A Corpus-based Approach for Spanish-Chinese Language Learning

Shuyuan Cao
Universitat Pompeu Fabra (UPF)
shuyuan.cao@hotmail.com

Iria da Cunha
Universidad Nacional de Educación a Distancia (UNED)
iriad@flog.uned.es

Mikel Iruskieta
University of Basque Country (UPV/EHU)
mikel.iruskieta@ehu.eus

Abstract

Due to the huge population that speaks Spanish and Chinese, these languages occupy an important position in the language learning studies. Although there are some automatic translation systems that benefit the learning of both languages, there is enough space to create resources in order to help language learners. As a quick and effective resource that can give large amount language information, corpus-based learning is becoming more and more popular. In this paper we enrich a Spanish-Chinese parallel corpus automatically with part of-speech (POS) information and manually with discourse segmentation (following the Rhetorical Structure Theory (RST) (Mann and Thompson, 1988)). Two search tools allow the Spanish-Chinese language learners to carry out different queries based on tokens and lemmas. The parallel corpus and the research tools are available to the academic community. We propose some examples to illustrate how learners can use the corpus to learn Spanish and Chinese.

1 Introduction[1]

As the most spoken two languages in the world, Spanish and Chinese are very important in the language learning field. Because of the different phonetics and written characters and extensive grammatical rules, syntactic structure and discourse elements between this language pair, it is not easy to carry out the Spanish-Chinese language learning tasks. Here we give some examples in order to show some morphological, syntactic and discourse similarities and differences between Spanish and Chinese that a language learner has to know and practice.

Among other issues, Chinese students that are learning Spanish as L2 need to know that Spanish language is not a gender neutral language, so the distinction of grammatical gender is crucial between masculine and feminine (among irregular constructions). There are some adjectives with a particular ending for feminine ('JJ+a'[2] such as *pública* 'feminine_public', *extranjera* 'feminine_foreigner', *china* 'feminine_chinese') and for masculine ('JJ+o' such as *moderno* 'masculine_modern', *chino* 'masculine_chinese', *rico* 'masculine_rich'). In Chinese, for example, the masculine *chino* and feminine *china* are translated as "zhongguo" (中国) ('China').

Ex.1[3]:

1.1.1 Sp: Aunque aún no contamos con resultados, intuimos que el modelo será más amplio que el del sintagma nominal.
[Aunque aún no contamos con resultados,]$Unit_1$ [intuimos que el modelo será más amplio que el del sintagma nominal.]$Unit_2$
[DM still no get results,] [we consider that the model will more extensive than the sentence group nominal.]

[1] This work has been supported by a Ramón y Cajal contract (RYC-2014-16935) to Iria da Cunha and it has been partially financed by TUNER (TIN2015-65308-C5-1-R) to Mikel Iruskieta.

[2] In the Stanford parser JJ means adjective.

[3] All the examples have been extracted from the corpus.

Proceedings of the 3rd Workshop on Natural Language Processing Techniques for Educational Applications, pages 97–106, Osaka, Japan, December 12 2016.

[Aunque aún no contamos con resultados,]$Unit_1$ [intuimos que el modelo será más amplio que el del sintagma nominal.]$Unit_2$

[DM[4] still no get results,] [we consider that the model will more extensive than the sentence group nominal.][5]

1.1.2 Sp: Intuimos que el modelo será más amplio que el del sintagma nominal, aunque aún no contamos con resultados.

[Intuimos que el modelo será más amplio que el del sintagma nominal,]$Unit_1$ [aunque aún no contamos con resultados.]$Unit_2$

[We consider that the model will more extensive than the sentence group nominal,] [DM still no get results.]

1.2 Ch: 尽管还没有取得最终结果，但是我们认为该模型已囊括了语段模型涉及的内容。

[尽管还没有取得最终结果，]$Unit_1$ [但是我们认为该模型已囊括了语段模型涉及的内容。]$Unit_2$

[DM1 still no get results,] [DM2 we consider that the model contains the sentence group nominal.]

