Cross-Linguistic Error Types of Misused Chinese Based on Learners' Corpora[1]

Keiko MOCHIZUKI[∗], Hiroshi SANO[∗], Ya-Ming SHEN[∗] and

Chia-Hou WU[+]

Abstract

This paper presents an empirical study on the difficulties in learning Chinese as a second language based on learners' corpora written by native English speakers and native Japanese speakers at CEFR-based A2 and B1 levels. The first part of this paper will discuss the procedures for how to collect learners' corpora, proofread, establish an error tag system and annotate errors. Later it will focus on a significant difference in the production of " 一 + Classifier" among the corpora of native English speakers and native Japanese speakers. The corpus of English native speakers   displays an overuse of " 一 + Classifier", even in an atelic context like a negative construction or a conditional construction where a " 一 + Classifier" should not occur. On the other hand, the corpus of Japanese native speakers displays a lack of " 一 + Classifier". This striking contrast is due to whether or not a determiner position exists in each language. Since English has a determiner position which accommodates an article, "a/an, the", "this/that/my/your/~'s", English-native learners tend to treat the " 一 + Classifier" as an article although it does not appear in an atelic event structure. On the other hand, Japanese does not have any determiner position before a Noun Phrase, therefore it is assumed that

[∗] Tokyo University of Foreign Studies

  E-mail: mkeiko@tufs.ac.jp; sano@tufs.ac.jp; yamingshen@lingua-house.jp

[+] National Taiwan Normal University

  E-mail: clothoray@gmail.com

Japanese learners find it difficult to learn the conditions where a " 一 + Classifier" is necessary.

Keywords: Learner's Corpus, Annotation System, Error Analysis, Online Dictionary of Misused Chinese based on Learners' Corpora, Interference of Mother Tongues.

1. Objectives of Constructing the Learners' Error Corpus

The purposes of constructing the Learners' Error Corpora can be divided into two categories. The first is to discover the errors made by advanced-level learners since we assume that these errors reflect grammatical difficulties, significant differences in conceptual representation between the target language and the native language, and a different focus of representation despite relatively easy sentence structures. We believe that lexical/syntactic areas that are difficult to learn are caused by cases where the natural language system itself is difficult and where translation is difficult due to negative transfer. Clarifying these differences will lead to improvements in language teaching materials.

The second purpose of the research is to obtain new findings for comparative linguistics. The error analysis of cross-linguistic learners' corpora will enable us to distinguish language-specific error types based on the learners' native language and universal error types which occur regardless of the learners' native languages. Distinguishing these two features will also lead to the improvement of language teaching methodologies, especially those based on comparative perspectives between the learners' native language and the target language.

2. Procedures

2.1 The 'Full Moon' Learner Corpus of Chinese at Tokyo University of Foreign Studies

The characteristics of the data set of the 'Full Moon' Learner Corpus of Chinese at Tokyo University of Foreign Studies (henceforth 'Full Moon Corpus') are as follows:

Table 1. Learner Corpus of Chinese 'Full Moon Corpus' at Tokyo University of Foreign Studies (TUFS), collected May 2013-August 2014.

Academic Year	Level Chinese Major Students	Number of essays	Approximate number of words	Number of students
2013	Advanced (4th year)	95	45,500	35
	Intermediate (2nd/ 3rd year)	132	51,200	58
2014	Advanced (4th year)	21	12,500	23
	Intermediate (2nd/ 3rd year)	34	25,100	69

These compositions are proofread by Chinese native speakers with an MA. or Ph.D in linguistics/language education and sufficient experience in teaching Chinese at university level. Proofread compositions clearly indicate errors and corrections so that the errors can be identified within the respective sentences.

The 'Full Moon Corpus' includes learner's information as shown in Table 2.

