Translation Divergences in Chinese–English Machine Translation: An Empirical Investigation

Dun Deng*
Tsinghua University

Nianwen Xue**
Brandeis University

In this article, we conduct an empirical investigation of translation divergences between Chinese and English relying on a parallel treebank. To do this, we first devise a hierarchical alignment scheme where Chinese and English parse trees are aligned in a way that eliminates conflicts and redundancies between word alignments and syntactic parses to prevent the generation of spurious translation divergences. Using this Hierarchically Aligned Chinese–English Parallel Treebank (HACEPT), we are able to semi-automatically identify and categorize the translation divergences between the two languages and quantify each type of translation divergence. Our results show that the translation divergences are much broader than described in previous studies that are largely based on anecdotal evidence and linguistic knowledge. The distribution of the translation divergences also shows that some high-profile translation divergences that motivate previous research are actually very rare in our data, whereas other translation divergences that have previously received little attention actually exist in large quantities. We also show that HACEPT allows the extraction of syntax-based translation rules, most of which are expressive enough to capture the translation divergences, and point out that the syntactic annotation in existing treebanks is not optimal for extracting such translation rules. We also discuss the implications of our study for attempts to bridge translation divergences by devising shared semantic representations across languages. Our quantitative results lend further support to the observation that although it is possible to bridge some translation divergences with semantic representations, other translation divergences are open-ended, thus building a semantic representation that captures all possible translation divergences may be impractical.

Note: Part of Section 2 of this paper has appeared in Deng and Xue (2014c). All the other contents are new.

* Department of Chinese Languages and Literature, Tsinghua University, Beijing, China.
E-mail: ddeng@tsinghua.edu.cn.

** Computer Science Department, Brandeis University, 415 South Street, Waltham MA 02453.
E-mail: xuen@brandeis.edu.

Submission received: 25 May 2015; revised version received: 16 September 2016; accepted for publication: 15 December 2016.

doi:10.1162/COLI_a_00292

1. Introduction

Statistical machine translation (SMT), currently the dominant approach to machine translation (MT), relies on large amounts of parallel text called **bitext** to learn translation patterns that can be used to translate new text. The statistical paradigm of MT started over two decades ago with word-based models described in the seminal work of Brown et al. (1993), where an MT system automatically learns translation correspondence between word-aligned sentence pairs and selects the most plausible translation for a source language sentence using the language model trained on monolingual data in the target language. Although this general paradigm has not changed, SMT approaches have evolved since then. Only several years after the introduction of word-based models, phrase-based models were proposed that better use local context and handle the translation of non-compositional phrases to yield superior translation accuracy (Och 1999; Koehn, Och, and Marcu 2003). Hierarchical phrase-based models (Chiang 2005, 2007) further advanced the state of the art by allowing non-terminals in phrase-based translation rules called **hierarchical phrase pairs**, which effectively capture long-distance lexical dependencies because the yields of the non-terminals are of variable lengths and can be arbitrarily long (Zollmann et al. 2008).

Along this main thread, in the last decade there has been intensive research on incorporating syntactic trees produced by syntactic parsers trained on human-annotated treebanks into an SMT model. The attempt to provide syntactic information for SMT models is driven by the widely accepted assumption that word order varies in systematic ways among languages and reordering in a sentence pair often involves syntactic constituents rather than individual words. It is natural then to expect that incorporating syntactic structures into SMT models would lead to improved MT accuracy. Various approaches have been proposed to incorporate syntactic structures, and their differences can be described along two dimensions: whether they use syntactic structures on the source side or the target side or both, and whether they use phrase structures or dependency structures. String-to-tree systems model the syntactic structures of target language sentences (Galley et al. 2004, 2006) and tree-to-string systems model the syntactic structures of source language sentences (Huang, Knight, and Joshi 2006; Liu, Liu, and Lin 2006; Liu et al. 2007; Liu and Gildea 2008). Tree-to-tree systems model the syntactic structures of both source and target language sentences (Eisner 2003; Ding and Palmer 2005; Cowan, Kučerová, and Collins 2006; Zhang et al. 2008; Liu, Lü, and Liu 2009). Early syntax-based models generally use phrase structure (or constituent structure) trees and later syntax-based systems also use dependency trees (Shen, Xu, and Weischedel 2008). The general observation about syntax-based models is that although incorporating syntactic structures has led to solid gains, there are also challenges that have prevented syntax-based systems from realizing their full potential. The first challenge is the inevitable errors acquired in automatic syntactic parsing. Syntax-based systems rely on syntactic parsers trained on manually annotated treebanks to automatically parse large quantities of parallel text in order to extract translation rules. Even though state-of-the-art parsers can parse English text at over 90% accuracy (Charniak 2000; Petrov and Klein 2007) when evaluated against standard benchmarks such as the Penn TreeBank (Marcus, Santorini, and Marcinkiewicz 1993; Marcus et al. 1994), syntactic parsing accuracy for other languages is considerably lower (Wang and Xue 2014). In addition, even for English, there is considerable performance degradation when the data that needs to be parsed is different from the *Wall Street Journal* newswire articles that the parsers are generally trained on. Syntax-based

SMT systems are only competitive when used in conjunction with techniques such as packed forests, which relax the need to use one-best parses (Mi and Huang 2008; Mi, Huang, and Liu 2008). The second challenge is that the constraints imposed by the syntactic structures have been shown to be too stringent, and prevent useful syntax-based translation rules that do not obey syntactic constituent boundaries from being used in SMT models and hurt MT performance (Koehn 2009). Effective techniques have been developed to address this issue by identifying heuristics to relax the constraints of syntactic structures when extracting translation rules (Zollmann and Venugopal 2006).

As the gain of incorporating syntactic information has plateaued, the field is poised to climb up the Vauquois Pyramid (Vauquois 1968) and start exploring the utility of semantic representations for SMT systems, mirroring the progression of earlier rule-based approaches in the previous incarnation of MT research. The hope is that more abstract semantic representations can better address translation divergences than syntactic structures, and this advantage will hopefully offset the potential harm caused by the expected drop in the accuracy of semantic analyzers that are more difficult to develop than syntactic parsers. The development of the AMR Bank (Banarescu et al. 2013) is the latest attempt in that direction, although it is important to point out that the significance of such semantically annotated corpora goes far beyond the narrow purpose of MT, and that the AMR Bank is neither the first nor the only attempt to develop semantically annotated resources that can be used for MT purposes. Similar efforts include a series of head-driven phrase structure grammar–based resources (e.g., LingGO Redwoods Treebank and the DeepBank) (Oepen et al. 2002; Flickinger, Zhang, and Kordoni 2012; Bender et al. 2015; Flickinger, Oepen, and Bender 2017), the semantic layers of which are based on Minimal Recursion Semantics (MRS) (Copestake et al. 2005), a semantic representation framework that has been adopted in modern semantic transfer-based MT systems (Lønning et al. 2004). As the field prepares to take this next step of incorporating semantic representation into the SMT paradigm, it is worth asking: i) Are translation divergences properly represented by the kind of syntax-based translation rules such as hierarchical phrase pairs used in current systems? ii) Can these rules be properly extracted from existing syntactically annotated parallel treebanks? iii) What are the advantages and challenges in building semantic representations that can bridge translation divergences?

We try to answer these questions by identifying and categorizing actual translation divergences in a parallel treebank, and extracting (an enhanced form of) hierarchical phrase pairs to see if the translation divergences can be captured by these hierarchical phrase pairs. We then look into whether the translation divergences pose any challenge for attempts at devising semantic representations that are supposed to bridge the divergences. To do this, we first manually align about 10,000 Chinese–English sentence pairs that have been manually parsed syntactically on both sides, and then semi-automatically extract and categorize translation divergences between the two languages. The manual alignment is hierarchical in that it is performed at both the word level (between terminal nodes) and the constituent level (between non-terminal nodes), and is done in a way that eliminates conflicts and redundancies between word alignments and syntactic trees. This is necessary in order not to generate spurious translation divergences. The Chinese–English language pair is chosen for this study because parallel treebanks for these two languages can be readily found. Manually constructed rather than automatically produced parse trees are used because we want to isolate genuine translation divergences resulting from different syntactic

realizations between the source and target languages from artificial ones caused by parsing errors.[1]

There are three main findings from our study: 1) The translation divergences are much more diverse than previously realized. Previous discussions of translation divergence are mostly qualitative in nature, covering a few widely recognized linguistic differences between languages without the necessary empirical support (Dorr 1994). Using an actual corpus allows us to not only extract all possible divergences that actually occur in the corpus but also quantify each type of translation divergence. This quantitative information can guide MT researchers to focus on high pay-off translation divergences when designing their MT systems. For example, some high-profile translation divergence such as "head-switching" that are frequently mentioned in MT work (e.g., Ding and Palmer 2005) turn out to be very rare in our corpus, whereas other translation divergences that are barely mentioned exist in large quantities. 2) For the most part, the translation divergences can be captured by the kind of hierarchical phrase pairs used in modern SMT systems. By that we mean that the translation divergences can be encapsulated in hierarchical phrase pairs that contain a small number of lexical items that are likely to repeat in a parallel corpus of the size that is typically used to train SMT systems. Our study also found, however, that existing treebanks are not optimally suited for extracting such phrase pairs in that some structures are too flat for the purpose of extracting minimal rules. As a result, there is a significant number of rules that are rather long and unlikely to repeat in an actual corpus. This suggests that for MT purposes it may be worth enriching existing treebanks with additional structures to the extent that they are linguistically justified. Such enhanced treebanks, together with a hierarchical alignment scheme as described in this article, will allow us to extract syntax-based translation rules that better capture translation divergences. 3) Although some translation divergences can be bridged by designing shared semantic representations across languages, other translation divergences are open-ended and building shared semantic representations for such translation divergences may be impractical. The problem would only become more severe in a multilingual setting than in a bilingual setting.

The rest of the article is organized as follows: In Section 2, we describe in detail the guiding principle and annotation procedure of our Hierarchically Aligned Chinese–English Parallel Treebank (HACEPT). We show that our hierarchical alignment approach harmonizes word alignment and syntactic structures, and prevents the extraction of spurious translation divergences. In Section 3, we systematically examine translation divergences between Chinese and English based on this corpus. Using the alignment between non-terminal nodes, we semi-automatically identify and categorize the translation divergences into seven types, and provide statistics for each type. In Section 4, we show that HACEPT can be used to extract hierarchical phrase pairs that are consistent with the constituent boundaries in the parse trees and able to capture the translation divergences identified in our corpus. We present a distribution of the extracted hierarchical phrase pairs by the number of lexical items to show that for the most part it is feasible to learn such phrases from a large parallel corpus. We also show, however, that the flat structures in existing treebanks are not optimal for extracting such phrase pairs and result in a significant number of large phrase pairs that are unlikely to repeat even in a large parallel corpus. In Section 5, we briefly discuss the

1 We did our best to minimize artificial divergences, but, as pointed out by a reviewer, "surely some remain because of annotation errors/differences in analytical choices."

implications of our research for efforts to bridge translation divergences by devising semantic representations that are shared across languages. We discuss related work in Section 6 and conclude the paper in Section 7.

2. A Hierarchical Approach to Aligning the Parallel TreeBank

In this section, we introduce in detail how we construct HACEPT for the study of translation divergences. But first, we explain why we chose to build HACEPT instead of directly making use of existing parallel corpora.

2.1 Issues with Word Alignment

The main reason why we do not use existing word-aligned parallel corpora to extract and study translation divergences is because they have many spurious word alignments, which cause two problems for the study of translation divergences. The first problem is that, as shown by Zhu, Li, and Xiao (2015) and Deng, Xue, and Guo (2015), they impede the extraction of legitimate phrasal translation equivalents, which may manifest translation divergences. The second problem is that they generate phrase pairs, which, if used to extract translation divergences, will provide artificial divergences that do not reflect real cross-linguistic differences. This subsection elaborates on the latter problem, which is much more harmful to the investigation of translation divergences. The discussion of the problem motivates our hierarchical alignment approach, which ensures that word alignments are harmonized with syntactic parses and avoids the generation of artificial translation divergences.

We begin with the discussion of different word alignment scenarios. Given a word in a source-language sentence, there are in total three logical possibilities about its translation counterpart in the target-language sentence:

- An equivalent exists in the target-language sentence, which matches the word in both lexical meaning and grammatical function.
- There is a candidate in the target-language sentence, which does not have the same lexical meaning and/or grammatical function as the word but could be used as the translation counterpart of the word in the given context.
- The word has no translation counterpart in the target-language sentence at all.

We now provide concrete examples to discuss each of these three possibilities. For all the examples, a word-by-word gloss is provided under the Chinese expression, which is on the left of the <> symbol used to connect the Chinese expression and its English translation. Word alignments between the two expressions are marked by numeric subscripts, so words with the same subscript digit form a word alignment. Irrelevant constituents in the expressions are replaced by phrase category variables.[2]

2 Abbreviations used in the gloss for the Chinese expressions are as follows: CL = Classifier, PERF = Perfective aspect marker, PRT = Particle, RS = Resultative Suffix.

First consider the following examples that illustrate the first possibility.

(1) 双边$_1$ 贸易$_2$ <> bilateral$_1$ trade$_2$
bilateral trade

(2) 不$_1$ 会$_2$ VP <> will$_2$ not$_1$ VP
not will

The two content words in the Chinese phrase in Example (1) both have the same lexical meaning as their English translations. Similarly, the two Chinese function words in Example (2), namely, the negation 不 and the modal 会, match their English translations in grammatical function. As a result, the words are aligned as indicated by the subscripts. In MT literature, word alignments illustrated by Examples (1) and (2) are called **sure** alignments. We will ignore sure alignments in our discussion, because they are straightforward and easy to deal with.

Next consider these examples:

(3) 这般 黑色 幽默 的 背后 究竟 有 几许 猫腻$_1$ 呢?
this black humor PRT back on-earth have how-many dirty-trick PRT
<> How many stories$_1$ are hiding behind that black humor?

(4) 离开$_1$ 的 时间$_2$ <> time$_2$ of leaving$_1$
leave PRT time

Let us first look at the two content words with subscripts in Example (3), which exemplify the second possibility. The Chinese noun 猫腻 literally means "something fishy or shady that probably involves some illegal deal or dirty trick." The "fishy/shady/illegal/dirty" component in the lexical meaning of the Chinese noun is missing in the lexical meaning of the English noun *story*. In other words, strictly speaking, 猫腻 and *story* have different lexical meanings. However, contextual information makes up for what is missing in the literal meaning of *story* (the context for Example (3) is a news story about a thief who left a note on the wall of a home he broke into to encourage the home owner to work harder), and we believe that *stories* is an appropriate translation for 猫腻 in that context. The fact is that it is usually not easy, sometimes even impossible, to find a perfect match in translation of words when it comes to lexical semantics. We think that in cases like Example (3), the two words in question should be aligned and we will not discuss them further.

The example in (4) illustrates the second possibility with two function words that have no lexical content, namely, 的 and *of*. The Chinese function word 的 is used after a constituent, a VP headed by the verb 离开/'leave' in the current case, to signal a modification relationship in a nominal phrase. The English translation of the Chinese phrase also contains a function word, namely the preposition *of*. The fact is that 的 and *of* do not have the same grammatical function because the former is not a preposition. Chinese prepositions behave like English ones in taking a following complement to form a PP. By contrast, 的 can never function as a head to take a complement after it. It always forms a constituent with a previous phrase, which can then modify a head or stand alone without a following head. Although 的 is not a preposition and has very different grammatical functions than *of*, they could, in practice, be treated as translation counterparts in this particular context for alignment purposes. Such context-dependent alignments are generally referred to as **possible** alignments in MT literature. We will

come back to issues involving possible alignments such as whether 的 and *of* should be aligned shortly.