1.3 Eng: Although we haven't got the results yet, we consider that the model will be more extensive than the nominal sentence group.

In Example 1, we can see that the Spanish passage has a similar discourse structure to the Chinese passage. Both passages start the text with a discourse marker in the first unit. However, the usage of discourse markers in both languages is different. To show same meaning, in Chinese, it is mandatory to include two discourse markers: one marker is "*jinguan*" (尽管), at the beginning of the first unit, and another marker is "*danshi*" (但是), at the beginning of the second unit. These two discourse markers are equivalent to the English discourse marker 'although'. By contrast, in Spanish, just one discourse marker "*aunque*" is being used at the beginning of the first unit, and this discourse marker is also equivalent to the English discourse marker *although*. Moreover, the order of the discourse units in the Spanish passage can be changed and it makes sense syntactically, but the order cannot be changed in the Chinese passage, because neither syntactically nor grammatically makes sense.

Ex.2:

2.1 Sp: La automatización de la gestión terminológica no es una mera cuestión de producir programas informáticos, aunque ésta sea una labor de por sí costosa.

[La automatización de la gestión terminológica no es una mera cuestión de producir programas informáticos,]$Unit_1$ [aunque ésta sea una labor de por sí costosa.]$Unit_2$

[The automation of the management terminological no is a merely question of producing programs informatics, DM this is a work which costly.]

2.2 Ch: 术语管理自动化不仅仅是建立开销巨大的信息系统的问题。

[术语管理自动化不仅仅是建立开销巨大的信息系统的问题。]$Unit_1$

[Terminology management automation is not only produce costly program informatics.]

2.3 Eng: Although this is a work which costly, the automation of the terminological management is not only a merely question of producing programs informatics.

In Example 2, there are two units in the Spanish passage meanwhile in the Chinese passage there is just one unit to show the same meaning. In the Spanish passage, the DM *aunque* 'although' is at the beginning of the second unit; in contrast, there is no corresponding DM in the Chinese passage and a translation strategy has been used. The Chinese phrase "*kaixiaojuda*" (开销巨大) ('great costly') includes the same meaning of the second Spanish unit and is part of the whole sentence in the Chinese passage.

[4] DM means discourse marker. In this work, we use the definition of DM by Portolés (2001). DMs are invariable linguistic units that depend on the following aspects: (a) distinct morph-syntactic properties, (b) semantics and pragmatics and (c) inferences that made in the communication.

[5] In this work, we give an English literal translation for both Spanish examples and Chinese examples in order to make the readers understand the content better.

These examples show that a comparative study could provide useful discourse information for language learning. Comparative or contrastive studies of discourse structures reveal information to identify properly equivalent discourse elements in the language pair and the information helps language learning.

An important source for language learning is corpus. As a large electronic library, a corpus can provide a large amount of linguistic information (Wu, 2014). In addition, Johns (2002) indicates that a corpus-based research could help the language learners get large amount of language information easily.

This paper aims to create a Spanish-Chinese parallel corpus annotated automatically with POS and annotated manually with discourse information in order to help Spanish-Chinese language learning with a friendly online environment to perform POS-based queries, as "it has been demonstrated that discourse is a crucial aspect for L2[6] learners of a language, especially at more advanced level" (Neff-van Aertselaer, 2015: 255).

In the second section, we mention some works related to our work. In the third section, we include information about our research approach. In the fourth section, we explain how to use our corpus for Spanish-Chinese language learning. In the last section, we present the conclusions and look ahead at our future work.

2 Related Work

Corpus-based studies for different language pairs learning exist, including some works for Spanish and Chinese. On the one hand, for example, we highlight the following corpus-based language learning studies:

i) In order to help language learning and translation tasks between English and Chinese, Qian (2005) created an English-Chinese parallel corpus with functions of sentence search, calculation of words, search of texts and authors.

ii) To compare the similarities and differences between English and Chinese from different aspects, such as aspect marking, temporal adverbials, passive construction, among other interesting topics, Xiao and McEnery (2010) used the FLOB corpus (Albert-Ludwigs, 2007)[7] and The Lancaster Corpus of Mandarin Chinese (LCMC) (McEnery and Xiao, 2004)[8], which is designed as a Chinese parallel corpus for FLOB. The study offers a great amount of language information that is useful for English-Chinese language learning.

iii) To compare both languages via different language activities, such as exploration of language differences, comparative discourse analysis and semantic analysis Lavid, Arús and Zamorano (2010) developed a small online English-Spanish parallel corpus. Then, based on the activity results, they give some linguistic suggestions for English-Spanish teaching, which can also help the English-Spanish language learners to comprehend the language differences between both languages.