Table 2. Example of Learner's Profile

1	Learner's ID	Th_Ch_001
2	Name	Tokyo Taro
3	Major	Chinese
4	Year	3
5	Gender	male
6	Age	21
7	Nationality	Japan
8	Residential History	Canada 4-9；Japan 0-4,9-21
9	Native Language	Japanese
10	Language of Education	Japanese, English
11	Length of Chinese study	3 years and 2 months
12	Institution	Tokyo University of Foreign Studies
13	Study Abroad Experience Institution / Period	Mandarin Center, National Taiwan Normal University, August1-31. 2014
14	Speaking with my family	Japanese
15	Speaking with friends	Japanese
16	Language used in Elementary School	5-9 English, 9-12 Japanese
17	Language used in Junior High School	Japanese, English
18	Language used in Senior High School	Japanese, English
19	Test of Chinese as a Foreign Language (TOCFL)	Band B(2014)
20	HSK 汉语水平考试	5 级　(2012)
21	English TOEFL(iBT)	108　(2013)
22	TOEIC	955　(2012)
23	IELTS (academic)	8.0　(2013)

The 'Full Moon Corpus' has four key features : 1) compositions are written by experienced learners majoring in Chinese in Japan, 2) compositions go through an appropriate proofreading process conducted by university teachers, 3) errors and corresponding corrections are recorded, and 4) the detailed profiles of the learners are also recorded.

2.2 Error Tag Categories

There are two tag categories for misuse: Error and Modify. The Error tag indicates grammatical errors while the Modify tag indicates inappropriate use of expressions ('expression' tag), punctuation and Chinese characters as shown in Figure 1.

Interpretation Framework (error tags)
Classification and in-text marking of syntactical, lexical, stylistic, rhetorical and notational misuses
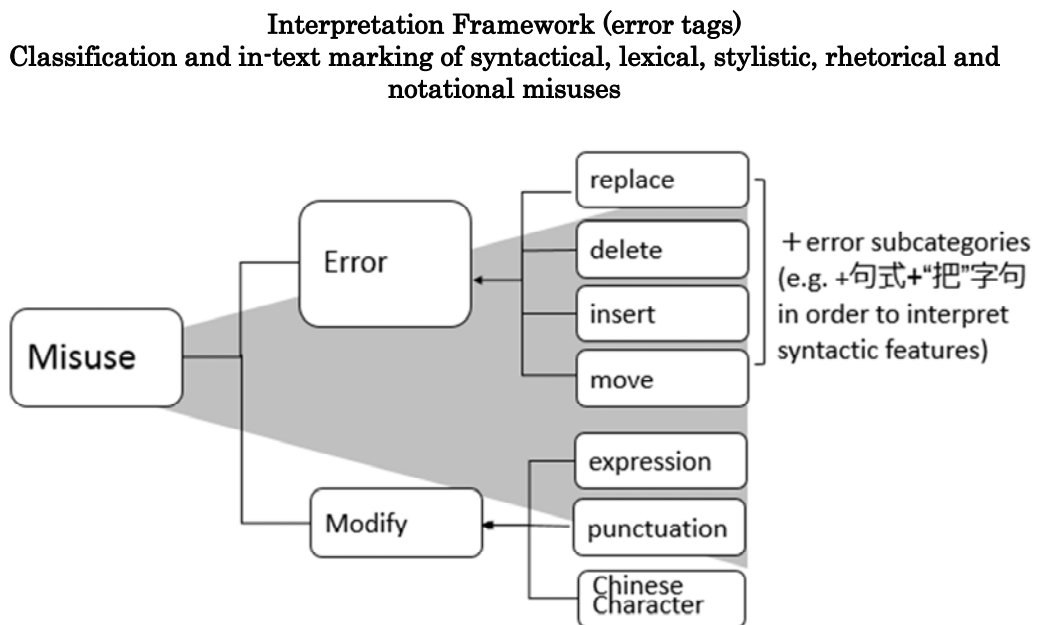

Figure 1. Misuse Tag System

The Error tag consists of the following four sub-categories: Replace, Delete, Insert and Move. The Replace tag indicates the need to replace an error with another correct expression. The Delete tag indicates that deleting an error will lead to a correct expression. The Insert tag indicates that inserting a new expressions will lead to a correct expression. The Move tag indicates a word order error.
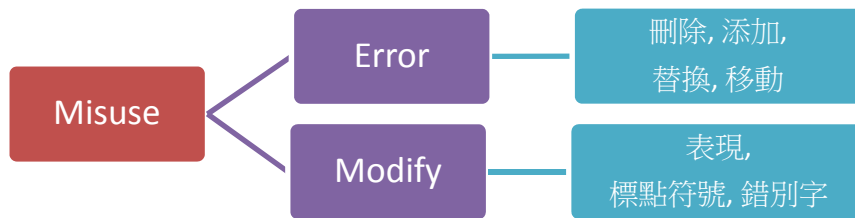
The Modify tag consists of the following three sub-categories: Expression, Punctuation and Chinese character. The Expression tag indicates that it is preferable to use another expression or that the misuse cannot be categorized as any one specific error. The Punctuation

tag indicates the need for correction in view of the style of writing. The Chinese character tag indicates the misuse of a Chinese character.

