TR－SLT－0057

Comparative Analysis of Chinese, Japanese and Korean Numeral Classifier

Tam Wai Lok　　Kyonghee Paik

January 27,2004

We will present our analysis of numeral classifiers extracted from Japanese, Korean, and Chinese corpora. We compare how numeral classifiers are matched with their referents in our corpora with the results produced by the algorithm given in Bond and Paik (2000) for generating classifiers using semantic classes from an ontology provided by Goi-Taikei. We also attempt at automatically analyzing the Japanese sentences containing classifiers by typing the classifiers contained following Bond (2001) and the syntactic construction following Asahioka et al (1990). We have identified some problematic constructions in Chinese and Japanese and point out the phenomenon that classifier types change in the course of translation. We have also shown that the anaphoric usage of numeral classifiers is problematic to machine translation. In conclusion, we point out the difficulty to predict the correct numeral classifiers to be used when translating between Chinese, Japanese and Korean as the domain covered by the same type of classifiers and the constructions containing numeral classifiers vary. For further work, we suggest analyzing classifier constructions using statistical model based on the data produced here and applying word sense disambiguation techniques to the referents.

（株）国際電気通信基礎技術研究所
音声言語コミュニケーション研究所
〒619－0288「けいはんな学研都市」光台二丁目2番地2 TEL：0774－95－1301
Advanced Telecommunication Research Institute International
Spoken Language Translation Research Laboratories
2-2-2 Hikaridai "Keihanna Science City" 619-0288,Japan
Telephone:+81-774-95-1301
Fax　　:+81-774-95-1308

1 Introduction

Chinese, Japanese and Korean have developed their own systems of classifiers. Japanese and Korean borrow some classifiers from Chinese. The three languages show some similarities and differences in the usage of classifiers. In this report, we examine how numeral classifiers are used in Chinese, Japanese, and Korean in our corpora and compare the data with the output of Bond and Paik (2000)'s algorithm for generating classifiers. We are interested to know if the corpus data is different from the results of generating classifiers using semantic classes from an ontology and what causes such difference. We hope this research will be useful for improving the quality of machine translation among the three languages.

We will start with giving an overview (Section 2) of our task, our data, and the programs we have written in the course of our work and specifying the goal of our work (Section 3). Section 4 is structured into four sub-sections (Section 4.1 to 4.4), each dedicated to one of the four sub-tasks mentioned in overview. In each of the sub-sections, we will start with explaining the method we use for completing our sub-tasks, such as extracting sentences including numeral classifiers or paring Japanese sentences containing classifiers with their Chinese and Korean translation. Also, we will give our algorithms and present the theoretical background and linguistic knowledge on which our programs are based. After presenting the algorithms and the linguistics knowledge that our algorithms are grounded in each of the sub-sections, we would present our results. Following the results, we will interpret the figures generated, describe the problems we have encountered and highlight their implications to machine translation. After going through all the sub-tasks, we will give our conclusion (Section 5) in which we will summarize the discussion made in Section 4, give an overall discussion of the results and suggest further work.

2 Overview

We break down our task of examining the use of classifiers in each of the three languages mentioned into several subtasks, namely, extraction, sorting, sentence pairing and referent-classifier pair comparison. For each of these subtasks, we write one or two programs to handle it.

We start with extracting Japanese sentences containing classifiers from the Japanese corpus and then provide an analysis of these sentences with the program, extraction.pl. The most important part of the analysis involves the extraction of the classifiers in them and the referents of the classifiers. The extracted sentences are sorted by the program, report_by_cla_type.pl and then every extracted sentence is paired with a Chinese sentence and two Korean sentences bearing the same meaning. Pairing of the Chinese sentences and the Japanese sentences is done by the program, match_cj.pl. The same program also extracts the classifiers from the Chinese sentences. The pairing of the Korean sentences and the Japanese sentences is done by another program, match_jk.pl. This program also extracts the classifiers from the Korean sentences. The Chinese, Japanese, and Korean classifiers used with each referent are finally compared with the results generated by Bond and Paik (2000) using semantic classes from the ontology provided by Goi-Taikei for the same referent. The comparison is done by the program, match_bp.pl.

Our data comes from four corpora: one Chinese corpus, one Japanese corpus and two Korean corpora. The Chinese corpus and one of the Korean corpora are translated from the same Japanese corpus. As there are two Korean corpora, we give each of them a name to distinguish them. This Korean corpus which is directly translated from the Japanese corpus is given the name: J-Korean. The other Korean corpus is made by translating an English corpus which were originally matched with the Japanese corpus. This latter Korean corpus is given the name: E-Korean.

3 Goal

The goal of the present study is to compare the referent-classifier pairs extracted from the corpora mentioned with the output of Bond and Paik (2000)'s algorithm for generating Chinese, Japanese and Korean classifiers based on the semantic classes of the referents. We hope that, not only the results generated, but also the discovery we make in the course of generating the results will give us insights on the issues we have to pay attention to when working on machine translation among the three languages.

4 Procedures and Findings

4.1 Extraction

The subtask of extraction in fact is not only about extracting Japanese sentences containing classifiers from the Japanese corpus. We have also attempt to provide an analysis of these Japanese sentences and prepare them for being processed by the following subtasks.

4.1.1 Extracting Japanese sentences containing classifiers

4.1.1.1 Method

As we have a tagged and segmented Japanese corpus, we can achieve this task by extracting sentences in which one of the words contained is tagged as *助数詞* (classifier). The default classifier *つ*, together with the numeral preceding it, is tagged as *名詞* (noun). So we have to match the orthography of every single word against a list of numeral-classifier-combination from *一つ* to *九つ*. Not all sentences containing the matches are extracted. Morphemes tagged as *助数詞* but used with ordinal numbers are filtered away. These morphemes include *号*, *丁目*, *ゲート*, *等*, *位*, *番*, *両目*, *時*, *月* and *年次*. Classifiers whose referents are always omitted are also filtered away. These classifiers include *食*, *コース*, *カ国*, *か国語*, *カ国語*, *ホール*, *階*, *チャンネル* and *段階*. We also ignore phrases which we doubt whether it is appropriate to be tagged as classifiers. Two examples of these words are *割引* and *年来*. *一割引き* in 1. is tagged as [一]$_{NUM}$[割引き]$_{CL}$ but it should be tagged as [一]$_{NUM}$ [割]$_{CL}$[引き]$_{AFF}$.

1. 一割引きなら、買うんですけど。
2. 彼は私の十年来の友人です。

As for the phrase *年来* in 2, it is tagged as [*年来*]$_{CL}$, but it should be tagged as [*年*]$_{CL}$[*来*]$_{AFF}$.

Two more classifiers are also filtered away. These two classifiers are *倍* and

重. We agree that they are classifiers because both of them can postfix to numerals and form quantifier phrases with numerals, as illustrated by the following sentence taken from our corpus:

3.

約	[[二]$_{NUM}$	[倍]$_{CL}$]$_{QP}$	[の]$_{ADN}$	[時間]$_{N}$	が	かかる	と	思います
About	2	fold	ADN	Time	SUB	spend	COMP	Think

It would spend about two times the normal time

倍 can be classified as a measure classifier (Bond, 2001) and *重* an arrangement classifier, (Bond and Paik, 2004). But *倍* is a bit different from the other measure classifiers in that it selects for some original amount of the referent and measures the referent in terms of this amount. Other measure classifiers simply measure the referent in terms of an arbitrary amount. *重* is also a bit different from the other arrangement classifiers in that it selects for the pattern in which the referent is arranged and says that it is arranged in the same pattern for a certain number of times. For other arrangement classifiers, the pattern in which the referent is arranged is inherent in the semantics of the classifiers. We have yet to decide on whether or not to classify these two special classifiers like the other measure classifiers and arrangement classifiers respectively. We do not include them for our analysis.

The morpheme *次*, which can be used as a classifier in phrases like [二]$_{NUM}$[次]$_{CL}$[会]$_{N}$, (This phrase is not found in any of the corpora we use) is filtered away as we found out that it is not used as a classifier in any of the sentence extracted from the Japanese corpus and containing the morpheme. Including it would only make our counts less accurate.

Figure 1 gives the pseudo code for our algorithm:

Figure 1 Algorithm for extracting sentences including classifiers

(1). For a noun found in a sentence

(a) If the noun contains a Japanese numeral less than 10 and the default classifier つ, extract the sentence

(2). For a classifier found in a sentence

(a) if the classifier is not one of the morphemes which are used with ordinal numbers,
and the classifier is not one of the anaphoric classifiers,
and the classifier is not one of the morphemes mistagged as classifiers,
and the classifier is not one of the untyped classifiers:
extract the sentence.

4.1.1.2 Results

In total, 10530 sentences containing classifiers are extracted from the Japanese corpus.

4.1.2 Analyzing Japanese sentences containing classifiers

4.1.2.1 Method

The extracted sentences are then matched against seven syntactic patterns, six of which are found in Asahioka et at (1990). These patterns are given in table 1.

Table 1 Major patterns of classifier usages

Type of Structure	Pattern
Prenominal	(numeral)[classifier][adnominal particle][noun]
Floating	[noun][case particle]/[topic marker] [numeral] [classifier]
Partitive	[noun][adnominal particle][numeral][classifier]
Predicative	[noun][topic marker][numeral][classifier][copular]
Verb-modifying	(numeral)[classifier][verb]
Appositive	[noun][numeral][classifier]
Attributive	(numeral)[classifier][noun]
[ ] denotes a constituent. The underlined words are the referents. ( ) denotes a constituent which is part of the numeral-classifier combination but not used for pattern-matching in the program.	

Here are example sentences for each of the given patterns:

4. そして一つの人形を[三]$_{num}$[人]$_{CL}$[の]$_{ADN}$[人形使い]$_{N}$があやつります。(Prenominal)
5. この[薬]$_{N}$[を]$_{OBJ}$[一]$_{NUM}$[錠]$_{CLA}$水で飲んでください。(Floating)
6. [日本人]$_{N}$[の]$_{ADN}$[四]$_{NUM}$[人]$_{CL}$に一人は六十五歳以上です。(Partitive)
7. [定員]$_{N}$[は]$_{TOP}$[何]$_{NUM}$[名]$_{CL}$[です]$_{COP}$か。(Predicative)
8. もう[一]$_{NUM}$[度]$_{CL}$[言っ]$_{V}$てください。(Verb-modifying)
9. この[薬]$_{N}$[一]$_{NUM}$[錠]$_{CL}$を水で朝飲んでください。(Appositive)
10. [三]$_{NUM}$[人]$_{CL}$[部屋]$_{N}$をお願いします。(Attributive)

Following Downing (1996), we only allow floating from subjects, marked by *が*, and direct objects, marked by *を*. This rules out an analysis of *船便*, which is marked by で in the following sentence as the referent of the classifier *日*:

11. [船便]$_{N}$ [で]$_{CASE}$ [何]$_{NUM}$ [日]$_{CL}$ ぐらい かかり ます か 。

When matching sentences to the appositive construction, we would avoid taking classifier, which belongs to a subclass of noun, as the referent of another classifier. This makes it possible to analyze sentences like 12 without taking *ドル* as the referent of *セント*:

12. [四十五]$_{NUM}$[ドル]$_{CL}$[五十]$_{NUM}$[セント]$_{CL}$になります。

A number of affixes tagged wrongly as nouns are also avoided when extracting the referent of a classifier. Most of these words share the quality of serving function like English preposition semantically. They include *以上, 以下, 以内, 以外, 以降, 以来, 以前, 前, 後, 手前, 奥, あたり, ごろ, 用, 間, 向こう, 掛け* and *続き*. To illustrate, we have paired example sentences containing these affixes with English translation of each of them.