On the other hand, corpus-based studies for Spanish-Chinese language learning are still few:

i) Yao (2008) uses film dialogues to elaborate an annotated corpus, and compares the Spanish and Chinese discourse markers in order to give some suggestions for teaching and learning Spanish and Chinese.

ii) Yang (2008) compares the discourse structure of proverbs between Spanish and Chinese based on the novel *Don Quijote* in order to give some conclusions for the Spanish-Chinese translation works, and language teaching and learning tasks.

iii) Taking different newspapers and books as the research corpus, Chien (2012) compares the Spanish and Chinese conditional discourse markers to give some conclusions of the conditional discourse marker for foreign language teaching between Spanish and Chinese.

iv) Wang (2013) uses *Pedro Almodóvar's* films *La mala educación* and *Volver* as the corpus to analyze how the subtitled Spanish discourse markers can be translated into Chinese, so as to make a guideline for human translation and audiovisual translation between the language pair.

[6] L2 means second language.

[7] The FLOB Corpus: http://www.helsinki.fi/varieng/CoRD/corpora/FLOB/ [Last consulted: 27 of July of 2016].

[8] The Lancaster Corpus of Mandarin Chinese: http://www.lancaster.ac.uk/fass/projects/corpus/LCMC/ [Last consulted: 27 of July of 2016].

v) Cao, da Cunha and Bel (2016) annotate all the cases of the Spanish DM *aunque* ('although') and their corresponding Chinese translations in The United Nations Multilingual Corpus (UN) (Rafalovitch and Dale, 2009; Eisele and Yu, 2010). They analyze the used translation strategies in the translation process and give some suggestions for how to translate this Spanish DM into Chinese.

vi) Several Spanish-Chinese parallel corpus exist and have been used for different research purposes, including Spanish-Chinese language learning, these corpora are: (a) *The Holy Bible* (Resnik, Olsen and Diab, 1999), (b) The United Nations Multilingual Corpus (UN) (Rafalovitch and Dale, 2009; Eisele and Yu, 2010) and (c) Sina Weibo Parallel Corpus (Wang et al., 2013).

The above mentioned works are great achievements that offer different approaches for language learning. However, comparing to our work, none of them gives a friendly environment to consult Spanish-Chinese parallel corpus based on POS and segmented discourse information, showing how foreign language learners can apply this information to improve or learn languages.

3 Research Approach

3.1 Theoretical Framework

In this study, we use the Rhetorical Structure Theory (RST), proposed by Mann and Thmopson (1988), which a widely used theory for discourse analysis. RST offers discourse information from two approaches: linear segmentation and rhetorical annotation. Under RST, linear segmentation is compased with Elementary Discourse Units (EDUs) (Marcu, 2000). But in linear segmentation different discourse phenomena can be studied, such as the position of the DM, the number of DMs, etc.

Discourse-annotated corpus can provide valuable insights into L2 discourse aspects, problems, and solutions (Vyatkina, 2016). Thus, information and examples of discourse segmentation are useful for those language learners who have an advanced level, that is, students that should be competent to solve complicated discourse questions.

3.2 Elaboration of the Research Corpus

The previous mentioned Spanish-Chinese parallel corpora are not adequate for language learning purposes between Spanish and Chinese from discourse level. For example, the texts in the Holy Bible cannot represent the modern language; the Spanish-Chinese UN Corpus is not a direct translated corpus, which affects the discourse structure; and the texts in the Sina Weibo Parallel Corpus are tweets that do not include a complex discourse structure. In order to use formal and natural expressed texts, we decide to use the Spanish-Chinese parallel corpus by Cao, da Cunha and Iruskieta (in press), which is especially designed for discourse studies with formal texts.

