

# Linguistic Analysis of Non-ITG word Reordering between Language Pairs with Different Word Order Typologies

HWIDONG NA and JONG-HYEOK LEE, Pohang University of Science and Technology (POSTECH)

The Inversion Transduction Grammar (ITG) constraints have been widely used for word reordering in machine translation studies. They are, however, so restricted that some types of word reordering cannot be handled properly. We analyze three corpora between SVO and SOV languages: Chinese-Korean, English-Japanese, and English-Korean. In our analysis, sentences that require non-ITG word reordering are manually categorized. We also report the results for two quantitative measures that reveal the significance of non-ITG word reordering. In conclusion, we suggest that ITG constraints are insufficient to deal with word reordering in real situations.

Categories and Subject Descriptors: I.2.7 [Artificial Intelligence]: Natural Language Processing—*Machine translation*

General Terms: Languages, Experimentation, Verification

Additional Key Words and Phrases: Machine translation, inversion transduction grammar, corpus analysis

## ACM Reference Format:

Na, H. and Lee, J.-H. 2014. Linguistic analysis of non-ITG word reordering between language pairs in different word order typologies. *ACM Trans. Asian Lang. Inform. Process.* 13, 3, Article 11 (September 2014), 12 pages.

DOI : <http://dx.doi.org/10.1145/2644810>

## 1. INTRODUCTION

In machine translation (MT), one of the most important issues is word reordering. Because word order often differs between languages, we need to rearrange words in the source language (or select which source word should be translated in which target position) so that the translation becomes a natural sentence in the target language. Formally, word reordering is a problem of finding a correct permutation of the source sentence, where the number of all possible permutation grows factorial to the number of the words. Hence, word reordering has been generally approximated in previous studies.

Inversion Transduction Grammar (ITG) constraints proposed by Wu [1997] have been widely used for word reordering in machine translation studies. ITG is a subclass of synchronous context-free grammar (SCFG) that allows the orientation of nonterminals to be straight or inverted. The ITG constraints have global reordering capacity that overcome the limitation of the simple distortion model, for example, which restricts the maximum range of movements [Koehn et al. 2003]. Because every ITG can

---

This work was partly supported by the IT R&D program of MSIP/KEIT (10041807), the CSLi corporation, the BK 21+ Project, and the National Korea Science and Engineering Foundation (KOSEF) (NRF-2009-0075211).

Authors' addresses: H. Na (corresponding author) and J.-H. Lee, Department of Computer Science and Engineering, Pohang University of Science and Technology, 77 Cheongam-Ro, Nam-Gu, Pohang, 790-784, R. of Korea; email: [leona@postech.ac.kr](mailto:leona@postech.ac.kr).

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).

2014 Copyright held by the Owner/Author. Publication rights licensed to ACM. 1530-0226/2014/09-ART11 \$15.00

DOI : <http://dx.doi.org/10.1145/2644810>

Table I.

An example of non-ITG word reordering with a gold-standard word alignment. A filled box denote a sure link, and an empty box a possible link, respectively.

Korean:	너	는	그녀	가	누구	를	초대했다고	생각하니	?
Gloss:	you	TOP	she	NOM	who	ACC	invite	think-do	?
who					■				
do								□	
you	■								
think								■	
she			■						
invited							■		
?									■

be expressed as an equivalent ITG in a two-normal form, there exist polynomial time algorithms for both monolingual  $O(n^3)$  and bilingual  $O(n^6)$  parsing. Thus, the ITG constraints provide an efficient approximation for restricting the search space of word reordering.

Unfortunately, the ITG constraints make it impossible to cover all possible word reordering needs. Under ITG constraints, a nonterminal symbol always covers adjacent words in both the source and target languages. In other words, ITG cannot describe a pair of languages if a nonterminal needs to cover discontinuous words in either the source or target language. In real situations, non-ITG word reordering is often required when translating between linguistically divergent language pairs. ITG constraints restrict certain kinds of word reordering, for example, wh-questions which have non-ITG word reordering, as shown in Table I.

We transcribe the gloss for each Korean word, where TOP is a topic marker, NOM a nominative marker, and ACC an accusative marker, respectively. The five words in English “who, you, think, she, invited” and their corresponding words in Korean “너/you, 그녀/she, 누구/who, 초대했다고/invited, 생각하니/think”<sup>1</sup> cannot be described with ITG constraints, because none of the continuous words in one language has its corresponding continuous words in other language.

We raise the following research questions that need to be answered. 1) Is it sufficient to cover word reordering phenomenon under ITG constraints in real situation? 2) If not, what are the characteristics of non-ITG word reordering, and how significant are they? 3) If there are significant portion of non-ITG word reordering phenomena, how to deal with them in an efficient manner?