Lastly, let us look at the following examples for the third possibility:

(5) 三$_1$ 本 书$_2$ <> three$_1$ books$_2$
three CL book

(6) 想$_1$ VP <> want$_1$ to VP
want

(7) 关注$_1$ 文化$_2$ 建设$_3$ 方面 的 内容$_4$ <>
pay-attention-to culture construction aspect PRT content
pay$_1$ attention$_1$ to$_1$ the contents$_4$ of cultural$_2$ construction$_3$

(8) 一九九五年$_1$, 国家$_2$ 开发$_3$ 银行$_4$ VP <>
1995 nation development bank
In 1995$_1$, the National$_2$ Development$_3$ Bank$_4$ VP

There are two situations where the third possibility takes place. The first scenario is that a word in the source language has no equivalent in the target language and therefore cannot possibly have a translation counterpart. The two cases in Examples (5) and (6) both illustrate this situation. Chinese is a so-called classifier language, where a function word called "classifier" such as 本 in Example (5) is required between a numeral and a count noun. English, by contrast, is not a classifier language, where numerals can directly modify count nouns. The Chinese classifier 本 lacks an equivalent in English and therefore has no translation counterpart in Example (5) at all. Similarly, in Example (6), Chinese lacks an infinitive marker and therefore the English *to* has no translation counterpart.

The other situation for the third possibility is that a word in the source language does have an equivalent in the target language, but because of linguistic or translation-related reasons, the word does not get translated and therefore has no translation counterpart. The two cases in Examples (7) and (8) are examples of this situation. The Chinese phrase 文化 建设 方面 的 内容 in Example (7) literally means "content of the aspects in cultural construction." The noun 方面/'aspect' is not translated and has no translation counterpart. Chinese temporal expressions such as 一九九五年/'1995' in Example (8) can directly function as an adverbial without the help of the preposition 在/'in'. As a result, the English preposition *in* can but does not have a translation counterpart in Example (8).

Cases illustrated by Examples (4) through (8) pose a big challenge for word alignment. For cases like Example (4) where two function words such as 的 and *of* have different grammatical functions, should we align them or not? For cases like Example (5) where words such as the classifier 本 in Example (5) simply have no translation counterparts, how should we do word alignment? To align, or not to align, is the dilemma we are in when dealing with words without translation counterparts in word alignment. In the literature, both the "align" and the "not to align" option have been put into practice, and we will first discuss the problems of aligning these words in the next subsection.

2.2 The Motivation for a Hierarchical Approach to Word Alignment

If we take the "align" option to deal with words without translation counterparts, we will need to align, say, 的 and *of* in Example (4), despite the fact that the two words do not match in grammatical function. But what about words such as 本 in Example (5), which simply has no translation counterpart at all? A common approach adopting the "align" option is to glue such a word to a neighboring host word that has a translation counterpart,[3] and then align the two as a whole with the counterpart of the host (Melamed 1998; Li, Ge, and Strassel 2009). For instance, the Chinese classifier 本 in Example (5) will be attached to, say, the previous numeral 三/'three', and the string 三 本 will be aligned to the English *three*. For ease of discussion, we will refer to this practice as the "glue-to-a-host" strategy (GTAHS).

There are two serious problems of aligning words without translation counterparts for our study of translation divergences. The first is that it creates spurious ambiguities and translation divergences, which we elaborate on herein.

For cases like 的 and *of* in Example (4), because the two words do not match in grammatical function, they only co-occur in the current particular example and the translation correspondence will change with the context. 的 is only one of many elements that co-occur with the English *of* in translation. Similarly, many other words besides the preposition *of* have been used to translate 的. Let us just take the translation of 的, for example:

(9) 约翰$_1$ 的 父亲$_2$ <> John$_1$'s father$_2$
John PRT father

(10) 北京$_1$ 的 天气$_2$ <> weather$_2$ in Beijing$_1$
Beijing PRT weather

(11) 我$_1$ 买$_2$ 的 书$_3$ <> books$_3$ which I$_1$ bought$_2$
I buy PRT book

(12) 外向型$_1$ 经济$_2$ 的 发展$_3$ <> outwardly$_1$ economic$_2$ development$_3$
outward economy PRT development

As shown here, if we are going to word-align 的, its alignment can be *'s* in Example (9), the preposition *in* in Example (10), the relative pronoun *which* in Example (11), and nothing in Example (12) among many other things.

As for cases like 本 in Example (5), the GTAHS also creates spurious ambiguities and translation divergences. Consider the following example, where the Chinese noun 苹果/'apple' is aligned to six English strings:

(13) 吃 苹果$_1$ <> eat apples$_1$/an$_1$ apple$_1$/the$_1$ apple$_1$
eat apple

(14) 喜欢 苹果$_1$ <> fond of$_1$ apples$_1$
like apple

3 This raises the question of which neighboring word should be chosen as the host, which is by no means easy to answer. To the best of our knowledge, this issue has never been explicitly discussed in the literature of word alignment practice that adopts this approach.

(15) 谈论 苹果$_1$ <> talk about$_1$ apples$_1$
discuss apple

(16) 给 他们 苹果$_1$ <> provide them with$_1$ apples$_1$
give they apple

The Chinese 苹果 and the English *apple* match in lexical meaning and are both unambiguous. In cases where the English noun is used with a determiner as in Example (13), because Chinese has no determiners and the bare noun 苹果 can be the appropriate translation for either *an apple* or *the apple*, given a context, the GTAHS attaches the determiner to *apple* and the whole string is aligned with 苹果. In other similar cases where an English element such as a preposition is absent in Chinese, as in Examples (14), (15), and (16), the GTAHS glues the preposition to *apple* and the whole PP is aligned with 苹果. With the GTAHS, the unambiguous Chinese 苹果 ends up being aligned with more than one English string. This kind of spurious ambiguity is very common given the GTAHS, which generates many word alignments where one source language word is aligned to multiple target language words or vice versa and causes artificial translation divergence.

The second issue of aligning words without translation counterparts is that it causes incompatibilities between word alignments and syntactic structures. The problem is more prominent with words that simply have no translation counterparts, such as 本 in Example (5). For these words, by attaching them to a host, the GTAHS effectively creates rudimentary syntactic structures that are often incompatible with the syntactic structures annotated based on existing treebanking annotation standards. Consider the following examples:

(17) 如果$_1$ 我 是 他 的话$_1$ <> If$_1$ I were him
if I be he PRT

(18) 他 正$_1$ 访问$_1$ 北京 <> He is$_1$ visiting$_1$ Beijing
he right visit Beijing

(19) 新 年$_1$ 伊始 <> the beginning of the$_1$ new year$_1$
new year start

(20) 迅速 有效地 解决$_1$ 问题 <>
to$_1$ quickly and efficiently solve$_1$ the problem
quick efficiently solve problem

Let us take Example (17) for instance. The first four words in the Chinese phrase each make a one-to-one correspondence with the four words in the English translation. The last word in the Chinese phrase, namely, the particle 的话, is left out with no translation counterpart. The GTAHS as implemented by Li, Ge, and Strassel (2009) attaches this word to 如果/'if' and aligns the discontinuous string 如果 ... 的话 with *If*.[4] However, the two words do not form a constituent in the parse tree of the sentence generated by the Chinese TreeBank (Xue et al. 2005). As a matter of fact, all the aligned multi-word strings in these four examples do not correspond to a constituent in a

4 The four examples discussed here are all quoted from Li, Ge, and Strassel (2009). As mentioned in footnote 3, there is no discussion as to how the host word is chosen. The reason why 的话 is attached to 如果 but not another word is presumably because the two usually co-occur.

Penn TreeBank (Marcus, Santorini, and Marcinkiewicz 1993) or Chinese TreeBank (Xue et al. 2005) parse tree. This means that the word alignments conflict with the syntactic structures of the sentences. This kind of conflict is very common in the GTAHS, and will introduce noise in our study of translation divergence.

Given that aligning words without translation counterparts has problems, the other option is not to align them. But if we are not going to align them, where and how should we capture them?

A deep linguistic reason for the problems of existent approaches to words without translation counterparts is that these approaches all try to represent syntactic information on the word level. Words without translation counterparts are mostly function words, especially language-particular ones such as Chinese classifiers or the English infinitive marker. A function word does not stand alone in the syntax. Rather, it involves other constituents in the syntax to signal important grammatical relations. For instance, *'s* in English indicates the possessive relation between two constituents within a noun phrase. Furthermore, every function word has its syntactic domain where it plays its grammatical role. For instance, the Chinese classifier could be viewed as a functional head in a nominal phrase, where it takes the projection of the following noun as its complement (Li 1999). This means that function words are syntactic in nature and should be represented syntactically. However, all the existing word alignment practice we know of handles function words on the word level, and treats word alignment as a stand-alone task without systematically considering its interaction with the syntactic structure of a sentence. The inevitable consequence of this practice is that incompatibilities between word alignments and syntactic structures will arise in many places as shown in these examples.[5]

To solve the problem, we should separate the word level and phrasal level and shift the burden of representing syntactic information to the latter level. Guided by this consideration, we propose hierarchical alignment, where alignment happens on both the word level and the phrasal level in a coordinated and harmonious way. Making use of parallel treebanks, we perform word-level and phrase-level alignments simultaneously on parallel phrase-based parse trees, attempting to construct a hierarchically aligned corpus where word alignments are harmonized with syntactic structures. Our main innovation is to leave words without translation counterparts unaligned on the word level, and capture them on the phrasal level with the alignment between the appropriate phrases that encapsulate the unaligned words. In the next subsection, we describe how we do this type of new annotation in detail, and show that our corpus is free of the problems discussed in this section.

2.3 Annotation Specification and Procedure of Hierarchical Alignment

We take the Chinese–English portion of the Parallel Aligned Treebank described in Li et al. (2012) for annotation. Our data consist of three batches: one batch is Web blogs, one batch is postings from online discussion forums, and one batch is newswire. The actual direction of translation in creating this data is this: Newswire articles and discussion forum postings are from Chinese to English, and blogs are from English

5 Attaching a function word to a head during word alignment does not necessarily lead to incompatibilities if the word alignment guidelines are compatible with the treebank guidelines. In practice, existing word alignment guidelines neither explicitly take the syntactic structure into consideration nor come up with aligning methods that are compatible with treebanks. The need to attach a function word to its head demonstrates the need to refer to syntactic structures, but this has never been systematically considered.

to Chinese. The English sentences in the data set are annotated based on the original Penn TreeBank (PTB) annotation stylebook (Bies et al. 1995) as well as its extensions (Warner et al. 2004), while the Chinese sentences in the data set are annotated based on the Chinese TreeBank (CTB) annotation guidelines (Xue and Xia 1998) and its extensions (Zhang and Xue 2012). The Parallel Aligned Treebank only has word alignments, which are done under the GTAHS, and no phrase alignments.

The main departure of our approach is that we loosen the requirement that every word in a sentence pair needs to be word-aligned.[6] On the word level, we only align words that have an equivalent in terms of lexical meaning and grammatical function.[7] For words that do not have a translation counterpart, we leave them unaligned and locate the appropriate phrases in which they appear to be aligned. This immediately eliminates the spurious ambiguities discussed for the GTAHS. Because phrase alignment is done between syntactic nodes on parallel parse trees, we also eliminate the incompatibilities between word alignments and syntactic structures. For an illustration of the points made here, see the discussion of the concrete example in Figure 1.

Next let us introduce our annotation procedure. Our annotators are presented with sentence pairs that come with parallel parse trees. The task of the annotator is to decide, first on the word level and then on the phrase level, if a word or phrase needs to be aligned at all, and if so, to which word or phrase it should be aligned. The decisions about word alignment and phrase alignment are not independent, and must obey the well-formedness constraints as outlined in Tinsley et al. (2007):

A. A non-terminal node can only be aligned once.

B. If node n_c is aligned to node n_e, then all the descendants of n_c can only be aligned to the descendants of n_e.

C. If node n_c is aligned to node n_e, then all the ancestors of n_c can only be aligned to the ancestors of n_e.

This means that once a word alignment is in place, it puts constraints on phrase alignments. A pair of non-terminal nodes (n_c, n_e) cannot be aligned if a word that is a descendant of n_c is aligned to a word that is not a descendant of n_e on the word level.

Let us use the concrete example in Figure 1 to illustrate the annotation process, which is guided by a set of detailed annotation guidelines (Deng and Xue 2014b). As shown in the figure, on the word level, only those words that are connected with a dashed line are aligned, because they have equivalents. Note that five Chinese words, namely, 把 (a function word used to prepose the object to the left of the verb), 一/'one', 个 (a classifier), 这样/'this way', 做/'do', and three English words, namely, the infinitive marker *to* and the two determiners *a* and *the*, are left unaligned on the word level. Aligning these words will generate spurious ambiguities and create incompatibilities between word alignments and syntactic structures, which we elaborate on next.

6 With the GTAHS, it is required that all the words should be aligned between two parallel sentences, as can be seen from the following statement quoted from page 5 of the LDC guidelines for Chinese–English word alignment (Li, Ge, and Strassel 2009): "All words in both source and target languages should be linked or marked. No single piece could be left unattended."

7 Note that this does not rule out one-to-many or many-to-many alignments, which are allowed as long as they are equivalent in both lexical meaning and grammatical function. For instance, the one-to-many alignment between the English compound noun *White House* and the Chinese simple noun 白宫 is considered legitimate.