As subcategories of the error tag, we have designed the 74 tags as shown in (1) referring to the grammatical system in『新編現代漢語』(張斌、齊滬揚等，2002: 273-467).

(1) Tag List in Chinese

   A. Subcategories of Misuse

```
                    ┌─── Error ───── 刪除, 添加,
                    │                 替換, 移動
        Misuse ─────┤
                    │
                    └─── Modify ──── 表現,
                                     標點符號, 錯別字
```

   B. Subcategories of Error

	大分類	小分類
1	名詞	時間名詞, 處所名詞, 方位詞
2	數詞	
3	量詞	
4	動詞	狀態動詞, 動作動詞, 存現動詞, 關係動詞, 能願動詞, 趨向動詞, 使令動詞
		及物動詞, 不及物動詞, 雙賓動詞
		重疊動詞
5	形容詞	
6	副詞	程度副詞, 範圍副詞, 時間副詞, 情態副詞, 否定副詞, 語氣副詞, 關聯副詞
7	代詞	人稱代詞, 指示代詞, 疑問代詞
8	連詞	
9	介詞	
10	助詞	結構助詞, 時態助詞, 時制助詞, 比況助詞, 表數助詞, 列舉助詞, 語氣助詞, 其他助詞
11	短語	量詞短語, 方位短語, 介詞短語, "的"字短語
12	主語	

13	賓語	雙賓語
14	補語	結果補語, 趨向補語, 可能補語, 程度補語, 情態補語, 數量補語, 介詞, 短語補語
15	疑問句	是非問句, 特指問句, 選擇問句, 正反問句
16	句式	主謂謂語句, "把"字句, "被"字句, 連動句, 強調句, 兼語句, 使役句, 存現句, 比較句, "連"字句
17	複句	並列複句：承接複句, 遞進複句, 選擇複句, 注解複句
		偏正複句：因果複句, 條件複句, 轉折複句, 讓步複句, 目的複句

2.3 Method of Proofreading and Annotation

We use the 'TNR_Chinese Writing Correction2014' and 'TNR_Chinese Error Corpus Tagger2014' (2014) tools developed by 于康(Yu Kang) and 田中良(Ryo Tanaka) for proofreading and annotation. The procedures are as follows. First, compositions written by learners in a WORD file are converted to text files. Next, errors and the corresponding corrections are added to the composition texts using the 'TNR_ Chinese Writing Correction 2014' system. The following figure 2 is an example of proofreading using 'TNR_ Chinese Writing Correction 2014'.

Figure 2. Proofreading System

The 'TNR_ Chinese Writing Correction 2014' system displayed in Figure 4 has two windows: the left window displays the composition text and the right window displays corrections. Each correction in the right window and its corresponding error expression in the left window are marked up in the same color for better visibility.

For annotation, 'TNR_Chinese Writing Correction2014' and 'TNR_Chinese Error Corpus Tagger2014' (2014) enable free creation of tags and the displaying of a tag list underneath the composition text as shown in Figure 3.

Figure 3. Annotation System: Tag Buttons

The first step in annotating a composition is to designate the region of each misused expression in the composition text. The second step is to choose one of 'Replace 替换, Delete 删除, Insert 添加, Move 移動, Expression 表現, Punctuation 標點符號, Chinese Character 錯别字' and click on the appropriate button. This procedure enables annotations to be made