As the answers of the first two questions,<sup>2</sup> this article is organized as follows: First, we briefly illustrate the characteristics of ITG for word reordering, and summarize related works (Section 2). Then, we investigate three corpora (Chinese-Korean, English-Japanese, and English-Korean) between subject-verb-object (SVO) and subject-object-verb (SOV) languages to analyze the characteristics of non-ITG word reordering (Section 3). Here, we identify four types of non-ITG word reordering, and find that they appear in approximately 4% to 10% of all sentences. It is possible to convert or rewrite sentences that require non-ITG word reordering, in order to satisfy ITG constraints (Section 4). Finally, we conclude that word reordering methods should deal

<sup>1</sup>(do, 생각하니) is a possible link, but excluded for brevity.

<sup>2</sup>We leave the third question as a future work.

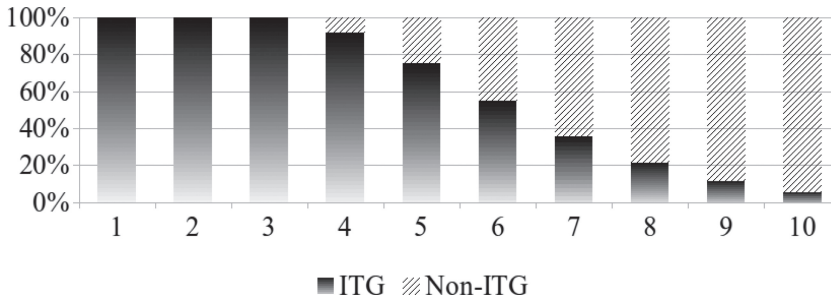


Fig. 1. Theoretical coverage of ITG and non-ITG permutations with respect to sentence length.

with such non-ITG cases, especially between linguistically divergent language pairs (Section 5).

## 2. RELATED WORKS

Previous studies have focused on investigating the capabilities and function of ITG constraints between English and Chinese [Wu 1997], other European languages [Søgaard and Wu 2009; Zens and Ney 2003], and Arabic [Wu et al. 2006]. Wellington et al. [2006] summarized the results for an English-Hindi corpus, where Hindi is an SOV language, as well as for English-European languages and English-Chinese. While Wu [1997] was “unable to find the real example” in their data, others [Wellington et al. 2006; Wu et al. 2006; Zens and Ney 2003] reported the coverage of ITG constraints at sentence level. In addition to the sentence-level coverage, Søgaard and Wu [2009] estimated empirical lower bounds for the full class of ITGs at the level of translation units. Xiong et al. [2010] analyzed reordering patterns in a phrase structure tree and investigated the three reasons for a non-reorderable case.

Although previous works reported the percentage of sentences or translation units violating ITG constraints, we think that it is more novel to perform a linguistic analysis that identifies types of non-ITG word reordering with concrete examples. We also report two quantitative measures that reveal the significance of non-ITG word reordering, which are inspired by Wellington et al. [2006] and Wu [1997]. French-English and Chinese-English pairs were already examined in Wellington et al. [2006] in a different way. The “failure rates for hierarchical alignment of bilingual bitexts under word alignment constraints only” (Table 2 in their paper) are 1% and 5% for French-English and Chinese-English, respectively, which are roughly comparable with the portion of non-ITG word reordering in this article (approximately 4%–10%). In addition, we suggest a simple heuristic to convert sentences belonging two of four non-ITG types into ITG-constrained ones.

## 3. ANALYSIS OF NON-ITG WORD REORDERING

Theoretically, ITG coverage decreases drastically as permutation length grows, as shown in Figure 1. It is possible for two non-ITG cases to appear in a length-4 permutation. For a length-5 permutation, ITG cannot describe 25% of cases, though it is questionable this may happen in real situations. Hence, we analyze three word-aligned corpora, and manually identify four types of non-ITG word reordering.

In this article, three aspects of non-ITG word reordering are investigated. First, we categorize the types of non-ITG word reordering and collect their statistics. Second, we collect statistics of the size of the word sequence causing non-ITG word reordering. Third, we measure the distribution of minimum SCFG rank to handle non-ITG word reordering.

Table II. Analysis of Non-ITG Word Reordering between SVO and SOV Languages

Corpus		Chinese-Korean	English-Japanese	English-Korean
# of sentences		367	1,235	1,895
# source/target words		8,546/9,677	30,822/34,366	22,594/29,710
average # words		23.28/26.36	24.95/27.82	11.92/15.67
Aligned by		Human		MGIZA
Alignment error		3(0.8%)	10 (0.8%)	202 (10.65%)
Total # of non-ITG w/o error		17 (4.67%)	123 (10.04%)	67 (3.95%)
Type	Wh-questions	0	0	16 (0.95%)
	Adverbials	3 (0.82%)	10 (0.82%)	14 (0.83%)
	Divergence	5 (1.37%)	61 (4.98%)	17 (1.00%)
	Others	9 (2.47%)	52 (4.24%)	20 (1.18%)
Size	~ 5	2 (0.55%)	13 (1.06%)	17 (1.00%)
	6 ~ 10	5 (1.37%)	31 (2.53%)	29 (1.71%)
	11 ~ 15	7 (1.92%)	35 (2.86%)	13 (0.77%)
	16 ~ 20	1 (0.27%)	15 (1.22%)	5 (0.30%)
	21 ~	2 (0.55%)	29 (2.37%)	3 (0.18%)
SCFG Rank	4	14 (3.85%)	82 (6.69%)	30 (1.77%)
	5	0	13 (1.06%)	14 (0.83%)
	6	3 (0.82%)	14 (1.14%)	9 (0.53%)
	7	0	3 (0.24%)	6 (0.35%)
	8 ~	0	11 (0.90%)	8 (0.47%)

