden City’ L1: 我們去參觀了故宮 <i>women qu/VERB cangan/VERB le gugong</i> ‘we’ ‘go’ ‘visit’ LE ‘Forbidden City’ “We went to visit the Forbidden City”</p>	
<p>(c) Locative Expressions (Location + V) L2: 你做什麼在這裡 <i>ni zuo/VERB shenme zai/ADP zheli/NOUN</i> ‘you’ ‘do’ ‘what’ ‘at’ ‘here’ L1: 你在這裡做什麼 <i>ni zai/ADP zheli/NOUN zuo/VERB shenme</i> ‘you’ ‘at’ ‘here’ ‘do’ ‘what’ “What are you doing here?”</p>	

Table 3: Examples of L1-L2 tree queries used in the case study (Section 3.1).

focused on the first three principles, namely the “Greenberg Pattern Principle”, which prescribes the canonical word order in Chinese; the “Principle of Modifier Before Head”; and “Temporal Sequence”.¹ These are the largest and more syntax-oriented principles, covering a majority of the errors attested in the dataset.

As development set, we used 58 sentence pairs from Level 1. We manually annotated the learner sentences with the Universal Dependencies (UD) scheme for Learner Chinese (Lee et al., 2017), and the target hypotheses with the UD scheme for standard Chinese (Leung et al., 2016); we then performed word alignment between each sentence pair. Based on the development set and on error definitions in Jiang (2009), we manually crafted 30 parse tree patterns for 10 error categories under the three principles mentioned above. Table 3 shows some example patterns.

As test set, we drew 114 sentences from Levels 2 and 3, and manually performed similar dependency annotation and word alignment. Com-

¹The interested reader is referred to Jiang (2009) for details about these principles.

pared with those in the development set, these sentences are linguistically more complex and likely contain more diverse errors, thus ensuring that the accuracy of the proposed approach is not overestimated.

3.2 Results

We applied the manually crafted tree queries on the test set, and measured their accuracy in retrieving and distinguishing between sentences with different kinds of word-order errors. As shown in Table 2, the highest precision was achieved for the categories “Question”, “DE position”, “Beneficiary” and “Subsidiary Relations” (all at 100%), since their parse structures are most distinct and predictable. Precision was lowest for “Modifiers + V” (50%). Because of unclear meaning in the L2 sentences, their parse trees are often prone to matching similar patterns from other error categories, such as “Time Expressions”. In certain cases, the L2 sentence contains multiple errors but the gold annotation marks only one.

Recall was highest for “Modifiers + N” (100%) and “Time Expressions” (92%), and lowest for

“DE position” (38%). Error analysis revealed that while the L1 parse patterns were mostly adequate, the L2 parse patterns were sometimes not sufficiently general to cover the variety of learner usage that could produce unexpected parse tree structures.

4 Conclusion

This opinion paper advocates the use of L1-L2 parallel treebank as learner corpus. We have argued that such a treebank can better facilitate corpus reuse and interoperability than a fixed error tagset. We have shown the feasibility of the proposed approach in a case study, by measuring the accuracy of tree queries in distinguishing between fine-grained categories of word-order errors in learner Chinese sentences. It is hoped that this paper will spur development of L1-L2 parallel treebanks. They in turn should lead to more accurate parsers for learner text, eventually enabling the proposed approach to be fully automated.

Acknowledgments

This work was partially supported by a Strategic Research Grant (Project no. 7004494) from City University of Hong Kong.