In their corpus, 100 texts are included: the longest text contains 1,774 words and the shortest one includes 111 words. The genres of the texts are: (a) Abstract of research paper, (b) News, (c) Advertisement and (d) Announcement. The corpus contains seven domains: (a) Terminology, (b) Culture, (c) Language, (d) Economy, (e) Education, (f) Art and (g) International affairs.

Firstly we enriched this corpus automatically with POS information by using Freeling (Carreras et al., 2004) for Spanish and the Stanford parser (Levy and Manning, 2003) for Chinese. Then, we segmented and harmonized the Spanish and Chinese texts into EDUs to obtain a gold standard segmented corpus. Two Chinese experts and two Spanish experts carried out the segmentation work manually, by using the RSTTool (O'Donnell, 2000) following Iruskieta, da Cunha and Taboada's (2015) segmentation and harmonization criteria. Finally, we developed a free online interface that allows students of Spanish or Chinese to do different linguistic queries that can help their language learning process[9].

3.3 Level Requirement for the Spanish-Chinese Language Learning with the Corpus

In this study, for the Chinese users who learn Spanish, we adopt the language level standardizations of *Instituto Cervantes*, the official Spanish organization to check the language level for L2 Spanish learn-

[9] The access of the interface is the following: http://ixa2.si.ehu.es/rst/zh/.

ing[10]; for the Spanish users who learn Chinese, we adopt the language level standardizations of *Hanban* (汉办), the official organization of Chinese government for L2 Chinese learning[11].

As we have mentioned, the corpus by Cao, da Cunha and Iruskieta (in press) includes specialized texts from different sources, which include terminology from several domains. Therefore, the users of our annotated-corpus and search tool should have an intermediate or advanced level of the language. As the webpage of *Instituto Cervantes* indicates, in the Spanish initial level only some basic expressions and vocabulary are learned. Also, the webpage of *Hanban* (汉办) offers similar information about the Chinese initial level.

In order to use our annotated-corpus and search tool, the appropriate levels for Spanish foreign language learners are level B2 (intermediate level) and level C (including C1 and C2) (advanced level). Level B2 requires language users to understand complex texts with different topics. Level C1 requires understanding a wide variety of long and demanding texts, and also writing and expressing well-structured texts in Spanish. Level C2 is a more advanced level and requires Spanish learners have enough linguistic competence to prove a spontaneous capacity of adaptation to any context, with a great deal of semantic and grammatical precision.

The appropriate level for Chinese foreign language learners is level 4 (intermediate level) and advanced level (level 5 and level 6). Level 4 requires language learners to know a certain amount of words and produce texts related to a wide range of topics, in order to maintain a fluent communication with native speakers. Level 5 requires learners to read magazines, newspapers, and films and give a full-length speech. Level 6 language learners should easily comprehend written and spoken information in Chinese.

4 Spanish-Chinese Language Learning with the Corpus

As we have mentioned, the aim of the annotated-corpus and the search tool is to help language learners of both languages by providing them with real examples that can be extracted by means of different linguistic queries including linguistic information: morpho-syntactic information and discourse-segmentation information.

On the one hand, regarding morpho-syntactic information, a Chinese foreign language-learning student can search any wanted Chinese tokens or lemmas, and a Spanish foreign language-learning student can carry out the search in an inverse way. Here we give a real example by using the search tool for Spanish. The word that we use is the Spanish word *profesor* ('teacher'). By using the search tool, we can search the token of profesor or the lemma of this word. We give some lemma search results of profesor as the example. The results are presented in Figure 1[12].

