Aligning Chinese-English Parallel Parse Trees: Is it Feasible?

Dun Deng and **Nianwen Xue**

Computer Science Department, Brandeis University
415 South Street, Waltham MA, USA
ddeng@brandeis.edu, xuen@brandeis.edu

Abstract

We investigate the feasibility of aligning Chinese and English parse trees by examining cases of incompatibility between Chinese-English parallel parse trees. This work is done in the context of an annotation project where we construct a parallel treebank by doing word and phrase alignments simultaneously. We discuss the most common incompatibility patterns identified within VPs and NPs and show that most cases of incompatibility are caused by divergent syntactic annotation standards rather than inherent cross-linguistic differences in language itself. This suggests that in principle it is feasible to align the parallel parse trees with some modification of existing syntactic annotation guidelines. We believe this has implications for the use of parallel parse trees as an important resource for Machine Translation models.

1 Introduction

Parallel treebanks have been proved to be a valuable resource in Machine Translation research (Gildea, 2003; Liu et al., 2009; Sun et al., 2010; Chiang, 2010; Xiao and Zhu, 2013), but one issue that hampers their utility is the incompatibility between the syntactic parse trees for a sentence pair (Chiang, 2010), as the trees are annotated based on independently developed monolingual syntactic annotation standards. For example, even though the Penn Chinese Treebank (Xue et al., 2005) and English TreeBank (Marcus et al., 1993) are often referred to collectively as the Penn series of treebanks and are both annotated with phrase structure trees in very similar annotation frameworks, different annotation decisions have led to divergent tree structures (Chiang, 2010). The purpose of this study is to investigate to what extent the divergences between Chinese-English parallel parse trees are caused by different annotation styles (and therefore can be avoided by revising the annotation guidelines), and to what extent they are caused by cross-linguistic differences inherent in language. The answer to this question would shed light on whether it is possible to align the parse trees in parallel treebanks, and on the feasibility of building Machine Translation systems based on these aligned parallel treebanks.

The question above cannot be answered without first having a concrete alignment specification and knowing what types of alignments are attempted. No incompatibility issue would arise for sentence-level alignment when sentences are aligned as a whole. By contrast, both word-level alignment (or the alignment of terminal nodes) and phrase-level alignment (or the alignment of non-terminal nodes) interact with syntactic structures, which could potentially cause incompatibility between the alignments and the tree structures. In the next section, we outline an alignment approach where we perform word alignments and phrase alignments simultaneously in a parallel Chinese-English treebank to prevent incompatibilities between word alignments and syntactic structures. The alignment approach alone, however, does not prevent incompatibilities between the two parse trees of a sentence pair, which are either due to inherent cross-linguistic divergences or differences in treebank annotation styles. In Section 3, we report three types of incompatibilities between the syntactic structures of a sentence pair that prevent proper

LAW VIII - The 8th Linguistic Annotation Workshop, pages 29–37,
Dublin, Ireland, August 23-24 2014.

phrase-level alignments. We analyze two of them and show how they make certain phrase alignments impossible. In Section 4, we discuss the third and also the most common type of incompatibility, which is caused by different annotation decisions as specified in the Penn Chinese and English Treebank syntactic bracketing guidelines (Xue and Xia, 2000; Bies et al., 1995). We propose modifications to the tree structures for the purpose of aligning the parse trees, which means that proper phrase alignment is possible if certain common patterns of incompatibility in syntactic parse trees are fixed. We conclude our paper in Section 5 and touch on the workshop theme. We argue that the quality and level of linguistic sophistication of an linguistic annotation project is tied to the purpose of the resource, and how it is going to be used.

2 Overview of the HACEPT Project

The purpose of the HACEPT (Hierarchically Aligned Chinese-English Parallel TreeBank) Project is to perform word-level and phrase-level alignments between parallel parse trees to develop a linguistic resource for Machine Translation models. We are currently in the process of aligning about 9,000 sentence pairs where syntactic parses already exist for sentences on both the Chinese and English side.