13a. [十]$_{NUM}$[分]$_{CL}$[以上]$_{AFF}$です
13b. It is [above]$_{P}$ [10]$_{QP}$ [minutes]$_{N}$.

14a. 予算は[千]$_{NUM}$[ドル]$_{CL}$[以下]$_{AFF}$です
14b. The budget is [below] $_{P}$[1000]$_{QP}$ [dollar]$_{N}$.

15a. 列車は[三十]$_{NUM}$[秒]$_{CL}$[以内]$_{AFF}$に出発します。
15b. The train will depart [within]$_{P}$ [30]$_{QP}$ [seconds]$_{N}$.

16a. [三]$_{NUM}$[軒]$_{CL}$[手前]$_{AFF}$です。
16b. It is [three]$_{QP}$ [blocks]$_{N}$ [ahead]$_{ADJ}$.

17a. [一]$_{NUM}$[日]$_{CL}$[あたり]$_{AFF}$の料金はいくらですか。
17b. How much is the fare [for]$_{P}$ [a]$_{QP}$ [day]$_{N}$?

18a. [四]$_{NUM}$[人]$_{CL}$[用]$_{AFF}$のテーブルをお願いします。
18b. I would like to book a table [for]$_{P}$ [four]$_{QP}$?

19a. [四]$_{NUM}$[両]$_{CL}$[向こう]$_{AFF}$です。
19b. It is[four]$_{QP}$ [carriages]$_{N}$ [ahead]$_{ADJ}$.

20a. あいにく込んでいて、[三]$_{NUM}$[人]$_{CL}$[掛け]$_{AFF}$の席しか取れませんが。
20b. Unfortunately, it is very crowded. I can only get [3]$_{QP}$ [seats]$_{N}$ [in a row]$_{PP}$.

21a. [四]$_{NUM}$[人]$_{CL}$[続き]$_{AFF}$の席がいいのですが。
21b. How about [four]$_{QP}$ [seats]$_{N}$ [in a row]$_{PP}$?

A sentence may contain more than one classifier. This means that an extracted sentence can be matched to more than one of the seven patterns as shown in table 1. For each of the classifier contained in an extracted sentence, any word in a position of no more than four words away from the concerned classifier is matched against any of the syntactic category which forms part of one of the constructions given in table 1. When a sequence of words matches any of the patterns, the sentence containing that sequence is extracted and the noun in that sequence is identified as the referent of the classifier contained in the same sequence, except in the case of the verb-modifying construction. In that case, the verb is extracted as the referent of the classifier. For a sequence of words to be matched to the verb-modifying construction, the classifier in that sequence has to be one of the few classifiers typed as event classifiers. We will describe how classifiers are typed later. If a classifier and all the words in positions of no more than four words away from it do not form a sequence that matches any of the patterns given in table 1, the classifier is considered to be used anaphorically, which means that its referent is omitted. No referent is identified in such case.

Our algorithm is given as follows:

Figure 2 Algorithm for matching sentences to major patterns

(1). For every classifier found in a sentence

(a) If the classifier is a event classifier,
and the word immediately following it is tagged as a verb:
extract the verb as the referent of the classifier,
assign to the sentence "verb-modifying" as the type of its construction.

(b) Else if the word immediately following the classifier is the adnominal particle,
and the word immediately following the adnominal marker is tagged as a noun:
extract the noun as the referent of the classifier.
assign to the sentence "prenominal" as the type of its construction.

(c) Else if the word immediately preceding the numeral is the topic marker,
and the word immediately preceding the topic marker is tagged as a

noun,
and the word immediately following the classifier is tagged as a copula:
extract the noun as the referent of the classifier.
assign to the sentence "predicative" as the type of its construction.
(d) Else if the word immediately preceding the numeral is the subject marker, the direct object marker or the topic marker,
and the word immediately preceding the subject marker, the direct object marker or the topic marker is tagged as a noun:
extract the noun as the referent of the classifier.
assign to the sentence "floating" as the type of its construction.
(e) Else if the word immediately following the classifier is tagged as a noun,
and the noun is not among the following list of words:*以上, 以下, 以内, 以外, 以降, 以来, 以前, 前, 後, 手前, 奥, あたり, ごろ, 用, 間, 向こう, 掛け,続き*:
extract the noun as the referent of the classifier.
assign to the sentence "attributive" as the type of its construction
(f) Else if the word immediately preceding the classifier is tagged as a noun,
and the noun is not tagged as a classifier at the same time:
extract the noun as the referent of the classifier.
assign to the sentence "appositive" as the type of its construction.
(g) Else if the word immediately preceding the numeral is the adnominal marker,
and the word immediately preceding the adnominal marker is tagged as a noun:
extract the noun as the referent of the classifier.
assign to the sentence "partitive" as the type of its construction.
(h) Else assign to the sentence "anaphoric" as the type of its construction.

Next, we examine how the classifiers found are typed. Based on the typology proposed by Bond (2001) and a slightly revised version of it in Bond and Paik (2004), we come up with eleven types: default, kind, shape, taxonomic, event, measure, group, container, arrangement, portion and temporal. The property of each type of classifiers and examples of classifiers found in the

Japanese corpus under each type are given in table 2:

Table 2 Examples and properties of classifiers of different types

DEFAULT	Property	● Can substitute all classifiers ● Possible with group reference and individual reference
	Example	つ
KIND	Property	● Select for targets' type ● Force individual reference
	Example	人、便、部屋、錠、席、尾、幕、曲、台、羽、名、通、冊、匹、艘、竿、軒、株、問、校、ゲーム、ヶ所、ヵ所、室、客、ページ、字、条、語、部、項目、足、両、 着、シート、アイテム、品、項、厘、区間
SHAPE	Property	Select physical properties of the target
	Example	個、本、面、枚、粒、巻、輪、段
TAXONOMIC	Property	Force subspecies reading
	Example	種類、種、類、通り
EVENT	Property	Quantify occurrences of an event
	Example	回、発、度、ラウンド、周、打、鞍、泊
MEASURE	Property	Measure a quality of the referent
	Example	メートル、平方メートル、キロメートル、平方キロ、センチ、ミリ、キログラム、グラム、キロ、リットル、シーシー、カロリー、ボルト、パウンド、ポンド、オンス、ガロン、クオート、パイント、マイル、ヤード、フィート、インチ、平方マイル、エーカー、ノット、カラット、金、ペニー、米ドル、香港ドル、セント、ドル、オーストラリアドル、‘ユーエスドル、円、銭、文、マルク、フラン、ペンス、カナダドル、斤、歩、駅、ユニット、丁、区画、回り、まわり、ブロック、畳
GROUP	Property	Make the referent represent a group
	Example	セット、群、組、クラス、ダース
CONTAINER	Property	Give information about container

	Example	缶、カートン、皿、箱、パック、パイ、びん、ビン、袋、ぱい、カプセル、瓶、杯、ケース
ARRANGEMENT	Property	Give information about how members are arranged
	Example	行、列
PORTION	Property	Divide the referent into portions
	Example	切れ、目盛、割、パーセント、ピース、厘
TEMPORAL	Property	Temporal expression
	Example	世紀、代、年、年間、歳、才、ヶ月、ケ月、か月、ヵ月、カ月、週間、週、日、晩、時間、分、秒

Although many classifiers given above can be assigned to more than one type or tagged wrongly as a classifier when it is not used as a classifier, we eventually assign only one type to each classifier. Classifiers that can be assigned to more than one type are given in table 3. The rightmost column gives the proportion of the usage of a classifier as the selected type to all usages.

Table 3 Classifiers that can be assigned to more than one types

CLASSIFIER	SELECTED TYPE	ALTERNATIVE TYPE	PROPORTION
着	Kind	Ordinal	1/1
部	Kind	Portion	15/16
ケース	Container	Kind	2/2
ユニット	Measure	Group	7/8
両	Kind	Measure	2/2
ブロック	Measure	Group	74/74
分	Temporal	Ordinal	607/969
丁	Measure	Kind	1/1
度	Event	Measure	306/348
点	Kind	Non-Classifier	17/37

Table 4 gives some examples of sentences which show how a classifier may be used in different ways.

Table 4 Examples of uses of multi-type classifiers

CLASSIFIER	TYPE	EXAMPLE
着	Kind	22. シャツが一着戻ってきていません。
着	Ordinal	23. 一着になると思う馬を選択して希望購入金額を入力して購入ボタンを押してください。(Not in the Corpus)
部	Kind	24. プログラムを一部ください。
部	Portion	25. 石川県外部監査契約に基づく監査に関する条 例の一部を改正する条例をここに公布する。(Not in the Corpus)
ケース	Kind	26. わたしは、近代日本は精神分裂病の一ケース、アメリカは強迫神経症の一ケースと見れば、それぞれの国家としての行動がよく理解できると考えている。(Not in the Corpus)
ケース	Container	27. 一日六百ケースぐらいの包装をしています。
ユニット	Measure	28. 百ユニット当たりの価格をお知りになりたいのですね。
ユニット	Group	29. 上下二段の寝台が両側にあって四寝台が一ユニットになっています 。
両	Kind	30. 三両前です。
両	Measure	31. 一両にこだわって一〇〇両に笑う。(Not in the Corpus)
ブロック	Group	32. 大会はＡとＢの二ブロックに分けられた勝ち抜き戦である。(Not in the Corpus)
ブロック	Measure	33. 二ブロック戻ってください 。
分	Temporal	34. 到着予定より三十分遅れます。
分	Portion	35. ここは二分咲きでした。.(Not in the Corpus)
丁	Measure	36. 二丁ばかり先です。
丁	Kind	37. 私は二丁拳銃を持って走るというシーンがあるんですけど、実はそれは台本には載ってなかったんですよ。(Not in the Corpus)

度	Measure	38. 水温は何度ですか。
度	Event	39. 子どもがふたりいるけど週に一度しか洗濯できないでしょ。
点	Kind	40. 私の批判は次の三点に絞られます。(Not in the Corpus)
点	Non-classifier	41. 一点五ボルトの乾電池を 四個ください。

With all the classifiers given in table 4 being assigned to only one type by the program, it is inevitable for some of them to be assigned to a wrong type when used in some sentences like sentence 38 (in table 4). In sentence 38, the classifier *度* is used as a measure classifier for measuring temperature. But our program would take it as an event classifier, as it is used in the case of sentence 39 (in Table 4). Such cases of mistyping are, however, minimized by selecting the more frequently used type, except in the case of *点*. We do not assign *点* to the more frequently used type, that is, non-classifier because doing so would filter away all sentences containing the classifier

4.1.2.2 Results

The number of sentences containing each type of classifiers, together with the percentage of every type among all extracted Japanese sentences, is given in table 5.

Table 5 Number and percentage of sentences according to types

TYPE	NUMBER OF SENTENCES	PERCENTAGE AMONG ALL SENTENCES
DEFAULT	626	5.9%
KIND	698	6.6%
SHAPE	1479	14%
TAXONOMIC	21	0.2%
EVENT	831	7.9%
MEASURE	2796	26.6%
GROUP	15	0.1%
CONTAINER	333	3.2%

ARRANGEMENT	6	0.05%
PORTION	284	2.7%
TEMPORAL	3441	32.7%
TOTAL	10530	100%

Forcing one type on each classifier, instead of assigning more than one type to a classifier, actually enables us to achieve a satisfactory level of accuracy in typing. We have verified all the 698 sentences marked by our program as containing kind classifiers and we find that the classifiers contained in 604 of them are used as the right type. This gives a 86.5% accuracy. While verifying the result, we discover that it fails to get the right type in the following cases:

i. Words that may be used as classifiers but found to form part of a numeral expression which shows the order of an object in a sequence; (An example of such word is *便* in "*ジャル三零便*" *"meaning "JAL 30"*)
ii. Classifiers that may be used as a kind classifier but found to be used as another type of classifier or not even used as a classifier in the concerned sentences; (An example of such word is *点* in "*一点五ボルト*" *meaning 1.5 volt"*)

With these sentences eliminated, we evaluated the 604 sentences left and found that 76.8% of our analysis is correct. Without taking away sentences containing words described in i. and ii above, the accuracy drops to 66.5%. We have not made any further attempt to repeat the process of verification with sentences containing other type of classifiers.