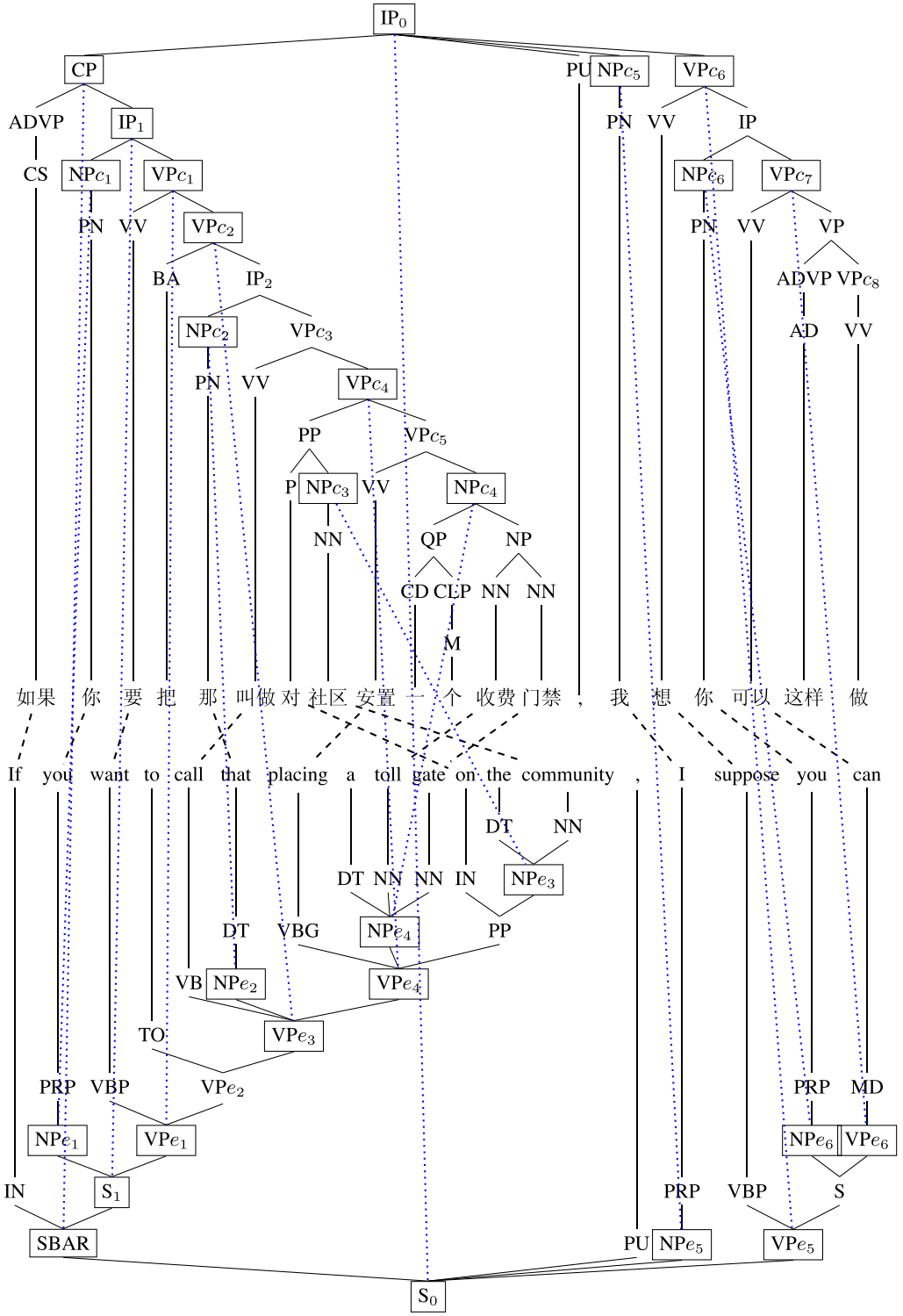


Figure 1
A hierarchically aligned sentence pair.

If 把 is to be word-aligned, it could be glued to the following demonstrative 那/'that' and the string 把 那 will be aligned to the English *that*. Note that 那 and *that* are unambiguous and form a one-to-one correspondence. With the word alignment between 把 那 and *that*, we make the unambiguous *that* correspond to both 那 and 把 那 (and possibly more strings), thus creating spurious ambiguity and translation divergence. Also note that the string 把 那 does not form a constituent in the Chinese parse tree, so the word alignment is incompatible with the syntactic structure of the sentence. By leaving 把 unaligned, we avoid both the spurious ambiguity and the incompatibility. The same logic applies to the English infinitive marker *to* if it were glued to, say, *want* and *want to* is aligned with 要/'want'. For the English determiner *the*, spurious ambiguity and translation divergence will arise if we attach it to a host, say *community*, and align *the community* with 社区/'community'. The incompatability issue will not arise though, because *the community* does form a constituent in the English parse tree. However, redundancy has been created between the word alignment and the syntactic structure since syntactic parsing has already associated *the* with *community* by grouping them to form an NP, and there is no need to repeat this information on the word level. What about *a*? One may suggest that *a* could be aligned with the string 一 个 since the English determiner and the Chinese numeral-classifier cluster are both about singularity, and aligning them takes care of the two function words *a* and 个 simultaneously. This alignment, however, will cause spurious ambiguity because *a* has been translated as many things in addition to 一 个, such as the word 某/'certain' with or without a classifier in "a teacher$_1$ <>某 (位) 老师$_1$", the word 每/'each' in "3 dollars$_1$ a gallon$_2$ <> 每 加仑$_2$ 3 美元$_1$", and so forth. Lastly, let us look at the two Chinese words 这样/'this way' and 做/'do'. The reason why these two words, both of which have lexical content, do not get translated is because of cross-linguistic differences about VP ellipsis between Chinese and English. The English grammar allows VP ellipsis, which is why it is possible to use *can* without repeating what has been said in the previous sentence, namely, the VP *call that placing a toll gate on the community*. By contrast, VP ellipsis in Chinese is more restricted and it is odd to use 可以/'can' alone in the second clause, which is why 这样 做/'do so' is used to refer back to the ellided VP 把 那 叫 做 对 社区 安置 一 个 收费 门禁/'call that placing a toll gate on the community'. To do word alignment with 这样 做, one needs to attach the two words to, say, 可以 and align the string 可以 这样 做 with *can*, which will create spurious ambiguity.

With all the word alignments in place, next the annotator needs to perform phrase alignments, which will capture the unaligned words. Note that word alignments place restrictions on phrase alignments. For instance, VP_{e3} cannot be aligned with VP_{c3}, because *that*, a descendant of VP_{e3}, is aligned to 那, which is not a descendant of VP_{c3}. By contrast, IP_2 is a possible alignment for VP_{e3} because the alignment does not violate the well-formedness constraints. The annotator then needs to decide whether this possible phrase alignment should actually be made. This is a challenging task because, for a given phrase, usually there is more than one candidate from which a single alignment needs to be picked. For instance, for VP_{e3}, besides IP_2, there is another possible phrase alignment, namely, VP_{c2}, which also obeys the well-formedness constraints. Because a non-terminal node is not allowed to be aligned to multiple non-terminal nodes on the other side, the annotator needs to choose one between the two candidates. This highlights the point that the alignment of non-terminal nodes cannot be deterministically inferred from the alignment of terminal nodes. This is especially true given our approach where some terminal nodes are left unaligned on the word level. For instance, VP_{e2} and VP_{e3} are both possible phrase alignments for VP_{c2}, and the reason why VP_{e2} is also a possible alignment for VP_{c2} is because the word *to* is left unaligned. If *to* were

glued to, say, *want*, and *want to* is aligned with 要/'want', VP_{e2} could not be aligned with VP_{c2} since 要/'want' is not a descendant of VP_{c2} and aligning the two nodes will violate Constraint B.

Given this discussion, Constraints A, B, and C together do not fully resolve the alignment of non-terminal nodes. Constraint A only requires that a non-terminal node should be aligned once; it says nothing about how to find the specific alignment for a non-terminal node. This is where manual annotation kicks in, and the decisions regarding the alignment of non-terminal nodes are based on linguistic considerations. One key consideration is to determine which non-terminal nodes encapsulate the grammatical relations signaled by the unaligned words so that the alignment of non-terminal nodes will effectively capture the unaligned words in their syntactic contexts. When identifying non-terminal nodes to align, we follow two seemingly conflicting general principles:

- Phrase alignment should not sever key dependencies involving the grammatical relation signaled by an unaligned word.
- Phrase alignment should be minimal, in the sense that the phrase alignment should contain only the elements involved in the grammatical relation, and nothing more.

The first principle ensures that the grammatical relation involving an unaligned word is properly encapsulated in the aligned non-terminal nodes. For example, in Figure 1, the reason why we align VPe_3 with VPc_2 but not IP_2 is that we need to include the function word 把 in the phrase alignment, because the very reason why the object of the verb 叫做/'call', namely, 那/'that', appears before the verb is due to the presence of 把. In other words, there is an important grammatical dependency between 那 and 把, and excluding 把 outside the phrase alignment that contains 那 will sever this dependency.

The first principle in and of itself is insufficient to produce desired alignment. Taken to the extreme, it can be trivially satisfied by aligning the two root nodes of the sentence pair. We also need the alignment to be minimal, in the sense that aligned non-terminal nodes should contain only the elements involved in the grammatical relation signaled by an unaligned word, and nothing more. These two requirements used in conjunction ensure that a unique phrase alignment can be found for each unaligned word.[8] Subsequently we show that all the unaligned words are properly captured by the alignment between the appropriate phrases that function as the syntactic contexts of the words.

As already mentioned, the grammatical function of 把 is to prepose the object to the left of the verb. VP_{c2} is the smallest constituent that contains 把, the object, and the verb, and the phrase alignment between VP_{c2} and VP_{e3} captures 把 in its syntactic context. Similarly, aligning VP_{c1} with VP_{e1} captures the English infinitive marker *to* in its smallest syntactic context because there is an idiosyncratic dependency between the word and

8 This means that, just like the word alignment practice adopting the GTAHS, every single terminal node in a sentence pair will be taken care of when the annotation is done in our practice (see footnote 6). The difference between our approach and the GTAHS is how and where the words without translation counterparts are captured. Note that there is no requirement that all non-terminal nodes should be aligned. The fact is that some non-terminal nodes may be unaligned, as shown by Figure 1.

Table 1
Statistics of IAA.

Chunk No.	Precision	Recall	F1-measure
1	0.91	0.86	0.89
2	0.92	0.80	0.86
3	0.89	0.89	0.89
4	0.88	0.88	0.88
5	0.89	0.89	0.86
micro-average	0.90	0.85	0.87

the verb *want*. The phrase alignment between NP_{c3} and NP_{e3} captures *the*, and the phrase alignment between NP_{c4} and NP_{e4} captures *a* and 个. The two unaligned Chinese words 这样/'this way' and 做/'do' are captured in the phrase alignment between VP_{c7} and VP_{e6} since VP_{c7} and VP_{e6} are the smallest syntactic contexts where VP ellipsis in the two languages takes place.

Following the principles and procedure introduced here, the annotator checks every non-terminal node in the Chinese parse tree to see if it can be aligned. They do the check from the topmost root node all the way down to the lowest non-terminal node(s). If a node can be aligned, the annotator makes an alignment. If a node cannot be aligned, they skip it and keep going. When the check finishes, they stop. This is how we have constructed HACEPT, which has 9,897 sentence pairs. Our purpose for constructing HACEPT is to provide a useful resource for MT research. In the rest of this article, we show how HACEPT could benefit MT research. But before that, we will first report some statistics about the inter-annotator agreement (IAA) of our annotation, which is a way of both evaluating the annotation quality and judging the intuitiveness of the annotation task.

An unintuitive annotation task would force the annotator to make subjective choices, which would result in low IAA. Because the annotation task involves parse trees, ideally we need annotators who are trained in syntax, but that would put a constraint on the pool of qualified annotators and make it difficult for the annotation to scale up. In our annotation experiments, we use four annotators who are fluent in both English and Chinese but have no prior linguistic training, led by a syntactician who performs the final adjudication.

The IAA statistics presented in Table 1 are based on 2,500 double-annotated sentence pairs, which are divided into five chunks of 500 sentence pairs each. The statistics are for phrase alignment only because that is the difficult part in the annotation (word alignments in our annotation are all sure alignments and comparatively easy to do). As can be seen from the table, the micro-average for the five chunks is 0.87 (F1-measure),[9] indicating that we are able to get good quality annotation for this task. In addition, the agreement statistics for the five chunks are very stable, even though they are performed by different pairs of annotators, indicating we are getting consistent annotation from different annotators.

9 Using chance-corrected measures such as *Kappa* is not straightforward in this annotation setting, so we only report raw agreement scores in terms of F1-measure.

In the next section, we rely on HACEPT to systematically study translation divergences between Chinese and English.

3. Extracting and Categorizing Translation Divergences using HACEPT

Translation divergence (TD) poses a great challenge to MT research because it is ubiquitous in parallel text and makes the straightforward transfer from source structures to target structures difficult or even impractical. With a hierarchically aligned parallel treebank in place, we are ready to semi-automatically extract and categorize the translation divergences between the two languages. In the first subsection, we provide a definition for TD and classify all the instances of TD we have found into seven types according to the linguistic causes of the divergence. In the second subsection, we provide key statistics about the TD instances, and compare our findings with previous work on TD.

3.1 Define and Classify TD Between Chinese and English

Based on our annotation, we define translation divergence as follows:

Definition 1
Suppose n_c and n_e are two aligned non-terminal nodes. There is a **translation divergence** between n_c and n_e if and only if at least one of the following two conditions is met:

- At least one of all the immediate daughters of either n_c or n_e is unaligned or aligned to more than one node.
- All immediate daughters of both n_c and n_e are one-to-one aligned, but the daughter nodes appear in different word order under n_c and n_e.

Figure 1 also provides a relevant example here. As shown by the blue dotted lines, there are in total 14 alignments between non-terminal nodes. Seven of these 14 phrase alignments do not have translation divergence and the other seven all have divergence. For instance, the alignment between IP_1 and S_1 do not have translation divergence because both nodes have two immediate daughter nodes, and there is a one-to-one alignment between the two immediate daughter nodes in the same word order. In other words, neither of the two conditions in the definition is met and therefore no translation divergence arises. Similarly, the alignment between the two subject NPs, namely, NP_{c1} and NP_{e1}, does not involve translation divergence for the same reason. Now let us look at the alignment between, say, VP_{c2} and VP_{e3}. The Chinese VP_{c2} has two immediate daughter nodes BA and IP_2, the first of which dominates the terminal node 把. Note that 把 is unaligned, which means that the first condition in the definition is met and therefore the phrase alignment between VP_{c2} and VP_{e3} is an instance of translation divergence.

Based on this definition, we have found in total 62,809 instances of TD. All the TD instances we have found in HACEPT arise either because of cross-linguistic differences between Chinese and English or because of non-literal translations. Non-literal

translations arise because it is a common practice in translation that the translator chooses a non-literal translation due to considerations about naturalness, discourse coherence, and so forth. Non-literal translations may cause a divergence between the source and target sentence. For instance, we have found in our corpus that *take glee in* instead of *like* is used to translate the Chinese verb 喜欢/'like', which, as a result, creates an instance of translation divergence. It is worth pointing out that previous research on TD, such as Dorr (1994), generally ignores TD caused by non-literal translations, presumably because it is avoidable if the non-literal translation is changed to a literal one. By contrast, we will take the TD instances caused by non-literal translations into consideration. This is because non-literal translations abound in real-life corpora like ours and the translation divergence that they cause cannot be ignored and needs to be dealt with.

Next we classify the TD instances into different patterns and describe these patterns. All the TD instances are classified according to the linguistic reason from which the divergence arises. We identified seven major linguistic reasons that cover all the TD instances:

- Lexical encoding (LE)
- Transitivity (TR)
- Absence of function words (AFW)
- Category mismatch (CM)
- Reordering (RE)
- Dropped elements (DE)
- Structural paraphrase (SP)

We now provide a definition and a few concrete examples for each of these seven types.

3.1.1 Lexical Encoding (LE). The definition of **lexical encoding** (LE) is as follows:

Definition 2
Because of a non-literal translation or cross-linguistic difference in lexicalization, a given lexical item in one language is translated or expressed by a continuous or discontinous string of words in the other language, and this causes translation divergence.

Here are some examples of LE from our corpus:

(21) 安排$_1$ NP 的 优先$_1$ 顺序$_1$ <> prioritize$_1$ NP
arrange NP PRT priority order

(22) NP 崛起$_1$ <> NP rise$_1$ to$_1$ prominence$_1$

(23) 每$_1$ 年$_1$ <> annually$_1$
each year

(24) 日前$_1$ <> a$_1$ few$_1$ days$_1$ ago$_1$

Examples (21) and (22) are LE instances due to differences in lexicalization between Chinese and English. Chinese lacks a lexical equivalent to the English verb *prioritize*, which has to be expressed by the discontinous string 安排 ... 的 优先 顺序 (i.e., to arrange the priority order of ...). Similarly, the Chinese verb 崛起 does not have an English counterpart and the lexical-semantic information encoded in the word needs to be conveyed by the multi-word expression *rise to prominence*. Examples (23) and (24) are both LE instances caused by non-literal translations. The Chinese phrase 每/'each' 年/'year' has the literal translation *each year* in English but is translated as *annually*. The English phrase *a few days ago* has the literal translation 几/'several' 天/'day' 前/'ago' in Chinese but is translated by the word 日前.

