A Corpus-Based Study of Restrictive Relative Clauses

Hui-Chuan Lu & Yun-Hui Chen

National Cheng Kung University, TAIWAN

Abstract

This paper aims to investigate the similarities & differences of Restrictive Relative Clauses (RRC) among 3 languages by comparing & contrasting parallel data extracted from a POS-tagged multilingual corpus. This research further provides examples for corpus-based language analysis & application of SLA.

This investigation consists of three major works. First, we construct a POS-tagged multilingual parallel corpus (CPEC) in order to search parallel translations of 3 languages. Then, based on Keenan & Comrie's Accessibility Hierarchy (1977), we study RRCs of Spanish, English & Chinese separately & in parallel by using the concordance of WordSmith & ParaConc to analyze data extracted from our created corpus. Finally, we apply the result to the area of SLA by contrasting the result of parallel data & that of LL.

This study reaches the following conclusions. In the analysis of translations, Spanish RRCs with *que* are mostly translated to other structures in English and to RRCs with *DE* in Chinese. In the contrastive analysis, the sequences of all AH are similar in Spanish and English (S>DO>PO>IO), but different from Chinese (DO>S). However, all 3 are different from Keenan and Comrie's AH. In the study of SLA of Spanish, the learner language shows that similar sequence of AH of Spanish and English of native language, and different from Chinese. We conclude that L1 doesn't affect the learner language while L2 might play a role.

1. Introduction

This paper aims to investigate the similarities & differences of Restrictive Relative Clauses (RRC) among 3 languages (Spanish, English and Chinese) by comparing & contrasting parallel data extracted from a POS-tagged multilingual corpus. This research further provides examples for corpus-based language analysis & application of SLA.

This investigation consists of three major works. First, we construct a POS-tagged multilingual parallel corpus (CPEC) in order to search parallel translations of 3 languages. Then, based on Keenan & Comrie's Accessibility Hierarchy (1977), we study RRCs (Restrictive Relative Clauses) of Spanish, English & Chinese separately & in parallel by using the concordance of WordSmith & ParaConc to analyze data extracted from our created corpus. Finally, we apply the result to the area of SLA by contrasting the result of parallel data & that of LL.

2. Previous studies

2.1. Creation of parallel corpus

To our knowledge, so far there is not any parallel corpus—that consists of Spanish-Chinese. However, we can find numerous parallel corpora either related to Spanish-English or English-Chinese. For instance, parallel corpora related to Spanish-English include: Reuters, MLCC, ECI, CRATER, Eur-LEX…and so on. But, compared to the quantity of English-related parallel corpora, Mandarin Chinese-related parallel corpora are much fewer: for example, LCMC, Multiple-Translation Chinese Corpus, Babel Chinese-English Corpus…etc.

Hence, the need of cross-linguistic research & the lack of existing parallel corpus motivate us to create an annotated corpus that compiles parallel translations of aligned texts of Spanish, English & Chinese. By experimenting with its construction, we will be able to provide examples

by sharing experience about the possibility & difficulties of creating such a corpus.

2.2. Corpus-based studies

With respect to the research dedicated to the studies on parallel corpus & translation, there are Altenberg et al. (2000), Schmied et al. (1996) and Santos (2004) among others, and most of them are related to the English language. Corpus-based contrastive studies of Spanish & Chinese translation are not so well-dedicated to the area of Corpus Linguistics. Compared to other areas, moreover, the themes & the amount of studies in the syntactic analysis are more limited, for example, Cermák and Klégr (2004), Uchida et al. (2002), and Santos (2004). Accordingly, the research related to syntactic analysis needs more attention and effort.

2.3. Relative clauses

Among the studies related to RRCs, the Accessibility Hierarchy (AH) proposed by Keenan & Comrie's (1977) is the most widely-discussed. The hierarchical order of relativization is Subject> Direct Object> Indirect Object> Genitive. Following AH, the study of Sheldon (1974), based on functional grammar indicating the interaction of antecedents & relative elements of RRCs. In addition, AH has been applied to account for NL as well as for LL. Furthermore, Oostdijk & De Haan (1994) investigated the word order of matrix & subordinate clauses of relative construction by adopting the corpus approach.

3. Study

3.1. Research questions:

In construction of RRCs corpus: in terms of native language, what are the differences among

3 languages, Spanish, English and Chinese? And how is Spanish translated to English & Chinese? As for learner language, how is the learner language of Spanish accounted for by contrasting 3 native languages?