4.1.2.3 Discussion

There are four major sources of errors. Some errors come from misidentifying temporal phrases as referents of classifiers. Here is an example which is analyzed inappropriately:

42. [[[[今晩]$_{NP}$[は]$_{TOP}$]$_{NP}$[[四人]$_{QP}$]$_{NP}$[です]$_{COP}$]$_{S}$...

Our program would identify the word *今晩* as the referent of the classifier *人*. There is no difference in syntactic structure between the sentence 42 and 43.

In 43, the topic marked noun phrase *荷物* is the referent of the classifier *個*.

43. [[[荷物]$_{NP}$ [は]$_{TOP}$]$_{NP}$[[三個]$_{QP}$]$_{NP}$[です]$_{COP}$]$_{S}$…

Our program would also identify any verbal noun adjacent to a classifier as its referent even if the classifier has no referent. For example, in example sentence 44, where the classifier *人* is used with no referent such that it forms a noun phrase with the numeral *三* preceding it, the verbal noun *予約* is identified as the referent of the classifier *人*.

44. [三]$_{NUM}$[人]$_{CL}$予約をお願い します

It makes no sense for *予約* (booking) to be counted by *人*(person). Hence, we get the wrong referent.

The possessive construction also creates difficulties for us. Sharing the same sequence of [NP] [ADN] [NP] with the prenominal construction given in table 1, the possessive construction is often confused with the prenominal construction in table 1. Our program would identify the noun *席* (seat) in sentence 45 as the referent of the classifier *人* (person) although it makes no sense for the noun *席* (seat) to be counted by the classifier *人* (person).

45. [四]$_{NUM}$[人]$_{CL}$の席はありますか

A parse that makes sense would treat the numeral-classifier combination *四人* (four person) as a noun phrase by itself and treat the noun *席* (seat) as being possessed by *四人*(four person).

Nouns pre-modified by adjectival nouns also cause problems to our program. Consider the following sentence.

46. 日本人 観光客 は 必ず この 海岸 と [二 千 五 百]$_{NUM}$ [羽]$_{CL}$[の]$_{ADN}$ [きれい]$_{ADJN}$ [な]$_{ADN}$ [鳥]$_{N}$ [の]$_{ADN}$ [飼育地]$_{N}$ で 知ら れる カルビアン 野鳥 保護区 を 訪れ ます

Our program would identify the adjectival noun *きれい* as the referent of the

classifier 匹, giving the following incorrect analysis:

[[二千五百]$_{NUM}$[匹]$_{CL}$[の]$_{ADN}$ [きれい]$_{ADJN}$]$_{NP}$.

The correct analysis would be:

[二千五百]$_{NUM}$[匹]$_{CL}$[の] $_{ADN}$[きれいな鳥]$_{NP}$]$_{NP}$.

The noun phrase きれいな鳥 (beautiful birds) should be the referent of the classifier 匹 (animal).

4.2 Sorting

The subtask of sorting is relatively straightforward. Every sentence taken from the Japanese corpus and processed by the extraction program is printed out together with its ID number, the classifier contained in it, the type of the classifier, the referent of the classifier, the type of its sentence construction and the head verb. What the sorting program does is to sort the sentences by the type of the classifiers contained. After being processed by the program, every sentence is then saved to a file bearing the name of the type of the classifier contained by it. Table 6 gives the sample output of the sorting program, which is taken as input by the pairing programs to be described in the next section.

Table 6 Sample output from sorting

ID	S	SENTENCE	N	C	T	REFERENT	CON
000104300	0	このチケットを一枚ください	一	枚	shape	チケット	Floating
S: sub-id N: numeral C: Classifier T: type of the classifier contained CON: sentence construction							

4.3 Sentence pairing

The subtask of sentence pairing is about paring each of the Japanese sentences extracted with its Chinese and Korean translation. We have also attempt to extract classifiers from the Korean and Chinese sentences found such that we can compare the classifiers used in these languages for conveying the same meaning. As it is relatively straightforward to pair sentences from the Korean and the Chinese corpora with the extracted Japanese sentences, we would focus our discussion on how to extract classifiers from the Korean sentences or Chinese sentence paired with a Japanese sentence containing a classifier.

4.3.1 Extraction of Korean Numeral Classifiers

4.3.1.1 Method

We start with aligning sentences bearing the same ID number from the single Japanese corpus and the two Korean corpora. Then we use tags such as NNC (cardinal number), NBU (unit-bounded noun) and NCN (common noun) provided by our parser to find out possible numeral-classifier combinations contained in these sentences. Although a classifier, if tagged correctly, would be a unit-bounded noun, the parser does not always analyze the sentences correctly. For this reason, we extract every word tagged as a common noun or a unit-bounded noun and preceded by a cardinal number.

4.3.1.2 Results

We assume that the classifier used for any Korean sentence has the same type as the classifier used in its Japanese source text. Table 7 shows how many kind classifiers are extracted from J-Korean and E-Korean.

Table 7 Kind classifiers in J-Korean and E-Korean

	KIND CLASSIFIERS IN KOREAN SENTENCES
J-Korean	499
E-Korean	477

The number of Korean sentences which are translation of Japanese sentences containing kind classifiers is the same as the number of Japanese sentence containing kind classifiers, 698. Table 8 tells us the percentage of sentences from which we have successfully extracted classifiers among all Korean sentences that are translation of Japanese sentences containing kind classifiers.

Table 8 Successful rate of kind classifier extraction

	SUCCESSFUL RATE OF KIND CLASSIFIERS EXTRACTION
J-Korean	71.5% (499/698)
E-Korean	68.3% (477/698)

The number of unique classifiers of a certain type found in each of the Korean corpora can also be generated. The figure for kind classifiers is given in table 9:

Table 9 Number of unique kind classifiers in J-Korean and E-Korean

	NUMBER OF UNIQUE CLASSIFIERS
J-Korean	37
E-Korean	41

We give all the unique classifiers found in J-Korean and E-Korean that correspond to some Japanese kind classifiers in table 10.

Table 10 Unique kind classifiers in J-Korean and E-Korean

CLASSIFIER IN BTEC	CLASSIFIER IN J-Korean	CLASSIFIER IN E-Korean
人	명, 인, 분, 사람	인, 사람, 명
錠	정, 알, 캡슐, 개	알
尾	마리	마리
台	대	차, 대, 아침
ヵ所	곳	

点	점, 가지	평방미터, 점
語	자	글자, 단어, 자
部	부, 패키지	부
通	통	통
席	석, 등석, 장	석, 자리
軒	집, 군데	칸, 채, 집
名	자리, 사람, 사이	
項	조	
ページ	페이지	페이지
羽	마리	
冊	권	권
アイテム	품목	아이템
便	편, 회	번, 편
部屋	방	밤, 층, 방, 개
足	켤레	켤레
着	벌	벌
名様	명	분
項目	품목	항목
字		자당
品		가지, 물건
両		량
条		조
室		박
ゲーム		게임
曲		곡

There are other types of classifiers and their distribution in the two corpora are shown in table 11. We have also included the counts for kind classifiers in the table to make it easier to compare the counts.

Table 11 Distribution of unique Korean classifiers of different types in J-Korean and E-Korean

	J - Korean		E-Korean	
TYPE	UNIQUE	ALL	UNIQUE	ALL
DEFAULT	22	363	34	311
KIND	37	499	41	477
SHAPE	46	1448	49	1121
TAXONOMIC	2	15	2	14
EVENT	17	592	24	433
MEASURE	61	2628	58	2495
GROUP	3	14	3	10
CONTAINER	19	305	31	201
ARRANGEMENT	1	3	2	5
PORTION	17	208	21	233
TEMPORAL	64	2232	69	2087
TOTAL	298	8307	310	7076

4.3.1.3 Discussion

Table 11 shows that more unique numeral classifiers are used in E-Korean corpus. In general, English does not use classifiers, whereas Japanese has to use numeral classifiers for counting objects. The reason we have less unique classifiers in J-Korean lies in translation strategy in part. When the numeral classifier is implicitly expressed in the source text of E-Korean, it will be more freely translated. That is why E-Korean has more unique numeral classifiers. This shows that the source language has a great effect on the human translation and we should consider this kind of characteristics of a corpus before we use it in the area of natural language processing (Paik et al: 2004).

4.3.2 Extraction of Chinese Numeral Classifiers

4.3.2.1 Method

To find the corresponding Chinese translation of a Japanese sentence

extracted from the corpus, we find in the Chinese corpus sentences with the same ID numbers as those Japanese sentences extracted earlier on. Classifiers have to be extracted from the Chinese sentences in a way very different from extracting classifiers from the Japanese sentences. This is because the Japanese sentences are tagged and segmented by a parser whereas the Chinese sentences are not. We start with using a program to construct the Chinese representation of the value of the numeral in the Japanese source text.

When constructing the Chinese representation of the value of a Japanese numeral, we have to consider two minor differences between the Sino-Japanese numeral system and the Chinese numeral system. The first difference lies in the representation of the value 2. In the Sino-Japanese numeral system, there is only one representation, that is, 二. In the Chinese numeral system, there are two representations. One is 二. Another is *两*. A second difference lies in the representation of zero digits. If a zero digit occurs between two non-zero digits, the Sino-Japanese numeral system would leave out the zero digit such that 601 would be *六百__一*. The Chinese numeral system would keep the zero digit such that 601 would be *六百零一*. Our algorithm used in constructing Chinese representation of the value of the Japanese numeral used in the source text is given as follows.

Figure 3 Algorithm for generating Chinese numerals

(1). If the last character of the Japanese numeral is a digit (*桁数*), that is, one of the following characters: 十,百,千,万,億:

(a) if the second last character of the Japanese numeral is a numeral below ten, that is, one of the following characters: 一,二,三,四,五,六,七,八,九:

take the numeral as a case of "number_digit", which means that the second last character of the numeral is a numeral below ten and the last character of the numeral is a digit.

(b) Else if the second last character of the Japanese numeral is also a digit (*桁数*)

take the numeral as a case of "digit_digit".

(c) Else the numeral is a case of "digit"

(2). Else if the last character of the Japanese numeral is a numeral below ten:

(a) if the second last character of the Japanese numeral is a digit:

(i) if the third last character of the Japanese numeral is a numeral below ten:

take the numeral as a case of "number_digit_numbert".

(ii) Else the numeral is a case "digit_number".

(b) Else the numeral is a case of "number"

(3) If the Japanese numeral is a case of "digi"t:

get a Chinese digit with the same value as the Japanese digit.

(4) Else if the Japanese numeral is a case of "number_digit":

get a Chinese numeral below ten with the same value as the second last character of the Japanese numeral in the source text.

get a Chinese digit with the same value as the Japanese digit

(a) If the value of the second last character of the Japanese numeral is equal to two,

concatenate the character class (二|两) with the Chinese digit such that both 二[Digit] and 两[Digit] would match the result of the concatenation.

