Annotation Guidelines for Chinese-Korean Word Alignment

Jin-Ji Li, Dong-Il Kim*, and Jong-Hyeok Lee

Department of Computer Science and Engineering,

Electrical and Computer Engineering Division,

Pohang University of Science and Technology (POSTECH),

San 31 Hyoja Dong, Pohang, 790-784, R. of Korea

E-mail: {ljj, jhlee}@postech.ac.kr

*Language Engineering Institute,

Department of Computer, Electron and Telecommunication Engineering,

Yanbian University of Science and Technology (YUST),

Yanji, Jilin, 133-000, P.R. of China

E-mail: {dongil}@ybust.edu.cn

Abstract

For a language pair such as Chinese and Korean that belong to entirely different language families in terms of typology and genealogy, finding the correspondences is quite obscure in word alignment. We present annotation guidelines for Chinese-Korean word alignment through contrastive analysis of morpho-syntactic encodings. We discuss the differences in verbal systems that cause most of linking obscurities in annotation process. Systematic comparison of verbal systems is conducted by analyzing morpho-syntactic encodings. The viewpoint of grammatical category allows us to define consistent and systematic instructions for linguistically distant languages such as Chinese and Korean. The scope of our guidelines is limited to the alignment between Chinese and Korean, but the instruction methods exemplified in this paper are also applicable in developing systematic and comprehensible alignment guidelines for other languages having such different linguistic phenomena.

1. Introduction

Word alignment is defined as an object indicating the correspondence between words in a parallel text (Brown et al., 1993), and usually it serves as an important source of knowledge for Statistical Machine Translation (SMT). However, the notion of "correspondence" between words is subjective (Och & Ney, 2003). For a language pair such as Chinese and Korean that belong to entirely different language families in terms of typology and genealogy, finding the correspondences is quite unclear in word alignment. Especially problematic is the difference in morpho-syntactic encodings of the two languages.

To achieve more objective, correct, and consistent evaluation results of word alignment, reasonable annotation guidelines are desired to resolve uncertain cases where correct counterparts of the languages are difficult to find. Relatively little research has been carried out on this issue, especially from the perspective of contrastive analysis of morpho-syntactic encodings.

There are several annotation guidelines for other languages such as Blinker project (Melamed, 1998), ARCADE project (Veronis & Langlai, 1999), PLUG project (Merkel, 1999), guidelines for Chinese-English word alignment of LDC, and (Lambert et al., 2005). However, these guidelines enumerate specific annotation rules classified by lexical categories such as Part of Speech (POS) of the source languages; these schemes cannot systematically describe linguistic phenomena occurring in morpho-syntactically distant language pairs such as Chinese and Korean.

This paper proposes annotation guidelines for Chinese-Korean word alignment. We analyze differences of morpho-syntactic encoding systems of Chinese and Korean. Korean is a typical agglutinative language, whose morphological form of verbs is much more complex than that of Chinese one. Most linking obscurities in word alignment between Chinese and Korean are caused by this difference.

We analyze the complex morpho-syntactic encoding system of Korean verbs to investigate the grammatical categories which the system conveys. Hence, the corresponding elements in Chinese are relatively easy to study. The perspective from grammatical categories provides a more comprehensible and general view for the alignment process. It improves the linking consistency in language pairs with highly different linguistic phenomena.

The scope of our guidelines is limited to the alignment between Chinese and Korean, but the instruction methods exemplified in this paper are also applicable in developing systematic alignment guidelines for other languages having such different linguistic phenomena.

We adopt Kappa statistic (Carletta, 1996) and perform a series of experiments to show the effectiveness of our proposed approach. The remainder of this paper is organized as follows. First, some general issues in annotation guidelines are mentioned in Section 2. Then we compare the linguistic differences between Chinese and Korean focusing on verbal systems in Section 3. Section 4 presents annotation guidelines for Chinese-Korean word alignment in detail. In Section 5

corpus data constructed for our experiments are described, Section 6 presents experimental results to show the effectiveness of our proposed guidelines. Finally, conclusion is given in Section 7.

2. Some issues in annotation guidelines

There are general alignment issues valid in most language pairs. We list general alignment instructions that are also reasonable in Chinese-Korean word alignment.

Two major rules for the word alignment are summarized well by Veronis & Langlais (1999).