### 3.1. Experimental Settings

We used three corpora consisting of different language pairs and genres. The Chinese-Korean corpus was collected from a newswire service, the English-Japanese corpus from Wikipedia articles [Neubig 2011], and the English-Korean corpus from example sentences in an electronic dictionary, respectively. Each corpus consists of sentence pairs between SVO and SOV languages, but with different characteristics such as average sentence length. The statistics of the corpora are illustrated in Table II.

For the Chinese-Korean and English-Japanese corpora, human-annotated word alignments are available, but not for the English-Korean corpus. As an alternative, an automatic word alignment was performed for the analysis of the English-Korean corpus. We used MGIZA [Gao and Vogel 2008] to obtain automatic word alignment of the development data together with the training data, consisting of approximately 600K sentences. For the high accuracy of word alignment, the intersection of bidirectional word alignment was analyzed. English sentences were segmented by `tokenizer.perl` included in Moses [Koehn et al. 2007], Japanese sentences by `KyTea` [Neubig et al. 2011], and Chinese and Korean sentences by in-house morphological analyzers.

A simple way to identify the sentences that require non-ITG word reordering is the use of a shift-reduce algorithm analogous to Huang et al. [2009]. Using a stack  $\sigma$  and a buffer  $\beta$ , each step performs one of the actions that follows:

- Reduce:  $([\sigma|i|j], \beta) \Rightarrow ([\sigma|i \cdot j], \beta)$  if  $i$  and  $j$  are adjacent in the target sentence,
- Shift:  $(\sigma, [k|\beta]) \Rightarrow ([\sigma|k], \beta)$  otherwise,

where  $i$  and  $j$  are top two elements in  $\sigma$ ,  $i \cdot j$  is a reduced element covering  $i$  and  $j$ , and  $k$  is the front element in  $\beta$ . After the termination (neither action is applicable), the ITG parsing is successful if the size of stack  $|\sigma| = 1$ : otherwise, the sentence requires non-ITG word reordering.

### 3.2. Types of Non-ITG Word Reordering

We categorize four types of non-ITG word reordering: wh-questions, adverbials, specific examples of divergences, and others. For the English-Korean corpus, we identify non-ITG word reordering based on an automatic word alignment, which may contain many errors. Thus, we manually exclude 202 sentences that contain word alignment error from the analysis. Even in human annotated word alignment, there exist minor errors of word alignment. We also manually exclude these sentences (around 0.8% of sentences) in the Chinese-Korean and English-Japanese corpora.

**3.2.1. Wh-questions.** A language pair with different word-order typologies often requires non-ITG word reordering to translate wh-questions. For a rigid word-order language like English, the order of constituents in wh-questions is almost completely fixed. For a relatively free-order language like Korean, in contrast, it is flexible. Therefore, ITG word reordering does not guarantee correct results for wh-questions between rigid and relatively free word-order languages.

Some examples of wh-questions are shown in Table IIIa. In (1), the word alignment is (why, 왜), (don't, 않느가), (police, 경찰), (raid, 단속하지), (schoolyard, 학교), and (bullies, 깡패들). The word order in English is rearranged into “police, why, schoolyard, bullies, raid, don't” in Korean, and none of adjacent words in one language is adjacent in the other language. The example (8) also requires non-ITG word reordering for the word order in Korean “Senator, China, trip, on, what, achieve.”

Because newswire and Wikipedia rarely utilize wh-questions,<sup>3</sup> we only found this type of non-ITG word reordering in the English-Korean corpus. We found a total of 16 wh-question sentences requiring non-ITG word reordering among 1,693 sentences with no errors of word alignment. They account for 39% of 41 wh-questions in this corpus.

We found that this type of non-ITG word reordering would be recognized as a generalized pattern. Let AUX be an auxiliary verb, S a subject, V a verb, and VP/PP a verb/preposition phrase, respectively. The example of Table I can be generalized as a pattern (Wh- AUX S<sub>1</sub> V<sub>1</sub> S<sub>2</sub> V<sub>2</sub> / S<sub>1</sub> S<sub>2</sub> Wh- V<sub>2</sub> V<sub>1</sub>). The example (1) of Table IIIa shows a pattern (Wh- AUX S<sub>1</sub> VP, S<sub>1</sub> Wh- VP AUX), and (2) a pattern (Wh- AUX S<sub>1</sub> V<sub>1</sub> PP, S<sub>1</sub> PP Wh- V<sub>1</sub>), respectively.