(b) Else form a Chinese numeral by concatenating the Chinese numeral below ten that correspond to the second last character of the Japanese numeral in the source text and the Chinese digit that corresponds to the last character of the Japanese numeral in the source text with the former preceding the latter.

(5) Else if the Japanese numeral is a case of "digit_digit":

get a Chinese digit with the same value as the second last character of the Japanese numeral in the source text.

get a Chinese digit with the same value as the last character of the Japanese numeral in the source text

form a Chinese numeral by concatenating the Chinese digit that correspond to the second last character of the Japanese numeral in the source text and the Chinese digit that corresponds to the last character of the Japanese numeral in the source text with the former preceding the latter.

(6) Else if the Japanese numeral is a case of "number":

(a) if the value of the last character of the Japanese numeral is equal to two,

take the character class (二|两) with the Chinese representation of the Japanese numeral in the source text such that both 二 and 两 would

match the representation .

(b) Else form a Chinese numeral by getting a Chinese numeral below ten with the same value as the last character of the Japanese numeral in the source text.

(7) Else if the Japanese numeral is a case of "digit_number":

get a Chinese digit with the same value as the second last character of the Japanese numeral in the source text.

get a Chinese numeral below ten with the same value as the last character of the Japanese numeral in the source text.

(a) if the value of the second last character of the Japanese numeral, that is, the digit, is larger than ten:

put the zero digit "零" between the Chinese digit and the Chinese numeral below ten.

(b) Else form a Chinese numeral by concatenating the Chinese digit that correspond to the second last character of the Japanese numeral in the source text and the Chinese numeral below ten that corresponds to the last character of the Japanese numeral in the source text with the former preceding the latter.

(8) Else if the Japanese numeral is a case of "number_digit_number":

get a Chinese numeral below ten with the same value as the third last character of the Japanese numeral in the source text.

get a Chinese digit with the same value as the second last character of the Japanese numeral in the source text.

get a Chinese numeral below ten with the same value as the last character of the Japanese numeral in the source text .

(a) if the value of the last character of the Japanese numeral is equal to two:

(i) if the value of the second last character of the Japanese numeral, that is, the digit, is larger than ten:

concatenate the character class (二|两) , the Chinese digit that corresponds to the second last character of the Japanese numeral in the source text, the zero digit "零", and the Chinese numeral below ten that corresponds to the last character of the Japanese numeral in the source text in the order they are mentioned here such that both 二[Digit]零[Numeral] and 两[Digit]零[Numeral] would match the result of the concatenation

(ii) Else concatenate the character class (二|两) , the Chinese digit that corresponds to the second last character of the Japanese numeral in the source text, and the Chinese numeral below ten that corresponds to the last character of the Japanese numeral in the source text in the order they are mentioned here such that both 二[Digit] [Numeral] and 两[Digit][Numeral] would match the result of the concatenation

(b) Else if the value of the second last character of the Japanese numeral, that is, the digit, is larger than ten:
concatenate the Chinese numeral below ten that correspond to the third last character of the Japanese numeral in the source text, the Chinese digit that corresponds to the second last character of the Japanese numeral in the source text, the zero digit "零", and the Chinese numeral below ten that corresponds to the last character of the Japanese numeral in the source text in the order they are mentioned here.

(c) Else form a Chinese numeral by concatenating the Chinese numeral below ten that correspond to the third last character of the Japanese numeral in the source text, the Chinese digit that correspond to the second last character of the Japanese numeral in the source text and the Chinese numeral below ten that corresponds to the last character of the Japanese numeral in the source text in the order they are mentioned here.

We find patterns that match the representation in the Chinese sentence having the same ID number as a Japanese source text. After we find the numeral, we adopt a two pass approach so as to improve the accuracy of our extraction. First, the first character immediately following the Chinese numeral found is regarded as a classifier and is tested for uniqueness. Any unique Chinese character extracted this way is printed out in a report with the Chinese sentence from which it is extracted and the Japanese classifier used in the source text. Second, we correct our program based on this result. Not many of the characters extracted by this method are classifiers. This is because Chinese classifiers can be multi-character. We replace those characters that are not classifiers with the correct classifiers by looking at the sentences ourselves and pair each of the Chinese classifiers with its corresponding Japanese classifier. These pairs of classifiers are put in a glossary where all

alternatives of Japanese classifier that corresponds to a Chinese classifier are listed out. The glossary is included as part of the rewritten program. When processing a Chinese sentence, the new program would first check whether the first character after a Chinese numeral is found in the glossary and any of the Japanese classifiers corresponds to it in the glossary is the classifier used in the source text of the concerned sentence. If the first character after the Chinese numeral is not found in the glossary or the classifier used in the source text is not among the Japanese classifiers corresponding to the character in the glossary, the program would go on with checking whether the first two characters after the same numeral are in the glossary and whether any of the Japanese classifiers corresponds to the two characters as a word in the glossary is the classifier used in the source text of the concerned sentence. If this check fails again, we will try the first three characters after the numeral. We would make no further attempt after trying the first four characters after the numeral. The following algorithm shows how we extract Chinese classifiers.

Figure 4 Algorithm for extracting Chinese classifiers

(1) For each fragments of a Chinese sentence that match the pattern constructed by the algorithm given in figure 3

(a) If the first character immediately follows the Chinese numeral is found in the glossary:

(i) if the Japanese classifier used in the sentence with the same ID number is among the list of Japanese classifiers corresponding to the Chinese character (classifier):
extract the character

(b) Else if the first two characters immediately follows the Chinese numeral form a phrase which can be found in the glossary:

(i) if the Japanese classifier used in the sentence with the same ID number is among the list of Japanese classifiers corresponding to the Chinese phrase (classifier):
extract the two characters

(c) Else if the first three characters immediately follows the Chinese numeral form a phrase which can be found in the glossary:

(i) if the Japanese classifier used in the sentence with the same ID number is among the list of Japanese classifiers

corresponding to the Chinese phrase (classifier):
extract the three characters

(d) Else if the first four characters immediately follows the Chinese numeral form a phrase which can be found in the glossary:

(i) if the Japanese classifier used in the sentence with the same ID number is among the list of Japanese classifiers corresponding to the Chinese phrase (classifier):
extract the four characters

(e) Else fail.

4.3.2.1 Results

The glossary is both the result produced by our program and the means by which it uses to produce more accurate results. In table 12, we give in the first two columns entries taken from our glossary of Chinese classifiers that correspond to Japanese kind classifiers. The two columns in the right hand side give a list of Chinese kind classifiers and their corresponding Japanese classifiers as found in the Chinese and Japanese corpora.

Table 12 Japanese *kind* classifiers and Chinese *kind* classifiers

CHINESE CLASSIFIER	JAPANESE CLASSIFIER	REFERENT	CHINESE CLASSIFIER	JAPANESE CLASSIFIER
个	人 客 席 名 字 校 問 ヶ所 ヵ所 部屋 語 便 点	People, Seat, Word, School, Question, Place, Room, Flight		
人	人 名	People	人	人 名
名	人	People	名	人
口	人	People	口	人
位	人 名	People	位	人 名
班	便	Flights	班	便
次	便	Flights	次	便
航班	便	Flights	航班	便
间	部屋 室	Rooms	间	部屋 室

条	尾 条 区画 丁	Streets	条	尾 条 区画 丁
首	曲	Songs	首	曲
台	台	Cameras	台	台
架	台	Cameras	架	台
辆	台 両	Cars, Carriages	辆	台 両
只	羽	Birds		
封	通	Letters	封	通
本	冊 部	Books	本	冊 部
册	冊	Books	册	冊
家	軒	Shops	家	軒
幢	軒	Buildings	幢	軒
座	軒	Buildings	座	軒
股	株	Shares	股	株
场	ゲーム	Games	场	ゲーム
页	ページ	Pages	页	ページ
句	語	Sentences	句	語
份	部	Copies	份	部
项	項目	Entries	项	項目
节	両	Carriages	节	両
丸	錠	Pills		
粒	錠	Pills		
片	錠	Pills		
件	着 アイテム 品 点	Commodities, Shirts, Artworks	件	着 アイテム 品 点 個 ピース
张	名 台	Tickets, Beds		
双	足	Shoes		
段	区間	Sections		
		Racquets	副	本
		Drawings	幅	本

1st column: Chinese classifiers corresponding to Japanese kind classifiers
2nd column: Japanese kind classifiers
4th column: Chinese kind classifiers
5th column: Japanese classifiers corresponding to Chinese kind classifiers

As we can see from the table 12, not all the Chinese classifiers in the first column are kind classifiers. Some of the Japanese kind classifiers in the source text (Japanese) are found to be replaced by classifiers of a different type in the target text (Chinese). For example, 个 and 只 are the default classifiers. 节, 丸, 粒, 片, 件,张 and 段 are shape classifiers. Also, 双 is a group classifier that can be used with everything in pairs. Not all Japanese classifiers in the fifth column are kind classifiers. 個, and 本 are shape classifiers. ピース is a portion classifier. 区画 and 丁 are measure classifiers. As our main goal is to compare the use of classifiers with the same referent in Chinese, Japanese and Korean, we try not to complicate our task by assigning Chinese classifiers that correspond to a certain type of Japanese classifiers types other than that of the Japanese classifiers. However, we would like to point out that the incompatibility between the types of the Chinese classifier in the target text and the type of the Japanese classifier in the source text is a difficult issue in machine translation.

In table 13, we give in the first two columns entries taken from our glossary of Chinese classifiers that correspond to Japanese shape classifiers. The two columns in the right hand side give a list of Chinese shape classifiers and their corresponding Japanese classifiers.

Table 13 Japanese *shape* classifiers and Chinese *shape* classifiers

CHINESE CLASSIFIER	JAPANESE CLASSIFIER	REFERENT	CHINESE CLASSIFIER	JAPANESE CLASSIFIER
个	個 面	Eggs, Courts		
根	本	Sausages	根	本
颗	本	Teeth	颗	本
枝	本	Pencils	枝	本
		Pills	丸	錠
副	本	Racket		
只	本	Rackets		
次	本	Dives		
趟	本	Trains		
粒	粒	Pills	粒	錠 粒

		Pills, Ham, Piccata	片	錠 切れ 切
张	枚	Tickets	张	枚 名 台
条	枚	Blankets	条	枚 カートン
枚	枚	Coins	枚	枚
块	枚	Napkins	块	枚 切れ
卷	卷	Paper Towels	卷	卷
轮	輪	Wheels	轮	輪
级	段	Shifts	级	段
层	段	Beds	层	段
件	個	Luggage	把	つ
罐	本	Coke		
支	本	Pen		
瓶	本	Beer		
小瓶	本	Sake		
大瓶	本	Beer		
床	枚	Blankets		
版	シート	Stamps		
1st column: Chinese classifiers corresponding to Japanese shape classifiers 2nd column: Japanese shape classifiers 4th column: Chinese shape classifiers 5th column: Japanese classifiers corresponding to Chinese shape classifiers				

Among the 25 Chinese classifiers in the first column of table 13, 10 of them are not shape classifiers. For example, *个* is a default classifier, *件* and *副* are kind classifiers, and *床*, *瓶*, *小瓶*, *大瓶*, *罐*‘are container classifiers. *趟* and *次* are event classifiers.

Among the 13 Japanese classifiers in the fifth column of table 13, 7 of them are not shape classifiers. *つ* is the default classifier. *錠*, *名* and *台* are kind classifiers. *切れ* and *切* are portion classifiers. *カートン* is a container classifier.