1) Mark as many words as necessary on both the target and source side.
2) Mark as few words as possible on both the target and source side.

In general, parallel texts are translated non-literally. Hence, using only word-to-word links is not sufficient enough to contain all the information conveyed in the given sentence pair. Therefore, sometimes it is better to use a group of words as alignment units but making it as small as possible according to the above rules.

We allow S(ure) and P(ossible) links in our annotation guidelines (Och & Ney, 2003). S and P link mean unambiguous and ambiguous link respectively. There exist many ambiguities of manual alignment because of non-literal translation and systematic differences between the language pair. We think various alignments are acceptable for annotating P links and no need to reach an agreement on this.

Unlinked cases are also needed when the corresponding parts are 'not translated' in the target language. To judge the 'not translated' case, we adopt the judgment rule defined by Blinker project as follows: "when you can answer 'Yes' to the following question: If the seemingly extraneous words were simply deleted from their verse, would the two verses become more similar in meaning?"

Unlike above rules that generally apply to all languages, there exist some language pair-specific issues. Following section describes such characteristics of Chinese-Korean language pair and proposes guidelines utilizing the contrastive analysis of the given language pair.

3. Contrastive analysis of morpho-syntactic encodings between Chinese and Korean

This section provides the contrastive analysis of Korean-Chinese and presents the difficulties of organizing guidelines for linguistically distant language pairs.

3.1 General comparison

Chinese is a typical isolating language while Korean is a highly agglutinative one. The morphological form of Korean is much more complex than that of Chinese.

As an isolating language, it is generally true that in Chinese, each word consists of only one morpheme and cannot be further analyzed into component parts. In other words, Chinese has a very little morphological complexity (Li & Thompson, 1989). Grammatical functions are expressed by means of word order and some independent morphemes. Usually, an alignment unit in Chinese is a segmented word from a morphological analyzer.

For each *eojeol*[1] in Korean, it consists of one or more base forms (stem morphemes or content morphemes) and their inflections (function morphemes) that have a very productive inflectional system. Inflections usually include postpositions and verbal endings (verbal affixes) on verbs and adjectives. After morphological analysis has been performed, the basic unit in a given sentence is a Korean morpheme. We generally consider it as an alignment unit. Korean function morphemes occupy 41.3% of all Korean morphemes in our corpus mentioned in Section 5. Aligning Korean function morphemes to Chinese should be dealt very carefully because Chinese has a very poor system of function morphemes.

Postpositions in Korean are either aligned to the corresponding Chinese function words or null-aligned in case of nominative case markers because such linguistic function do not exist in Chinese; these morphemes are relatively easily managed. However, verbal endings in Korean inflect in diverse forms and are mapped to various linguistic phenomena in Chinese. Verbal endings consist of 40.3% of all Korean function words, and the average number of function morphemes inflected by a verb is 1.94 while that of a content morphemes is 0.7 implies that verbal endings causes uncertain alignment cases.

In our annotation study, we found that most uncertain links are induced by the cases mentioned above. Understanding the organization of Korean verb is crucial in maintaining consistency throughout the proposed guidelines.

3.2 Verbal phrase in Korean

The complex form of Korean verbs (verbal phrases)[2] frequently causes annotation ambiguities in Chinese-Korean word alignment. A verbal phrase in Korean consists of a series of verbal affixes along with a verb stem. A verb stem cannot be used by itself but should take at least one affix to form a verbal complex. Verbal affixes in Korean are ordered in a relative sequence within a verbal complex (Lee, 1991) and express various modality information[3] viz. tense, aspect, mood, negation, and voice as Figure 1 shows. These five grammatical categories are the major constituents of modal expression in Korean. Table 1 shows the modality encodings of Korean verbal phrases.

Order	Type
1	Verb Stem
2	Causative & Passive
3	Honorific
4	Aspect Tense *Modality*
5	Negation
6	*Modality* - Evidential
7	*Mood* - Illocutionary Force

[1] An *eojeol* refers to a fully inflected lexical form separated by a space in a sentence.

[2] 'Korean verb' or 'verbal phrase' in this paper refers to Korean predicates in a sentence.

[3] Modality system refers to five grammatical categories such as tense, aspect, mood (*modality & mood*), negation, and voice. The definition of these categories is described in (Li, 2005) in detail.