References

- Yevgeni Berzak, Jessica Kenney, Carolyn Spadine, Jing Xian Wang, Lucia Lam, Keiko Sophie Mori, Sebastian Garza, and Boris Katz. 2016. Universal Dependencies for Learner English. In *Proc. ACL*.
- Barli Bram. 1995. *Write Well, Improving Writing Skills*. Kanisius.
- Daniel Dahlmeier, Hwee Tou Ng, and Siew Mei Wu. 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*.
- Ana Diaz-Negrillo, Detmar Meurers, Salvador Valera, and Holger Wunsch. 2010. Towards Interlanguage POS Annotation for Effective Learner Corpora in SLA and FLT. *Language Forum* 36(1-2):139–154.
- Mariano Felice, Christopher Bryant, and Ted Briscoe. 2016. Automatic Extraction of Learner Errors in ESL Sentences Using Linguistically Enhanced Alignments. In *Proc. COLING*.
- Sylviane Granger. 2015. Contrastive Interlanguage Analysis: A Reappraisal. *International Journal of Learner Corpus Research* 1(1):7–24.
- Emi Izumi, Kiyotaka Uchimoto, and Hitoshi Isahara. 2005. Error Annotation for Corpus of Japanese Learner English. In *Proc. 6th International Workshop on Linguistically Interpreted Corpora*.
- Wenying Jiang. 2009. Acquisition of Word Order in Chinese as a Foreign Language. In Peter Jordens, editor, *Studies on Language Acquisition* 38. De Gruyter Mouton.
- M. Junczys-Dowmunt and R. Grundkiewicz. 2014. The AMU system in the CoNLL-2014 shared task: grammatical error correction by data-intensive and feature-rich statistical machine translation. In *Proc. 18th Conference on Computational Natural Language Learning: Shared Task*.
- T. J. Ko. 1997. *Acquisition of Word Order in Chinese as a Foreign Language*. PhD Dissertation, Rutgers University.
- John Lee, Herman Leung, and Keying Li. 2017. Towards Universal Dependencies for Learner Chinese. In *Proc. Workshop on Universal Dependencies*.
- Lung-Hao Lee, Li-Ping Chang, and Yuen-Hsien Tseng. 2016a. Developing Learner Corpus Annotation for Chinese Grammatical Errors. In *Proc. International Conference on Asian Language Processing (IALP)*.
- Lung-Hao Lee, Gaoqi Rao, Liang-Chih Yu, Endong Xun, Baolin Zhang, and Li-Ping Chang. 2016b. Overview of NLP-TEA 2016 Shared Task for Chinese Grammatical Error Diagnosis. In *Proc. 3rd Workshop on Natural Language Processing Techniques for Educational Applications*.
- Herman Leung, Rafaël Poiret, Tak sum Wong, Xinying Chen, Kim Gerdes, and John Lee. 2016. Developing Universal Dependencies for Mandarin Chinese. In *Proc. Workshop on Asian Language Resources*.
- Anke Lüdeling, Seanna Doolittle, Hagen Hirschmann, Karin Schmidt, and Maik Walter. 2008. Das lernerkorpus falko. *Deutsch als Fremdsprache* 2:67–73.
- Beáta Megyesi, Bengt Dahlqvist, Éva Á. Csató, and Joakim Nivre. 2010. The English-Swedish-Turkish Parallel Treebank. In *Proc. Seventh International Conference on Language Resources and Evaluation (LREC)*. pages 55–60.
- Ryo Nagata and Keisuke Sakaguchi. 2016. Phrase Structure Annotation and Parsing for Learner English. In *Proc. ACL*.
- Diane Nicholls. 2003. The Cambridge Learner Corpus - error coding and analysis for lexicography and ELT. In *Proc. Computational Linguistics Conference*.
- Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. Universal Dependencies v1: A Multilingual Treebank Collection. In *Proc. LREC*.

- Marwa Ragheb and Markus Dickinson. 2014. Developing a Corpus of Syntactically-Annotated Learner Language for English. In *Proc. 13th International Workshop on Treebanks and Linguistic Theories (TLT)*.
- Marc Reznicek, Anke Lüdeling, and Hagen Hirschmann. 2013. Competing Target Hypotheses in the Falko Corpus: A Flexible Multi-Layer Corpus Architecture. In Ana Diaz-Negrillo, editor, *Automatic Treatment and Analysis of Learner Corpus Data*. John Benjamins, Amsterdam, pages 101–123.
- Alla Rozovskaya and Dan Roth. 2016. Grammatical Error Correction: Machine Translation and Classifiers. In *Proc. ACL*.
- Rico Sennrich. 2015. Modelling and Optimizing on Syntactic N-Grams for Statistical Machine Translation. *Transactions of the Association for Computational Linguistics* 3:169–182.
- Sebastian Sulger, Miriam Butt, Tracy Holloway King, Paul Meurer, Tibor Laczkó, György Rákosi, Cheikh Bamba Dione, Helge Dyvik, Victoria Rosén, Koenraad De Smedt, Agnieszka Patejuk, Özlem Cetinoğlu, I Wayan Arka, and Meladel Mistica. 2013. ParGramBank: The ParGram Parallel Treebank. In *Proc. ACL*.
- Martin Čmejrek, Jan Cuřín, Jiří Havelka, Jan Hajič, and Vladislav Kuboň. 2004. Prague Czech-English Dependency Treebank Syntactically Annotated Resources for Machine Translation. In *Proc. EAMT*.
- Martin Volk, Torsten Marek, and Yvonne Samuelsson. 2017. Building and Querying Parallel Treebanks. In Silvia Hansen-Schirra, Stella Neumann, and Oliver Čulo, editors, *Annotation, Exploitation and Evaluation of Parallel Corpora*. Language Science Press, Berlin, pages 7–30.
- Helen Yannakoudakis, Ted Briscoe, and Ben Medlock. 2011. A New Dataset and Method for Automatically Grading ESOL Texts. In *Proc. ACL*.
- Shuhuai Yu. 1986. Word Order and Topic Prominence in the Interlanguage of an Australian Learner of Chinese. *Australian Review of Applied Linguistics* 9:83–91.
- Baolin Zhang. 2009. The Characteristics and Functions of the HSK Dynamic Composition Corpus. *International Chinese Language Education* 4(11).