The case of taxonomic classifier is less complicated. Table 14 gives the Chinese classifiers that correspond to Japanese taxonomic classifiers and the

Japanese classifiers that correspond to Chinese taxonomic classifiers:

Table 14 Japanese *taxonomic* classifiers and Chinese *taxonomic* classifiers

CHINESE CLASSIFIER	JAPANESE CLASSIFIER	REFERENT	CHINESE CLASSIFIER	JAPANESE CLASSIFIER
种	種類 種 類 通り	Typed Entities (Tobacco, Perfume)	种	種類 種 類 通り つ
1st column: Chinese classifiers corresponding to Japanese taxonomic classifiers 2nd column: Japanese taxonomic classifiers 4th column: Chinese taxonomic classifiers 5th column: Japanese classifiers corresponding to Chinese taxonomic classifiers				

The only Chinese classifier used for translating the Japanese taxonomic classifiers is the classifier *种*. But this taxonomic classifier can also be used for translating the default classifiers. This is because the Japanese default classifier つ can be used in a way such that the referent is regarded as a subtype of its kind.

Table 15 gives the Chinese classifiers that correspond to Japanese event classifiers and the Japanese classifiers that correspond to Chinese event classifiers:

Table 15 Japanese *event* classifiers and Chinese *event* classifiers

CHINESE CLASSIFIER	JAPANESE CLASSIFIER	REFERENT	CHINESE CLASSIFIER	JAPANESE CLASSIFIER
个	泊 ラウンド	Stay, Games		
次	回 度 ラウンド	Intake, Go, Games	次	回 度 ラウンド 本
		Trains	趟	本
轮	回 ラウンド	Games, Rounds	轮	回 ラウンド

回	回	Go	回	回
场	回 ラウンド	Games, Shows	场	回 ラウンド
班	回	Buses	班	回
局	回 ラウンド	Games, Rounds	局	回 ラウンド
遍	度 回	Times, Check	遍	度 回
看	度	Look	看	度
查	度	Investigation	查	度
找	度	Search	找	度
面	度	Meeting	面	度
下	度	Investigation	下	度
回合	ラウンド	Rounds	回合	ラウンド
圈	周 鞍 ラウンド	Ride, Rounds	圈	周 鞍 ラウンド
杆	打	Hits	杆	打
天	泊	Stay		
晚	泊	Stay		
晚上	泊	Stay		
1st column: Chinese classifiers corresponding to Japanese event classifiers 2nd column: Japanese event classifiers 4th column: Chinese event classifiers 5th column: Japanese classifiers corresponding to Chinese event classifiers				

The major difference between the first column and the fourth column lies in the inclusion of the Japanese shape classifier *本* which correspond to the Chinese event classifiers *次* and *趟*. The Chinese temporal classifiers *天*, *晚* and *晚上* used for counting the days of stay, which correspond to the Japanese event classifier *泊* for counting stay, are missing in the fourth column.

Table 16 gives the Chinese classifiers that correspond to Japanese measure classifiers and the Japanese classifiers that correspond to Chinese measure classifiers:

Table 16 Japanese *measure* classifiers and Chinese *measure* classifiers

CHINESE CLASSIFIER	JAPANESE CLASSIFIER	REFERENT	CHINESE CLASSIFIER	JAPANESE CLASSIFIER
个	ブロック	District		
米	メートル	Distance (Walk)	米	メートル
平方米	平方メートル	Area (Room)	平方米	平方メートル
公里	キロメートル キロ	Distance (Drive)	公里	キロメートル キロ
平方公里	平方キロ	Area (Japan)	平方公里	平方キロ
厘米	センチ	Size (Shoes)	厘米	センチ
公分	センチ	Size (Hip)	公分	センチ
毫米	ミリ	Length (Film)	毫米	ミリ
公斤	キログラム	Weight (Luggage)	公斤	キログラム
克	グラム	Weight (Meat)	克	グラム
公斤	キロ	Weight (Body)	公斤	キロ
升	リットル	Volume (Gasoline)	升	リットル
CC	シーシー	Volume (Cylinder)	CC	シーシー
卡路里	カロリー	Energy (Herb Tea)	卡路里	カロリー
伏	ボルト	Voltage (Battery)	伏	ボルト
度	度	Temperature	度	度
磅	パウンド ポンド	Weight (Meat), Value (Money)	磅	パウンド ポンド
盎司	オンス	Weight (Fragrance)	盎司	オンス
加仑	ガロン	Volume (Gasoline)	加仑	ガロン
夸脱	クオート	Volume (Milk)	夸脱	クオート
品脱	パイント	Volume (Beer)	品脱	パイント
英里	マイル	Distance (Drive)	英里	マイル
码	ヤード	Length (Cloth)	码	ヤード

英尺	フィート	Depth (Lake)	英尺	フィート
英寸	インチ	Size (Caps)	英寸	インチ
平方英里	平方マイル	Area (Japan)	平方英里	平方マイル
英亩	エーカー	Area (Park)	英亩	エーカー
节	ノット	Speed (Ship)	节	ノット
克拉	カラット	Size (Diamond)	克拉	カラット
开	金	Purity (Ring)	开	金
K	金	Purity (Necklace)	K	金
港币	香港ドル	Value (Charge)	港币	香港ドル
分	セント 銭	Value (Money)	分	セント 銭
美分	セント	Value (Postal)	美分	セント
美元	ユーエスドル ドル 米ドル	Value (Money)	美元	ユーエスドル ドル 米ドル
澳元	オーストラリアドル	Value (Tax)	澳元	オーストラリアドル
日元	円	Value (Money)	日元	円
马克	マルク	Value (Money)	马克	マルク
法郎	フラン	Value (Money)	法郎	フラン
便士	ペンス ペニー	Value (Stamp)	便士	ペンス ペニー
加拿大元	カナダドル	Value (Money)	加拿大元	カナダドル
步	歩	Distance (Walk)	步	歩
站	駅	Distance (Rail)	站	駅
号	回り まわり	Size (Overall)	号	回り まわり
单元	ユニット	Unit (Compartment)	单元	ユニット
套	ユニット	Unit (Goods)	套	ユニット
叠	畳	Area (Room)	叠	畳

1st column: Chinese classifiers corresponding to Japanese measure classifiers
2nd column: Japanese measure classifiers
4th column: Chinese measure classifiers
5th column: Japanese classifiers corresponding to Chinese measure classifiers

The difference between the first column and the fourth column is obvious. They are the same except in the first row. That is to say, Chinese measure classifiers almost always correspond to Japanese measure classifiers. In the

first row, we give the Japanese classifier for counting districts *区間* but there is no corresponding measure classifier for counting districts in Chinese. In the Chinese translation, the default classifier is used instead. We will give a more detailed discussion of the use of default classifier in similar situations in the next section.

Table 17 gives the Chinese classifiers that correspond to Japanese portion classifiers and the Japanese classifiers that correspond to Chinese portion classifiers:

Table 17 Japanese *portion* classifiers and Chinese *portion* classifiers

CHINESE CLASSIFIER	JAPANESE CLASSIFIER	REFERENT	CHINESE CLASSIFIER	JAPANESE CLASSIFIER
个	目盛	Medicine		
件	ピース	Tableware		
块	切れ 切	Pizza		
		Presentation	部分	部

1st column: Chinese classifiers corresponding to Japanese portion classifiers
2nd column: Japanese measure classifiers
4th column: Chinese measure classifiers
5th column: Japanese classifiers corresponding to Chinese portion classifiers

None of the Chinese classifiers in the first column is a portion classifier. *部分* is a Chinese portion classifier that can be found in the corpus. But the corresponding Japanese classifier used in the source text *部* is taken as a kind classifier by our program. We would also like to point out that all the Japanese compound classifiers are not given in table 17. A compound classifier is formed by combining affixes such as *分* or *前* with a preceding classifier. Most of these compound classifiers are portion classifiers. Examples of compound classifiers are *人分*, *枚分*, *ドル分* and *人前*. There is no corresponding Chinese compound classifier. Therefore, in the Chinese translation of a Japanese sentence containing such classifier, either one of the constituents forming the compound classifier will not have a translation counterpart, as illustrated in 47 and 48.

47a.

[紅茶]$_{N}$ [を]$_{OBJ}$ [三]$_{NUM}$ [[人]$_{CL}$ [分]$_{CL}$]$_{CL'}$ [ください]$_{V}$

Tea Obj 3 People Portion Please

Three people's portion of tea, please.

47b.

[请]$_{V}$ [给]$_{V}$ [我]$_{N}$ [三]$_{NUM}$ [份]$_{CL}$ [红茶]$_{N}$

Please Give Me 3 Portion Tea

Please give me three portion of tea

48a.

[十]$_{NUM}$ [[ドル]$_{CL}$ [分]$_{CL}$]$_{CL'}$ [の]$_{P}$ [フェアカード]$_{N}$ [を]$_{OBJ}$ [ください]$_{V}$

10 Dollars Portion AND Farecard OBJ Please

Ten Dollars' Portion of Farecard, please

48b.

[请]$_{V}$ [给]$_{V}$ [我]$_{N}$ [十]$_{NUM}$ [美元]$_{CL}$ [的]$_{AND}$ [票卡]$_{N}$

Please Give Me 10 US Dollars AND Farecard

Please give me a ten-US dollar-farecard.

In 47b, only the Chinese portion classifier *份* is used. The classifier *人* for counting people who will consume tea in the context of 47 is omitted. In 48b, the Chinese classifier *美元* (dollar) corresponds to the *ドル* (dollar) part of the compound classifier *ドル分*. The *分* part is left untranslated.

Next we look at some Chinese classifiers that correspond to Japanese container classifiers and some Japanese classifiers corresponding to Chinese container classifiers

Table 18 Japanese *container* classifiers and Chinese *container* classifiers

CHINESE CLASSIFIER	JAPANESE CLASSIFIER	REFERENT	CHINESE CLASSIFIER	JAPANESE CLASSIFIER
罐	缶	Peanuts	罐	缶 本
瓶	びん ビン 瓶	Jelly	瓶	本 びん ビン 瓶
		Sake	小瓶	本
		Beer	大瓶	本
条	カートン	Tobacco		
盘	皿	Sausages	盘	皿
盒	パック 箱 ケース	Tobacco	盒	パック 箱
杯	杯 ぱい パイ	Water	杯	杯 ぱい パイ
板	袋	Batteries		
袋	袋	Peanuts	袋	袋
粒	カプセル	Medicine		
勺	杯 ぱい パイ	Medicine	勺	杯 ぱい パイ
汤匙	杯 ぱい パイ	Medicine	汤匙	杯 ぱい パイ
茶杯	杯 ぱい パイ	Water	茶杯	杯 ぱい パイ
扎	杯 ぱい パイ	Beer	扎	杯 ぱい パイ
壶	杯 ぱい パイ	Coffee	壶	杯 ぱい パイ
箱	ケース	Wine		
听	缶	Peach		
		Bed Sheet	床	枚

1st column: Chinese classifiers corresponding to Japanese container classifiers
2nd column: Japanese measure classifiers
4th column: Chinese measure classifiers
5th column: Japanese classifiers corresponding to Chinese container classifiers

Table 18 shows that speakers of both languages may use a shape classifier that describes a property of a container for counting objects in the container but the domain covered by such shape classifiers in each of the language may be different. To illustrate, Chinese speakers can count boxes of tobaccos by using the shape classifier 条, which selects for the length of the boxes, but in the

source text the Japanese container classifier カートン is used. Chinese speakers can also count powder contained in a capsule by using the shape classifier *粒*, which selects for the small size of the capsule, but in the source text the Japanese container classifier カプセル is used. In Japanese, one can count bed sheets by using the shape classifier *枚*. In Chinese, it is possible to count bed sheets by using a container classifiers *床*. Chinese speakers can command a variety of container classifiers for counting bottles of liquid: *小瓶*,*大瓶*, and *瓶*, which have the literal meaning of small bottle (for containing sake), big bottle (for containing beer) and bottle (for containing jelly) respectively. When counting liquid contained in bottles, Japanese speakers have the shape classifier *本* in command but they cannot add any adjective to the common shape classifier *瓶* in the way Chinese speakers do. Sung (1996) has pointed out this property.