Table 1: Relative orderings of verbal affixes in Korean

1. 먹(stem)고_있(aspect)었(aspect)었(tense)다(*mood*)
(had been eating)

2. 잡(stem)히(passive)었(aspect)겠(*modality*)다(*mood*)
(may have been captured)

Figure 1: Verbal phrases in Korean

The prominence and correlations of modality system is different from language to language, and such difference increases the annotation ambiguity.

The modality of Korean is expressed intensively by verbal affixes of complex inflectional forms. However, as a typical isolating language, Chinese expresses modality using discontinuous morphemes around lexical verbs.

Modal expression in Korean is much more various than that of Chinese. Many-to-few assignment of modality expression causes linking obscurity in word alignment.

Languages, in general, do not give equal prominence to modality information. Chinese is an aspect- and topic-prominent language. As an aspect-prominent language, Chinese does not have grammatical markings for expressing tense. Unlike Korean, Chinese does not have a specific grammatical form in the voice system, which is natural for a typical topic-prominent language.

There exist some correlations among the grammatical categories, and such categories tend to share affixes for conveying modality information. For example, tense and aspect are interconnected as they both are involved with the 'temporal structure' of an event. In Korean, '었(eoss)' can be used as temporal and aspectual marker as Figure 1 shows.

Chinese has different ways of expressing modality: modal information is scattered throughout a sentence. We locate such modal expressing elements to provide correct candidate words.

4. Annotation guidelines for Chinese-Korean word alignment

In this section, we dedicate much space to explaining how Korean verbal phrases are linked to corresponding Chinese words because they are where most linking obscurities occur.

Since Korean is a verb-final language, identification of verbal phrases is much easier than Chinese. For efficiency, consistency, and accuracy, we propose an annotation principle that first judge Korean verbal phrases, then match the correspondent words in Chinese. The correspondences in Chinese are mainly composed of features used to display Chinese modality information.

Because of linguistic differences and liberal translations in parallel corpora, there exist phrasal correspondences and different link types: S-link, P-link, and not-translated. Explicit and unambiguous correspondences are S-linked and implicit correspondences are P-linked. As mentioned before, annotators may have disagreements on P-links.

4.1 Guidelines based on Korean verbal system

We propose special guidelines based on Korean verbal system as follows. To find the correspondences of Korean verbal phrases, we need to clarify the method for expressing modality information in Chinese. Such clarification can help link the verbal affixes in Korean because verbal affixes in Korean also convey modal expression.

We will give an explanation based on five grammatical categories such as tense, aspect, mood, negation, and voice, which, in Chinese, compose most of the modal expression.

4.1.1 Tense

Chinese does not have a grammatical category of tense, because the concept of tense is indicated by content words such as temporal adverbs, times[4], and auxiliary verbs. It is also inferred using aspect markers and attribute of main predicate (MP) by inspecting whether it is a motion verb of instantaneity or not. Table 2 shows the features which provide temporal information in Chinese.

Tense Marker	**Examples**
Time	明天(MingTian), 去年(QuNian), 下星期(XiaXingQi)
MP_motion verb of instantaneity	送(Song), 告诉(GaoSu)
Temporal adverb	将(Jiang), 将要(JiangYao), 已经(YiJing), 总是(ZongShi)
Aux. verb	会(Hui), 要(Yao)
Aspectual particle	了(Le), 着(Zhe), 过(Guo)

Table 2: Tense markers in Chinese

4.1.1.1 Time & MP-motion verb of instantaneity

Times and motion verbs of instantaneity indicate tense information and also their counterparts in Korean can be found readily. In Ex 2, the attribute of main verb '送(Song)' indicates past tense. In this case, we can simply link the counterparts together.

Ex 1.[5]
[cn] *明天(tomorrow)*/我(I)/*去(go)*/北京(Beijing)/。
[kr] **나**(I)+는 ***내일****(tomorrow)* **북경**(Beijing)+에 *가(go)+ㄹ* ***것****+ 이+ 다*
[en] I will go to Beijing tomorrow.

Ex 2.
[cn] 老王(Mr. Wang)/*送(give)*/我(me)/一(one)/本(Cls.)[6]/书(book)/。
[kr] **왕**(Wang)+**씨**(Mr.)+는 **나**(me)+에게 **책**(book)+을 ***선물****(give)+ 하+ 였+ 다*.
[en] Mr. Wang gave me a book as a present.