In our project, the annotator is presented with a pair of parallel Chinese-English sentences which have parse trees. The task of the annotator is to do both word and phrase alignments between the two parse trees. The reason for doing word alignments and phrase alignments simultaneously is to make sure word alignments and syntactic structures are harmonized to avoid both redundancies and incompatibilities. Let us use the concrete example in Figure 1 to illustrate the point.

A big challenge to word alignment comes from language-particular function words that do not have counterparts in the translation language. Take the sentences in Figure 1 for instance, the Chinese prenominal modification marker 的 has no English counterpart. Similarly, the English infinitive marker *to* has no Chinese counterpart. Word alignments done without taking syntactic structures into consideration generally glue a function word such as 的 and *to* here to a neighboring content word which has a counterpart and align the two words together to the counterpart of the content word (Li et al., 2009). Under this practice, the first 的 will be glued to 国家/country, and the two words 国家/country 的 as a whole will be aligned to *countries*. Similarly, *to* will be glued to *weigh in* and the whole string *to weigh in* will be aligned to 品评/weigh in. In our project, we take a different approach to word alignments: we leave all the words without a counterpart unaligned on the word level and mark them as "extra". For each unaligned word, we locate the appropriate phrase which contains the unaligned word and has a phrasal counterpart on the other side. By aligning the two phrases, the unaligned word is captured in its appropriate context. Under this new strategy, the Chinese 的 and the English *to* are both left unaligned on the word level. For 的, we align the NP 所有/all 国家/country 的 人民/people with the NP *people in all countries*, because the Chinese NP is the relevant context where 的 appears (的 is used in the NP to indicate that 所有/all 国家/country is the modifier of the noun 人民/people) and matches in meaning with the English NP. For *to*, we align the VP *use their own methods of expression to weigh in on this* with the VP 利用/use 自己/own 的 表达/expression 方式/method 品评/weigh in 此/this 事/thing, because *to* is used in the English VP to connect *use their own methods of expression* and *weigh in on this* and also because the English VP and the Chinese one matches in meaning.

Under our approach, word alignments and syntactic structures are harmonized, and both redundancies and incompatibilities between the two are avoided. For example, the phrase alignment between the two NPs 所有/all 国家/country 的 人民/people and *people in all countries* specifies the context for the occurrence of the function word 的. There is no need to glue 的 to the previous noun 国家/country on the word level. As a matter of fact, the host of 的 (namely the modifier signaled by it) is not the noun 国家/country but the NP 所有/all 国家/country. Similarly, the phrase alignment between *use their own methods of expression to weigh in on this* and 利用/use 自己/own 的 表达/expression 方式/method 品评/weigh in 此/this 事/thing captures the syntactic environment in which *to* appears. The phrase alignment also avoids an incompatibility issue caused by attaching *to* to *weigh in* and aligning the