There is a special case in this category that is worth mentioning, which involves a lexical category in Chinese called localizers (Chao 1968). Two examples are given here to illustrate the translation divergence caused by localizers:

(25) 在$_1$ NP 下$_1$ <> under$_1$ NP
at NP underneath

(26) 在$_1$ IP 后$_1$ <> after$_1$ S
at IP back

About Chinese localizers such as the word 下 in Example (25), three facts are relevant for the issue under discussion. First, localizers were historically nouns and are still nominal in nature in contemporary Chinese. Second, a localizer cannot stand on its own and needs to be used after an NP to form a localizer phrase, which refers to a place that stands in a specific spatial relation with the denotation of the NP. For instance, the localizer phrase 桌子 下 literally means *table underneath* and refers to the place under some table. Third, a localizer phrase is usually used as the object of the preposition 在 to form a preposition phrase (PP). For instance, the Chinese PP 在 桌子 下 corresponds to the English PP *under the table*. These three facts indicate that, in cases where an English PP is translated as a Chinese PP, the English preposition corresponds to a discontinuous string of words consisting of the Chinese preposition and a localizer. Take Example (25) for instance—the Chinese counterpart of the English preposition *under* is 在 ... 下 in terms of grammatical function and lexical semantics. The first piece in the Chinese discontinuous string, namely, 在, has the same grammatical function as *under* in that they are both prepositions. However, 在 does not have the same lexical semantics as *under*. As a matter of fact, this Chinese preposition has little, if any, lexical semantics. The Chinese localizer 下 expresses the same lexical meaning as *under* in terms of spatial relation, but a localizer like 下 is nominal in nature and therefore differs with *under* in terms of grammatical function. In other words, the Chinese equivalent of the English *under* is neither 在 nor 下 but the string 在 ... 下 if we take both grammatical function and lexical semantics into consideration.[10]

In addition to typical localizers such as 下 that express spatial relations, there are also some localizers that express temporal relations such as the one in Example (26). The localizer 后 literally means *back*, so the English word *after* is expressed in Chinese as *at the back of ...*, which causes translation divergence.

10 For cases like the example discussed here, we word-align the discontinuous string on the Chinese side with the English preposition and then align the two PPs. Besides the example here, there are other cases involving Chinese localizers that cause complications; interested readers are referred to Deng and Xue (2014b) to see how word and phrase alignments are done in those cases.

3.1.2 Transitivity (TR). The definition of **transitivity** (TR) is as follows:

Definition 3
A verb in one language and its lexical equivalent or actual translation in the other language differ in transitivity,[11] which causes translation divergence.

Here are some examples from our corpus:

(27) 抱怨$_1$ NP <> complain$_1$ about NP
complain

(28) 对 NP 有利$_1$ <> benefit$_1$ NP
for NP benefit

(29) 了解$_1$ 到 IP <> learn$_1$ that S
learn RS IP

(30) 讨论$_1$ NP <> argue$_1$ about NP
discuss

The Chinese verb 抱怨 in Example (27) is transitive and can take an object directly whereas its English lexical counterpart *complain* is intransitive and needs the preposition *about* to introduce the object. In Example (28), the English verb *benefit* is transitive whereas the Chinese 有利 is intransitive and relies on the preposition 对 to introduce its argument. In Example (29), the English verb *learn* takes a clausal object directly whereas the Chinese 了解 needs the suffix 到 to do that. Sometimes, the divergence may be caused by non-literal translations. For instance, in Example (30), the Chinese verb 讨论, which is transitive, is translated as *argue*, which is an intransitive verb and needs the preposition *about* to introduce its object.

We also include cases like the following in this category:

(31) 责怪$_1$ NP VP <> blame$_1$ NP for VP

(32) 阻止$_1$ NP VP <> prevent$_1$ NP from VP

Take Example (31), for instance. We view the Chinese 责怪 as a ditransitive verb that takes two complements: an NP denoting the person/entity that was blamed and a VP denoting the event for which the person/entity was blamed. The English *blame* also has these two semantic arguments, but unlike in Chinese where the verb can take the two arguments directly, the VP complement needs to be introduced by *for* to the verb. The same logic also applies to the case in Example (32). We view this as a difference in transitivity.

11 By transitivity, we focus on the ability of the verb to directly take an object without the help of a preposition or some other kind of particle such as a suffix-like morpheme in a Chinese resultative compound verb.

3.1.3 Absence of Function Words (AFW). The definition of **absence of function words** (AFW) is as follows:

Definition 4
Language-particular function words in one language do not exist in the other, which brings about translation divergence.

Here are some examples from our corpus:

(33) 首府$_1$ <> the capital$_1$

(34) 成为$_1$ NP <> has become$_1$ NP

(35) 一$_1$ 个 月$_2$ <> one$_1$ month$_2$
one CL month

(36) 起到$_1$ 了 NP <> played$_1$ NP
play PERF

In Example (33), the English determiner *the* has no counterpart in Chinese. In Example (34), the English auxiliary verb *has* does not have a counterpart. In Example (35), the Chinese classifier 个 does not have a counterpart. In Example (36), the Chinese perfective aspect marker 了 lacks a counterpart.

Note that both TR and AFW involve function words but for different reasons. The majority of function words involved in TR are prepositions, many of which exist in both Chinese and English. By contrast, all the function words involved in AFW are language-particular items that only exist in one language but not the other. For the sake of distinction, we list here the most common language-particular function words from both English and Chinese.

The following English function words do not exist in Chinese:

- Determiners: *the, a, an*;
- Auxiliary verbs: *do, have, be*;
- Complementizers: *that, if,*[12] *whether, for*;
- Relative pronouns: *which, that, who, when, where, how*;
- Other: *to* (infinitive marker), *of, 's*, *and* (discourse connective).

The following Chinese function words do not have lexical equivalents in English:

- Classifiers: 个, 只, 棵, 匹, 头 ...;
- Structural particles: 的, 得, 地, 所, 之, 把, 被 ...;
- Aspectual particles: 了, 着, 过, 起来, 下去 ...;
- Sentence-final particles: 吧, 呢, 吗, 啊 ...

12 We distinguish the *if* used to introduce an adverbial conditional clause, which has an equivalent in Chinese, and the *if* used to introduce a complement clause as in *John wondered if Mary has left*, which has no equivalent in Chinese.

3.1.4 Category Mismatch (CM). The definition of **category mismatch** (CM) is as follows:

Definition 5
A phrase and its translation differ in phrase types, which brings about translation divergence.

Some examples from our corpus:

(37) 进一步$_1$ 扩大$_2$ NP <> the further$_1$ expansion$_2$ of NP
further expand

(38) 主管$_1$ NP <> in$_1$ charge$_1$ of NP
oversee

(39) 一九八八年$_1$ <> in 1988$_1$

(40) 在 国际$_1$ 上 <> internationally$_1$
at international top

In Example (37), the Chinese phrase is a VP meaning *to further expand NP* whereas its English translation is an NP. In Example (38), the Chinese phrase is a VP whereas its translation is a PP. In Example (39), the Chinese phrase is an NP functioning as an adverbial, and its translation is a PP. In Example (40), the Chinese phrase is a PP, but its translation is an adverb phrase.

One thing to note is that we exclude category mismatch caused by notational differences.[13] For example, a mono-clausal sentence is represented as an IP in CTB whereas it is represented as an S in PTB. If CM is taken at face value, the alignment between a Chinese IP and an English S is an instance of CM. However, we do not count such cases as CM. Real CM cases generally involve the use of function words and change of structure. For instance, the CM case in Example (37) involves the use of the determiner *the* and the preposition *of*.

3.1.5 Reordering (RE). The definition of **reordering** (RE) is as follows:

Definition 6
A phrase and its translation differ in word order, which brings about translation divergence.

Some examples from our corpus:

(41) 再$_1$ 伤害$_2$ 我们$_3$ <> hurt$_2$ us$_3$ again$_1$
again hurt we

(42) 往前$_1$ 去$_2$ <> go$_2$ forward$_1$
forward go

13 A complete list of all the notational differences can be obtained by a comparison of the category labels used in the PTB annotation guidelines (Bies et al. 1995) and those used in the CTB annotation guidelines (Xue and Xia 1998). To save space, we will not provide the list here.

(43) 发展$_1$ 速度$_2$ <> the rate$_2$ of development$_1$
development speed

(44) IP$_1$ 的 NP$_2$ <> NP$_2$ SBAR$_1$

In both Examples (41) and (42), the adverbial modifier appears post-verbally in English whereas their Chinese counterparts appear before the verb. This is a very typical word order difference between the two languages. As for the word order differences in Examples (43) and (44),[14] that is because Chinese is head-final whereas English is head-initial in the nominal domain. In Example (43), the Chinese compound has its head 速度/'speed' on the right of the modifier 发展/'development' whereas in the English NP the head *rate* comes before *development*. In Example (44), the head NP is after the relative clause IP in Chinese whereas the head NP in English is before the relative clause SBAR.

3.1.6 Dropped Elements (DE). The definition of **dropped elements** (DE) is as follows:

Definition 7
A constituent in a phrase has no overt manifestations in its translation due to a non-literal translation or some independent grammatical reason rather than AFW, which brings about translation divergence.

Some examples from our corpus:

(45) 在$_1$ NP 方面 <> in$_1$ NP
at NP aspect

(46) 我$_1$ 喜欢$_2$ <> I$_1$ like$_2$ it
I like

(47) 停止$_1$ VP <> They had stopped$_1$ VP
stop

(48) 预计$_1$ IP <> It is estimated$_1$ that S
estimate

Example (45) is a case where the divergence under discussion is caused by non-literal translation. The Chinese phrase literally means "in (在) the aspect (方面) of VP." Note that the noun 方面 does not get translated. This is presumably because the translator thinks that "in VP" is more natural than the literal translation "in the aspect of VP" in the context where the translation appears. The other three examples are all cases where the divergence is caused by some grammatical difference between Chinese and English. For both Examples (46) and (47), Chinese is a pro-drop language that allows the omission of both subject and object pronouns. By contrast, English is not a pro-drop language. Because of this difference, in Example (46), the object of the verb 喜欢 is dropped, which causes translation divergence since the English object pronoun *it* cannot be dropped and lacks a counterpart. In Example (47), the subject pronoun on the Chinese side is dropped. Because *they* on the English side cannot be dropped, a translation divergence arises. As for Example (48), the subject pronoun *It* on the English

14 Word order difference is not the only divergence here. These two examples also involve AFW. It is common for an example to involve more than one divergence; see Example (53) and its discussion.

side is an expletive pronoun, which is used to satisfy the grammatical requirement in English that every clause needs a subject. Chinese differs from English in that Chinese does not require expletive subjects, which is why the Chinese sentence lacks a subject before the verb 预计.[15]

Note that in both the current case and the case of AFW, a word/constituent present on one side is absent on the other side. However, the absence is due to different reasons. In AFW, what is missing is a language-particular function word that exists in one language but not the other. In the case here, a missing constituent in a phrase of a language may actually exist in the language, and it is absent in that particular phrase because of either non-literal translation or some independent grammatical reason.

Let us compare an AFW example with Example (33) to see the difference. For AFW, the English determiner *the* is missing in 首府$_1$ <> the capital$_1$. We cannot argue that Chinese has an equivalent to *the* that gets dropped because such an element does not exist in the language. By contrast, we can reasonably say that Chinese has a counterpart for *They* that gets dropped in Example (47) because Chinese does have the pronoun 他们/'they'. The pronoun is dropped in Example (47) because Chinese allows pro-dropping for independent reasons that we will not pursue here.

3.1.7 Structural Paraphrase (SP). The definition of **structural paraphrase** (SP) is as follows:

Definition 8
A phrase and its translation are structural paraphrases of each other. The structural difference and lack of word alignments between the two cause translation divergence.

Some examples from our corpus:

(49) 是 农村$_1$ 的 孩子 <> grew up in the countryside$_1$
be countryside PRT kid

(50) 拿出 责任心 <> Be responsible
take-out sense-of-responsibility

(51) 面 朝 黄土 背 朝 天 <> toiling on the land
front face land back face sky

(52) 大家 公正 行事 <> no one has a finger on the scale
everybody justly act

In Example (49), the Chinese phrase 是 农村 的 孩子 literally means "is a child from a rural area." The English translation is a paraphrase of the literal meaning of the Chinese phrase. Note that the two have different structures and also that only the word alignment between 农村 and *countryside* is available. In Example (50), the Chinese phrase 拿出 责任心 literally means "to take out one's sense of responsibility." The English translation, again, is a paraphrase of the Chinese phrase. Note that the structures of the two are different and no word alignment is possible.

The most extreme case in this category is idiomatic expressions. It is rare, if not impossible, for two idioms, one from Chinese and the other from English, to mean

15 A reviewer pointed out that the divergence caused by the absence of the expletive on the Chinese side can also be AFW because Chinese lacks the expletive. We agree, but the focus here is the fact that Chinese allows a dropped subject whereas English does not, whether the English subject is expletive or not.

the same thing and at the same time have the same structure. The translation of an idiom in one language is usually a literal expression in the other language, which has very different structure than that of the idiom. In other words, a divergence arises. For instance, in Example (51), the Chinese idiom 面 朝 黄土 背 朝 天 literally means "one's front faces the yellow soil and one's back faces the sky," which is a description of the position of a person when he does farm work such as hoeing in a field. The idiom is truthfully translated by the English translation, which has a different structure and no common words with respect to the Chinese phrase. In Example (52), the translation for the English idiom *no one has a finger on the scale* is 大家 公正 行事, which literally means *everyone does things justly*. Again, the two have different structures and no word alignment is possible between them.

3.1.8 Summary. The seven types introduced here cover all the TD instances we have found in HACEPT. Note that the categories are not a partition of the TD instances but rather features that can be assigned to TD instances satisfying the requirement(s) specified in their definitions. A TD instance may have more than one feature, and, as a matter of fact, many TD instances fit in more than one category.

For instance, an English *wh*-question and its Chinese translation will involve at least reordering and absence of function words. It involves reordering because Chinese is a *wh*-in situ language that does not move the *wh*-word to the beginning of the sentence whereas English does. It involves absence of function words because English *wh*-questions undergo subject–auxiliary inversion that requires an auxiliary such as *do*, which does not have a lexical equivalent in Chinese. So an aligned node pair between two *wh*-questions at least has the two features AFW and RE. The point here is illustrated by the following example:

(53) 你$_1$ 吃$_2$ 了 什么$_3$? <> What$_3$ did you$_1$ eat$_2$?
you eat PERF what

Here, there is a word order difference because the Chinese question word 什么 is at the end of the sentence whereas its English counterpart is at the beginning of the sentence. In addition, both the Chinese perfective aspect marker 了 and the English auxiliary verb *did* have no counterparts, which causes AFW.

In this subsection, we defined and classified translation divergence between Chinese and English. In the following subsection, we provide important statistics for the TD types and compare our findings with previous work on TD.