4.1.1.2 Temporal adverb

Some temporal adverbs such as '将(Jiang)' and '已经(YiJing)' in Chinese almost behave as function words since they only provide tense information. The actual translation is usually omitted in target sentence. Ex 3 shows this phenomenon. In this case, '已经(YiJing)' is

[4] Time is a category of Part of Speech in Chinese, which shows the temporal information.

[5] *Eojeols* are separated by a space. For each *eojeol*, bold-faced content morphemes followed by functional ones with + sign. Corresponding morphemes in each language are italicized and main predicates are underlined. Italicized morphemes in each language have high chances to be linked each other.

[6] Prt.: Particle; Prep.: Preposition; Cls.: Classifier;

glued to main verb '回家(HuiJia)' and linked to the verbal phrase in Korean.

Ex 3.
[cn] 他(he)/*已经(already)*/*回家(go home)*/ *了(Prt.)* /。
[kr] 그(he)+는 *집(home)+ 에 가(go)+ 았+ 다*.
[en] He went home.

Ex 4.
[cn] 我(I)/*将(soon)* /*去(go)*/北京(Beijing)/
[kr] 나(I)+ 는 북경(Beijing)+ 에 *가(go)+ ㄹ 것+ 이+ 다.*
[en] I will go to Beijing.

4.1.1.3 Aux. verb & Aspectual particle
Auxiliary verbs and aspectual particles are completely translated into verbal affixes in Korean. These two markers also convey the modal and aspectual information.

Ex 5.
[cn] *明天(tomorrow)*/*会(will)*/ *下(fall)*/雨(rain)/。
[kr] *내일*(tomorrow) 비(rain)+가 *오(fall)+ ㄹ 것+ 이+ 다*.
[en] It will rain tomorrow.

Ex 6.
[cn] 我(I)/*去(go)*/*过(Prt.)*/北京(Beijing)/。
[kr] 나(I)+는 북경(Beijing)+에 *가(go)+ 았+ 었+ 다*.
[en] I have been to Beijing.

4.1.2 Aspect
Chinese is recognized as an aspect prominent language with a complete set of markers to express aspectual distinctions. Conforming to the aspect system classified by (Xiao, 2002), we see that there are several types of aspect markers as Table 3 shows.

Aspect Marker	Examples
Aspectual Particle	了(Le)，着(Zhe)，过(Guo)
Adverbs	在(Zai)，正在(ZhengZai)，正(Zheng)，曾经(CengJing)，曾(Ceng)
Reduplication	笑一笑(Xiao)，看看(Kan)，讨论讨论(TaoLun)，过过瘾(GuoYin)，看了一看(Kan)
RVC[7]	(跳)下去(XiaQu)，(交)上来(ShangLai)，(携)起(手)来(QiLai)，(写)清楚(Qingchu)

Table 3: Aspect markers in Chinese

4.1.2.1 Aspectual particle & Adverb
As mentioned before, aspectual particle indicates temporal information as well as aspectual one. In Korean, tense and aspect also share verbal affixes to express temporal structures such as tense and aspect.

Ex 7.
[cn] 他(he)/*在(now)*/*写(do)*/作业(homework)/。
[kr] 그(he)+는 숙제(homework)+를 *하(do)+ 고 있+ 다*.
[en] He is doing homework.

Ex 8.
[cn] 我(I)/*曾(already)*/*去(go)*/*过(Prt.)*/北京(Beijing)/。
[kr] 나(I)+는 북경(Beijing)+에 *가(go) 보+ ㄴ 적+ 이 있+ 다*.
[en] I have been to Beijing.

4.1.2.2 Reduplication
Verb reduplication is an idiosyncratic linguistic form in Chinese. Some verbs can be reduplicated to convey delimitative aspect in a sentence. There are several formats for verb copying such as VV, V 了(Le)V, V 一(Yi)V and V 了(le)一(Yi)V.

Ex 9.
[cn] 给(prep.)/我(me)/*看*/*看(see)*/报纸(newspaper)/*吧*(Prt.)/。
[kr] 저(me)+에게 신문(newspaper) *좀 보(see)+ 여 주+ 세+ 요*.
[en] Let me see the newspaper, please.

Ex 10.
[cn] 我(I)/*看*/*了(Prt.)*/*看(read)*/报纸(newspaper)/。
[kr] 나(I)+는 신문(newspaper)+을 *보(read)+ 았+ 다*.
[en] I glanced at the newspaper.