**3.2.2. Adverbials.** Adverbials have a higher degree of freedom in word order than other constituents. “In reality, adverbials are very free in their placement, appearing in different positions in the sentence, not just sentence final” [Brinton 2000] in English. Generally, the relatively free order of adverbials is language universal. Therefore, adverbials often result in non-ITG word reordering phenomena during translation.

Some examples of adverbials are shown in Table IIIb. In (3), an adverb “主要/mostly” is located between the subject “脂肪/fat” and the main predicate “堆积/accumulate” in Chinese, while its corresponding word “주로/mostly” appears at the beginning in Korean. The example (4) also shows that an adverb phrase “in 1594” at the beginning of the English sentence is located between “として/as” and “伏見/Fushimi” in Japanese.

Regardless of domains and language pairs, adverbials incur non-ITG word reordering. After excluding the erroneous word alignment, there remains around 0.8% of sentences that need non-ITG word reordering because of adverbials. This type takes the smallest portion among those we found for each corpus as shown in Figure 2. We also found that this type of non-ITG word reordering can be generalized. For (3), a pattern

<sup>3</sup>We cannot find any wh-questions in the Chinese-Korean and English-Japanese corpora.

Table III.

Examples of non-ITG word reordering. English glosses are given below non-English sentences with grammatical markers such as TOP (topic), NOM (nominative), ACC (accusative), LOC (locative), or GEN (genitive).

(a) Wh-questions	
(1) English: <i>why don't the police raid the school-yard bullies?</i>	
Korean: 경찰은 왜 학교 강제들을 단속하지 않는가 ?	
Gloss: police TOP why school-yard bullies ACC raid not-do ?	
(2) English: <i>what did the Senator achieve on his trip to China?</i>	
Korean: 상원의원은 중국 여행에서 무엇 을 얻었는가 ?	
Gloss: senator TOP china trip on 에서 무엇 ACC 을 achieve-do ?	
(b) Adverbials	
(3) Chinese: 脂肪 主要在肠与肠之间	
Gloss: fat mostly accumulate at intestine and intestine between	
Korean: 주로 창자와 와 사이 에 지방 이 쌓인다	
Gloss: mostly intestine and intestine between LOC fat NOM accumulate	
(4) English: <i>in 1594, Hideyoshi began construction of Fushimi Castle for his residence</i>	
Japanese: 秀吉 は 居所 として 1594年 には 伏見 城 の 築城 を 始め ている	
Gloss: Hideyoshi TOP residence as 1594 in Fushimi castle GEN construct-castle ACC begin did	
(c) Specific examples of divergence	
(5) English: <i>he spent his life in Africa helping the poor.</i>	
Korean: 그는 아프리카 에서 가난한 사람들을 을 평생 도왔다	
Gloss: he NOM Africa LOC 에서 가난한 poor 사람들 people ACC ACC lifelong helped	
(6) English: <i>a unit called Hokoshu or Bugyoshu was organized to handle practical matters.</i>	
Japanese: 奉公衆 や 奉行衆 と 呼ばれる 実務 官僚 を 整備 した	
Gloss: Hokoshu or Bugyoshu as called handle-practical-matter 実務 bureauacracy 官僚 ACC organize did	
(7) English: <i>the construction of Jurakudai began in February 1586.</i>	
Japanese: 聚楽第 は 1586年 2月 に 着工 された	
Gloss: Jurakudai TOP 1586 February in construction-begin was	
(d) Others	
(8) English: <i>Fujiwara became one of the Emperor Uda's wives after his enthronement.</i>	
Japanese: 藤原 は 宇多 帝 即位 後に 入内 した	
Gloss: Fujiwara TOP Uda emperor enthronement after become-queen did	
(9) Chinese: 精制油 的 下调 率 是 调和油 的 下调 率 的 8%	
Gloss: refined-oil GEN GEN tariff-rate decrease to mixed-oil decrease to 8%	
Korean: 정제유 관세율은 은 혼합유 관세율 的 8% 인 하 된다	
Gloss: refined-oil tariff-rate TOP TOP 10% decrease to mixed-oil TOP 8% decrease to 8% respectively	

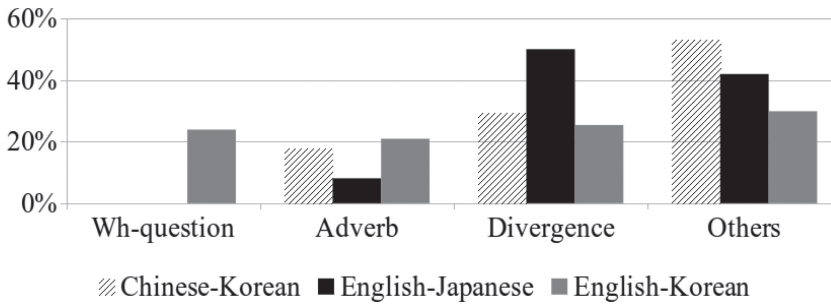


Fig. 2. The proportions of each type of non-ITG word reordering across the three corpora.

would be (S AdvP V PP / AdvP PP S V), and for (4), it is (AdvP S V PP / S PP AdvP V), where AdvP is an adverb phrase.