3.2. Data

The corpus that we compile for the following analysis includes 3 sub-corpora containing the parallel and translated texts of 3 different languages, Spanish, English and Chinese. The main source is from *BibleGateway.com* (http://www.biblegateway.com/). Why do we choose *the Bible*? We choose *the Bible* as our data source for its being one of the richest & most accessible corpus. It is inevitable to deny that *the Bible*, as a data source, is controversial due to its comparably archaic language. Since the Bible was originally written in old Hebrew & ancient Greek, its language is admittedly different from nowadays. However, given the circumstances that the access to Spanish-Chinese translation texts is limited, the Bible is a valuable data resource. With the above considerations in mind, we choose *the Bible* as data base for the present research. Also in the area of corpus linguistics, the research value of *the Bible* does not go unrecognized. As a consequence, we decide to use texts from *the Bible* for our research.

In terms of *the Bible* version, in order to make a quasi-parallel comparison, we use the New International Version for English, la Nueva Versión Internacional for Spanish and the Union version for Chinese. In dealing with the texts, we extract from the New Testament the 4 gospels, Matthew, Mark, Luke, and John, along with the Acts of the Apostles; and process them into electronic format for 3 different sub-corpora. The total words or characters of each sub-corpus: 103,267 words for Spanish, 105,128 words for English and 133,078 characters for Chinese.

3.3. Methodology

In the process of creating the Spanish-English-Chinese parallel corpus, we collect data of multilingual translations from the internet and POS-tag the texts with linguistics-specialized tools. *WordSmith* & *ParaConc* are used to search appropriate data separately for analyzing RRCs of each language and contrasting 3 languages in parallel.

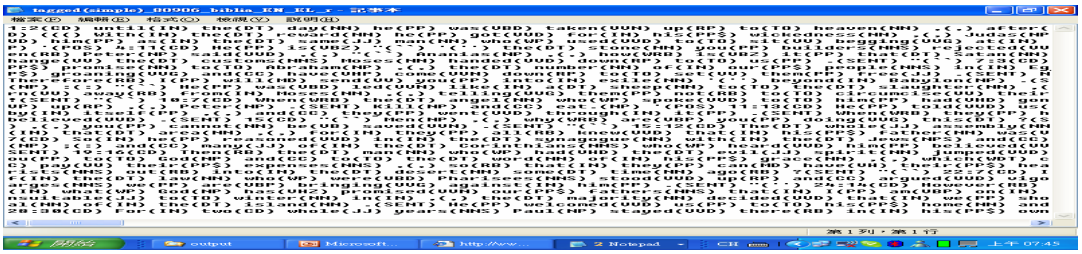
3.3.1. Construction of corpus

In compiling the cross-linguistic corpus, we first incorporate free and publicly accessible Spanish, English and Chinese electronic texts of *the Bible*, taken from on-line multilingual corpora *Bible Gateway*. By doing so, it spares us not only the time & labors of manual inputting but also helps facilitate the operation of Concordance when aligning the texts. And all the texts are saved in the format of .txt for the following research.

3.3.2. POS-tagging

In the process of POS-tagging, on one hand, Spanish and English texts are POS-tagged by *Tree Tagger* separately and the tagged results are simplified by our self-developed programming (Figure 1) in order to facilitate the visualized analysis after processing by *WordSmith* & *ParaConc*.

Figure 1: Simplification of tagged result



On the other hand, Chinese texts are to be word-segmented and POS-tagged with the aids of Chinese annotated system of Academia Sinica of Taiwan (http://rocling.iis.sinica.edu.tw/CKIP/).

Then, we use the relative pronouns, *que, that & DE*, as keywords to search in the concordance tool. However, different from previously untagged corpus which only allows the search for specific words, now we are able to filter inappropriate sentences more effectively—by setting the POS-tags of these keywords and their antecedents or words nearby as conditions. By using auxiliary tools, we search for the combination of structure 'N+*que/that* V' for English and Spanish, and '*DE* N' for Chinese.

After extracting the appropriate data, we annotate the relativized elements in the embedded clause according to their grammatical functions (Examples (1)-(4)). Then, we analyze the similarities and differences among 3 languages in the parallel contexts in order to answer our research questions.

Examples:

Spanish: Direct object.