3.2 Statistics and Discussion

We present the statistics first and then provide some comparative discussion.

3.2.1 Statistics. Before getting to the numbers, let us first briefly introduce how we find the TD instances in each of the seven categories introduced earlier. As we have mentioned, we treat the seven categories as features carried by TD instances. Here is how we semi-automatically assign relevant features to each TD instance:

Given a sentence pair, we recursively search from the root of the parse tree down to the terminal nodes on each side. Any descendant terminal or non-terminal nodes under a pair of aligned nodes that are unaligned or aligned to several nodes would be considered as triggers for an instance of translation divergence. These triggers will

Table 2
Proportion of the TD types.

TD type	No. of TD instances	Percentage of TD type
LE	10,871	17.31
TR	2,251	3.58
AFW	51,165	81.46
CM	9,183	14.62
RE	19,997	31.84
DE	5,083	8.09
SP	5,967	9.50

then be used to classify this aligned node pair into different TD categories. After the TD triggers are discovered, we design a heuristics-based approach to set up a preliminary classification:

- LE: One-to-many word alignment exists on either side.
- TR: An unaligned PP or VRD is found under a VP.
- AFW: An unaligned terminal node is found.
- CM: The labels of the two aligned root nodes are different.
- RE: Alignment of terminal nodes is not monotonically increasing.
- DE: None of the terminal nodes within an NP or VP is aligned.[16]
- SP: Many-to-many word alignment exists.

These definitions will be checked one by one. If the condition of a TD category is satisfied, then this aligned node pair would be tagged with that TD category. If none is satisfied, then this aligned node pair is considered not divergent.

We then manually check the preliminary results and filter out errors in each category. Take CM for instance: Automatic search includes many cases that have category mismatch caused by notational differences mentioned above. These cases are trivial and will be excluded during the checking process. After filtering out trivial and incorrect cases, we obtain 62,809 aligned node pairs with translation divergence out of a total of 103,796 aligned node pairs. The TD rate is 0.61. We now report some important statistics about the seven categories.

Table 2 below shows the proportion of each of the seven TD categories.

Note that the numbers in the second column add up to 104,517, which is more than the total number of the TD instances (62,809). In addition, the total percentage in the third column is 166.40%—over 100%. This is because when doing statistics for each category, instances that fit in multiple categories are counted whenever the requirement of a category is met. As noted at the end of the previous subsection, many TD instances involve more than one TD category. As a result, many instances are counted more than once, giving rise to a total percentage above 100%. There are, however, instances that

16 Only NP and VP are considered because this divergence is mainly caused by dropped arguments, which are NPs, and VP ellipsis.

Table 3
TD instances with only one feature.

TD type	Number of instances	Relative percentage	Absolute percentage
LE	2,552	23.48	4.06
AFW	24,796	48.46	39.48
CM	2,600	28.31	4.14
RE	2,197	10.99	3.50
DE	606	11.92	0.96
SP	668	11.19	1.06
Total	33,419	N/A	53.21

carry only one feature. The statistics about TD instances that have only one feature tag are given in Table 3.

The Relative percentage column provides the proportion of the instances from a TD category that only carry the TD feature to all the instances in that category. The Absolute percentage column provides the proportion of the instances from a TD category that only carry the TD feature to all the TD instances (namely, 62,809). For instance, out of the 10,871 instances in the LE category, 2,552 of them carry only the LE feature, which gives us a relative percentage of 23.48% and an absolute percentage of 4.06%. Note that TR is missing in the table. This is because a TR instance necessarily carries the AFW tag because of the definition of TR. Also note that the relative percentage in all the categories except AFW is rather low (below 30%). This indicates that an instance in one of these categories generally involves another category.

Given the fact that a TD feature rarely occurs all by itself, next we review the statistics about the co-occurrence of the features. Table 4 shows the statistics of feature combinations that cover more than 1,000 instances. "Relative percentage" in this table refers to the proportion of the instances carrying the feature combination in question to all the instances that have the same number of co-occuring features as the instances in question. Absolute percentage in this table is defined as the proportion of the instances carrying the feature combination in question to all the TD instances (namely, 62,809). For instance, among all the TD instances that have two co-occurring features, there are 8,975 TD instances that have the two features AFW and RE, which accounts for a relative percentage of 45.11%. The 8,975 TD instances with AFW and RE take up an absolute percentage of 14.29% in all the 62,809 TD instances.

What is most notable in Table 4 is that all the feature combinations involve AFW except the last one (LE and CM). This indicates that function words play an important role in defining the translation divergence between the two languages and therefore deserve more detailed investigation. Because of this consideration, in Table 5 we provide the statistics about the number of TD instances in each category that carry the AFW feature:

"Relative percentage" in the table refers to the proportion of the instances in a particular TD category that have the AFW feature to all the instances in that TD category. Absolute percentage refers to the proportion of the instances in a particular TD category that have the AFW feature to all the TD instances (namely, 62,809). For instance, among the 10,871 instances in the LE category, 6,626 of them carry the AFW feature, and this accounts for a relative percentage of 60.95% and an absolute percentage of 10.55%. As can be seen from the table, the lowest relative percentage is 46.76% and the highest is

Table 4
Co-occurrence of TD features.

Co-occurring features	No. of instances	Relative percentage	Absolute percentage
AFW and RE	8,975	45.11	14.29
AFW and LE	2,778	13.96	4.42
AFW and CM	2,115	10.63	3.37
AFW and DE	1,229	6.18	1.96
AFW, LE, and RE	1,620	22.89	2.58
AFW and SP	1,199	6.03	1.91
AFW, RE, and SP	1,118	15.80	1.78
LE and CM	1,100	5.53	1.75
Total	20,134	N/A	32.06

100%, indicating that function words play an important role in generating all types of translation divergences.

Of the 46,669 TD instances that involve AFW, 26,445 of them (56.7%) have unaligned function words only on the English side; 11,480 of them (24.6%) have unaligned function words on both sides; and 8,744 of them (18.7%) have unaligned function words only on the Chinese side. This means that English function words account for most of the TD occurrences. Among all the unaligned function words, 12 of them have occurred more than 1,000 times in the corpus. Table 6 shows a ranked list of 12 most frequent function words with their frequencies.

As can be seen from the list, only 3 of the 12 words come from Chinese—namely, the prenominal modification marker 的; the aspect-related particle 了, which appears either right after the verb or at the end of the sentence; and the 是, which is sometimes used as a copula and sometimes as an emphatic marker.

3.2.2 Discussion. There is a large body of literature on qualitative analysis of translation divergence. However, quantitative analysis of translation divergence seems to be rare. HACEPT makes it possible to study translation divergence both qualitatively and quantitatively. Our findings of translation divergence between Chinese and English confirm major TD types identified in other language pairs by previous qualitative research, and, more importantly, make up an aspect missing in previous work, namely, a quantification of each identified TD type that can be used to inform and direct MT research.

Table 5
TD instances with the AFW feature.

TD Type	No. of instances with AFW	Relative percentage	Absolute percentage
LE	6,626	60.95	10.55
TR	2,251	100.00	3.58
CM	4,294	46.76	6.84
RE	16,281	81.42	25.92
DE	4,160	81.84	6.62
SP	4,901	82.14	7.80
Total	38,513	N/A	61.32

Table 6
Top 12 function words with frequencies.

word	freq	word	freq	word	freq	word	freq
(1) *the*	12,595	(2) 的	11,231	(3) *of*	5,459	(4) *and*	3,301
(5) *a*	3,277	(6) *to*	2,864	(7) 了	2,295	(8) *is*	1,631
(9) *'s*	1,560	(10) *that*	1,457	(11) *in*	1,238	(12) 是	1,047

Here we show that our classification system covers the seven types of translation divergence with respect to English, Spanish, and German reported by Dorr (1994), where the following classification is provided:

(I) Thematic divergence:
E: I like Mary ⇔ S: María me gusta a mí
'Mary pleases me'

(II) Promotional divergence:
E: John usually goes home ⇔ S: Juan suele ir a casa
'John tends to go home'

(III) Demotional divergence:
E: I like eating ⇔ G: Ich esse gern
'I eat likingly'

(IV) Structural divergence:
E: John entered the house ⇔ S: Juan entró en la casa
'John entered in the house'

(V) Conflational divergence:
E: I stabbed John ⇔ S: Yo le di puñaladas a Juan
'I gave knife-wounds to John'

(VI) Categorial divergence:
E: I am hungry ⇔ G: Ich habe Hunger
'I have hunger'

(VII) Lexical divergence:
E: John broke into the room ⇔ S: Juan forzó la entrada al cuarto
'John forced (the) entry to the room'

Types (IV) to (VII) all have equivalents in our classification system. The fourth type corresponds to transitivity (TR) in our system, because the English verb *enter* takes its object directly whereas the Spanish verb *entró* needs the preposition *en* to introduce its object. The fifth type belongs to lexical encoding (LE) in our system. It seems that Spanish does not have a lexical item corresponding to the English verb *stab* and needs to express the lexical meaning of *stab* with the multi-word expression *dar/'give' puñaladas/'knife-wounds' a/'to'*. The sixth type is category mismatch (CM) in our classification, because in the example the predicate is adjectival (*hungry*) in English but nominal (*Hunger*) in German. The seventh type also falls in the LE category in our system. The English *break into* can be viewed as a single lexical entry (i.e., a phrasal

verb) because the meaning of the whole expression cannot be compositionally derived from the literal meaning of its two components. Spanish does not have a lexical item corresponding to the English phrasal verb and expresses the meaning analytically using the expression *forzar*/'*force*' *la*/'*the*' *entrada*/'*entry*' *a*/'*to*'.[17]

As for the first three types of divergence found by Dorr (1994), they are rare between Chinese and English and there seem to be few, if any at all, instances of these three divergences in our corpus. Even if such instances do exist, they can be captured in our classification system. To be specific, thematic divergence can be captured by reordering (RE) because it involves word order change: The two arguments of the English psych verb *like* and those of the Spanish *gustar* occupy different positions in the sentence. The two head switching divergences, namely, promotional and demotional divergence, can both be captured by SP, because the main predicate has been changed from the source language to the target language: in (II), *usually* functions as an adverbial in English and its counterpart is the main verb *soler* in Spanish. In (III), *like* is the main verb in English whereas its counterpart is an adverbial *gern* in German. In other words, the original meaning is structurally paraphrased by the translation.

An aspect missing in qualitative analysis of translation divergence such as Dorr (1994) is the statistics of each identified TD type that specifies the extent of the types. It is unclear how frequently, say, the two head switching divergences occur in parallel texts. With the statistics reported in the previous subsection, we now have a clear notion about the distribution of each of the seven types of translation divergence between Chinese and English. It is worth pointing out that AFW and RE, which were not listed in the previous classification, turn out to be the most frequent among all TD types. SP, which covers the two head switching divergences, is actually the least frequent and therefore far less pressing than AFW and RE.

To summarize, in this section we rely on HACEPT to systematically investigate the translation divergence between Chinese and English. In the next section, we look into the question of whether the divergences reported in this section can be syntactically captured.

4. Can the Translation Divergences Be Captured by Syntax-Based Translation Rules?

Having extracted and characterized the translation divergences empirically using HACEPT, in this section we try to answer the question whether these translation divergences can be captured by the kind of syntax-based rules used in modern SMT systems. Given a pair of hierarchically aligned parse trees, we show that a simple algorithm can be followed to extract Hiero-style rules described in Chiang (2005). In comparison with Chiang's approach, which relies on word-aligned sentence pairs to extract translation rules, the procedure of translation rule extraction is much simpler for us because we have hierarchically aligned parse trees.

This is how we extract a hierarchical translation rule from a phrase alignment: Given a pair of hierarchically aligned trees T_s and T_t, find all aligned node pairs (n_s, n_t). For

17 Our interpretation of the divergence here is slightly different from that of Dorr (1994). According to Dorr, "in (VII), the event is lexically realized as the main verb *break* in English but as a different verb *forzar* (literally *force*) in Spanish." So Dorr is focusing on the difference between *break* and *forzar* and treating it as the origin of the divergence. In our view, it does not seem accurate to say that "the event is lexically realized as the main verb *break* in English" because an event denoted by the verb *break* and that denoted by *break into* are quite different, and the English sentence is an event of breaking into rather than breaking (the room). In other words, the divergence here involves not only the two verbs but also other elements such as the preposition *into* and the object of *forzar*.

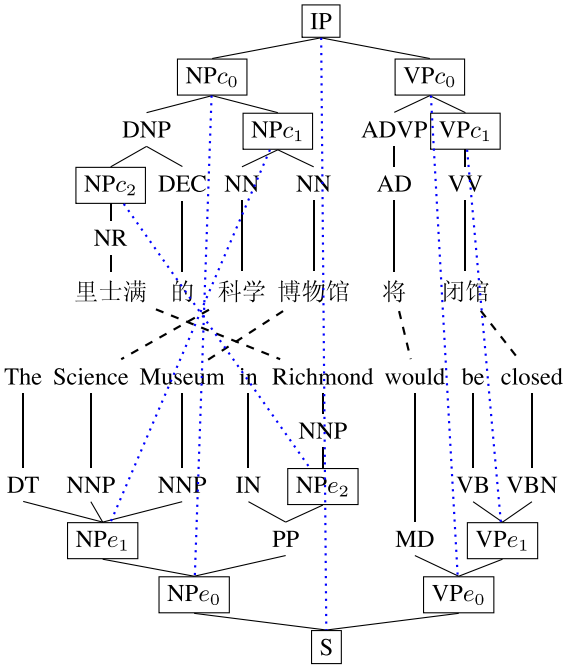


Figure 2
An example of rule extraction.

each pair of aligned nodes (n_s, n_t), if some of the immediate daughter nodes are also aligned, replace the aligned daughter nodes with variables; otherwise, take the yield of the node pair. The result is a translation rule that consists of any number of lexical items and variables. Let us use the aligned sentence pair in Figure 2 to illustrate the translation rules that can be extracted from HACEPT.

Following the procedure as spelled out here, we can obtain the following translation rules based on the phrasal alignments in Figure 2 (before the colon is the pair of phrases that have been aligned and after the colon is the translation rule extracted based on the phrase alignment):

(a) IP $\Leftrightarrow$ S: NPc_0 VPc_0 <> NPe_0 VPe_0

(b) NPc_0 $\Leftrightarrow$ NPe_0: NPc_2 的 NPc_1 <> NPe_1 in NPe_2

(c) VPc_0 $\Leftrightarrow$ VPe_0: 将 VPc_1 <> would VPe_1

(d) NPc_1 $\Leftrightarrow$ NPe_1: 科学 博物馆 <> The Science Museum

(e) NPc_2 $\Leftrightarrow$ NPe_2: 里士满 <> Richmond

(f) VPc_1 $\Leftrightarrow$ VPe_1: 闭馆 <> be closed

As we can see, a translation rule may consist of only lexical items such as those in (d), (e), and (f); or may be composed of only phrase variables such as the one in (a); or may contain both lexical items and variables such as those in (b) and (c). In the last case, the lexical items in a rule may be continuous or discontinuous. When the lexical items are discontinuous, that is, separated by variables, it is a case of

long distance lexical dependency,[18] which means that the lexical items tend to co-occur. This is precisely what makes Hiero-style translation rules capable of capturing long-distance dependencies. Another way of looking at this is that the Hiero-style rules make a hierarchical partition of a sentence pair instead of a linear partition as phrase-based translation rules make. The question we are interested in answering is whether Hiero-style translation rules extracted from hierarchically aligned parse trees can capture the translation divergences we identified in Section 3.