4.1.2.3 RVC
RVCs not only convey the aspectual values, but also retain their original lexical meanings. Therefore, they can be translated into auxiliary predicates, as well as independent lexical verbs in Korean. In the latter case, we link RVCs to the correspondent verbs in Korean such as Ex 12. The RVC '清楚(QingChu)' is translated into an adverb "똑바로(ddok-ba-ro)" in Korean.

Ex 11.
[cn] 大家(everybody)/把(Prep.)/作业(homework)/*交(submit)*/*上来(RVC)*/。
[kr] 모두(everybody) 숙제(homework)+를 *내(submit) 주+ 세+ 요*.
[en] Everybody, submit your homework.

Ex 12.
[cn] *写(write)*/*清楚(clearly)*/你(your)/的(Prt.)/名字(name)/。
[kr] 당신(your)+의 이름(name)+을 *똑바로(clearly) 적(write)+ 어 주+ 세+ 요*.
[en] Please write down your name clearly.

4.1.3 Mood
Mood refers to a general linguistic term: a grammatical category signaling the expression of the speaker's attitude towards a proposition. It includes the concepts of both '*mood*' and '*modality*'.
Usually the category of *mood* is defined as a morphological verbal category which indicates the modal value of a sentence. It is usually expressed by inflection in most languages. In a broader category, it covers so-called sentence-moods. However, as an isolating language, *mood* system of Chinese is not expressed by verbal inflection.
'*Modality*' is expressed by various means of modal encoding ranging from lexical to highly grammaticalized ones. In particular, as an isolating language, Chinese mainly uses modal auxiliaries to express the *modalities*. The correlation between future tense and *modality* makes

[7] RVC is an acronym of Resultative Verb Complement like *open* in *push the door open* (Xiao, 2002).

it possible that future events also can be expressed temporally or modally. In fact, auxiliary verbs for future tense are developed from the modal auxiliaries in Chinese.

Mood Marker	Examples
Aux. verb	应该(YingGai), 能(Neng), 可以(KeYi), 必须(BiXu), 得(Dei)
Sentence-final particle	呢(Ne), 呀(Ya), 吗(Ma), 了(Le)

Table 4: Mood markers in Chinese

4.1.3.1 Auxiliary verbs

In some cases, the auxiliary verbs can translate into adverbs in Korean as well as indicate modal information. Such auxiliary verb should have links to both of the counterparts as in Ex 14.

Ex 13.
[cn] 你(you)/*应该(should)*/先(first)/*做(do)*/作业(homework)/。
[kr] **너**(you)+는 **먼저**(first) **숙제**(homework)+를 *하(do)+어야 하+ㄴ+다*.
[en] You should do your homework first.

Ex 14.
[cn] 你(you)/*必须(ought to)*/先(first)/*做(do)*/作业(homework)/。
[kr] **너**(you)+는 ***반드시(ought to)*** **먼저**(first) **숙제**(homework)+를 *하(do)+어야 하+ㄴ+다*.
[en] You ought to do your homework first.

4.1.3.2 Sentence-final particle

Sentence-final particle shows the information of sentence-type mood. In Korean, it is expressed by inflection of verbal affixes with respect to honorific information.

Ex 15.
[cn] 您(you)/*明天(tomorrow)*/*去(go)*/北京(Beijing)/*吗*(Prt.)/?
[kr] **당신**(you)+은 ***내일(tomorrow)*** **북경**(Beijing)+에 *가(go)+시+ㅂ니까*?
[en] Are you going to Beijing tomorrow?

4.1.4 Negation

The negation systems in Chinese and Korean are very similar. In general, there are standard negation, double negation, and imperative/propositive negation. There are four negative forms commonly use in Chinese: '不(Bu)', '别(Bie)', '没(Mei)', and '没有(MeiYou)'. The most general and neutral form of negation is '不(Bu)'.

There are also special negative formats in Chinese. One is a negative particle '不(Bu)' before an RVC and the other is '不了(BuLiao)/不得(BuDe)' after main predicates to show negative view. Besides these two formats, some negative particles such as '未能(WeiNeng)' and '别(Bie)' also show other modality information like aspect and mood.