(1) Le hemos oído decir que ese Jesús de Nazaret destruirá este lugar y cambiará las tradiciones que nos dejó Moisés . (Acts 6:14)

English: Subject

(2) Enter through the narrow gate. For wide is the gate and broad is the road that leads to destruction, and many enter through it. (Matthew 7:13)

Chinese: Subject

(3)他說：我祖亞伯拉罕哪，不是的，若有一個從死裡復活的，到他們那裡去的，他們必要悔改。(路加16:30)

Chinese: Direct object

(4)我卻不以性命爲念，也不看爲寶貴，只要行完我的路程，成就我從主耶穌所領受的職事，證明神恩惠的福音。(使徒行傳20:24)

3.3.3. ParaConc

By conducting the methodology in 3.3.2., we are able to observe how 3 languages assimilate and differ from each other with respect to the RRCs in general. Going one step further, we are interested in knowing how the structures of RRCs are translated to other languages. Thus we use ParaConc to facilitate the analysis of the translated structures.

Focusing on the Sp. RRCs data processed by *WordSmith*, we look for their translations in English and Chinese, and align sentences of 3 languages for further analysis.

4. Result and discussion

4.1. Restrictive relative clauses in 3 languages

Table 1

Distribution of RRCs in 3 languages

	Sentences	RRCs	%
Sp.	5594	443	7.9%
Eng.	6102	120	2.0%
Ch.	55864	1014	1.8%

From Table 1, we can observe that the difference between Spanish and Chinese (7.9% vs. 1.8%) shows the contrast between 2 languages and it might imply a certain difficulty of learning Spanish for Taiwanese learners.

4.1. Accessibility Hierarchy

Table 2

Distribution of relativized elements in 3 languages

	S	DO	IO	PO	Total
Sp.	317 (71.6%)	119 (26.7%)	1 (0.2%)	6 (1.4%)	443
Eng.	95 (79.1%)	23 (19.1%)		2 (1.7%)	120
Ch.	325 (32%)	672 (66.2%)	17 (1.7%)		1015

According to Table 2, we derive the following order in which the accessibility hierarchy are Prepositional Object> Indirect Object both in Spanish RRC with *que* and in English RRCs with *that*, contrary to Keenan & Comrie's AH. The points of view in the descriptive grammar have

been examined by the corpus approach and our result provides further evidence to modify the proposed argument at least for the data analyzed here. On the other hand, our attention has been drawn to the high accessibility of DO in Chinese RRCs with *DE* (DO>S>IO). Although Spanish and English behave differently with respect to their occurrences, they are similar in the sequence of the AH (S>DO>PO>IO).

The differences between Spanish and Chinese show the contrast between 2 languages and it might also imply a certain degree of difficulty of learning Spanish if these 2 structures are compared. What is more, the similarities between English and Spanish and the differences between Spanish and Chinese indicate that English RRCs with *that* can be assimilated to Spanish RRCs with *que* while they differ from Chinese RRCs with *DE*.

4.2. Translation

From the previous section, we see that how RRCs behave differently among languages. In this section, we will focus on the parallel translation of 3 languages.

Table 3

Distribution of Spanish RRCs translated in English

Eng.	who	that	which	what	where	VP	IS	∅	others	
#	24	9	1	1	0	11	5	20	35	106
%	22.6%	8.5%	0.9%	0.9%	0.0%	10.4%	4.7%	18.9%	33.0%	

With respect to the translation among different languages, Table 3 shows that (1) RRCs with *que* are translated to *who* most of the time (22.6%) and less frequently to *that* (8.5%) in English. (2)

More than half of the time (51.9%), RRCs in Spanish are translated to other structures in English.

Table 4

Distribution of Spanish RRCs translated in Chinese

Ch.	Suo…DE	DE	Conj	IS	others	
#	20	36	22	13	15	106
%	18.9%	34.0%	20.8%	12.3%	14.1%	

Furthermore, in Table 4, we can observe that there are at least 2 major patterns. (1) RRCs with *DE*: The word *DE* functions as connecting the subordinate clause and the nuclear element. The result shows that 52.9% of Spanish RRCs have been translated to Chinese RRCs with *DE* and 18.9% of them contains *SUO*. (2) The other pattern has changed the Spanish RRCs into different structures in Chinese. The result of analysis shows 33.1% of the Spanish RRCs have been translated to 2 different clauses or sentences with pronouns to replace the repeated noun in Chinese sentences.