Before answering the question, we first need to have a more precise definition of "capturing." By "capturing," we mean that the translation divergences are properly encapsulated in the translation rules, which are holistic mappings of the source side to the target side, and nothing extra needs to be done about the translation divergences in the translation process. This requires that the number of lexical items in the translation rules is small so that they can realistically repeat in a typical parallel corpus used to extract translation rules. In the rest of this section we first show that the translation divergences can be captured in Hiero-style translation rules. We then present a distribution of the translation rules by the number of lexical items that they contain. The distribution statistics show that most of the translation rules contain only a small number of lexical items, but there are also a significant number of translation rules that contain a large number of lexical items. A closer look reveals that the latter kind of rules are due to flat structures in the parallel parse trees. This indicates that the existing treebanks are not optimal for purposes of extracting translation rules. More articulated structures are preferred to flatter structures.

Capturing translation divergence caused by lexical encoding. Figure 3a provides an example for the translation divergence caused by LE. The English verb *desensitize* does not have a lexical counterpart in Chinese and is translated by the discontinuous string 让/'make' ... 变得/'become' 漠然/'indifferent' (i.e., to make ... become indifferent). The phrase alignment between the two root VPs is the syntactic context where the divergence appears. Note that the two aligned VPs are semantically equivalent and contain the words that are involved in the divergence. There are three phrase alignments contained within the divergence context. We replace these three phrase alignments with the appropriate phrase variables and we obtain the following translation rule, which effectively captures the translation divergence:

- 让$_1$ NP$_2$ PP$_3$ 变得$_1$ 漠然$_1$ <> desensitize$_1$ NP$_2$ PP$_3$

Capturing translation divergence caused by dropped elements. Chinese is a language that allows the omission of different kinds of elements given a context. In the example in Figure 3, both the subject and the object pronoun are dropped on the Chinese side whereas the English sentence keeps both. In our view, identifying the location and the lexical-semantic content of a Chinese dropped element is a different NLP task than capturing the translation divergence caused by the dropped element (interested readers are referred to Baran, Yang, and Xue [2012] for annotation work on Chinese dropped pronouns, and Yang and Xue [2010] for results of experiments done to automatically recover Chinese dropped pronouns). When the dropped elements are recovered and put back on the parse tree, the divergence may disappear. This is shown in Figure 3b. Without the subject on the Chinese side, there is translation divergence between IP and

18 An example for this is the translation rule: VPc ⇔ VPe: 阻止$_1$ NP VPc_1 <> prevent$_1$ NP from VPe_1, where there is a long distance lexical dependency between "prevent" and "from."

S. Similarly, the absence of the object on the Chinese side causes translation divergence between the two VP_2. If the two dropped pronouns are recovered and represented on the parse tree by pro_0 and pro_1, respectively, the two instances of translation divergence no longer exist.

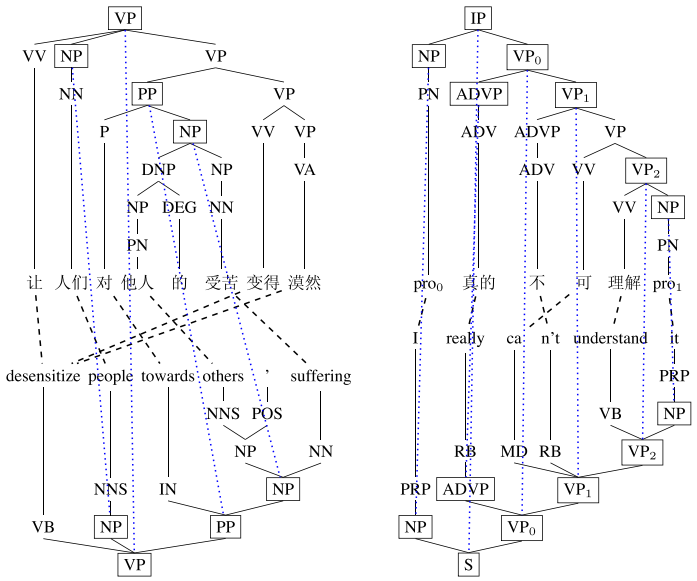
Capturing translation divergence caused by transitivity. The example in Figure 4a is an instance of the translation divergence caused by transitivity. The English verb *fund* is transitive and can directly take an object denoting the entity that receives the funding. By contrast, the Chinese translation of the verb, 出资, is intransitive and needs the preposition 对/'to' to introduce the semantic argument to the verb. After replacing the phrase alignments contained in the divergence context with a phrase variable, we obtain the translation rule (this rule also involves a word order difference, which is due to the fact that PPs must appear before the verb in Chinese):

- 对 NP 出资$_1$ <> fund$_1$ NP

This rule captures the translation divergence in question.

Capturing translation divergence caused by absence of function words. Figure 4b provides an example for the translation divergence caused by AFW. The infinitive marker *to* in English used to introduce a purpose clause in the example does not exist in Chinese. With the phrase alignments given in the figure and after the abstraction, we obtain the translation rule that captures the divergence:

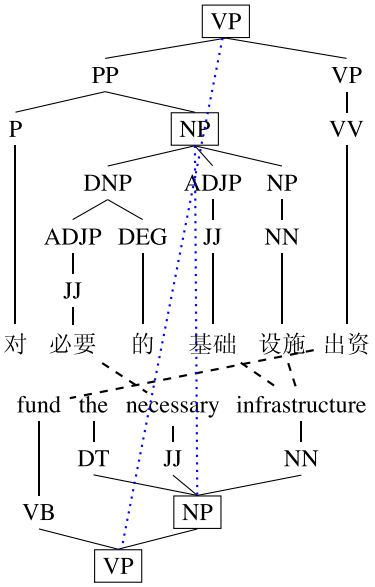
- 拿$_1$ NP VP <> take$_1$ NP to VP



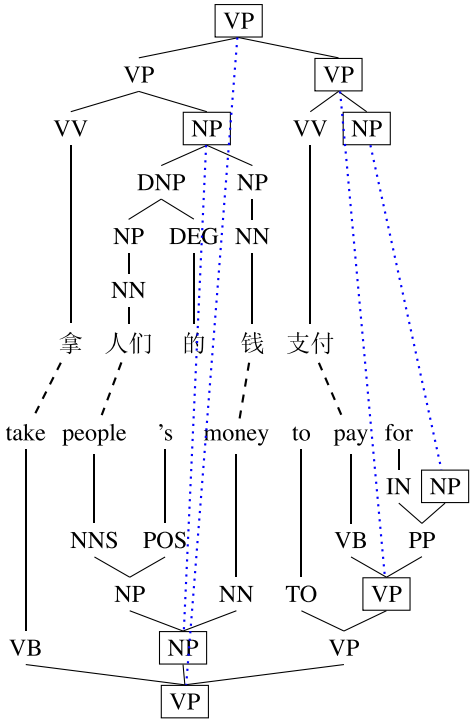
(a) Translation divergence caused by lexical encoding

(b) Translation divergence caused by dropped elements

Figure 3
Translation divergence examples: LE and DE.



(a) Translation divergence caused by Transitivity



(b) Translation divergence caused by Absence of Function Words

Figure 4
Translation divergence examples: TR and AFW.

Capturing translation divergence caused by category mismatch. Now we turn to the translation divergence caused by CM, which is illustrated in Figure 5, where an English NP is translated by a Chinese sentence (IP). Given the phrase alignments in the figure, we can obtain this translation rule to capture the divergence:

- 他们$_1$ 完全$_2$ 遵循$_3$ NP <> their$_1$ utter$_2$ conformity$_3$ to NP

Capturing translation divergence caused by reordering. There are quite a few word order differences between Chinese and English. One representative difference is the relative order between a verb and its modifier. In general, verbal modifiers like PPs appear before the verb in Chinese whereas they generally appear after the verb in English. This is shown in Figure 6a. We can obtain the following rule for this example:

- PP 引起$_1$ <> driven$_1$ PP

Capturing translation divergence caused by structural paraphrase. Now we discuss the last type of translation divergence caused by SP. The challenge posed by SP, especially idiomatic expressions, is that there can be no phrase alignments contained inside the alignment between the source phrase and its translation. This is not surprising because the source phrase and its translation (which is a structural paraphrase of the source phrase) are usually very different both lexically and structurally. As shown by

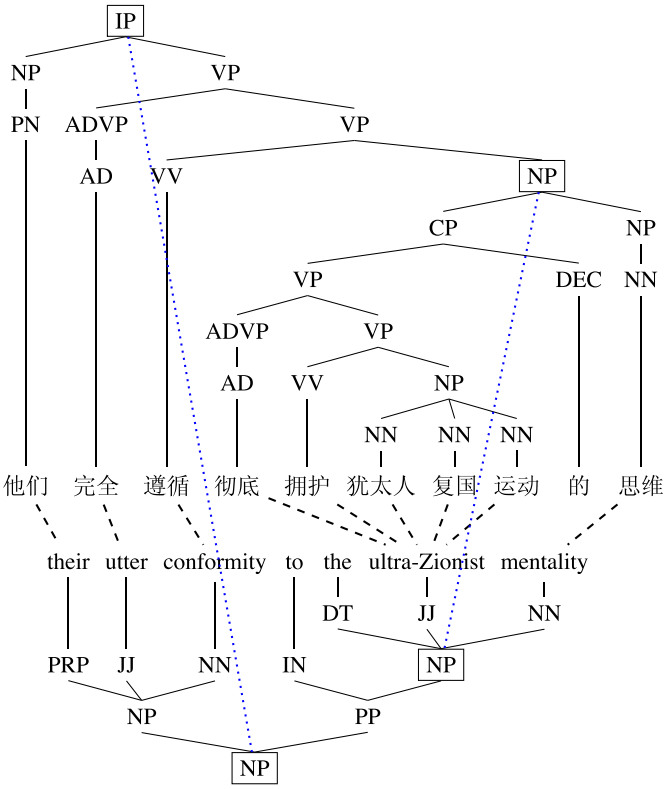


Figure 5
Translation divergence caused by CM.

the example in Figure 6b, the literal translation of the English idiom *think outside the box* is 跳出/'jump out' 常规/'regular' 思维/'thinking', where not a single word alignment is possible. For extreme cases like this, we get a translation rule that consists of only words and no phrase variables like the one below:

- 跳出 常规 思维 <> think outside the box

4.1 Statistics About the Length of Extracted Rules in HACEPT

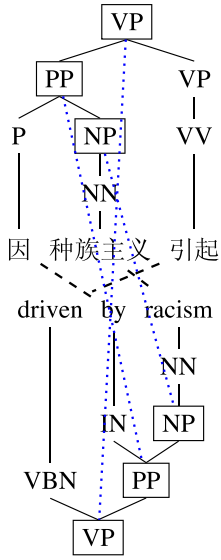
In this subsection, we provide statistics with respect to terminal nodes (lexical items) contained in the hierarchical translation rules extracted from HACEPT. In total, we have extracted 103,796 rules. As illustrated by the example in Figure 2, a rule may contain only lexical items, or only phrase variables, or both. Here is the distribution of the three types of rules:

(a) Total number of rules: 103,796 (100%)

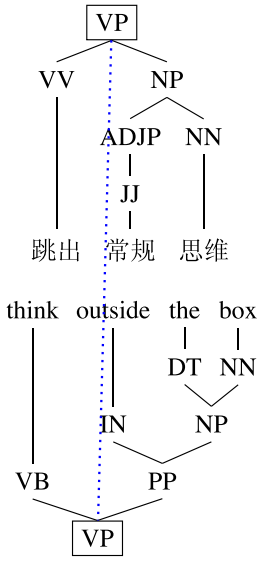
(b) Number of rules that contain only lexical items: 52,379 (50.46%)

(c) Number of rules that contain only phrase variables: 2,621 (2.53%)

(d) Number of rules that contain both lexical items and phrase variables: 48,796 (47.01%)



(a) Translation divergence caused by Reordering



(b) Translation divergence caused by Structural Paraphrase

Figure 6
Translation divergence examples: RE and SP.

Table 7
Distribution of terminal nodes in rules.

No. of terminal nodes per rule	Rule count	Percentage	Cumulative Percentage
0	6,974	6.72	6.72
1	4,017	3.87	10.59
2	30,829	29.70	40.29
3	18,780	17.09	58.38
4	12,897	12.43	70.81
5	9,387	9.04	79.85
6	6,079	5.86	85.71
7	4,404	4.24	89.95
More than 7	10,429	10.05	1

Among all the rules that contain both lexical items and phrase variables, 19,766 contain discontinuous lexical items, which accounts for 40.51% of these rules and indicates the importance of hierarchical rules.

Now let us focus on the number of lexical items contained in the rules. The statistics are given in Table 7. As shown in the table, 89.95% of the rules contain seven or fewer words.[19] Still, there are 10% of the rules that are quite long.

19 We chose seven as the cut-off as 90% of the rules have seven or fewer lexical items, well within the maximum number of lexical items allowed in a Hiero-style system in the default setting, which is five on each side (i.e., 10 total per rule).

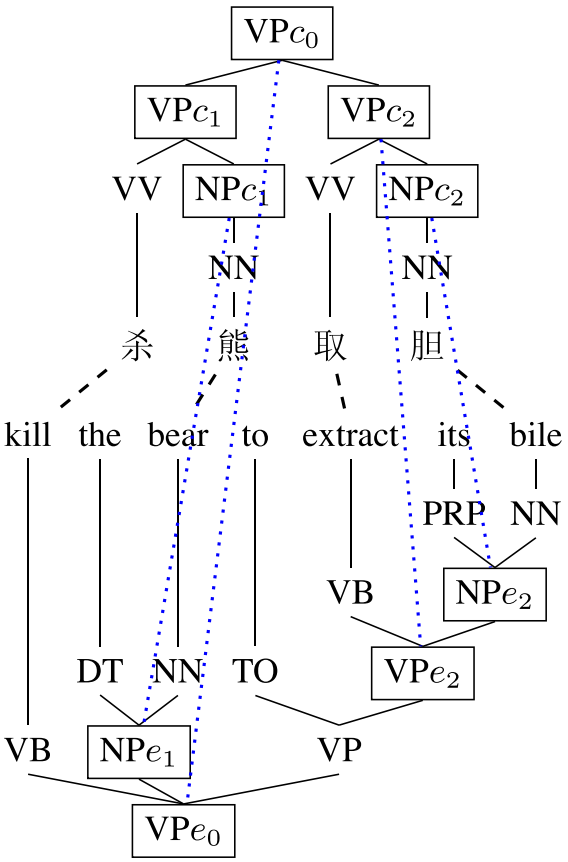


Figure 7
Unalignable phrases due to flat structures

One primary factor that increases the number of terminal nodes in a rule is that some parts of the parse trees are flat, which places some legitimate phrase alignments out of reach. This situation happens both within a clause and across clause boundaries in a multi-clausal sentence. Consider the example given in Figure 7.