Next, we will compare the group classifiers of the two languages. The Chinese classifiers corresponding to Japanese group classifiers and Japanese classifiers corresponding to Chinese group classifiers are given in table 19:

Table 19 Japanese *group* classifiers and Chinese *group* classifiers

CHINESE CLASSIFIER	JAPANESE CLASSIFIER	REFERENT	CHINESE CLASSIFIER	JAPANESE CLASSIFIER
个	クラス	Class		
组	セット	Movements	组	セット
套	セット，ユニット	Golf Clubs	套	セット，ユニット
对	組	Shoes	对	組
打	ダース	Oranges	打	ダース
		Shoes	双	足

1st column: Chinese classifiers corresponding to Japanese group classifiers
2nd column: Japanese measure classifiers
4th column: Chinese measure classifiers
5th column: Japanese classifiers corresponding to Chinese group classifiers

The Chinese default classifier 个 which can be used for counting classes is missing in the fourth column. The Chinese group classifier *双*, which

corresponds to the Japanese kind classifier 足 is found in the fourth column but not in the first column. Notice that the Chinese group classifier 双 is a more general classifier for counting everything in pairs whereas the Japanese kind classifier 足 is specifically used for counting things that people put on their feet.

The case of arrangement classifiers is also a straightforward one. All the Chinese classifiers in column one that correspond to Japanese arrangement classifiers are also arrangement classifiers themselves. This means that Chinese arrangement classifiers have the same domain as the Japanese arrangement classifiers.

Table 20 Japanese *arrangement* classifiers and Chinese *arrangement* classifiers

CHINESE CLASSIFIER	JAPANESE CLASSIFIER	REFERENT	CHINESE CLASSIFIER	JAPANESE CLASSIFIER
行	行	Words	行	行
排	列	Seats	排	列
队	列	Queues	队	列

1st column: Chinese classifiers corresponding to Japanese arrangement classifiers
2nd column: Japanese measure classifiers
4th column: Chinese measure classifiers
5th column: Japanese classifiers corresponding to Chinese arrangement classifiers

The case of temporal classifier is, however, not straightforward. Let us have a look of table 21, which gives the Chinese classifiers corresponding to Japanese temporal classifiers and Japanese classifiers corresponding to Chinese temporal classifiers.

Table 21 Japanese *temporal* classifiers and Chinese *temporal* classifiers

CHINESE CLASSIFIER	JAPANESE CLASSIFIER	REFERENT	CHINESE CLASSIFIER	JAPANESE CLASSIFIER
个	世紀 ヶ月 ケ月 か月 ヵ月 カ月 週間 週 晩 時間	Centuries, Months, Weeks, Nights, Hours		
年	年 年間	Years	年	年 年間
年间	年間	Years	年间	年間
岁	歳 才	Ages	岁	歳 才
周	週間	Weeks	周	週間
星期	週間	Weeks	星期	週間
日	日	Days	日	日
天	泊 日	Days	天	泊 日
晚	晩 泊	Nights	晚	晩 泊
晚上	晩 泊	Nights	晚上	晩 泊
小时	時間	Hours	小时	時間
分	分	Minutes	分	分
分钟	分	Minutes	分钟	分
秒钟	秒	Seconds	秒钟	秒
秒	秒	Seconds	秒	秒

It is possible to use the default classifier 个 to count some temporal entities in Chinese but it is not possible to do so in Japanese. The default classifier 个 carries very little information about what it counts. Therefore, the referent, that is, a period of time, has to be explicitly mentioned in the Chinese target text. Consider the following.

49a.

[是]$_{COP}$	[好]$_{ADV}$	[几]$_{NUM}$	[个]$_{CL}$	[世纪]$_{N}$	[前]$_{P}$	[的]$_{ADN}$	[东西]$_{N}$	[了]$_{N}$
Is	Quite	Several	Object	Century	Before	ADN	Thing	P

It was built several centuries ago.

*49b.

[是]$_{COP}$ [好]$_{ADV}$ [几]$_{NUM}$ [个]$_{CL}$ [前]$_{P}$ [的]$_{ADN}$ [东西]$_{N}$ [了]$_{N}$

Is Quite Several Object Before ADN Thing P

Without the referent *世纪* (Century) for the default classifier *个*, the sentence 49b is semantically unacceptable.

Whether the referent-default-classifier combination is used for translating a Japanese temporal classifier is determined in an arbitrary way. Some of the temporal nouns in Chinese can be used as classifiers without adding the default classifier *个*, as in example 50a to 57a: The corresponding Japanese sentences are also listed as 50b to 57b.

50a. 我在东京住了[十]$_{NUM}$[年]$_{CL}$。(Chinese)50b. 東京に[十]$_{NUM}$[年間]$_{CL}$住んでいます。(Japanese)

51a. 打算停留[两]$_{NUM}$[周]$_{CL}$。 (Chinese)

51b. [二]$_{NUM}$[週間]$_{CL}$滞在する予定です。(Japanese)

52a. 我想租[一]$_{NUM}$[天]$_{CL}$。(Chinese)

52b. [一]$_{NUM}$[日]$_{CL}$借りたいのですが。(Japanese)

53a. [一]$_{NUM}$[晚上]$_{CL}$多少钱 (Chinese)

53b. [一]$_{NUM}$[晩]$_{CL}$いくらですか。(Japanese)

54a. 大约[两]$_{NUM}$[小时]$_{CL}$后到达。(Chinese)

54b. だいたい[二]$_{NUM}$[時間]$_{CL}$ほどで着きます。(Japanese)

55a. 请等[十]$_{NUM}$[分钟]$_{CL}$。(Chinese)

55b. [十]$_{NUM}$[分]$_{CL}$待っていてください 。(Japanese)

56a. 我的手表一天快[三十]$_{NUM}$[秒]$_{CL}$。(Chinese)

56b. 私の時計は一日[三十]$_{NUM}$[秒]$_{CL}$進んでしまうのです。(Japanese)

57a. 先生，就[一]$_{NUM}$[秒种]$_{CL}$。 (Chinese)

57b. お客様[一]$_{NUM}$[秒間]$_{CL}$だけ。 (Japanese)

The above examples show that Chinese have specific temporal classifiers for years (*年*), weeks (*周*), days (*天*), nights (*晚上*), hours (*小时*), minutes (*分钟*) and seconds(*秒种*, *秒*). Alternatively, we can analyze these temporal classifiers as countable nouns. However, the following examples show that some of these temporal nouns can also be counted by the default classifier *个*, together with

other temporal entities.

58a. 是好[几]$_{NUM}$[个]$_{CL}$[世纪]$_{N}$前的东西了。(Chinese)
58b. [何]$_{NUM}$[世紀]$_{CL}$も前のものです 。(Japanese)

59a. 大概[一]$_{NUM}$[个]$_{CL}$[月]$_{N}$。(Chinese)
59b. 約[一]$_{NUM}$[ヶ月]$_{CL}$です。(Japanese)

60a. 至少住[一]$_{NUM}$[个]$_{CL}$[星期]$_{N}$。(Chinese)
60b. 最低[一]$_{NUM}$[週間]$_{CL}$滞在します。(Japanese)

61a. [一]$_{NUM}$[个]$_{CL}$[晚上]$_{N}$多少钱? (Chinese)
61b. [一]$_{NUM}$[晩]$_{CL}$いくらですか。(Japanese)

62a. 请等[一]$_{NUM}$[个]$_{CL}$[小时]$_{N}$左右。(Chinese)
62b. [一]$_{NUM}$[時間]$_{CL}$ほどお待ちください 。(Japanese)

These examples show that centuries (*世纪*), months (*月*), weeks (*星期*), nights (*晚上*) and hours (*小时*) can take the default classifier *个*. In these examples, the representations of temporal entities, that is, *世纪*, *月*, *星期*, *晚上* and *小时*, must be analyzed as uncountable nouns. We can divide the temporal expressions into two kinds. One group of temporal expressions (examples 58-62) is analyzed as nouns, not classifiers and the other group (examples 50-57) can be analyzed either as specific classifiers or nouns that take no classifiers. Comparing sentences 58-62 with sentences 50-57, we can tell that weeks, nights and hours can both be used with or without the default classifier *个*. This means that these temporal entities are multi-kind.

The last type of classifiers that we would consider is the default classifiers. Table 22 gives the Chinese classifiers corresponding to Japanese default classifiers

Table 22 Chinese classifiers corresponding to Japanese *default* classifier

CHINESE CLASSIFIER	JAPANESE CLASSIFIER	REFERENT
个	つ	Hamburger, Puppets, Festivals,

		Islands, Soup, Requests
把	つ	Keys, Knives
件	つ	Parcels, Shirts
盒	つ	Tobacco
份	つ	Sandwiches, Whisky
家	つ	Companies, Restaurants
顶	つ	Tents
辆	つ	Camping Cars
张	つ	Chairs, Beds
杯	つ	Tea
只	つ	Hamburgers, Eggs, Glasses
卷	つ	Films
块	つ	Erasers, Sugar
种	つ	Ceremonies, Trips
间	つ	Rooms
条	つ	Skirts, Streets, Ties
双	つ	Socks
号	つ	Size (Caps)
颗	つ	Planets
罐	つ	Beer
大杯	つ	Wine
站	つ	Stations
副	つ	Rackets
封	つ	Mails
床	つ	Bedding
门	つ	Classes
本	つ	Pamphlets
场	つ	Concerts
班	つ	Flights
幅	つ	Murals

Among the Chinese classifiers of table 22, only 个 and 只 are default classifiers. We treat both 个 and 只 as the default classifiers in Chinese since both of them can take referents of more than one animacy. 个 can be used for counting inanimate objects and people. 只 can be used for counting inanimate

objects and animals. It can sometimes be used for counting people in a derogative manner. Their referents have very little in common in terms of physical properties. Table 23 shows the referents taken by the two classifiers and their corresponding Japanese classifiers as given in tables 12 to 22:

Table 23 Japanese classifiers corresponding to Chinese *default* classifiers

CHINESE CLASSIFIER	JAPANESE CLASSIFIER	REFERENT
个	つ 世紀ヶ月 ケ月 か月 ヵ月 カ月 週間 週 晩 時間 クラス 目盛 ブロック 個 面 人 客 席 名 字 校 問 ヶ所 ヵ所 部屋 語 便 点	Hamburger, Puppets, Festivals, Islands, Soup, Requests, Centuries, Months, Weeks, Nights, Hours, Classes, Medicine, Districts, Eggs, Courts, People, Seats, Words, Schools, Questions, Places, Rooms
只	つ 羽	Hamburgers, Eggs, Glasses, Birds

The referents given in table 23 are by no means exhaustive. The difference between the two classifiers is that 个 is never used with animals and 只 is not used with people in our Chinese corpus. 个 is used with a larger number and variety of referents. It occurs 1914 times in the Chinese corpus. It can be used for counting referents counted by Japanese kind classifiers, shape classifiers, measure classifiers, event classifiers, group classifiers, temporal classifiers and the default classifier. 只 occurs only 191 times in the same corpus. It is only used with referents counted by Japanese kind classifiers and the default classifier.