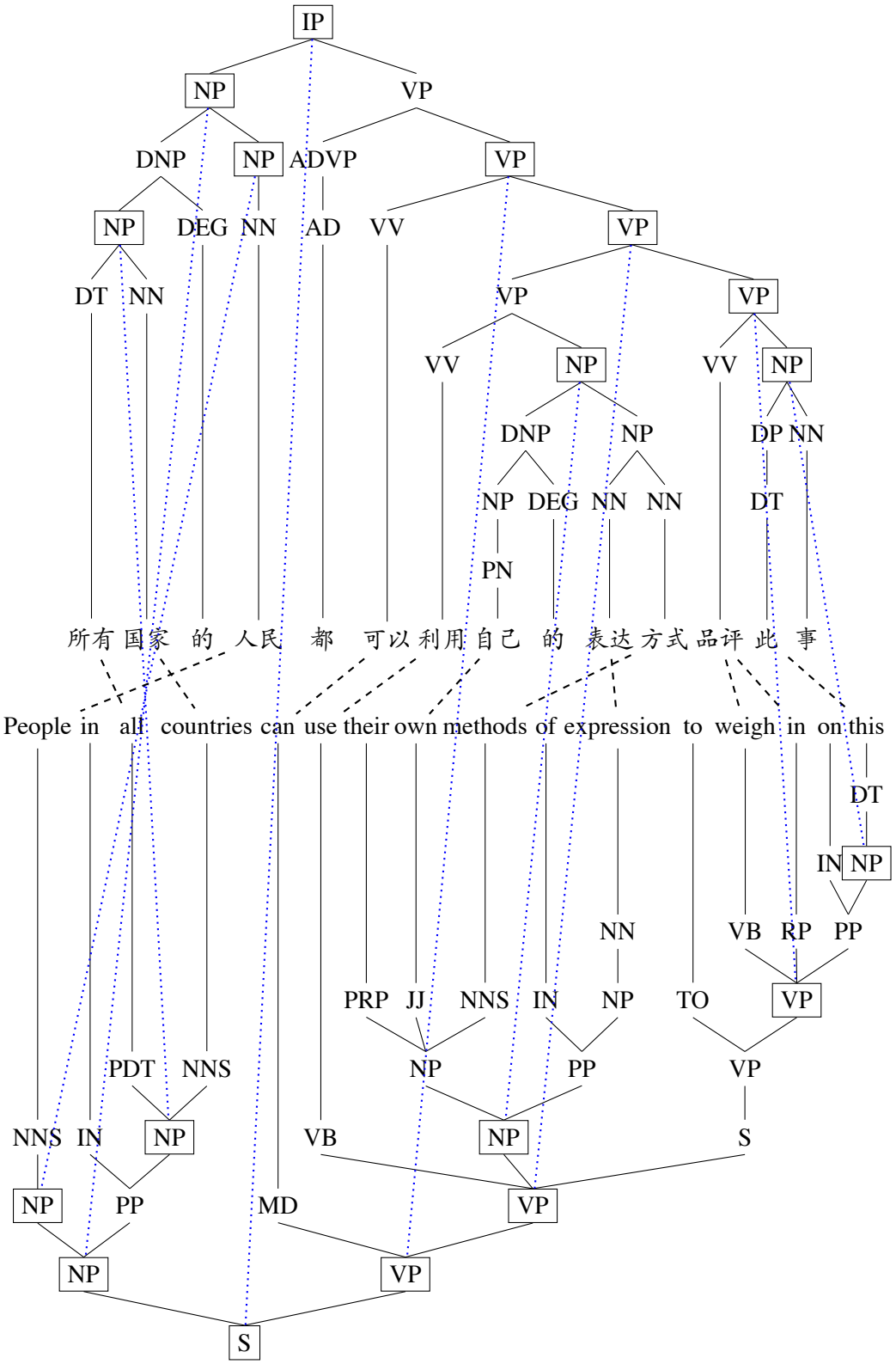


Figure 1: A hierarchically aligned sentence pair

string to 品评/weigh in since *to weigh in* is not even a constituent in the English parse tree. For a more comprehensive and detailed description of the HACEPT project, see (Deng and Xue, 2014).

A natural question arises for our approach: cross-linguistic divergences between languages may cause parse tree incompatibilities to arise, which calls into question the possibility of doing phrase alignments to a useful extent. The fact is that we did find incompatibilities between parse trees in our annotation. In the next section, we report three types of parse tree incompatibilities we have encountered.