It is not hard to see from the example that the hierarchical translation rule based on the phrase alignment between VP_{c0} and VP_{e0} is: "杀 NP VP <> kill NP to V", which has three terminal nodes in it (the Chinese verb 杀, the English verb *kill*, and the English infinitive marker *to*). Note that the Chinese VP_{c1} 杀/'kill' 熊/'bear' corresponds in meaning with the English string *kill the bear*. However, VP_{c1} cannot be aligned with its meaning correspondence because the English parse tree is flat and does not group the verb *kill* and its object *the bear* to form a VP.[20] If such a VP exists, it could be aligned with VP_{c1} and the hierarchical rule will be "VP VP <> VP to VP", which is a much shorter rule and contains only one terminal node. Whenever a legitimate pair of phrases cannot be aligned, the terminal nodes that they dominate will accumulate until an aligned node pair higher up in the tree is reached. As a result, we will obtain translation rules that

20 See Deng and Xue (2014a) for discussion about both the linguistic and practical engineering reasons for the fact that legitimate phrase alignments cannot be made in some places.

contain more lexical items. The situation illustrated by Figure 7 also happens across clause boundaries in a multi-clausal sentence. Because of considerations of space, we will not provide more examples.

Generally speaking, hierarchical alignment prefers more articulated structures than flat ones so that more nodes are available for phrase alignment. In the meantime, it is also clear from the examples that we have provided that not all non-terminal nodes in syntactic parses need to be aligned for the purpose of extracting Hiero-style rules. This suggests that the syntactic trees in existing treebanks simultaneously have too much and too little structure and are not optimal for the purpose of translation rule extraction, as they are not designed for any particular natural language applications. This observation is consistent with the findings of other researchers—in particular, those of Lavie, Parlikar, and Ambati (2008) and Ambati and Lavie (2008). However, when there is sufficient structure in the parse trees on both sides that are hierarchically aligned, Hiero-style translation rules that capture the translation divergences between Chinese and English can be extracted. We suspect that this observation generalizes to other language pairs as well. This suggests that one way to advance MT research is to build hierarchically aligned treebanks that systematically consider the interaction of word and phrase alignment with the syntactic structure of each sentence in a sentence pair. This is in contrast with the current state of affairs where tools and resources for word alignment and syntactic parsing are built independently.

5. Implications of Translation Divergences for Cross-Lingual Semantic Representations

In this section, we briefly discuss the implications of the translation divergences we identified from HACEPT for building cross-lingually valid semantic representations. The Vauquois Pyramid (Vauquois 1968) has had tremendous influence in shaping our conception of MT as a problem. The Vauquois Pyramid states that as we appeal to more abstract translation representations, the gap between the representations of the source and target language sentences narrows and the cost of mapping the source sentence representation to the target sentence representation will be lower, at the expense of a higher cost of analyzing a source language sentence into the abstract source representation and generating the target sentence from the abstract target translation representation. "Cost" here can be understood as the level of processing difficulty. For interlingua approaches, the goal is to achieve an abstract universal semantic representation that is shared by both the source and target language. Between the interlingua approach and those approaches that directly map words in the source language sentence to those of the target language sentence, there is a whole range of different representations that make different tradeoffs between the analysis/generation step and the mapping or transfer step. Syntactic representations, for example, are generally considered to be more abstract than word-based translation but less abstract than semantic representations. Prior to SMT, rule-based techniques for mapping between the source and target language representations were limited in their capability, and finding the right level of abstraction in the translation representation was all-important.

SMT approaches are capable of much more sophisticated mapping or transfer models, so the burden of finding the right translation representations has been eased. In fact, early IBM models (Brown et al. 1993) are word-based and rely on the direct translation between words. The introduction of syntactic trees into SMT models (Galley et al. 2004, 2006) has improved MT performance (Zollmann et al. 2008), when they are

used as constraints on the extraction of translation rules. The natural question that arises is whether there will be additional benefits if semantic structures are incorporated into SMT systems. The argument for using semantic representations remains the same as it was in the pre-SMT era, namely, that semantic representations abstract away from surface characteristics that differentiate the languages and bring the languages together. Given the translation divergences between Chinese and English that we have presented, it is worth asking if semantic representations such as Abstract Meaning Representation (AMR) can better bridge these translation divergences than syntactic representations.[21] AMR tries to develop semantic representations for MT, which form an inventory of abstract concepts and relations that can generalize over morphosyntactic variations. As a result, it has no difficulty bridging translation divergences caused by reordering, category mismatch, or the absence of function words on either side of a language pair. However, coming up with a shared semantic representation for translation divergences that involve alternative lexicalizations for the same semantic content (e.g. translation divergences that we characterize as SP and LE) may be impractical as these divergences are arbitrary and open-ended. This is consistent with findings in Xue et al. (2014), which show that structural divergences prevent AMRs for Czech and Chinese from being fully compatible with those in English. In light of our discussion in the previous section where we show that Hiero-style translation rules can encapsulate these translation divergences, and the challenges in building cross-linguistically valid semantic representations, the underlying assumptions of the Vauquois Pyramid may need to be re-examined in the context of SMT where translation equivalence can be modeled on linguistic units much larger than single words or concepts.

6. Related Work

In this section, we discuss the literature that is related to our work. We will first discuss the literature on translation divergence, and then that on alignment.

6.1 Related Work on Translation Divergence

Translation divergence was at the forefront of interlingua-based MT research in the early 1990s. Dorr (1993, 1994) discussed the issue extensively and described an elaborate scheme aimed at representing translation divergences in the Lexical Conceptual Structure framework (Jackendoff 1983, 1992) that she uses as the interlingua for MT. Since then, research on translation divergence has expanded from European languages such as German, French, and Spanish to languages spoken in Asia such as Urdu (Saboor and Khan 2010), Hindi (Dave, Parikh, and Bhattacharyya 2001; Gupta and Chatterjee 2001, 2003; Sinha, Mahesh, and Thakur 2005b, 2005a), and Sanskrit (Mishra and Mishra 2009). Translation divergence is also a challenge for MT approaches based on semantic transfer. An early survey of translation divergences in the context of transfer-based MT is provided by Lindop and Tsujii (1991). Most of the MT field has shifted to the SMT paradigm since the pioneering effort at IBM, but a few transfer-based MT efforts persevered. One recent such effort is the LOGON system (Lønning et al. 2004; Oepen et al. 2007), which uses MRS (Copestake et al. 1995, 2005) as a semantic representation framework. MT systems based on semantic transfer are designed to systematically

21 We use AMR as an example, but similar remarks could also apply to MRS (Copestake et al. 2005) and Discourse Representation Structures (Kamp and Reyle 1993).

handle "syntactic divergences" such as word order differences, but "lexical-semantic divergences" (caused by how certain meaning is encoded lexically) often need to be tackled with additional semantic transfer rules. The challenges that translation divergences pose for interlingua- and semantic transfer–based approaches are similar, and so are the solutions. For example, Dorr (1994) introduced a set of markers in her Lexical Conceptual Structure lexicon to indicate how specific lexical semantic concepts should be mapped or realized. Stymne and Ahrenberg (2006) also introduced additional rules within the MRS framework to handle certain lexical-semantic divergences between English and Swedish.

We investigate translation divergence in a very different time, when the use of large-scale parallel corpora in SMT is the standard practice, and we semi-automatically extract instances of translation divergence from a parallel corpus annotated based on a carefully designed hierarchical alignment scheme that preserves the integrity of lexical dependencies (which can alternatively be viewed as constructions or patterns). That makes it possible for us to exhaustively examine all possible translation divergences that appear in naturally occurring data, without limiting ourselves to a predefined set of translation divergences gathered from linguistic knowledge. The translation divergences that Dorr examined are all related to verbs and the realization of their arguments. Although they reflect important cross-lingual differences, they are certainly not the only translation divergences, as we have demonstrated. The use of an annotated corpus also allows us to automatically compute the distribution of the translation divergences. Although Dorr et al. (2002) also attempt to quantify the translation divergences in their data, their computation is based on a predetermined set of translation divergences, and does not necessarily cover all possible translation divergences. We do not distinguish "syntactic" and "lexical-semantic" translation divergences, because, as we show in Section 4, the Hiero-style translation rules used in statistical translation models can encapsulate both "syntactic" and "lexical-semantic" translation divergences in a uniform manner, making such a distinction largely irrelevant.

An empirical investigation of the translation divergences of the kind we describe here also sheds light on the feasibility of developing synchronous grammars for MT. Early work such as Wu (1997) and (Alshawi, Bangalore, and Douglas 2000) assumes a form of context-free grammar that has very strict restrictions on the types of grammatical rules that are allowed. Eisner (2003) argues for the need to accommodate non-isomorphic syntactic structures in MT systems and proposes using synchronous tree substitution grammars that are collections of pairs of aligned elementary trees as basic units of the grammar. Ding and Palmer (2005) implemented a statistical MT model that uses synchronous dependency insertion grammars, the basic units of which are also elementary trees, but their elementary trees are subgraphs of a dependency tree rather than a phrase structure tree. However, none of this previous work attempts to demonstrate that their form of synchronous grammar can accommodate the major types of translation divergences. We have shown that the translation divergences we have characterized can be encapsulated in Hiero-style stochastic context-free grammar rules extracted from a parallel treebank annotated with a carefully designed hierarchical alignment scheme if there is sufficient structure in the syntactic trees. In practice, Hiero-style translation rules are typically extracted from word-aligned parallel corpora without using syntactic structures in existing work, but we believe a hierarchically aligned treebank can be used to extract translation rules that can better preserve lexical dependencies or constructions in language. Of course, in order to achieve competitive performance, the hierarchically aligned corpus needs to be either automatically reproduced or acquired on a large scale.

6.2 Related Work on Alignment

There is also a wealth of work that addresses the issue of "cohesion" or compatibility between word alignments and syntactic trees. Cherry and Lin (2003) developed a probabilistic model to improve word alignment using the dependency tree structure as features to influence word alignment decisions. Cherry and Lin (2006) develop an approach that shrinks the search space for word alignment by bringing cohesion constraints imposed by the dependency structure inside an Inverse Transduction Grammar framework (Wu 1997). DeNero and Klein (2007) demonstrate how word alignments that do not respect the constituent structure of the target sentence hinder the extraction of generalizable translation rules and propose an unsupervised word alignment model that takes into account the constituent structure of the target sentence. May and Knight (2007) use rules extracted from a syntax-based MT model to re-align the words in a sentence pair. Fossum (2010) reports work in which she uses syntactic features to correct word alignment errors, and uses word alignment information to correct syntactic parsing errors. Riesa and Marcu (2010) and Riesa, Irvine, and Marcu (2011) use a large number of syntactic features from source and target language syntax to perform word alignment in a discriminative machine learning framework. Wang and Zong (2013) use dependency structure to constrain word alignment in a generative framework. All of these works recognize the need to use syntactic structure to influence word alignment (or vice versa), but, typically, work on automatic word alignment focuses on statistical modeling and does not discuss the linguistic basis of word alignment and how it interacts with syntactic structure, perhaps with the exception of Hermjakob (2009), who recognizes the difficulty of aligning what he calls "orphan" function words, words that do not have an equivalent in the other language. He ultimately adopts a G-TAHS, which seems to be the only plausible option short of performing hierarchical alignment. Our work differs from this line of research in that we systematically consider the interaction between word and phrase alignment, and propose an alignment scheme that operates on syntactic parses rather than on just words in a sentence pair.

There have been previous attempts to align non-terminal nodes in a tree (sometimes referred to as "subtree alignment") in the context of syntax-based SMT. Tinsley et al. (2007) proposes an algorithm that automatically aligns the phrase structure trees of a sentence pair, as well as a set of well-formedness constraints that such alignments have to obey. Hearne et al. (2007) automatically align a parallel English–French treebank using this algorithm and study translation divergences between the two languages as reflected in the aligned subtrees. However, their study does not go beyond translation divergences that have been previously observed such as those described in Dorr (1994) and does not attempt to quantify them. Lavie, Parlikar, and Ambati (2008) propose an algorithm to automatically align the nonterminal nodes between pairs of word-aligned phrase structure trees in a parallel corpus, and extract aligned subtrees from this corpus to create a syntax-based phrase table for use in an MT system. They show that although phrase pairs extracted this way are precise, they suffer from low coverage due to the non-isomorphic nature of the parallel trees. Ambati and Lavie (2008) extend this work and propose an approach to automatically restructure a target tree and make it more isomorphic with its corresponding source tree. Sun, Zhang, and Tan (2010) describe a machine-learning based model to align subtrees between Chinese and English sentences. Our work departs from all the previous work in that we propose a hierarchical alignment scheme that clearly separates words that should be aligned at the word level from those that should be aligned at the phrase level, and explicitly propose

the preservation of lexical dependencies as a criterion for the alignment of non-terminal nodes.

7. Conclusions

In this article, we conduct an empirical investigation of translation divergences between Chinese and English using a parallel treebank. In order to semi-automatically identify and categorize the translation divergences, we first devise a hierarchical alignment scheme between Chinese and English parse trees that eliminates conflicts and redundancies between word alignments and syntactic parses to prevent the generation of spurious translation divergences. Using this hierarchically aligned Chinese–English parallel treebank that we call HACEPT, we are able to semi-automatically identify translation divergences, classify them into seven types, and quantify each type of translation divergence. Our results show that the translation divergences are much broader than previously described in studies that are largely based on anecdotal evidence and linguistic knowledge. Our results also quantitatively demonstrate that some high-profile translation divergences that motivate previous research are actually very rare in our data, whereas other translation divergences that have previously received little attention actually exist in large quantities. We show that the type of syntax-based translation rules currently used in state-of-the-art SMT systems can be automatically extracted from HACEPT and they are expressive enough to capture the translation divergences. We also point out that existing treebanks are not optimal for extracting such translation rules. We also discuss the implications of our study to attempts to bridge translation divergence by devising shared semantic representations across languages. We show that although it is possible to bridge some translation divergences with semantic representations, other translation divergences are open-ended and building a semantic representation that captures all possible translation divergences may be impractical.

Acknowledgments

This collaborative work is supported by the IBM subcontract no. 4913014934 under DARPA prime contract no. 0011-12-C-0015 entitled "Broad Operational Language Translation." Dun Deng is also supported by Tsinghua University School of Social Sciences and Institute for Data Science. 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 and Loretta Bandera for developing and maintaining the annotation tool. 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 here.

References

Alshawi, Hiyan, Srinivas Bangalore, and Shona Douglas. 2000. Learning dependency translation models as collections of finite-state head transducers. *Computational Linguistics*, 26(1):45–60.

Ambati, Vamshi and Alon Lavie. 2008. Improving syntax driven translation models by re-structuring divergent and non-isomorphic parse tree structures. In *Proceedings of the Eighth Conference of the Association for Machine Translation in the Americas*, pages 235–244, Waikiki Beach, HI.

Banarescu, L., C. Bonial, S. Cai, M. Georgescu, K. Griffitt, U. Hermjakob, K. Knight, P. Koehn, M. Palmer, and N. Schneider. 2013. Abstract meaning representation for sembanking. In *Proceedings of LAW VII - The 7th Linguistic Annotation Workshop*, pages 178–186, Sofia, Bulgaria.

Baran, Elizabeth, Yaqin Yang, and Nianwen Xue. 2012. Annotating dropped pronouns in Chinese newswire text. In *Proceedings of LREC-2012*, pages 2795–2799, Istanbul, Turkey.

Bender, Emily M., Dan Flickinger, Stephan Oepen, Woodley Packard, and Ann Copestake. 2015. Layers of interpretation: On grammar and compositionality. In *Proceedings of the 11th International Conference on Computational Semantics*, 239–249, London.