**3.2.3. Specific Examples of Divergence.** Translation divergences are a well-known problem in machine translation [Dorr 1994]. We found that certain specific examples of divergence, such as head-switching, structural, and categorial divergences, cause non-ITG word reordering. Head switching divergence refers to cases where the dependency relation between two constituents has switched after translation. For example, the example (5) of divergence in Table IIIc shows that the dependency relation between a main verb “spent” and its argument “helping” in English is switched in Korean.

The structural or categorial divergences refer to the syntactic structure or category changes during translation, respectively. For example, (6) in Table IIIc shows that the to-infinitive phrase “to handle practical matters” in English becomes the one-word noun phrase “実務/handle-practical-matters” in Japanese. The example (7) also shows that the argument “Jurakudai” of the preposition “of” in English becomes the subject “聚楽第/Jurakudai” in Japanese.<sup>4</sup>

In general, the percentage of non-ITG word reorderings caused by translation divergences is larger than by wh-questions and adverbial cases. Especially, in the English-Japanese corpus, translation divergences are the most significant reason for non-ITG word reordering. For the Chinese-Korean corpus, head-switching accounts for 40% and categorial 60%; For the English-Japanese corpus, head-switching 27.8%, structural 32.8%, and categorial 39.4%; For the English-Korean corpus, head-switching 23.5%, structural 35.3%, and categorial 41.2%, respectively. They show a wide variety of non-ITG word reorderings, which are hard to fully describe using simple patterns or rules.

**3.2.4. Others.** The fourth type of non-ITG reordering includes discontinuous translation units, natural translation, and so on. Here, discontinuous translation units mean that they have discontinuity in either the source or target language only. In such cases, discontinuous translation units in one language are translated into a word or adjacent words in the other language, so that correct word reordering of discontinuous translation units (DTUs) can improve machine translation. In Table IIIId, the example (8) shows this type of non-ITG word reordering. A translation units “became one of the Emperor’s ... wives” in English is translated into two adjacent words “入内/become-queen した/did”.

As opposed to literal translation, natural translation often drops some constituents during translation. In (9), the first main verb “下调/decrease” in the Chinese sentence is

<sup>4</sup>There is also a conflation divergence, that a lexical gap “construction” and “began” is translated in a Japanese word “着工/begin-construction”.

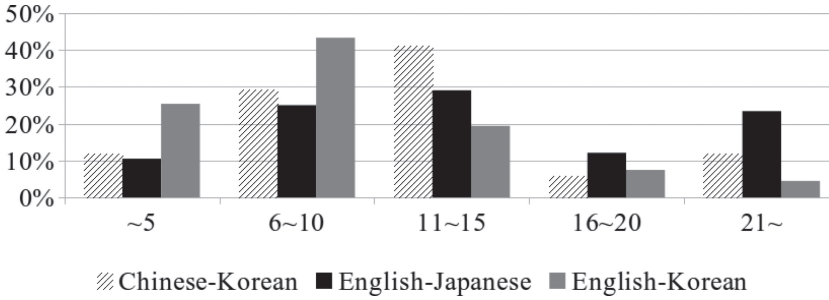


Fig. 3. The distribution of the size of non-ITG word reordering across the three corpora.

omitted in the Korean sentence. Hence, two Chinese words “下调/decrease” correspond to one Korean word “인하된다/decrease”, which causes non-ITG word reordering. In addition, there are difference in writing style between languages. For example, a list of pairs in one language is written in two separated lists in the other language as follows: “A is a, B b, and C c, respectively” in English “A, B, C 는/TOP 각각/respectively a, b, c 이다/is” in other language.

This type of non-ITG word reordering accounts for the largest percentages of the Chinese-Korean and English-Korean corpora. As the corpora have different characteristics, the causes for this type are different. For the Chinese-Korean corpus, the omission of counterparts in one language accounts for 66.6% and the different style of writing 33.3%; For English-Japanese corpus, the omission 63.5%, DTUs 23%, and the style 13.5%; For English-Korean corpus, the omission 44%, DTUs 44%, and the style 12%, respectively. Similar to the third type, this type depends on so specific lexical pattern that we cannot find any generalized/structural patterns.