3 Three types of parse tree incompatibilities

During the annotation process, we encountered three types of parse tree incompatibilities that make some phrase alignments impossible. The three types are distinguished by the sources of their occurrence and are listed below:

Three types of incompatibilities between parallel parse trees:

a. Incompatibilities caused by lexical-semantic differences between the two languages

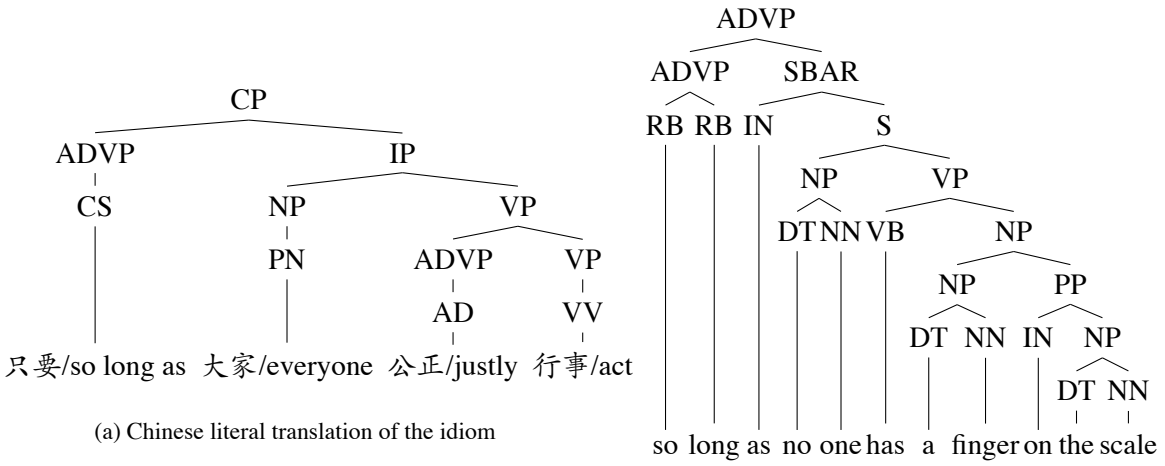
b. Incompatibilities caused by translation-related reasons

c. Incompatibilities caused by different annotation standards

Let us look at the first type. On the lexical level, languages differ in terms of whether or not a piece of semantic information is encoded in a lexical item. For instance, Chinese does not have a verb that expresses the meaning of the English verb *prioritize*, which needs to be translated using a phrase. This does not necessarily cause problems for phrase alignments. Taking *prioritize* for instance, the English phrase *prioritize transportation projects* is translated as 安排/arrange 交通/transportation 项目/project 的 优先/priority 顺序/order (literally *arrange transportation projects' priority order*, i.e., *prioritize transportation projects*). Note that a phrase alignment can be made between the two VPs and also the two NPs *transportation projects* and 交通/transportation 项目/project despite the fact that the meaning of *prioritize* is expressed by a discontinuous phrase in Chinese (安排/arrange ... 的 优先/priority 顺序/order, i.e., *arrange the priority order of* ...). The most extreme case in this category which usually causes incompatibilities and makes phrase-level alignment impossible is idiomatic expressions. An idiom is a single lexical item just like a word and its meaning generally has to be expressed literally in another language. For instance, the idiomatic part in *Markets function best so long as* ***no one has a finger on the scale*** is translated as (只要/so long as) 大家/everyone 公正/justly 行事/act (市场/market 运作/function 最/most 好/good), which literally is *everyone justly acts*. The parse tree for both the English idiom and its Chinese translation is given in Figure 2. No phrase alignment is possible between the idiom and its translation except that between the two root nodes that dominate each string. Phrase alignments are reduced to a minimal extent in cases like this.

Now let us discuss the second type. Consider this example, where the Chinese sentence 他/he 没有/not 提到/mention 这/this 一/one 点/point (*He did't mention this point*) is translated as *There was no mention made of this by him*. Given this particular translation, it is impossible to make a phrase alignment between the Chinese VP 没有/not 提到/mention 这/this 一/one 点/point and *no mention made of this* although the two strings match in meaning. This is because, as shown in Figure 3, the NP node that dominates the English string also dominates the PP *by him*. Note that *him* in the PP corresponds to 他/he, which is outside the Chinese VP. The issue here is caused by the translation. Note that the Chinese sentence is in active voice, but the given translation is in passive voice, which is why the PP *by him* appears at the end of the sentence and causes the problem. If the more literal translation *He didn't mention this point* were provided, 没有/not 提到/mention 这/this 一/one 点/point could be aligned with *didn't mention this point*, and 提到/mention 这/this 一/one 点/point could be aligned with *mention this point*, which is also impossible with the given translation. Phrase alignments are reduced by some extent in cases like this.