automatically. The third step is to choose one of the error subcategories, e.g. 'Resultative Complement 結果補語'. This click-annotation system greatly reduces the burden of annotation. 'TNR_ Chinese Writing Correction 2014' also has the function to convert annotated data into XML data.
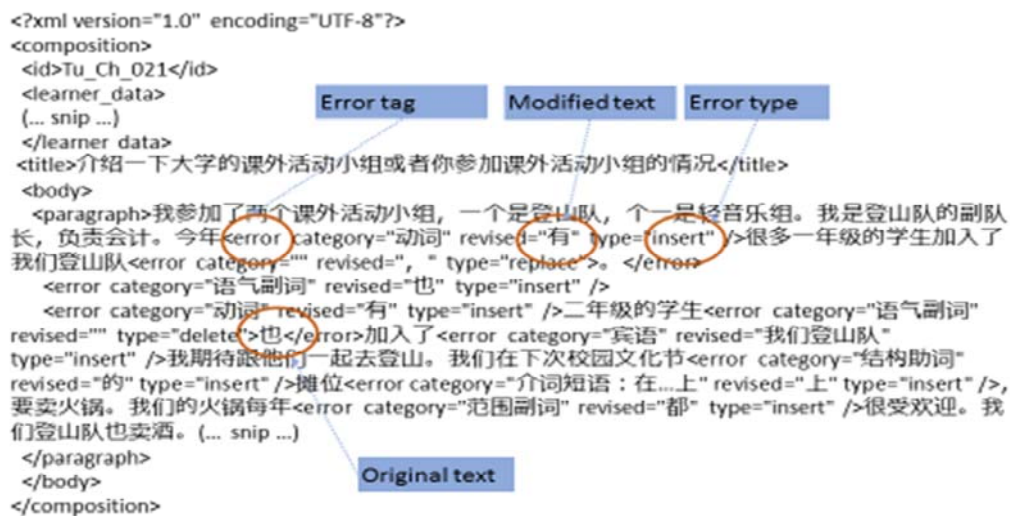

Figure 4. Digitization Framework (XML Data)

3. Cross-linguistic Analysis of Errors

We will discuss two significant error types in two learners' corpora by comparing The Full Moon Corpus written by Japanese native speakers at TUFS with the TOCFL learners' corpus of Chinese written by English native speakers (henceforth, TOCFL corpus)[2]. (張莉萍 Chang Li-Ping:2013)

Table 3. the TOCFL English-Native Learners' Corpus of Chinese

TOCFL (CEFR)	Number of Compositions	Number of Chinese characters	Number of Students
基礎(A2)	223	119,971	223
進階(B1)	344	31,852	344

3.1 Classifier Phrase(量詞短語) " 一 + Classifier(量詞)"

One of the most significant error categories observable in The Full Moon Corpus is the lack of " 一 + Classifier(量詞)" while the TOCFL Corpus displays an overuse of " 一 + Classifier(量詞)".張莉萍 Chang Li-Ping(2014:68) also indicates the same contrast between English-Native learners and Japanese-Native learners.

     Table 4 compares the frequency of " 一 + Classifier '-ge 個' " in The Full Moon Corpus and the TOCFL Corpus.

Table 4. the Frequency of " 一 + Classifier '-ge 個' "

	CEFR Level	Number of Chinese characters	Occurrence of "一個"
The TOCFL English-Native Learners' Corpus	B1	119,971	586 tokens
	A2	31,852	159 tokens
	Total	151,823	745 tokens 1,490 Chinese characters
The Full Moon Japanese-Native Learners' Corpus	A2-B1	134,094	385 tokens 770 Chinese characters

     Table 4 shows an interesting contrast in the frequency of "一個" between The TOCFL English-Native Learners' Corpus and The Full Moon Japanese-Native Learners' Corpus. The TOCFL English-Native Learners' Corpus displays a higher frequency than The Full Moon Japanese-Native Learners' Corpus. Upon conducting a chi squared test, a significant difference between the data sets was discovered (0.1%, x^2=150.03, p=0.000).

3.2 Lack of " 一 + Classifier" : Japanese Learners

Let us examine the lack of " 一 + Classifier(量詞)" in The Full Moon Japanese-Native Learners' Corpus. The following examples (2) to (18) show that each sentence lacks the bracketed " 一 + Classifier" in The Full Moon Japanese-Native Learners' Corpus. There are almost no examples of overuse of " 一 + Classifier" in The Full Moon Japanese-Native Learners' Corpus.