### 3.3. Size of Non-ITG Word Sequence

We report the smallest size of the word sequence causing non-ITG word reordering, which is computed from failed ITG parsing results. After parsing, if a sequence of elements in the stack  $\sigma$  forms a continuous span in the target, we regard its corresponding word sequence in the source language as a non-ITG word reordering sequence. Many different sub-sequences in a sentence may require non-ITG word reordering. Among certain sequences, the minimum size is regarded to be equivalent to the size of non-ITG word reordering of the sentence. For example, the size of non-ITG word reordering of the example of Table I is 6, because the stack is [who, do, you, think, she, invited, ?] after the ITG parsing and “who, do, you, think, she, invited” is the minimum size of the word sequence forming a continuous span in the target sentence.

The histogram of the size of non-ITG word reordering is illustrated in Figure 3. The Chinese-Korean and English-Japanese corpora have longer sentences on average than English-Korean one, and the size of non-ITG word sequence from 11 to 15 accounts for the largest portion. For the English-Korean corpus, the largest portion ranges from 6 to 10 and the average is 9.25, which almost reaches the average sentence length of 11.92 for this corpus. This indicates that non-ITG word reordering is more likely to be global reordering than local one.

### 3.4. SCFG rank of non-ITG word reordering

It is also valuable to inspect the minimum SCFG rank to deal with non-ITG word reordering. The rank of an SCFG is the maximum number of the nonterminal symbols in the right hand side of the grammar. The expressive power of an SCFG with arbitrary



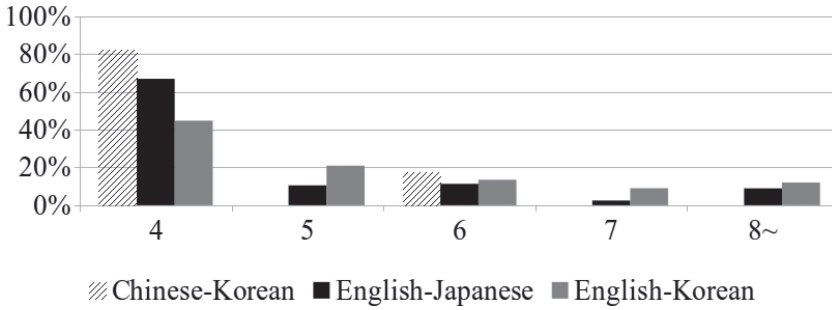


Fig. 4. The distribution of the minimum SCFG rank for non-ITG word reorderings across the three corpora.

rank, which generates all possible permutations for a given sequence, is generally much greater than that for ITG. The minimum rank of non-ITG word reordering is also computed from the failed ITG parsing result. After parsing, if a sequence of elements in  $\sigma$  forms a continuous span in the target, the number of elements in the sequence is the rank. Among certain sequences consisting of at least 4 elements, the minimum rank is regarded as the minimum SCFG rank of non-ITG word reordering of the sentence.

The distribution of the minimum rank is illustrated in Figure 4. Most non-ITG word reordering requires a rank-4 SCFG, while there exist long tails of larger ranks. Obviously, ITG constraints, which has an expressive power of a rank-2 SCFG, is insufficient because it misses a certain amount of non-ITG word reordering. In next section, we explore the possibility of transforming non-ITG word reordering to satisfy ITG constraints.

#### 4. TRANSFORMATION OF NON-ITG WORD REORDERING

We found many non-ITG word reordering cases between SVO and SOV languages. One major reason is the degree of freedom in word order. The SOV languages have a general tendency of having flexible word order, and they are morphologically-rich languages to indicate the roles of constituents. Adverbials are another source of freedom of word order regardless of languages.

In a head-final language with a flexible word order such as Korean or Japanese, nearly all constituents except the main predicate can be located anywhere before its head. The examples (2) to (4) are such cases that constituents except the main predicate can be scrambled. A simple rule for handling the freedom of word order is moving the *wh*-words and adverbials in front of the clause they belong to. This is a possible solution converting the word order without harming the meaning of the source sentence. In Table I, for example, the following two Korean sentences have the same meaning, but Korean<sub>1</sub> causes non-ITG word reordering, while Korean<sub>2</sub> does not. The brackets denote the clause boundaries.

*Converting the wh-question type.*

English: who do you think she invited?

Korean<sub>1</sub>: [너/you 는/TOP [그녀/she 가/NOM 누구/who 를/ACC 초대했다고/invite] 생각하니/think-do ?]

Korean<sub>2</sub>: [너/you 는/TOP [누구/who 를/ACC 그녀/she 가/NOM 초대했다고/invite] 생각하니/think-do ?]

For the English-Korean corpus,<sup>5</sup> we inspect the portion of acceptable conversions using the simple rule, and 30 sentences of 16 wh-questions and 14 adverbials types are indeed successful, accounting for 100%. This simple heuristic is also applicable to sentences including more than one adverbials as follows.<sup>6</sup>

*Converting the adverbial type.*

English: I really want to<sub>1</sub> speak to<sub>2</sub> her directly.