Negation Marker	Examples
Negative particle	不(Bu), 没(有)(Mei(You)), 别(Bie), 未能(WeiNeng), 从未(CongWei), 甭(Beng)
MP_bu_RVC	(吃)不(Bu)(下去), (看)不(Bu)(过去)
MP_buliao/bude	(开)不了(BuLiao), (听)不了(BuLiao), (吃)不得(BuDe)

Table 5: Negation markers in Chinese

4.1.4.1 Negative particle

Although the main usage of negative particles is to show negative view in a sentence, it negates different modal situations in Chinese. For example, '没(有)(Mei(You))' negates completion of an event and '别(Bie)' is a negative imperative.

Ex 16.
[cn] 他(he)/*不(not)*/*在(Prt.)*/*学习(study)*/。
[kr] **그**(he)+는 ***공부(study)+하+지 않(not)+고 있+다.***
[en] He is not studying now.

Ex 17.
[cn] 我(I)/*没有(not)*/*吃(eat)*/饭(meal)。
[kr] **나**(I)+는 **밥**(meal)+을 ***먹(eat)+지 않(not)+았+다.***
[en] I did not eat a meal.

Ex 18.
[cn] *别(not)*/*让(Prt.)*/她(she)/*出去(go out)*/。
[kr] **그녀**(she)+가 ***나가(go out)+게 하+지 말+라.***
[en] Do not let her go out.

4.1.4.2 MP_bu_RVC & MP_buliao(bude)

These two formats not only indicate negation information, but also give *modality* information.

Ex 19.
[cn] 我(I)/实在(really)/是(be)/*吃(eat)*/*不(not)*/*下去*(RVC)/。
[kr] **나**(I)+는 **정말**(really) **더**(more) **이상**(over) ***먹(eat)+을 수 없+다.***
[en] I can not eat anymore.

Ex 20.
[cn] 我(I)/*听(hear)*/*不了(can not)*/音乐(music)/。
[kr] **나**(I)+는 **음악**(music)+을 ***듣(hear)+을 수 없+다.***
[en] I can not hear the music.

4.1.5 Voice

Generally there are two kinds of voice construction in Chinese: with, or without voice markers.

The typical passive marker is '被(Bei)', however the non-adversity usage of passive sentence makes it possible to express passive voice without any markers. Usually topic–comment structure in Chinese can function as a passive sentence as in Ex 23.

A variety of notional causative forms are adopted in Chinese to express the causative voice. Some RVC constructions in Chinese convey the causative meaning as in Ex 24 and Ex 25. The typical causative markers are '使(Shi)', '让(Rang)' and '叫(Jiao)'.

Generally, sentences without any passive/causative markers are more productive than sentences with markers in Chinese. Although they help to convey voice information using special constructions such as '被字(BeiZi)' phrases or '使字(ShiZi)' phrases, these voice markers do not directly provide voice information of the lexical verbs.

Voice Marker	Examples
Passive/Causative particle	被(Bei), 让(Rang), 使(Shi), 叫(Jiao), 令(Ling), 给(Gei)
Topic-comment construction	那 本 书 (Topic) 已 经 出 版 了 (Comment).
RVC_causative	(写)好(Hao), (搞)清楚(QingChu)
MP_causative	放沉(FangChen), 加强(JiaQiang), 弄醒(NongXing)

Table 6: Voice markers in Chinese

4.1.5.1 Passive/Causative marker

Ex 21.
[cn] 他(he)/*被(Prt.)*/老师(teacher)/*骂(scold)*/*了(Prt.)*/。
[kr] 그(he)+는 **선생님**(teacher)+께
야단(scold)+ 맞+ 았+ 다.
[en] He was scolded by the teacher.

Ex 22.
[cn] 这(this)/件(Cls.)/事(thing)/*使(Prt.)*/我(me)/非常(very)/*高兴(happy)*/。
[kr] **이**(this) **일**(thing)+은 **나**(me)+로 **하여금 매우**(very)
기쁘+ 게 하+ 였+ 다.
[en] This thing makes me very happy.

4.1.5.2 Topic-Comment construction

Ex 23.
[cn] *那(that)/本(Cls.)/书(book)/已经(already)/出版(publish)/了(Prt.)*/。
[kr] 그(that) **책**(book)+은 **이미**(already)
출판(publish)+ 되+ 었+ 다.
[en] That book has already been published.