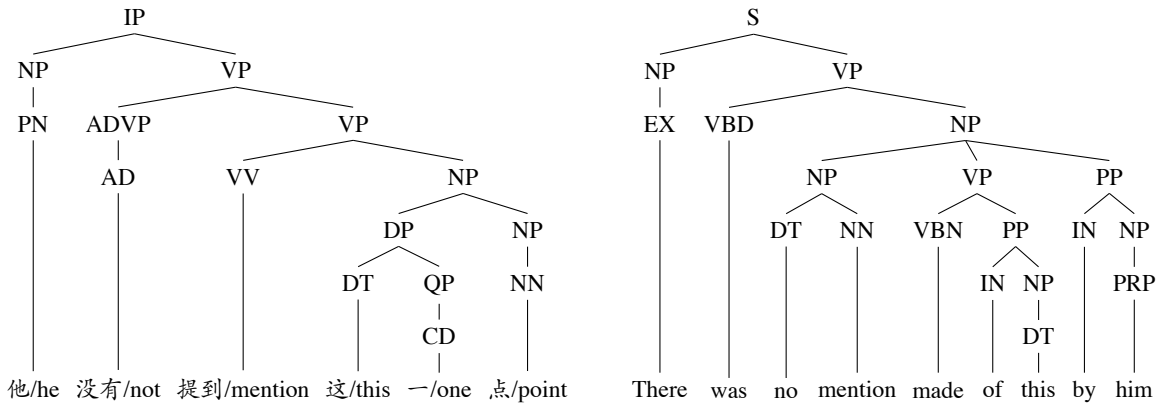
For the first two types of incompatibilities already discussed, the negative impact of them on phrase alignments can be reduced by the enlargement of the corpus, which currently has 8, 932 sentence pairs.



(a) Chinese literal translation of the idiom

(b) An English idiom

Figure 2: Structural divergence caused by idiomatic expressions



(a) Chinese sentence

(b) Non-literal English translation

Figure 3: Structural divergence caused by non-literal translations

(a) PTB structure for VP

(b) CTB structure for VP

Figure 4: Bracketing decisions for VP made by PTB and CTB. XP = {PP, ADVP, S}

Idioms which make phrase-level alignment impossible are rare in our corpus. On average, there are about 5 cases in a file of 500 sentence pairs. As for the incompatibilities caused by translation, it is possible for the phrase alignments missed in those cases to be made up if the phrases involved reappear in a more literal translation. These two issues do not pose a real threat to our approach. As annotators, we cannot do much about these two issues, especially the latter one, since our data is got as is. Due to these two reasons, we will not discuss them further in this paper.

Next let us turn to the last type of incompatibility. Use the sentence pair in Figure 1 for instance. Note that the Chinese VP 利用/use 自己/own 的 表达/expression 方式/method matches the English string *use their own methods of expression* in terms of both grammaticality and meaning. However, the English parse tree has no phrasal node for the string that could form an alignment to the Chinese VP. Similarly, the Chinese NP 表达/expression 方式/method corresponds to the English string *methods of expression*, but again, no phrasal node is present in the English parse tree that could be aligned with the Chinese NP. Our statistics shows that, in a file with 500 sentence pairs, there are approximately 50 instances of the incompatibility in VPs illustrated here and 20 in NPs (an instance is a case where a legitimate phrase alignment cannot be made). These are both quite high frequency. In the next section, we discuss the reason for the incompatibility and give a solution to fix the issue.