   (2) Copula "是 Shi" Construction:

     'Topic(Old Information) + "是 Shi"+ Comment(New Information)'

     a. 我認爲這是(**一種**)有益的愛好。

     b. 但是，生孩子是(**一件**)不簡單的事。

c. 從前我去過京都，京都<u>是</u>**(一個)**很美麗的地方。

d. 但是，找到好工作並不<u>是</u>**(一件)**好的事情。

e. 小學生跟他們交流<u>是</u>**(一個)**好機會，但是，孩子們聽得懂他們的課嗎？

f. 現在，環境問題<u>是</u>世界的**(一個)**很大的課題。

(3) Existential "You 有" Construction

a. 東大和<u>有</u>**(一個)**很大的公園---東大和南公園，附近也有**(一條)**小河。

b. 這台電視還<u>有</u>**(一個)**功能,那就是聽音樂。

(4) Perfective Construction with "-le 了"

這幾年，一位很有名的漫畫家畫<u>了</u>**(一部)**跟帶廣的挽曳賽馬有關的漫畫。

(5) "Give" Construction and "Become" Construction

a.比如，開始工作掙錢以後，我想<u>送給</u>父母**(一份)**禮物，例如，海外旅行。

b.我也在留心這些事情，希望能<u>成為</u>**(一個)**很好的領導!!

(6) Presentative Construction

最近他在車站附近<u>開了</u>**(一家)**中餐館。

(7) Resultative/Directional Verb Compound

其他同學也<u>舉出</u>**(一些)**有意思的食物，比如納豆，豆漿等等。

(8) "Modifier +的 DE+ Noun"

a.原來有很多溫泉的日本的**(一個)**特色就是飯店旅館業很發達。

b.在我的印象裡很深的**(一件)**事是小學 5 年級的時候媽媽幫助我練習跑步。

c.去年網絡上的**(一篇)**文章"中國女性和日本女性的一生"引人註目。

(9) 'Source' with New Information:

那個名字來源于**(一條)**從南到北延伸的坡道。

The reason why it is very difficult for Japanese learners of Chinese to learn the principle of " 一 + Classifier" is because Japanese grammar is insensitive to 'Boundedness' (有界性) which controls the occurrence of " 一 + Classifier".

Shen(沈家煊) (1995) discusses the interaction between " 一 + Classifier" and the concept of 'bounded' and 'unbounded' events. Shen (1995) indicates that a " 一 + Classifier" is necessary before a 'bounded' Noun Phrase(NP) in 'Telic' events as follows:

(10) Indirect Object in a Move Construction:

    a. 盛碗裡兩條魚。

    b. *盛碗裡魚。

(11) Resultative Object (結果賓語)

    a. 蚊子叮了小王兩個大包。

    b. *蚊子叮了小王大包。

(12) Resultative Complement (結果補語)

    a. 打破兩塊玻璃。

    b. *打破玻璃。

(13) Directional Complement(趨向補語)

    a. 飛進來一個蒼蠅。

    b. *飛進來蒼蠅。

(14) Verb+ "-le 了" construction

    a. 吃了一個蘋果。

    b. *吃了蘋果。

Shen (1995)'s "bounded/unbounded" theory can explain why the following types, (4) Perfective Construction with "-le 了", (5) GOAL in "Give" Construction and "Become" Construction, (6) Presentative Construction and (7) Resultative/Directional Verb Compound require " 一 + Classifier" since all cases in (4)(5)(6)(7) have "telicity", the subcategory of "bounded" concept in the temporal structure.

In (2) Copula "是 Shi" Judgement Construction and (3) Existential "You 有" Construction , " 一 + Classifier" often appears after "是 Shi" /"You 有". Both constructions have the following informational structure:

(15)

"是 Shi"/ "You 有"Construction	Topic	"是 Shi" "You 有"	" 一 + Classifier" NP
1)Informational Structure	Old Information		New Information
2) Boundedness			Bounded

It is supposed that the NP with new information is a bounded entity, because the NP with new information is a focus in terms of cognition.

3.3 Overuse of " 一 + Classifier(量詞)" :English-Native Learners

We find the reverse phenomenon in The TOCFL English-Native Learners' Corpus: the overuse of " 一 + Classifier". The following examples (16) to (23) show that the bracketed " 一 + Classifier" should be deleted .