Korean<sub>1</sub>: [나/I 는/TOP [그녀/her 와/to<sub>2</sub> 반드시/really 직접/directly 통화-하/speak 고/to<sub>1</sub>] 싶다/want].

Korean<sub>2</sub>: [나/I 는/TOP [반드시/really 직접/directly 그녀/her 와/to<sub>2</sub> 통화-하/speak 고/to<sub>1</sub>] 싶다/want].

The third and fourth types of non-ITG word reordering would be transformed by rewriting the sentence. For (7) of Table IIIc, it is possible to make Japanese<sub>1</sub> as close to its literal translation as possible like Japanese<sub>2</sub>.

*Rewriting the divergence type.*

English: the construction of Jurakudai began in February 1586.

Japanese<sub>1</sub>: 聚楽第/Jurakudai は/TOP 1586年/1586 2月/February に/in 着工/construction-beginされた/was.

Japanese<sub>2</sub>: 聚楽第/Jurakudai の/GEN 建設/construction は/TOP 1586年/1586 2月/February に/in 始ま/beginった/was.

For (9) of Table III d, it is similarly possible to rewrite Korean<sub>1</sub> as Korean<sub>2</sub>, as follows.

*Rewriting the other type.*

Chinese: 精制油/refined-oil 的/GEN 关税/率/tariff-rate 下调/decrease 至/to 10%, 调和油/mixed-oil 下调/decrease 至/to 8%.

Korean<sub>1</sub>: 정제유/refined-oil 관세율/tariff-rate 은/TOP 10% 로/to, 혼합유/mixed-oil 는/TOP 8% 로/to 각각/respectively 인하된다/decrease.

Korean<sub>2</sub>: 정제유/refined-oil 관세율/tariff-rate 은/TOP 10% 로/to 인하되고/decrease-and, 혼합유/mixed-oil 는/TOP 8% 로/to 각각/respectively 인하된다/decrease.

Among 17 divergences and 20 other types in the English-Korean corpus, 20 sentences are successfully converted by moving words properly, and 17 sentences by rewording, which account for 54% and 46%, respectively. Although we cannot find any non-ITG sentences that are genuinely impossible to rewrite in a way that allows the ITG constraint. In practice, it is hard to manually correct all sentences in parallel corpora, and an automatic method of this process would be necessary as a future work.

## 5. CONCLUDING REMARKS

With three parallel corpora, we manually categorized the source of non-ITG word reordering between three SVO and SOV languages. We measured the size of non-ITG word reordering and the corresponding distribution of the minimum SCFG ranks. As a result, significant amount of non-ITG word reordering can be found. Although the portion of non-ITG word reordering in practice is much smaller than that in theory, it

<sup>5</sup>We focused on the English-Korean corpus because the numbers of the adverbial types in the Chinese-Korean and English-Japanese corpora are too small to inspect.

<sup>6</sup>It would be controversial whether the clause boundary include the shared subject “나/I 는/TOP” or not. Our analysis is based on the sentence structure as simple as possible, which is more like to be practical.

is nevertheless considerable. Therefore, ITG constraints is not adequate to the modelling of word reordering in machine translations in real situation. It remains difficult to correct sentences that violate ITG constraints automatically.

Dealing with non-ITG word reordering in an efficient manner is still a challenging problem. The search space beyond ITG constraints is much larger than that within ITG constraints, and “unconstrained reordering is only helpful if we are able to estimate the reordering probability reliably [Zens et al. 2004].” One possible solution is to improve the distortion model used in statistical machine translation [Al-Onaizan and Papineni 2006; Green et al. 2010; Goto et al. 2013]. With a large amount of distortion or unlimited distortion, these approaches achieved translation quality over a simple distance-based distortion model. Utilizing translation units with gaps (or discontinuous translation units) is also another approach [Crego and Yvon 2009; Galley and Manning 2010; Simard et al. 2005]. Some preordering approaches cast word reordering problem to a well-known mathematical problem such as the Travelling Salesman Problem [Khapra et al. 2013; Visweswariah et al. 2011; Zaslavskiy et al. 2009] or the Linear Ordering Problem [Tromble and Eisner 2009]. Their methods have potential to deal with non-ITG word reordering as well, though Tromble and Eisner [2009] restrict the search space with ITG constraints for efficiency. In future, we will deal with the problem of non-ITG word reordering to overcome the limitation of ITG constraints, to enhance the capabilities of word reordering, and in this way eventually to improve translation quality.