4 A common incompatibility pattern and its solution

There is a pattern for the incompatibility illustrated at the end of Section 3. The cause for the incompatibility is the bracketing annotation of the complement-adjunct distinction made by the Penn Treebank (PTB) bracketing guidelines (Bies et al., 1995). The pattern is found in both VPs and NPs.

Let us discuss VPs first. To see the pattern, we need some background information about the internal composition of both English and Chinese VPs and how VPs are parsed according to PTB and CTB annotation standards. Let us start with the English VP. Besides the verb, there can be both preverbal and postverbal constituents in an English VP. Preverbal constituents are much more restricted than postverbal constituents in terms of both phrase types and the number of constituents allowed. Most commonly seen in our corpus, an ADVP is present before the verb if there is a preverbal constituent at all. By contrast, various kinds of constituents (NP, PP, ADVP, S) can appear post-verbally and more than one of these phrases can co-occur. When there is more than one post-verbal constituent, quite often one of them is the complement of the verb and the others are adjuncts. Due to engineering considerations, the PTB bracketing guidelines decided on a flat structure for the English VP, where preverbal and postverbal constituents and the verb are treated as sisters that are directly attached to the VP-node (Bies et al., 1995). A general structure for the English VP is given in Figure 4a, where it can be seen that the complement-adjunct distinction is not made.

Now let us turn to the Chinese VP. In a Chinese VP, there can also be both preverbal and postverbal constituents, but the situation is quite different from that in English. Unlike in English VPs where postverbal constituents are freer, postverbal constituents in Chinese VPs are restricted and can only be

the complement of the verb or one particular kind of phrase, namely QP, which includes counting phrases such as *three times* as in *went there three times*, and duration phrases such as *for three years* as in *lived there for three years*. Adjuncts including ADVP, PP, and different kinds of adverbial clauses come before the verb. The second difference is that Chinese strongly favors no more than one constituent after the verb. In theory, a complement phrase and a QP can co-occur after the verb, but in reality, if the two co-occur in a sentence, the complement will most likely be preposed to the left of the verb by either topicalization or the introduction of the function word 把, leaving QP the only post-verbal element. The structure of a Chinese VP stipulated by the CTB bracketing standards (Xue and Xia, 2000) is provided in Figure 4b.

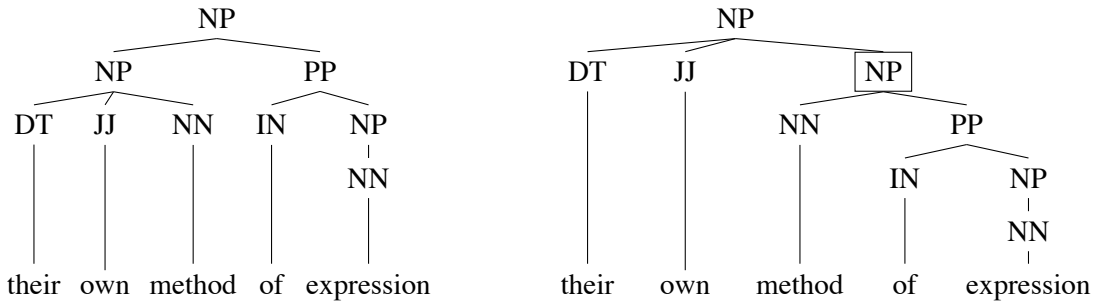
Now let us compare the two structures in Figure 4. Note that in the English VP there is no phrasal node that dominates the verb and its immediate sister on the right, which, in many cases, is the complement of the verb. By contrast, there is a node in the Chinese VP (the boxed VP) that groups together the verb and a post-verbal constituent, which could be either the complement or a QP (some QPs are complements and some others are adjuncts, an issue that does not need to bother us here). This is where the incompatibility arises: the boxed VP-node in the Chinese tree has no node-counterpart to align with in the English tree, but the string dominated by that boxed VP has a match in the English sentence. The example in Figure 1 illustrates the issue, where the Chinese VP dominating the string 利用/use 自己/own 的 表达/expression 方式/method has no possible phrase alignment although the string corresponds in meaning to the English string *use their own methods of expression*.