4.1.5.3 RVC_causative

Ex 24.
[cn] 信(letter)/*写(write)*/*好(good)*/了(Prt.)/。
[kr] **편지**(letter)+를 **다**(all) *쓰(write)+ 었+ 다*.
[en] The letter was written.

Ex 25.
[cn] 把(Prep.)/问题(problem)/*搞(make)*/*清楚(clear)*/。
[kr] **문제**(problem)+를 *명확(clear)+ 하+ 게 하+ 라*.
[en] Make the problem clear.

4.1.5.4 MP_causative

Ex 26.
[cn] 噪音(noise)/*弄醒(wake)*/*了(Prt.)*/我(me)/。
[kr] **소음**(noise)+은 **나**(me)+를 *깨(wake)+ 게 하+ 였+ 다*.
[en] The noise made me wake up.

4.1.6 Other cases

There are two special constructions and both of them are not appropriate to be classified into the five grammatical categories we have discussed.

	Examples
Nominalization	看书*的*(De)
Separated verb	理(Li)(了)发(Fa), 上(Shang)(了)风(Feng)

Table 7: Exceptional cases in Chinese

4.1.6.1 Nominalization

In Korean, verbal phrase can be in a nominalization form. It is commonly related to the special construction of '的字(DeZi)' phrase in Chinese.

Ex 27.
[cn] 这(this)/篇/(Cls.)论文/(paper)/*是(be)*/我们(our)/*发表(publish)*/*过(Prt.)*/*的*(Prt.)/ 。
[kr] **이**(this) **논문**(paper)+은 **우리**(our)+가
발표(publish)+ 하+ 였+ 던 것+ 이+ 다.
[en] This paper was published by us.

Ex 28.
[cn] *看(read)*/书(book)/*的(Prt.)*/*是(be)*/我(I)/的(Prt.)/朋友(friend)/。
[kr] **책**(book)+을 *보(read)+ 는* ***이****(person)*+ 는
나(my)+의 *친구(friend)+ 이+ 다*.
[en] The person who is reading a book is my friend.

4.1.6.2 Separated verb

Chinese has a kind of verbs whose internal construction is a verb-object compound. The first constituent, like a verb in a sentence, can take aspect markers. Also, it can be separated by a measure phrase, modifiers of object constituents and so on.

Ex 29.
[cn] 他(he)/*昨天(yesterday)*/*理(cut)*/*了(Prt.)*/*发(hair)*/。
[kr] 그(he)+ 는 **어제**(yesterday) ***이발****(cut hair)*
+ 하+ 였+ 다.
[en] He had his hair cut yesterday.

Ex 30.
[cn] 他(he)/*昨天(yesterday)*/*理(cut)*/*了(Prt.)*/一(one)/次(Cls.)/*发(hair)*/。
[kr] 그(he)+ 는 **어제**(yesterday) ***이발****(cut hair)*
+ 하+ 였+ 다.
[en] He had his hair cut yesterday.

5. Corpus Data

We automatically collected and constructed a sentence-aligned parallel corpus from the DongA newspaper [8]. Strictly speaking, it is a non-literally translated Korean-to-Chinese corpus. The corpus consists of 101,226 sentence pairs and we randomly selected 50 sentence pairs as test data. The corpus profile is shown in Table 8.

	Chinese	Korean
# of sentences	50	50
# of words	1,323	1,502
# of singletons	741	645
Avg. length	26.5	30.4

Table 8: Statistics for test corpus

6. Experiment

Our aim is to examine the effectiveness of proposed guidelines. Usually it is measured by agreements between

[8] http://www.donga.com/news/ (Korean) and http://chinese.donga.com/gb/index.html (Chinese)

annotators with the same test corpus. We adopt the Kappa statistic to measure the agreements between annotators. Although our paper presents guidelines regarding verbal systems, the experiment is performed to evaluate the effectiveness of the whole annotation guidelines for Chinese-Korean word alignment. The experimental scenario is as follows:

1. Kappa value between two skilled annotators (A1 and A2) who are very familiar with the annotation guidelines;
2. Kappa values between each skilled annotator and a beginner (B) who was never involved in corpus annotation;
3. Kappa values between each skilled annotator and the beginner acquainted (B_acquainted) with the annotation guidelines;

Table 9 shows Kappa values according to our proposed experimental scenario.