Bies, Ann, 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. Tech report MS-CIS-95-06, University of Pennsylvania.

Brown, Peter F., Vincent J. Della Pietra, Stephen A. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311.

Chao, Yuenren. 1968. *A Grammar of Spoken Chinese*, University of California Press.

Charniak, Eugene. 2000. A maximum-entropy-inspired parser. In *Proceedings of the 1st North American Chapter of the Association for Computational Linguistics Conference*, pages 132–139, Seattle, Washington.

Cherry, Colin and Dekang Lin. 2003. A probability model to improve word alignment. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics-Volume 1*, pages 88–95, Sapporo, Japan.

Cherry, Colin and Dekang Lin. 2006. Soft syntactic constraints for word alignment through discriminative training. In *Proceedings of the COLING/ACL on Main Conference Poster Sessions*, pages 105–112, Sydney, Australia.

Chiang, David. 2005. A hierarchical phrase-based model for statistical machine translation. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, pages 263–270, Ann Arbor, MI.

Chiang, David. 2007. Hierarchical phrase-based translation. *Computational Linguistics*, 33(2):201–228.

Copestake, Ann, Dan Flickinger, Rob Malouf, Susanne Riehemann, and Ivan Sag. 1995. Translation using minimal recursion semantics. In *Proceedings of the Sixth International Conference on Theoretical and Methodological Issues in Machine Translation*, pages 15–32, Leuven, Belgium.

Copestake, Ann, Dan Flickinger, Carl Pollard, and Ivan A Sag. 2005. Minimal recursion semantics: An introduction. *Research on Language and Computation*, 3(2-3):281–332.

Cowan, Brooke, Ivona Kučerová, and Michael Collins. 2006. A discriminative model for tree-to-tree translation. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 232–241, Sydney, Australia.

Dave, Shachi, Jignashu Parikh, and Pushpak Bhattacharyya. 2001. Interlingua-based English–Hindi machine translation and language divergence. *Machine Translation*, 16(4):251–304.

DeNero, John and Dan Klein. 2007. Tailoring word alignments to syntactic machine translation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 17–24, Prague.

Deng, Dun and Nianwen Xue. 2014a. Aligning Chinese-English parallel parse trees: Is it feasible? In *Proceedings of LAW VIII - The 8th Linguistic Annotation Worshop*, pages 29–37, Dublin.

Deng, Dun and Nianwen Xue. 2014b. Annotation guidelines for hierarchically aligning Chinese-English parallel parse trees (version 1.0). Technical report, Brandeis University.

Deng, Dun and Nianwen Xue. 2014c. Building a hierarchically aligned Chinese-English parallel treebank. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics*, pages 1511–1520, Dublin.

Deng, Dun, Nianwen Xue, and Shiman Guo. 2015. Harmonizing word alignments and syntactic structures for extracting phrasal translation equivalents. In *Proceedings of the Ninth Workshop on Syntax and Structure in Statistical Translation (SSST-9)*, pages 1–9.

Ding, Yuan and Martha Palmer. 2005. Machine translation using probabilistic synchronous dependency insertion grammars. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, pages 541–548, Ann Arbor, MI.

Dorr, Bonnie J. 1994. Machine translation divergences: A formal description and proposed solution. *Computational Linguistics*, 20(4):597–633.

Dorr, Bonnie J., Lisa Pearl, Rebecca Hwa, and Nizar Habash. 2002. Duster: A method for unraveling cross-language divergences for

statistical word-level alignment. In *Machine Translation: From Research to Real Users: Proceedings of the 5th Conference of the Association for Machine Translation in the Americas, AMTA 2002*, page 31, Tiburon, CA.

Dorr, Bonnie Jean. 1993. *Machine Translation: A View from the Lexicon*. MIT Press.

Eisner, Jason. 2003. Learning non-isomorphic tree mappings for machine translation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics-Volume 2*, pages 205–208, Sapporo, Japan.

Flickinger, Dan, Stephan Oepen, and Emily M. Bender. 2017. Sustainable development and refinement of complex linguistic annotations at scale. In Nancy Ide and James Pustejovsky, editors, *Handbook of Linguistic Annotation Science*. Springer, Netherlands.

Flickinger, Dan, Yi Zhang, and Valia Kordoni. 2012. Deepbank. A dynamically annotated treebank of the *Wall Street Journal*. In *Proceedings of the 11th International Workshop on Treebanks and Linguistic Theories*, pages 85–96, Lisbon, Portugal.

Fossum, Victoria L. 2010. *Integrating parsing and word alignment in syntax-based machine translation*. Ph.D. thesis, University of Michigan.

Galley, Michel, Jonathan Graehl, Kevin Knight, Daniel Marcu, Steve DeNeefe, Wei Wang, and Ignacio Thayer. 2006. Scalable inference and training of context-rich syntactic translation models. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*, pages 961–968, Sydney, Australia.

Galley, Michel, Mark Hopkins, Kevin Knight, and Daniel Marcu. 2004. What's in a translation rule? In *HLT-NAACL 2004: Main Proceedings*, pages 273–280, Boston, MA.

Gupta, Deepa and Niladri Chatterjee. 2001. Study of divergence for example based English-Hindi machine translation, Kanpur, India. In *STRANS-2001*, pages 43–51, Kanpur.

Gupta, Deepa and Niladri Chatterjee. 2003. Identification of divergence for English to Hindi EBMT. In *Proceedings of MT Summit-IX*, pages 141–148, New Orleans, LA.

Hearne, Mary, John Tinsley, Ventsislav Zhechev, and Andy Way. 2007. Capturing translational divergences with a statistical tree-to-tree aligner. In *Proceedings of the 11th International Conference on Theoretical and Methodological Issues in Machine Translation (TMI '07)*, pages 114–121, Skövde, Sweden.

Hermjakob, Ulf. 2009. Improved word alignment with statistics and linguistic heuristics. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1*, pages 229–237, Singapore.

Huang, L., K. Knight, and A. Joshi. 2006. Statistical syntax-directed translation with extended domain of locality. In *Proceedings of AMTA*, pages 66–73, Cambridge, MA.

Jackendoff, Ray. 1983. *Semantics and cognition*. MIT Press.

Jackendoff, Ray. 1992. *Semantic structures*. MIT Press.

Kamp, Hans and Uwe Reyle. 1993. *From Discourse to Logic: Introduction to Model Theoretic Semantics of Natural Language, Formal Logic and Discourse Representation Theory*. Kluwer, Dordrecht.

Koehn, Philipp. 2009. *Statistical machine translation*. Cambridge University Press.

Koehn, Philipp, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, volume 1, pages 48–54, Edmonton, Canada.

Lavie, Alon, Alok Parlikar, and Vamshi Ambati. 2008. Syntax-driven learning of sub-sentential translation equivalents and translation rules from parsed parallel corpora. In *Proceedings of the Second Workshop on Syntax and Structure in Statistical Translation*, pages 87–95, Columbus, OH.

Li, Audrey Y.-H. 1999. Plurality in a classifier language. *Journal of East Asian Linguistics*, 8:75–99.

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

Li, Xuansong, Stephanie Strassel, Stephen Grimes, Safa Ismael, Mohamed Maamouri, Ann Bies, and Nianwen Xue. 2012. Parallel aligned treebanks at LDC: New challenges interfacing existing infrastructures. In *Proceedings of LREC-2012*, Istanbul.

Lindop, Jeremy and Jun-ichi Tsujii. 1991. *Complex Transfer in MT: A Survey of Examples*. Centre for Computational Linguistics, UMIST.

Liu, Ding and Daniel Gildea. 2008. Improved tree-to-string transducer for machine translation. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 62–69, Columbus, OH.

Liu, Yang, Yun Huang, Qun Liu, and Shouxun Lin. 2007. Forest-to-string statistical translation rules. In *Proceedings of ACL*, pages 704–711, Prague.

Liu, Yang, Qun Liu, and Shouxun Lin. 2006. Tree-to-string alignment template for statistical machine translation. In *Proceedings of COLING/ACL*, pages 609–616, Sydney, Australia.

Liu, Yang, Yajuan Lü, and Qun Liu. 2009. Improving tree-to-tree translation with packed forests. In *Proceedings of ACL/IJCNLP*, pages 558–566.

Lønning, Jan Tore, Stephan Oepen, Dorothee Beermann, Lars Hellan, John Carroll, Helge Dyvik, Dan Flickinger, Janne Bondi Johannessen, Paul Meurer, Torbjørn Nordgård, et al. 2004. Logon. A Norwegian MT effort. In *Proceedings of the Workshop in Recent Advances in Scandinavian Machine Translation*, 6, Uppsala, Sweden.

Marcus, Mitchell, Grace Kim, Mary Ann Marcinkiewicz, Robert MacIntyre, Ann Bies, Mark Ferguson, Karen Katz, and Britta Schasberger. 1994. The Penn treebank: Annotating predicate argument structure. In *Proceedings of the workshop on Human Language Technology*, pages 114–119, Plainsboro, NJ.

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

May, Jonathan and Kevin Knight. 2007. Syntactic re-alignment models for machine translation. In *EMNLP-CoNLL*, pages 360–368, Prague.

Melamed, I. Dan. 1998. Annotation style guide for the blinker project. University of Pennsylvania. Institute for Research in Cognitive Science Technical Report No. IRCS-98-06.

Mi, Haitao and Liang Huang. 2008. Forest-based translation rule extraction. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 206–214, Honolulu, HI.

Mi, Haitao, Liang Huang, and Qun Liu. 2008. Forest-based translation. In *Proceedings of ACL-08: HLT*, pages 192–199.

Mishra, Vimal and R. B. Mishra. 2009. Divergence patterns between English and Sanskrit machine translation. *Journal of Computer Science*, 8(3):62–71.

Och, Franz Josef. 1999. An efficient method for determining bilingual word classes. In *Proceedings of the Ninth Conference of the European Chapter of the Association for Computational Linguistics (EACL '99)*, pages 71–76, Bergen, Norway.

Oepen, Stephan, Kristina Toutanova, Stuart Shieber, Christopher Manning, Dan Flickinger, and Thorsten Brants. 2002. The lingo redwoods treebank motivation and preliminary applications. In *Proceedings of the 19th International Conference on Computational Linguistics-Volume 2*, pages 1–5, Taipei, Taiwan.

Oepen, Stephan, Erik Velldal, Jan Tore Lønning, Paul Meurer, Victoria Rosén, and Dan Flickinger. 2007. Towards hybrid quality-oriented machine translation—On linguistics and probabilities in MT. In *Proceedings of the 11th Conference on Theoretical and Methodological Issues in Machine Translation*, pages 144–153, Skövde, Sweden.

Petrov, Slav and Dan Klein. 2007. Improved inference for unlexicalized parsing. In *Proceedings of HLT-NAACL*, pages 404–411, Rochester, NY.

Riesa, Jason, Ann Irvine, and Daniel Marcu. 2011. Feature-rich language-independent syntax-based alignment for statistical machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 497–507, Edinburgh, Scotland.

Riesa, Jason and Daniel Marcu. 2010. Hierarchical search for word alignment. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 157–166, Uppsala, Sweden.

Saboor, A. and M. A. Khan. 2010. Lexical-semantic divergence in Urdu-to-English example based machine translation. In *6th International Conference on Emerging Technologies (ICET), 2010*, pages 316–320, Chengdu, China.

Shen, Libin, Jinxi Xu, and Ralph M. Weischedel. 2008. A new string-to-dependency machine translation algorithm with a target dependency language model. In *Proceedings of ACL-08:HLT*, pages 577–585, Columbus, OH.

Sinha, R., K. Mahesh, and Anil Thakur. 2005a. Divergence patterns in machine translation between Hindi and English. In *Proceedings of the 10th Machine Translation Summit (MT Summit X)*, pages 346–353, Phuket.

Sinha, R., K. Mahesh, and Anil Thakur. 2005b. Translation divergence in English-Hindi MT. In *Proceeding of EAMT Xth Annual Conference*, pages 245–254, Budapest.

Stymne, Sara and Lars Ahrenberg. 2006. A bilingual grammar for translation of English-Swedish verb frame divergences. In *Proceedings of EAMT*, pages 9–18, Oslo, Norway.

Sun, Jun, 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, Uppsala, Sweden.

Tinsley, John, Ventsislav Zhechev, Mary Hearne, and Andy Way. 2007. Robust language pair-independent subtree alignment. In *Proceedings of Machine Translation Summit XI*, pages 467–474, Copenhagen.

Vauquois, Bernard. 1968. A survey of formal grammars and algorithms for recognition and transformation in machine translation. In *Proceedings of the IFIP Congress-6*, pages 254–260, Edinburgh, Scotland.

Wang, Zhiguo and Nianwen Xue. 2014. Joint POS tagging and transition-based constituent parsing in Chinese with non-local features. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 733–742, Baltimore, MD.

Wang, Zhiguo and Chengqing Zong. 2013. Large-scale word alignment using soft dependency cohesion constraints. *Transactions of the Association for Computational Linguistics*, 1:291–300.

Warner, Colin, Ann Bies, Christine Brisson, and Justin Mott. 2004. Addendum to the Penn treebank II style bracketing guidelines: Biomedical treebank annotation. Technical report, University of Pennsylvania.

Wu, Dekai. 1997. Stochastic inversion transduction grammars and bilingual parsing of parallel corpora. *Computational Linguistics*, 23(3):377–403.

Xue, Nianwen, Ondrej Bojar, Jan Hajic, Martha Palmer, Zdenka Uresova, and Xiuhong Zhang. 2014. Not an interlingua, but close: Comparison of English AMRS to Chinese and Czech. In *LREC*, pages 1765–1772, Reykjavik, Iceland.

Xue, Nianwen and Fei Xia. 1998. The bracketing guidelines for Penn Chinese treebank project. Technical Report IRCS-00-08. University of Pennsylvania.

Xue, Nianwen, Fei Xia, Fudong Chiou, and Martha Palmer. 2005. The Penn Chinese treebank: Phrase structure annotation of a large corpus. *Natural Language Engineering*, 11(2):207–238.

Yang, Yaqin and Nianwen Xue. 2010. Chasing the ghost: Recovering empty categories in the Chinese treebank. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pages 1382–1390, Beijing, China.

Zhang, Min, Hongfei Jiang, Aiti Aw, Haizhou Li, Chew Lim Tan, and Sheng Li. 2008. A tree sequence alignment-based tree-to-tree translation model. In *Proceedings of ACL-08:HLT*, pages 559–567, Columbus, OH.

Zhang, Xiuhong and Nianwen Xue. 2012. Extending and scaling up the Chinese treebank annotation. In *Proceedings of the Second CIPS-SIGHAN Joint Conference on Chinese Language Processing*, pages 27–34, Tianjin.

Zhu, Jingbo, Qiang Li, and Tong Xiao. 2015. Improving syntactic rule extraction through deleting spurious links with translation span alignment. *Natural Language Engineering*, 21(2):227–249.

Zollmann, Andreas and Ashish Venugopal. 2006. Syntax augmented machine translation via chart parsing. In *Proceedings of the SMT Workshop HLT-NAACL*, New York, NY.

Zollmann, Andreas, Ashish Venugopal, Franz Josef Och, and Jay Ponte. 2008. A systematic comparison of phrase-based, hierarchical and syntax-augmented statistical MT. In *Proceedings of COLING 2008, the 22nd International Conference on Computational Linguistics*, volume 1, pages 1145–1152, Manchester, UK.