To eliminate the incompatibility, an extra layer of projection is needed in the English tree. To be specific, we need to combine the verb and its complement to create a VP node, which then can be aligned to the boxed VP in the Chinese tree. Still using the example in Figure 1 for instance, we need to create a VP node by combining the English verb *use* and its object NP *their own methods of expression*, so that the Chinese VP 利用/use 自己/own 的 表达/expression 方式/method can be aligned with the resultant VP. This can be done through binarization.

Now let us turn to the pattern in NPs. We will look at the English NP first. There can be constituents both before and after the head in an English NP. Post-nominal constituents can be either a PP or an S whereas pre-nominal constituents can be one or more than one of the following kinds of elements: determiners (*the*/*a*/*an*), demonstratives (*this*/*that* etc.), quantifiers (*some*, *many* etc.), numerals and adjectives. The PTB bracketing guidelines make the decision that all pre-nominal elements and the head be grouped together using a flat structure to form a NP, which then is treated as a sister of a post-nominal constituent, be it a complement or an adjunct. As for the Chinese NP, the major difference between a Chinese NP and an English one is that there can only be pre-nominal constituents in Chinese NPs. In other words, the head noun is the rightmost element in a Chinese NP and nothing comes after it.

The incompatibility has to do with the complement-adjunct distinction. The complement of an English noun can be either a PP or an S, which always comes after the noun. Due to space limit, we only discuss PP below. An English noun and its PP complement, because of the close semantic relationship between the two, are usually translated as a compound noun in Chinese. For instance, *student of linguistics* is translated as the N-N compound 语言学/linguistics 学生/student. A compound is treated by the CTB bracketing standard as an NP dominating all its components. Unfortunately, the English head noun and its complement do not form a constituent, which, if present, can be aligned with the node for the Chinese compound. This causes incompatibility to arise. Take Figure 1 for instance, the English string *methods of expression* is translated as the Chinese compound noun 表达/expression 方式/method. As shown by the structure, the noun *method* and its PP complement do not form a constituent. As a result, the Chinese compound noun has no alignment.

To remove the incompatibility, we need to change the existing structure of the English NP. Still using the example in Figure 1 for instance, if the English noun phrase has the structure in Figure 5, then we can align the English NP *methods of expression* with the Chinese NP 表达/expression 方式/method. The structure in Figure 5 is different from what is given by the PTB standards in that the head noun (such as



(a) Current structure

(b) Proposed new structure

Figure 5: A proposed revision for the existing structure of English NPs

method) is combined with its complement (such as the PP *of expression*) first to create an NP, which then is modified by, say, an adjective (such as *own*) and a determiner (such as *their*). From the semantic point of view, a pre-nominal adjective is an adjunct to the head noun that is not as closely related to the head noun as its complement. The new structure given in Figure 5b reflects this semantic fact by combining the complement with the head before the adjective.

5 Conclusion

In this paper, we argue that it is feasible to align Chinese-English parallel parse trees despite incompatibility issues. We show that the most common incompatibility is caused by bracketing guideline design, which can be fixed by changing the existing structures stipulated by the current annotation standards. The revised structures we proposed to avoid the incompatibility are deeper than the existing PTB structures and respect the complement-adjunct distinction, which is a well-established notion in linguistics that has been shown to manifest itself in different kinds of phenomena cross-linguistically. In syntax, the distinction is made by combining the head and its complement first to form a constituent, which then is combined with an adjunct. This way of representing the distinction is standard and gives arise to a structure that is binary-branching and deep. In syntactic annotation, linguistic sophistication which requires the parse tree to reflect well-established linguistic notions such as the complement-adjunct distinction is an important consideration and generally gives arise to deeper structures. In addition to linguistic sophistication, another important consideration in syntactic annotation is engineering economy, which requires the annotation to be economical in the sense that it can be carried out in a convenient and efficient manner to save annotation effort and time. This means that the parse tree needs to be as flat as possible since shallow structures are much easier to annotate than deep ones. These two competing considerations interact to influence the establishment of bracketing standards.