	Document	Sent Id	Word(s)	Sentence	
1	BMCS_ESP2.txt	sent1	PROFESORES	PROFESORES Y MÉTODOS Todos nuestros profesores son nativos , han recibido una formación específica en la enseñanza de español como lengua extranjera (ELE) y tienen experiencia docente en China .	View context
2	BMCS_ESP2.txt	sent4	profesor	El profesor no impone conocimientos : ayuda a sus alumnos a comunicarse en español desde el primer día , animándolos a que participen activamente en el aula .	View context
3	BMCS_ESP2.txt	sent5	profesores	Los profesores cuentan siempre con el punto de vista de sus alumnos en la toma de decisiones de la clase , fomentando la autonomía del estudiante mediante el uso de las estrategias de aprendizaje más adecuadas para cada uno .	View context
4	BMCS_ESP2.txt	sent8	profesores	El aprendizaje del léxico y de la gramática está acompañado del valor del uso comunicativo que los profesores nativos pueden y saben transmitir .	View context

Figure 1: Search result of the Spanish lemma *profesor* with the result of two different forms *profesores* 'teachers' and *profesor* '(masculine) teacher'

A Chinese learner can find different POS structures in our corpus, for example, all the words which end with '*a*' that are adjectives (*española* 'feminine_Spanish', *pública* 'feminine_public', *his-*

[10] A detailed explanation of the Spanish level for L2 learners can be consulted: http://dele.cervantes.es/en/information/levels/spanish_levels.html [Last consulted: 17 of September of 2016]

[11] The detailed explanation of the Chinese level for L2 learners can be consulted: http://english.hanban.org/node_8002.htm [Last consulted: 17 of September of 2016]

[12] Due to the limitation of the required pages, here the space doesn't allow us to show the whole lemma research result of the Spanish word *profesor*. Also, we only give the partial results in the following figures.

panoamericana 'feminine_hispanicamerican', etc.) or feminine words ended with '*ora*' that are nouns (*directora* 'feminine_director', *coordinadora* 'feminine_coordinator', *editora* 'feminine_editor' etc.) to learn how feminine is used in real texts.

Also, a language learner can search the wanted token with "exact match", "start with" or "ends with". This function can help students of Chinese to learn different phrases by searching just one character. We use the Chinese word *fa* (发)[13] to explain how to search those phrases related with the character *fa* (发). Figure 2 shows some of the search results: the words starting with *fa* (发) are *fasong* (发送) ('to send'), *fayangguangda* (发扬光大) ('to flourish'), and *fazhan* (发展) ('to develop'). With different match functions, a Spanish student can learn different words including a specific character, in this case *fa* (发).

	Document	Sent Id	Word(s)	Sentence	
1	BMCS_CHN5.txt	18	发送	如有 任何 查询 请 发送 邮件 至 prof1sha@cervantes.es	View context
2	ICEG_CHN1.txt	2	发扬光大	格拉纳达 大学 的 汉语 教学 始自 1987年， 二十多 年 来 已 开设 了 各 种 与 中国 历史 、 文学 、 思想 相关 的 课程， 而 格 大孔 院 要 将 这 种 深厚 的 汉学 研究 传统 继续 发扬光大 ， 与 格拉纳达 大学 中 像 亚洲 研究会 这 样 的 学术 组织， 以及 其它 国内外 个 相关 机构 精诚 合作， 努力 开展 各 种 活动， 以 满足 社会 各界 不断 增 长 的 需求 。	View context
3	EEP_CHN4.txt	1	发展	第二 届 " 丝路 国际 论坛 2015年 会 " 在 马德里 召开 10月 28日 和 29日， 由 国务院 发展 研究 中心 、 国际 关系 和 可持续 发展 中心 、 中国 驻 西班牙 大使馆 和 和 托雷 多 国际 和平 中心 共同 主办 的 第二 届 " 丝路 国际 论坛 2015年 会 " 在 马德里 召开 。	View context
4	EEP_CHN5.txt	1	发展	第一 届 中 西 交流 发展 论坛 西班牙 工业 、 能源 与 旅游 大臣 索里亚 与 中国 驻 西班牙 大使 吕凡于 10月 27日 在 马德里 共同 出席 了 第一 届 中 西 交流 发展 论坛 的 开幕式 。	View context

Figure 2: Chinese words starting with the word *fa* (发)

Moreover, language learners can also search by POS information for both Spanish and Chinese. Based on the character *fa* (发), we give another real example in the corpus. A Spanish student can search the Chinese lemma that start with *fa* (发) but play as verb. Figure 3 shows partial results that match the search requirement.