We also notice that the referents taken by the two Chinese default classifiers hardly coincide with the referents taken by the Japanese default classifier. For example, the Chinese default classifier 个 can be used with a number of temporal expressions whereas the Japanese default classifier つ does not co-occur with any of the temporal expressions in the corpus. There are some referents that can be taken by the Japanese default classifier but not the

Chinese default classifiers. The Japanese default classifier つ can be used for replacing the container classifiers but the Chinese default classifiers 个 and 只 are not used in such manner.

4.3.2.3 Discussion

We have further examined the structure of sentences containing default classifiers and noticed an interesting difference between cases where a Japanese (specific) classifier in the source text is replaced by a Chinese default classifier and cases where the Japanese default classifier in the source text is replaced by a Chinese (specific) classifier in the target text. See the following examples.

63a. 我想预订今晚七点，[两]$_{NUM}$[个]$_{CL}$[人]$_{N}$。(Chinese)
63b. 今晚七時[二]$_{NUM}$[名]$_{CL}$の予約をしたいのですが。(Japanese)

64a. 到日本的电报[一]$_{NUM}$[个]$_{CL}$[字]$_{N}$多少钱？(Chinese)
64b. 日本への電報は[一]$_{NUM}$[字]$_{CL}$いくらですか。(Japanese)

65a. 那种班的话，[一]$_{NUM}$[个]$_{CL}$[班]$_{N}$上有多少人？(Chinese)
65b. そのコースは[一]$_{NUM}$[クラス]$_{CL}$何人ですか。(Japanese)

66a. 你想一想，如果向每个人再多问[一]$_{NUM}$[个]$_{CL}$[问题]$_{N}$，那么要花费多长时间。(Chinese)
66b もし一人にあと[一]$_{NUM}$[問]$_{CL}$ずつ質問してたら、どんなに仕事が長くなるかかんがえてもごらん。(Japanese)

Notice that in all four pairs of sentences, the Japanese sentences contain kind classifiers which are used anaphorically without any referent being counted by it. The kind classifier in each of the Japanese sentences carries very specific information about the omitted referent. The Japanese classifier in the source text carries most of the information carried by the referent in the target text. To put it in another way, the Japanese classifier in the source text is so specific such that its omitted referent cannot possibly be anything other than the referent explicitly given in the target text. To illustrate, the domain of the Japanese classifier 問 in sentence 66b contain only questions, the referent of the Chinese

target text 66a. The Japanese classifier *問* can be analyzed as a full noun and a contracted from of its referent. This would mean that the Japanese sentence would receive the following analysis where the full noun becomes countable and an absent classifier, denoted by a 0, is used.

67 もし一人にあと[一]$_{NUM}$[0]$_{CL}$[問]$_{N}$ずつ質問してたら、どんなに仕事が長くなるかかんがえてもごらん。(Japanese)

Both sentence 64b and 65b can be analyzed in a similar manner such that the classifier *字* (word) in 64b and *クラス* (class) in 65b are both analyzed as full nouns used with an absent classifier.

Among 229 cases where a Japanese (specific) classifier in the source text is replaced by a Chinese default classifier, we cannot identify the referent of the Japanese specific classifier in 133 cases (58%). We can, however, identify the referent for all except one of the Chinese sentences containing a Chinese default classifier.

In 294 cases where the Japanese default classifier in the source text is replaced by a Chinese (specific) classifier in the target text, as illustrated below:

68a. 日本人の男性はほとんどの人は[一]$_{NUM}$[つ]$_{CL}$[の]$_{ADN}$[会社]$_{N}$で引退するまで働くの。(Japanese)
68b. 几乎所有的人到退休都在[一]$_{NUM}$[家]$_{CL}$[公司]$_{N}$工作。(Chinese)

The referent of the Japanese default classifier cannot be identified in 56 (19%) of the cases. As for the Chinese sentences containing specific classifiers, only 31 (10.5%) of them are found to leave off the referents.

Although we may have to further verify these figures, we get from these figures the impression that the use of a specific classifier without a referent is more common in Japanese and it may be a significant difference between Chinese and Japanese of the same register in the use of classifiers. More examples of using a Japanese specific classifier without a referent in the source text and explicitly mentioning the referent in the target text are given in the following:

69a. 就在[两]$_{NUM}$[条]$_{CL}$[街]$_{N}$的前面。(Chinese)
69b. [二]$_{NUM}$[丁]$_{CL}$ばかり先です。(Japanese)

70a. 一直走到路口，向左转，再一直走[两]$_{NUM}$[条]$_{CL}$[街]$_{N}$，对吧？(Chinese)
70b. 角まで行って信号のところを左にまだりまっすぐ[二]$_{NUM}$[区画]$_{CL}$いきなさい。(Japanese)

It is found that two constructions in the Chinese target text are problematic. One such construction is the absent classifier partitive construction, illustrated by sentence 71a:

71a.

[是]$_{COP}$	[瑞士阿尔卑斯山脉]$_{NP}$	[的]$_{ADN}$	[最高峰]$_{N}$	[之]$_{ADN}$	[一]$_{NUM}$
Is	the Alps	ADN	highest mountain	ADN	1

"It is one of the highest mountains of the Alps."

Such construction expresses a partitive relation between the noun preceding the adnominal marker and the numeral-absent classifier combination. The absent classifier construction is also found in Korean. Paik and Bond (2001) gives the following example:

72.

이	마을에는	학교가	하나도	없다
This	town-LOC	school-NOM	one-even	has-not

"This town does not have a single school."

This example given by Paik and Bond (2001) is a negative sentence. Apparently there is a relation between the omission of the classifier and the negation of the numeral in this example. But the omission in sentence 71a appears to be the result of a stylistic decision.

71b.是瑞士阿尔卑斯山脉其中的[一]$_{NUM}$[个]$_{CL}$[最高峰]$_{N}$。

[是]$_{COP}$	[瑞士阿尔卑斯山脉]$_{NP}$	其中	[的]$_{ADN}$	[一]$_{NUM}$	[个]$_{CL}$	[最高峰]$_{N}$
Is	the Alps	Among	ADN	1	CL	Highest

Mountain

"It is one of the highest mountains of the Alps."

We can add back the classifier, as shown in 71b and end up with a sentence bearing the same meaning as 71a.

Another problematic construction is the absent numeral construction, illustrated by sentence 73a:

73a. 我来拿[件]$_{CL}$[行李]$_{N}$吧。(Chinese)
73b. [荷物]$_{N}$[を]$_{OBJ}$[一]$_{NUM}$[つ]$_{CL}$お持ちしましょう。(Japanese)

Such construction is used only when the quantity of the referent is equal to one, represented as 一 in the source text 73b, and this quantity is not considered important in the discourse.

The constructions exemplified in 71 and 73 are problematic because omitting either the classifier or numeral in the Chinese target text would make it difficult to find a starting point for us to analyze its syntactic structure and look for the referent of any classifier contained.

4.4 Referent-Classifier Pair Comparison

The subtask of referent-classifier pair comparison is about comparing the referent-classifier pairs found in the corpus with the output of the algorithm for generating classifiers by using semantic classes mentioned in Bond and Paik (2000). The algorithm is reproduced in figure 5, with the pseudo codes for handling coordinate noun phrase deleted.

Figure 5 Bond and Paik (2000)'s algorithm for generating classifiers

(1) If the head noun has a default classifier in the lexicon:
use the noun's default classifier
(2) Else if it exists, use the default classifier of the head noun's most salient semantic class (the class's default classifier)
(3) Else use the residual classifier

4.4.1 Method

The extracted Japanese sentences are then read one by one by a program which extracts the ID number of each sentence, the referents in each of the sentences and the classifier used with each of the referents. Then we find in the files in which the extracted Chinese and Korean sentences are stored Chinese sentences and Korean sentences with the same ID numbers as those Japanese sentences. For each of the referent extracted from a Japanese sentence, the Korean classifier used in a Korean sentence from E-Korean with the same ID number as the Japanese sentence, the Korean classifier used in a Korean sentence from J-Korean with the same ID number as the Japanese sentence and the Chinese classifier used in a Chinese sentence with the same ID number as the Japanese sentence are extracted. So now we have the Chinese, Japanese and Korean classifiers used with every entity denoted by every referent in every Japanese sentence. Notice that the entity denoted by the referent in a Japanese sentence is the same as the entity denoted by the referent in the Chinese sentence bearing the same ID number and the entity denoted by the referent in the Korean sentences bearing the same ID number. For every entity, the number of times of every classifier used with it is counted. And for each language, the classifier most frequently used with every entity is obtained by looking at the counts. The list of classifiers used with every entity, the Chinese classifier most frequently used with it, the Japanese classifier most frequently used with it and the Korean classifier most frequently used with it is paired with the results produced by Bond and Paik (2000)'s algorithm for generating classifiers using semantics classes. The results produced by Bond and Paik (2000) is given as a word list for my program to match every entity denoted by the referents of all sentences extracted from the corpora with the nouns in the word list. For every noun found in the word list, the list of

classifiers found to be possible for being used with every entity, the Chinese classifier found to be most frequently used with it, the Japanese classifier found to be most frequently used with it and the Korean classifier found to be most frequently used with it are extracted together with the noun.

After pairing what is extracted from the corpora for an entity with what is extracted from the word list that gives Bond and Paik's results for the same entity, the classifier found to be used most frequently with the entity in each language is compared with the classifier found to be most frequently used with the entity by Bond and Paik using semantic classes. And we take down the number of cases in which the classifier most frequently used is different from the classifier found to be most frequently used by Bond and Paik's algorithm for generating classifiers using semantic classes.

Figure 6 gives the algorithm described above:

Figure 6 Algorithm for pairing extracted referent-classifier pairs with Bond and Paik's results

(1) For every extracted sentence in every file outputed by the sorting program
    (a) For each fragment of a Japanese sentence containing a classifier
        (i) If the referent has been identified:
            add one to the counter for storing the total number of referents found in the BTEC corpus (counter 1).
            get the ID number
            use the ID number to search for a Chinese sentence from the extracted Chinese sentences (output of match_cj.pl) and two Korean sentences from the extracted Korean sentences, one for each of the Korean corpora (output of match_jk.pl)
            (一) For each of the extracted sentences (one Chinese sentence, one Japanese sentence, two Korean sentence)
                get the extracted classifier
                (あ) If the counter for counting the frequency (value of the counter) of using the classifier

(level two key of the counter) with the referent (level one key of the counter) is undefined:
define the counter
(い) Else add one to the counter.

(2) For every referent
(a) For every language
(i) For every counter created in (1.a.i.一.あ)
(一) If a store for saving the counts of the most frequently used classifier and a store for saving the orthography of the most frequently used classifier are undefined:
define the two stores.
(二) Else if the value of the counter > the value of the store that saves the counts of the most frequently used classifier:
assign the level two key of the counter to the value of the store for saving the orthography of the most frequently used classifier.
assign the value of the counter to the store for saving the counts of the most frequently used classifier
(三) Else if the value of the counter = the value of the store that saves the counts of the most frequently used classifier:
concatenate the level two key of the counter with the value of the store for saving the orthography of the most frequently used classifier.