	Kappa Value
A1 vs. A2	0.892
A1 vs. B	0.799
A2 vs. B	0.805
A1 vs. B_acquainted	0.858
A2 vs. B_acquainted	0.844

Table 9: Kappa values between annotators

The Kappa values between a beginner who is not familiar with the annotation process of Chinese-Korean and each skilled annotator is relatively low; some literature adopts assessment scales with Kappa values between 0.67 and 0.8 as only allowing tentative conclusion (Krippendorff, 1980). After acquainted with proposed guidelines, the Kappa values between the beginner and skilled annotators improves by about .05, in the range of definite conclusion of the assessment scale defined by (Krippendorff, 1980). We deduce that a novice annotator is able to achieve high agreements with skilled annotators with our suggested annotation guidelines

The improvement ratio of A1 vs. B and A1 vs. B_acquainted is greater than A2 vs. B and A2 vs. B_acquainted. B acquired the annotation guidelines through Question-and-Answering period with skilled annotator A1. We speculate that B could be influenced by the annotation style of annotator A1. This is fairly possible because many cases, especially regarding P links, are open to different interpretations according to the linguistic intuitions of annotators.

7. Conclusion

We presented annotation guidelines for Chinese-Korean word alignments through contrastive analysis of morpho-syntactic encodings. We discuss the differences in verbal systems that cause most linking obscurities in Chinese-Korean annotation process. Systematic comparison of verbal systems is conducted by analyzing morpho-syntactic encodings. Such approach from the viewpoint of grammatical category allowed us to define consistent and systematic instructions for linguistically distant languages such as Chinese and Korean. The proposed approach is also applicable to other language pairs with different morpho-syntactic encodings.

To validate the reliability of proposed guidelines, we adopted Kappa statistic. We achieved high Kappa value of 0.892 between two skilled annotators. 0.858 and 0.844 are also achieved between each skilled annotator and a beginner. Therefore, we believe the proposed guidelines produce consistent annotation results.

Word alignment and SMT system can also employ contrastive analysis of verbal system between Chinese and Korean, and our future work will focus on constructing a word alignment and SMT systems utilizing such analysis.

8. Acknowledgement

This work was supported in part by MKE & IITA through IT Leading R&D Support Project and also in part by the BK 21 Project in 2008.

9. References

Brown, P., Della Pietra, S., Della Pietra, V., and Mercer, R. (1993). The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2), 263--311.

Carletta (1996) Assessing agreement on classification tasks: the kappa statistics. *Computational Linguistics*, 22(2), 249--254.

Krippendorff, Klaus, (1980). Content Analysis: an Introduction to its Methodology. *Sage Publications*, Beverly Hills, CA.

Kruijff-Korbayova, Chvatalova, and Postolache (2005). Annotation Guidelines for Czech-English Word Alignment, *Proceedings of LREC 2006* (pp. 1256--1261). Genova.

Lambert, P., Gispert, DE A., Banchs, R., and Marino, B. J. (2005). Guidelines for Word Alignment Evaluation and Manual Alignment. *Language Resources and Evaluation*. 39(3), 267-285.

Lee, H.-S. (1991). Tense, aspect, and modality: A discourse-pragmatic analysis of verbal affixes in Korean from a typological perspective, PhD thesis, *Univ. of California*, Los Angeles.

Li, Charles N. and Thompson A. S. (1996). Mandarin Chinese: A functional reference grammar, *University of California Press*, USA.

Li, J.-J., Roh, J-E., Kim, D.-I., and Lee, J.-H. (2005). Contrastive Analysis and Feature Selection for Korean Modal Expression in Chinese-Korean Machine Translation System. *International Journal of Computer Processing of Oriental Languages*, 18(3), 227--242.

Melamed, I. D. (1998). Annotation Style Guide for the Blinker Project. IRCS Technical Report #98-06, *University of Pennsylvania*.

Merkel, M. (1999). Annotation Style Guide for the PLUG Link Annotator. Version 1.0, PLUG report, *Magnus Merkel Linkping university*.

Och, F. and Ney, H. (2003). A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1), 19--51.

Veronis, J. and Langlais, P. (1999). Evaluation of parallel text alignment systems, *In Parallel Text Processing* (ed. J. Veronis), Kluwer.

Xiao, R. Z. (2002). A corpus-based study of aspect in Mandarin Chinsese, PhD thesis, *University of Lancaster*.