(16) Conditional:

有什麼問題就跟我打(**一通**)電話吧！

(17) Plan:

我們游完泳我<u>計畫</u>我們去電影院看(**一部**)電影。

(18) Potential:

a.我們<u>也</u><u>可以</u>去「西門町」看電影，打撞球，或去(**一個**)茶店談天說笑。

b.你看我已經<u>可以</u>用中文寫(**一封**)信，…

(19) Future Activity:

我記得你說過你喜歡丟飛盤，所以我<u>會</u>把(**一張**)飛盤帶來。

(20) Topic Noun in "是 Shi" construction:

我媽媽上上個週末來台灣看我。我們去的(**一個**)地方是花蓮。

(21) "When" Clause: Old Information

你開(一個)慶祝會的時候我不能參加是因為我在外國工作。

(22) Negation: "沒(有)"

a.我在台北沒有發生(一個)大問題，…….

b.他們有一個農場，我去他們的家以前，還沒去(一個)農場…

(23) Missed Action:

今天他不但忘了帶手機，也忘了帶(一瓶)水。

It seems that the interlanguage of Chinese created by English native speakers displays the following incorrect overgeneralization:

(24) Overgeneralization by English-native learners of Chinese

a/an   NP   =   " 一 + Classifier" NP

Shen (1995)'s "bounded/unbounded" theory can also explain why " 一 + Classifier" cannot appear in (16) to (23): all cases express atelic events and an entity in an atelic event should be unbounded. Shen (1995) indicates that a " 一 + Classifier" cannot appear in the following atelic structures.

(25) Verb Reduplication(動詞重疊式):

a. (*) 今天要談談兩個問題。

b. *星期天在家洗洗一件衣服。

(26) Durative Aspect Marker "-Zhe 著"

a. Progressive Aspect: *他正吃著三碗飯。

b. Resultative State: *山上架著兩門炮。

(27) Negation:

a.*今天不談兩個問題。

b.*這個月不演三場電影。

3.4 Comparative Analysis of Error Types by Japanese Learners and English-Native Learners

The contrast between the lack of " 一 + Classifier" in The Full Moon Japanese-Native Learners' Corpus and the overuse of " 一 + Classifier" in The TOCFL English-Native Learners' Corpus suggests a difference in Noun Phrase Structures in Chinese, English, and Japanese.

Japanese syntax has no 'functional category', therefore there is no syntactic node (i.e. 'determiner') to accommodate a constituent like "a/an, the" while English has 'determiner' as Fukui (1995) proposes. This syntactic difference between English and Japanese causes the contrast between the lack and the overuse of " 一 + Classifier" in Japanese-native learners and English-native learners.

In addition, Ikegami（池上）(1981), (2007) and Kageyama（影山）(1997), (2002) suggest that Japanese is an "unboundedness-oriented" "less-individualization" type language in terms of having no grammatical category of number, ellipsis of subject/object, and no determiner node. This "unboundedness-oriented", "less-individualization" feature is reflected in second language acquisition of Chinese and English by Japanese learners. Since Japanese grammar has no syntactic strategy to individualize an entity/event, it is very difficult to acquire both the principle of " 一 + Classifier" NP which appears in an bounded/individualized noun, and the usage of the articles "a/an, the" in English. According to "NTNU/TUFS Sunrise Learners' Corpus of English", the most frequent error category in the Japanese-native learners corpus is articles "a/an, the" as shown in "TUFS Online Dictionary of Misused English" :

http://sano.tufs.ac.jp/lcshare/htdocs/?lang=english

On the other hand, English is a "boundedness-oriented" "high-individualization" type language in terms of having an obligatory grammatical category of number, determiner node, and an obligatory subject/object. The reason why the English-native TOCFL corpus displays an overuse of " 一 + Classifier" is because the principle of individualizing a noun is different between English and Chinese. Chinese cannot individualize a noun in an atelic unbounded event like a future event, a potential, a negation, a missed action or a conditional. On the other hand, in English, each noun is itself classified according to its property: countable or uncountable. The principle of individualization in English is not controlled by "Bounded/Unbounded" cognition.

Figure 5. TUFS Online Dictionary of Misused English

4. Conclusion

This paper introduced an empirical study on the difficulties in learning " 一 + Classifier(量詞)" in Chinese based on learners' corpora written by English-native learners and Japanese-native learners at CEFR-based A2 and B1 level. The interesting contrast between The TOCFL English-native learner's corpus and The Full Moon Japanese learners' corpus is the overuse and the lack of " 一 + Classifier".