	Document	Sent Id	Word(s)	Sentence	
1	BMCS_CHN5.txt	18	发送	如有 任何 查询 请 发送 邮件 至 prof1sha@cervantes.es	View context
2	ICEG_CHN1.txt	2	发扬光大	格拉纳达 大学 的 汉语 教学 始自 1987年， 二十多 年 来 已 开设 了 各 种 与 中国 历史 、 文学 、 思想 相关 的 课程， 而 格 大孔 院 要 将 这 种 深厚 的 汉学 研究 传统 继续 发扬光大 ， 与 格拉纳达 大学 中 像 亚洲 研究会 这 样 的 学术 组织， 以及 其它 国内外 个 相关 机构 精诚 合作， 努力 开展 各 种 活动， 以 满足 社会 各界 不断 增 长 的 需求 。	View context
3	CCICE_CHN3.txt	1	发布	西班牙 财政部 拟 拍卖 高达 50亿 欧元 短期 国债 经济 学家 报 11月 17日 消息： 据 西班牙 财政部 在 官网 发布 的 消息 显示， 该 机构 将 在 本 周二 拍卖 6 至 12月 到期 的 短期 国债， 预期 拍卖 40亿 至 50亿 欧元 。	View context
4	CCICE_CHN3.txt	2	发行	此 次 为 财政部 从 10月 以来 首次 发行 该 类型 国债， 当时 拍卖 金额 41.21亿 欧元， 而 本次 使用 的 利率 将 比 上 次 更低 。	View context

Figure 3: Search result of verbs that start with *fa* (发)

The partial results in Figure 3 gives us four different Chinese verbs starting with *fa* (发): (a) *fasong* (发送) ('to send'), (b) *fayangguangda* (发扬光大) ('to flourish'), (c) *fabu* (发布) ('to publish') and (d) *faxing* (发行) ('to issue').

The POS information also has another function in our corpus. In Chinese, some words have two categories; the category can be a verb and a noun at the same time (Yu, Duan and Zhu, 2005). Hence, under this circumstance, it is hard to choose the category of a word for L2 students of Chinese. POS information helps to distinguish the category of a word. For example, the Chinese noun *xuyao* (需要) ('requirement') can also be the verb 'to need to'. In the corpus, when including *xuyao* (需要) in the

[13] In Chinese, the verb *fa* (发) has various meanings, such as "to have over", "to express", "to expand", "to begin to", among others. [Consulted from: *Xiandai hanyu cidian* (现代汉语词典)]

lemma column and choosing "VV"[14] as a POS, seven results are obtained, as Figure 3 shows. Meanwhile, there is one result of *xuyao* (需要) as a noun in Figure 4.

	Document	Sent Id	Word(s)	Sentence	
1	TERM29_CN.txt	9	需要	我们需要找到一个能够在实际情况中有效应用的解决方案，这也促使我们在进行专项研究时，不仅要兼顾上述理论原则，还应考虑在术语和信息学方面采用不同的方法论。	View context
2	TERM34_CN.txt	2	需要	在很多情况下，要找到巴斯克语对应临近语种的关系形容词，需要经过多个步骤（Ensunza，1989；Loinaz，1995）。	View context
3	TERM38_CN.txt	2	需要	各种语言中与互联网相关的术语在以很快的速度诞生和传播，影响范围广，如同建造了一条需要与时间赛跑的跑道。	View context
4	TERM38_CN.txt	4	需要	对于科技进步来说，这种现象的产生并不稀奇，但需要注意的是，介于术语新词的特点，各领域的专业性要求又赋予了新词一定的特殊性。	View context
5	TERM31_CN.txt	2	需要	简介近年来，各语种都在开发科技类文章术语的自动构建工具，尽管如此，对于自动选出的术语条目还是需要人工进行最后一步筛选。	View context
6	TERM51_CN.txt	9	需要	例如安伯托（Anboto）山、拉蒙-卡哈尔（Ramón y Cajal）大街、伊拜萨巴（Ibaizabal）河、奥尔加山丘（Alto de la Horca）等，地名自身的定义引导我们判断地理术语的重要性，同时我们也注意到在进行地名标准化时需要提出两个版本（巴斯克语和西班牙语）。	View context
7	TERM51_CN.txt	10	需要	概括地讲，这意味着共有元素可通过翻译而来，而特定元素需要保持不变。	View context