(b) If the referent is found in the output of the algorithm proposed by Bond and Paik (2000):
add one to the counter for storing the total number of referents found in the output generated by the algorithm proposed by Bond and Paik (2000) (counter 2).
(i) If the classifier given as the most frequently used Japanese classifier for the referent in the output of Bond and Paik (2000) is different from the value of the store for saving the orthography of the most frequently used Japanese classifier:
add one to the counter for storing the number of cases in which the classifier found by Bond and Paik (2000)'s algorithm for

generating classifiers using semantic classes to be most frequently used Japanese classifier for a referent is different from the corpus data (counter 3)

(ii) If the classifier given as the most frequently used Korean classifier for the referent in the output of Bond and Paik (2000) is different from the value of the store for saving the orthography of the most frequently used Korean classifier:

add one to the counter for storing the number of cases in which the classifier found by Bond and Paik (2000)'s algorithm for generating classifiers using semantic classes to be the most frequently used Korean classifier for a referent is different from the corpus data (counter 4)

(iii) If the classifier given as the most frequently used Chinese classifier for the referent in the output of Bond and Paik (2000) is different from the value of the store for saving the orthography of the most frequently used Chinese classifier:

add one to the counter for storing the number of cases in which the classifier found by Bond and Paik (2000)'s algorithm for generating classifiers using semantic classes to be the most frequently used Chinese classifier for a referent is different from the corpus data (counter 5)

4.4.2 Results

The algorithm described in figure 6 gives us the following statistics.

Table 24 Comparing uses of classifiers in the corpora and the results of Bond and Paik's algorithm (Part 1)

	KIND	SHAPE	MEASURE	TAXONOMIC	GROUP
Counter 1	174	296	298	16	9
Counter 2	142	194	251	13	8
Counter 2 / Counter 1	81.6%	65.5%	84.2%	81.5%	88.8%
Counter 3	111	147	251	13	8
Counter 3 /Counter 2	78.2%	75.8%	100%	100%	100%
Counter 4	115	150	248	13	7

Counter 4 / Counter 2	80%	77.3%	98.8%	100%	87.5%
Counter 5	104	155	245	13	8
Counter 5 / Counter 2	73.2%	79.9%	97.6%	100%	100%

Table 25 Comparing uses of classifier in the corpora and the results of Bond and Paik's algorithm (Part 2)

	CONTAINER	ARRANGEMET	EVENT	TEMPORAL	PORTION
Counter 1	105	3	201	423	101
Counter 2	78	1	132	304	73
Counter 2 / Counter 1	74.3%	33.3%	65.7%	71.9%	72.3%
Counter 3	58	1	117	304	73
Counter 3 / Counter 2	74.4%	100%	88.6%	100%	100%
Counter 4	64	1	126	298	71
Counter 4 / Counter 2	82.1%	100%	95.5%	98%	97.3%
Counter 5	65	1	107	285	67
Counter 5 / Counter 2	83.3%	100%	81.1%	93.8%	91.8%

4.4.3 Discussion

The statistics given in table 24 and 25 do not serve our purpose of verifying the results generated by Bond and Paik (2000) by comparing it with corpus data. The difference between the corpus data and Bond and Paik's result has little to do with the accuracy of Bond and Paik's results of generating classifiers using semantic classes. We have problems with dealing with several types of constructions, which we have mentioned before. For example, temporal expressions are frequently identified as the referents of classifiers because temporal expressions are tagged as nouns and they fill positions filled by the referent of a classifier, as illustrated by sentence 42. We cannot find matches between the referent-classifier pairs extracted from the corpora and Bond and Paik's results not only because of these problematic constructions, which we would not repeat again, but also because of several other factors.

First, Bond and Paik's algorithm is meant to apply to kind classifiers and shape classifiers mainly. According to their typology, these two types are under a super-type *sortal*. Sortal classifiers and container classifiers count their referents as object units in a specific manner, as show in table 26:

Table 26 Japanese classifiers that count their referents as object units

REFERENT	CLASSIFIER	TYPE
飛行機	台	Kind
シャツ	着	Shape
ウイスキー	杯	Container

Classifiers that count other aspects of their referents are not included in the referent-classifier pairs given by Bond and Paik (2000). This can be shown by table 27:

Table 27 Cases where the results produced by Bond and Paik is different from the corpus data

Source	Referent	Classifier	Type	Quantification
BTEC	コイン	セント	Measure	Value
B + G	コイン	枚	Shape	Object Unit
BTEC	ビタミン	種類	Taxonomic	Type
B + G	ビタミン	個	Shape	Object Unit
BTEC	オレンジ	ダース	Group	Group
B + G	オレンジ	本	Shape	Object Unit
BTEC	ガソリン	ドル分	Portion	Portion
B + G	ガソリン	滴	Kind	Object Unit
BTEC	出張	日	Temporal	Time
B + G	出張	回	Event	Object Unit
B + G: the results produced by Bond and Paik (2000)'s algorithm using semantic classes from the ontology provided by Goi-Taikei				

This explains why we can find some matches between the results produced by

Bond and Paik (2000) and the data extracted from the corpora for referents counted by Japanese kind classifiers, shape classifiers and container classifiers in the corpus but we find almost no match between the results produced by Bond and Paik (2000) and the data extracted from the corpora for referents counted by Japanese measure classifiers, taxonomic classifiers, portion classifiers and temporal classifiers.

Another factor that accounts for the small number of matches found is the difference between the sense of a word found in a referent-classifier pair given by Bond and Paik (2000) and the sense of the same word found in the Japanese corpus.

Table 28 Words used in Bond and Paik with senses different from the senses in which they are used in the BTEC corpus

SOURCE	REFERENT	SENSE	CLASSIFIER
BTEC	飛行機	Flight	便
B + G	飛行機	Plane	台
BTEC	プログラム	Program Guide	冊
B + G	プログラム	Program	つ
BTEC	ファストクラス	First-class seats	席
B + G	ファストクラス	First-class	None
BTEC	オレンジ	Orange	個
B + G	オレンジ	Orange Juice	本
BTEC	寝台車	Sleeper Ticket	枚
B + G	寝台車	Sleeper	台
BTEC	ツイン	Twin room	部屋
B + G	ツイン	Twin brothers	人

Finally, there are some referent-classifier pairs produced by Bond and Paik (2000)'s algorithm using semantics classes which may be unacceptable. These pairs are given in table 29:

Table 29 Possibly unacceptable referent-classifier pairs produced by Bond and Paik using semantic classes

REFERENT	SENSE	CLASSIFIER	LANGUAGE
チーズバーガー	Cheeseburger	枚	Japanese
ハンバーガー	Hamburger	枚	Japanese
ホットドッグ	Hotdog	枚	Japanese
フィルム	Film	台	Japanese
車	Car	個	Japanese
靴	Shoes	副	Chinese
サイズ	Size	次	Chinese
便	Flight	块	Chinese

5. Conclusion

In this paper, we have presented our analysis of numeral classifiers extracted from Japanese, Korean, and Chinese corpora. We have compared how numeral classifiers are matched with their referents in our corpora with the results produced by the algorithm given in Bond and Paik (2000) for generating classifiers using semantic classes from an ontology provided by the Goi-Taikei.

First of all, we have typed each of the Japanese classifiers found in the Japanese corpus according to the categorization used in Bond (2001) and Bond and Paik (2004). We divide the numeral classifiers into ten types and also count the frequency for multi-type Japanese classifiers to be assigned to each of its possible types and the number of Japanese sentences containing a certain type of classifiers. Such information will be useful for developing statistics model for the same purpose. But we believe the significance of our work lies in providing a test bed for the typology. We have shown that classifying the classifiers according to the typology proves to be useful for analyzing the problems we encounter when translating Japanese classifiers. We believe that we can use the typology to design more powerful algorithms that match the semantic class of a referent to the type of a Japanese classifier for constructing semantics of or parsing Japanese sentences containing classifiers.

We have pointed out a number of problematic syntactic structures that prevent us from identifying the correct referent of a classifier found in a Japanese sentence. They include temporal expressions filling the position of the referent in a predicative construction, the possessive construction, the absent referent construction and noun phrases pre-modified by adjectival nouns. We believe that they are loopholes that future work on constructing semantics of or parsing Japanese sentences containing classifiers must take note of.

We have also obtained a list of the Korean classifiers used in sentences taken from each of the Korean corpora for translating the Japanese sentences containing classifiers in the Japanese corpus. The number of classifier tokens and unique classifiers found in each of the Korean corpora is obtained in the course of this work. These figures and the lists of Korean classifiers will be

useful for examining the quality of the Korean translation.

The most important finding of our work is the tables that encode the relation between Chinese classifiers and Japanese classifiers. We aim to make these tables exhaustive, although we may have left out a few pairs because of bugs in our programs. We have made some attempts at theorizing the relations. An important observation that we have made is that the classifier used in the source text can be of a type different from that of the classifier used in the target text. Looking at the list of referents taken by the default classifier(s) of each of the languages, we also come to be aware that the gaps left by the specific classifiers for the default classifiers to fill in can be very different in each of the languages. In addition, we get the impression that Japanese is more likely to use a specific classifier without a referent whereas the Chinese equivalent for sentences constructed this way would very likely involve the use of the default classifier with a referent. Another interesting difference we have found between the two languages is that Japanese has a compound classifier construction which has no equivalent in Chinese.

Like Japanese, Chinese also gets some problematic constructions that prevent us from effectively analyzing sentences containing classifiers. Two such constructions are the absent numeral construction and the absent classifier construction. We have given a description of them.

After a comparison of the data extracted from the corpora with the results generated by Bond and Paik (2000), using semantic classes from an ontology provided by Goi-Taikei, we suggest several factors that explain why such comparison is not appropriate. One such factor is, needless to say, the many problematic constructions mentioned in this paper that have prevented us from identifying the correct referents for a number of classifiers. Another factor is that Bond and Paik's algorithm is built for generating sortal classifiers but not all types of classifiers. A third factor is a small number of possibly unacceptable referent-classifier pairs (Bond and Paik (2000)). The last factor is the difference between a referent-classifier pair extracted from the corpora and a referent-classifier pair from Bond and Paik (2000) in the sense that the referent is used. Such difference is due to the polysemies as shown in table 28 and this is likely to be attributed to the characteristics of travel dialogues. For example,

counting flights is more likely than counting planes in travel situation. This points to the possibility of using some statistical word sense disambiguation techniques for resolving the referent of a classifier.

For further work, we suggest analyzing classifier constructions using statistical method based on the data produced here and applying word sense disambiguation techniques to the referents.

References

Yoshimi Asahioka, Hideki Hirakawa, and Shin-ya Amano. 1990. Semantic classification and an analyzing system of Japanese numerical expressions. *IPSJ SIG Notes*, 90-NL-78, pp.129-136

Francis Bond. 2001. *Determiners and Number in English, contrasted with Japanese, as exemplified in Machine Translation*. Ph.D. thesis, University of Queensland, Brisbane

Francis Bond and Kyonghee Paik. 2000. Re-using an ontology to generate numeral classifiers. In *18th International Conference on Computational Linguistics: COLING-2000*, pp. 90-96. Saarbruecken, Germany.

Francis Bond and Kyonghee Paik. 2004. Yet another classification of numeral classifiers, Unpublished paper

Pamela Downing. 1996. *Numeral Classifier Systems: The Case of Japanese*. John Benjamins, Philadelphia

Kyonghee Paik and Francis Bond. 2001. Multilingual Generation of Numeral Classifiers using a Common Ontology. In *International Conference on Computer Processing of Oriental Languages: ICCPOL-2001*, pp.141-147. Seoul, Korea.

Kyonghee Paik, Kiyonori Ohtake, Francis Bond, and Kazuhide Yamamoto. 2004. Source Language Effect on Translating Korean Honorifics. In CICLing-2004. pp. 334-337. Seoul, Korea.

Kuo-Ming Sung. 1996. Classifier incorporation in Japanese and Korean partitive constructions. In Noriko Akatsuka, Shoichi Iwasaki and Susan Strauss (eds.) Japanese/Korean Linguistics. 5: pp. 369-385. CSLI