5. Application to teaching

Based on the results from previous section, we would like to make a connection between the native language and the learner language. According to Lu (2007), the sequence concluded from data of CATE (Corpus de Aprendices Taiwaneses de Español) is: S (60.45%) > DO (34.09%) > PO (5.46%) > IO (0%). This sequence does not completely agree with the sequence of AH: S > DO > IO > PO > G (Keenan & Comrie, 1977).

It shows a similarity between Taiwanese learners of Spanish and Spanish native speakers in

this research in terms of syntactical functions within RRCs: S＞DO＞PO＞IO. However, the sequence is different from the result concluded from RRCs with *DE* in the parallel texts of Chinese. Hence, we might want to argue that the RRC with *DE* of L1 (Chinese) does not play a role in language learning while the RRCs with *that* in L2 (English) can be a positive transfer for the L3 Spanish learner.

6. Conclusion

In the analysis of translations, Spanish restrictive relative clauses with *que* are more frequently translated to other structures in English and to restrictive relative clauses with *DE* in Chinese. In the contrastive analysis, the sequences of AH of Spanish and English are similar (S>DO>PO>IO), but they are different from that of Chinese (DO>S). However, all 3 Accessibility Hierarchy sequences are different from Keenan and Comrie's. In the study of SLA of Spanish, the learner language shows the similar sequences of AH of Spanish and English of native language, and different from Chinese. We conclude that L1 doesn't affect the learner language while L2 might be more influential.

References

Altenberg, B. and Aijmer, K. (2000), 'The English-Swedish parallel corpus: A resource for contrastive research and translation studies', in C. Mair and M. Hundt (eds.) *Corpus Linguistics and Linguistic Theory*, 15-33. Amsterdam: Rodopi.

Cermák, F., & Klégr, A. (2004). 'Modality in Czech and English: Possibility particles and the conditional mood in a parallel corpus'. *International Journal of Corpus Linguistics* 9(1), 83-95.

Keenan, E. L. & Comrie, B. (1977). 'Noun phrase cccessibility and universal grammar.' *Linguistic Inquiry* 8, 63-99.

Lu, Hui-Chuan. 2007，「以語料庫為本之台灣西班牙語教學研究」，*外國與文研究專刊：歐洲語言文化在台灣*，1-22頁。台北：政治大學。

Oostdijk, N. & De Haan, P. (1994). 'Clause patterns in modern British English: A corpus-based (quantitative) study.' *ICAME Journal* 18, 41-79.

Santos, D. (2004). *Translation Based Corpus Studies: Contrasting English and Portuguese Tense and Aspect System*. Amsterdam: Rodopi.

Schmied, J. and Schäffler, H. (1996), 'Approaching translationese through parallel and translation corpora', in C. Percy, C. Meyer and I. Lancashire (eds.) *Synchronic Corpus Linguistics*, 41-56. Amsterdam: Rodopi.

Sheldon, A. (1974). 'The role of parallel function in the acquisition of relative clauses in English.' *Journal of Verbal Learning and Verbal Behavior* 13(3), 272-81.

Uchida, M. (2002), 'From participles to conjunctions: A parallel corpus study of grammaticalization in English and French', in T. Saito, J. Nakamura and S. Yamazaki (eds.) *English Corpus Linguistics in Japan*, 131-146. Amsterdam: Rodopi.

Babel Chinese-English Corpus

http://icl.pku.edu.cn/icl_groups/parallel/concordance.asp

Bible

BibleGateway.com (http://www.biblegateway.com/)

CKIP

http://ckipsvr.iis.sinica.edu.tw/

CRATER

http://www.comp.lancs.ac.uk/linguistics/crater/corpus.html

ECI

http://www.elsnet.org/eci.html

EUR-Lex

http://eur-lex.europa.eu/

LCMC

http://bowland-files.lancs.ac.uk/corplang/lcmc/

MLCC

http://catalog.elra.info/product_info.php?products_id=46&osCsid=66...

Multiple-Translation Chinese Corpus

http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2002T01

ParaConc

http://www.athel.com/para.html

Reuters

http://www.reuters.com/

Tree Tagger

http://www.cele.nottingham.ac.uk/~ccztk/treetagger.php

WordSmith Tools

http://www.lexically.net/wordsmith/