The overuse of " 一 + Classifier" in the English-native TOCFL corpus is due to the overgeneralization by English-native learners of Chinese that "a/an NP" is equivalent to " 一 + Classifier" NP. On the other hand, the lack of " 一 + Classifier" in The Full Moon Japanese learners' corpus is due to the lack of individualization in terms of cognition in Japanese. The different features of the three languages are summarized below:

(28) **Different Features in Number, Classifier and Degree of Individualization**

	1)Number(Singular/Plural)	2) Classifier	3) Degree of Individualization
English	Obligatory	No Classifier	High
Chinese	None except for 我們/這些	Rich system	Moderate " 一 + Classifier" occurs in a "bounded" cognition
Japanese	None except for Watashi-**tachi**(we), kore-**ra** (these)	Not as rich a system as in Chinese	Low No article No determiner in syntax

This comparative research into cross-linguistic learners' corpora suggests that it is indispensable to explore the pedagogy of Chinese based on leaners' native language to develop more efficient and advanced learning science.

References

Fukui, N. (1995). *Theory of projection in syntax.* Stanford: CSLI Publications, Tokyo: Kurosio Publishers.

Huang, C. T. J., Li, Y. H. A., & Simpson, A. (2014) *The Handbook of Chinese Linguistics*, John Wiley &Sons.

Ikegami, Y. (1981). *Suru (DO)" vs "Naru (BECOME)"in the Typology in Language and Culture*. Tokyo: Taishukan Publishers. [池上嘉彦 (1981)。「する」と「なる」の言語学 ： 言語と文化のタイポロジーへの試論。東京:大修館書店]

Ikegami, Y. (2007). *Japanese and Japanese Typology*. Tokyo: Chikuma Publishers. [池上嘉彦 (2007)。日本語と日本語論。東京:筑摩書房]

Kageyama, T. (1996). *Verb Semantics: the Interface between Language and Cognition.* Tokyo: Kuroshio Publishers. [影山太郎 (1996)。動詞意味論―言語と認知の接点。東京: Kuroshio Publishers]

Kageyama, T. (2002). *Japanese as a unbounded language.* Tokyo: Iwanami Publishers. [影山太郎 (2002). けじめのない日本語。東京：岩波書店]

Li, C．N. & Thompson. S. A. (1981). *Mandarin Chinese: A Functional Reference Grammarc(漢語語法).* Taipei:Crane Publishing Co. Lid.

Mochizuki, K. (2004). *Causative and Inchoative Alternation: Comparative Studies on Verbs in Chinese and Japanese.* Ph.D dissertation, National Tsing Hua University, Taiwan. http://140.113.39.130/cgi-bin/gs32/hugsweb.cgi/ccd=lJoInl/record?r1=2&h1=0

Shen, Y.-M. (2009). *Resultative Compound Verbs in Chinese- From a Viewpoint of Comparative Analyses with Resultative Compound Verbs in Japanese and English Resultative Constructions.* Ph.D dissertation, Tokyo University of Foreign Studies.

http://repository.tufs.ac.jp/handle/10108/56738

Tai, J. H.-Y. (1984). Verbs and Times in Chinese: Vendler's Four Categories. *Lexical Semantics,* 289-296, Chicago Linguistics Society.

Vendler, Zeno. (1967). *Linguistics in Philosophy*. Cornell University Press.

沈家煊 (1995)。"有界"与"无界"。中国语文,第 5 期,367-380。 [Shen, J.-X. (1995). Boundedness and unboundedness. *Chinese Language and Writing*, 5, 367-380]

鄧守信 (2009)。*對外漢語教學語法修訂二版*。台北:文鶴出版有限公司。[Deng, S.-X. (2009). *A Pedagogical Grammar of Chinese.* Taipei:Crane Publishing Co. Lid.]

張斌、齊滬揚等 (2002)。*新編現代漢語*。復旦大學出版社。

張莉萍(2013)。TOCFL 作文語料庫的建置與應用。載於崔希亮、張寶林(主編),*第二屆漢語中介語語料庫建設與應用國際學術討論會論文選集*(頁 141-152)。北京:北京語言大學出版社。

張莉萍(2014)。不同母語背景華語學習者的用詞特徵:以語料庫為本的研究。*中文計算語言學期刊(IJCLCLP)*,19(2),53-72。