Due to engineering pressure caused by the fact that it is not easy to make a consistent distinction between complements and adjuncts in annotation, the PTB bracketing guidelines chose a shallow structure for both VPs and NPs as shown above. The decision is understandable since no incompatibility ever arises in the construction of a monolingual treebank like PTB. With the advent of new use cases of monolingual treebanks such as hierarchically aligned parallel treebanks, new issues like incompatibility emerge and call for adjustments to some decisions that have been made without such issues. As shown in Section 4, some decisions made in existing bracketing annotation cause incompatibilities and make legitimate phrase alignments impossible. For the purpose of aligning parallel parse trees, deeper and linguistically motivated structures are needed. This raises the interesting question whether we should have a deeper and linguistically motivated structure to start with when constructing a monolingual treebank. Based on what we have seen in this paper, a positive answer to the question seems reasonable at least in some cases such as VPs and NPs for the sake of better serving uses cases like constructing parallel

treebanks with hierarchical alignments.

Acknowledgements

The HACEPT project, under which the work presented in this paper is done, is supported by the IBM subcontract No. 4913014934 under DARPA Prime Contract No. 0011-12-C-0015 entitled "Broad Operational Language Translation". We would like to thank Libin Shen and Salim Roukos for their inspiration and discussion during early stages of the project, Abe Ittycheriah and Niyu Ge for their help with setting up the data, Loretta Bandera for developing and maintaining the annotation tool, and three anonymous reviewers for their helpful comments. We are grateful for the hard work of our four annotators: Hui Gao, Shiman Guo, Tse-ming Wang and Lingya Zhou. Any opinions, findings, conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect those of the sponsor or any of the people mentioned above.

References

Ann Bies, Mark Ferguson, Karen Katz, Robert MacIntyre, Victoria Tredinnick, Grace Kim, Mary Ann Marcinkiewicz, and Britta Schasberger. 1995. Bracketing guidelines for Treebank II style Penn Treebank project. Technical report, University of Pennsylvania.

David Chiang. 2010. Learning to translate with source and target syntax. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1443--1452.

Dun Deng and Nianwen Xue. 2014. Building a Hierarchically Aligned Chinese-English Parallel Treebank. In *Proceedings of the 25th International Conference on Computational Linguistics (COLING 2014)*.

Daniel Gildea. 2003. Loosely tree-based alignment for machine translation. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*, pages 80--87.

Xuansong Li, Niyu Ge, and Stephanie Strassel. 2009. Tagging guidelines for Chinese-English word alignment. Technical report, Linguistic Data Consortium.

Yang Liu, Yajuan Lü, and Qun Liu. 2009. Improving tree-to-tree translation with packed forests. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 558--566.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a Large Annotated Corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313--330.

Jun Sun, Min Zhang, and Chew Lim Tan. 2010. Exploring syntactic structural features for sub-tree alignment using bilingual tree kernels. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 306--315.

Tong Xiao and Jingbo Zhu. 2013. Unsupervised sub-tree alignment for tree-to-tree translation. *Journal of Artificial Intelligence Research*, 48:733--782.

Nianwen Xue and Fei Xia. 2000. The bracketing guidelines for Penn Chinese Treebank project. Technical report, University of Pennsylvania.

Nianwen Xue, Fei Xia, Fu dong Chiou, and Martha Palmer. 2005. The Penn Chinese Treebank: Phrase Structure Annotation of a Large Corpus. *Natural Language Engineering*, 11(2):207--238.