## REFERENCES

- Yaser Al-Onaizan and Kishore Papineni. 2006. Distortion models for statistical machine translation. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Stroudsburg, PA, 529–536.
- Laurel J. Brinton. 2000. *The Structure of Modern English*. Benjamins, John Publishing Co.
- Josep M. Crego and François Yvon. 2009. Gappy translation units under left-to-right SMT decoding. In *Proceedings of the 13th Annual Conference of the European Association for Machine Translation*. 66.
- Bonnie J. Dorr. 1994. Machine translation divergences: a formal description and proposed solution. *Comput. Linguist.* 20, 4, 597–633.
- Michel Galley and Christopher D. Manning. 2010. Accurate non-hierarchical phrase-based translation. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics (HLT'10)*. Association for Computational Linguistics, Stroudsburg, PA, 966–974.
- Qin Gao and Stephan Vogel. 2008. Parallel implementations of word alignment tool. In *Software Engineering, Testing, and Quality Assurance for Natural Language Processing (SETQA-NLP'08)*. Association for Computational Linguistics, Stroudsburg, PA, 49–57.
- Isao Goto, Masao Utiyama, Eiichiro Sumita, Akihiro Tamura, and Sadao Kurohashi. 2013. Distortion model considering rich context for statistical machine translation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 155–165.
- Spence Green, Michel Galley, and Christopher D Manning. 2010. Improved models of distortion cost for statistical machine translation. In *Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, 867–875.
- Liang Huang, Hao Zhang, Daniel Gildea, and Kevin Knight. 2009. Binarization of synchronous context-free grammars. *Comput. Linguist.* 35, 4, 559–595.
- Mitesh M. Khapra, Ananthakrishnan Ramanathan, and Karthik Visweswariah. 2013. Improving reordering performance using higher order and structural features. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, 315–324.
- Philipp Koehn, Hieu Hoang, Alexandra Birch, et al. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*. 2.

- Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*-Volume 1. Association for Computational Linguistics, 48–54.
- Graham Neubig. 2011. The Kyoto free translation task. <http://www.phontron.com/kftt>.
- Graham Neubig, Yosuke Nakata, and Shinsuke Mori. 2011. Pointwise prediction for robust, adaptable Japanese morphological analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, 529–533.
- Michel Simard, Nicola Cancedda, Bruno Cavestro, Marc Dymetman, Eric Gaussier, Cyril Goutte, Kenji Yamada, Philippe Langlais, and Arne Mauser. 2005. Translating with non-contiguous phrases. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing (HLT'05)*. Association for Computational Linguistics, Stroudsburg, PA, 755–762. DOI:<http://dx.doi.org/10.3115/1220575.1220670>.
- Anders Søgaard and Dekai Wu. 2009. Empirical lower bounds on translation unit error rate for the full class of inversion transduction grammars. In *Proceedings of the 11th International Conference on Parsing Technologies (IWPT '09)*. Association for Computational Linguistics, Stroudsburg, PA, 33–36.
- Roy Tromble and Jason Eisner. 2009. Learning linear ordering problems for better translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 1007–1016.
- Karthik Visweswariah, Rajakrishnan Rajkumar, Ankur Gandhe, Ananthkrishnan Ramanathan, and Jiri Navratil. 2011. A word reordering model for improved machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 486–496.
- Benjamin Wellington, Sonjia Waxmonsky, and I. Dan Melamed. 2006. Empirical lower bounds on the complexity of translational equivalence. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Stroudsburg, PA, 977–984.
- Dekai Wu. 1997. Stochastic inversion transduction grammars and bilingual parsing of parallel corpora. *Comput. Linguist.* 23, 3 (1997), 377–403.
- Dekai Wu, Marine Carpuat, and Yihai Shen. 2006. Inversion transduction grammar coverage of Arabic-English word alignment for tree-structured statistical machine translation. In *Proceedings of the IEEE/ACL Workshop on Spoken Language Technology*.
- Deyi Xiong, Min Zhang, Aiti Aw, and Haizhou Li. 2010. Linguistically annotated reordering: Evaluation and analysis. *Comput. Linguist.* 36, 3, 535–568. DOI:[http://dx.doi.org/10.1162/coli\\_a.00009](http://dx.doi.org/10.1162/coli_a.00009).
- Mikhail Zaslavskiy, Marc Dymetman, and Nicola Cancedda. 2009. Phrase-based statistical machine translation as a traveling salesman problem. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP (ACL'09)*. Association for Computational Linguistics, Stroudsburg, PA, 333–341.
- Richard Zens and Hermann Ney. 2003. A comparative study on reordering constraints in statistical machine translation. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1 (ACL'03)*. Association for Computational Linguistics, Stroudsburg, PA, 144–151.
- Richard Zens, Hermann Ney, Taro Watanabe, and Eiichiro Sumita. 2004. Reordering constraints for phrase-based statistical machine translation. In *Proceedings of the 20th International Conference on Computational Linguistics (COLING'04)*. Association for Computational Linguistics, Stroudsburg, PA, Article 205.

Received March 2014; revised June 2014; accepted July 2014