Figure 3: Search results of *xuyao* (需要) as verb

	Document	Sent Id	Word(s)	Sentence	
1	ICP_CHN8.txt	12	需要	证书分以下几个级别：·西班牙语水平证书A1级别级别证明拥有者的语言水平足以应对简单的交流、即时性需要和非常日常性的话题·西班牙语水平证书A2级别证明拥有者能够理解日常表达和其所涉及领域相关的习惯用法，尤其是一些与自身相关的基本信息，比如自己、家庭、购物、景点、职业等等。	View context

Figure 4: Search result of *xuyao* (需要) as noun

A Spanish student who uses the corpus to learn Chinese can distinguish the words that have more than one category easily by using the combination of lemma and POS, and also check their contexts of use.

The interface we created allows the search of maximum four tokens/lemmas at the same time, that is, it is possible to do complex queries. This is useful to obtain different language information, such as the use of adjectives in a phrase. For example, if a Spanish student knows the phrase *xibanyayu ketang* (西班牙语课堂), 'Spanish class' in Chinese, and wants to search for more adjectives associated to the word *ketang* (课堂) ('class'), which could be inserted in the middle of the two units of the phrase, he could do the following complex query: i) lemma *xibanyayu* (西班牙语) ('Spanish'), ii) POS information "JJ"[15], and iii) lemma *ketang* (课堂) ('class') (see Figure 5).

Figure 5: Example of the search for an adjective in a phrase

Figure 6 includes the search results of the example. Two results are obtained, including the same adjective related to the noun *ketang* (课堂) ('class'): *xuni* (虚拟) ('virtual').

	Document	Sent Id	Word(s)	Sentence	
1	BMCS_CHN3.txt	4	西班牙语 / 虚拟 / 课堂	该课程适合学生在课堂学习之外同时积极的开展自主学习，在每周不上课的日子里学生可以在家完成许多教材中的课外活动，此外也可以使用我们网络学习平台的资源西班牙语 虚拟 课堂。	View context
2	BMCS_CHN5.txt	1	西班牙语 / 虚拟 / 课堂	西班牙语 远程 课程：西班牙语 虚拟 课堂（AVE） 西班牙语 虚拟 课堂 （AVE） 是塞万提斯学院专门为对外西班牙语教学和学习设计的，以互联网为媒介的虚拟环境。	View context

Figure 6: Result of the search for an adjective between *xibanyayu*（西班牙语）and *ketang* (课堂)

[14] In the Stanford parser VV means verb and NN means noun.
[15] In parsing research, J means adjective.

Another similar search function is illustrated in Figure 9. The Chinese word *kecheng* (课程) ('course') is a noun and, by using the search of POS information, different adjectives related with *kecheng* (课程) are extracted. Figure 9 shows three results of the adjective that can be combined with *kecheng* (课程): *changgui* (常规), *yiban* (一般) and *putong* (普通). All of them mean 'regular' in English. In this case, three different words with the same meaning are extracted.

	Document	Sent Id	Word(s)	Sentence	
1	BMCS_CHN3.txt	3	常规 / 课程	常规 课程 常规 课程 时 长 3 个 月， 每 周 5 小时 课时 分为 2 节 课， 每节 课 2.5 小时， 一般 是 周一 、 周三 或者 周二 、 周四 各 上 一 课 。	View context
2	BMCS_CHN3.txt	6	常规 / 课程	和 常规 课程 一样， 每 周 5 小时 课时， 不同 的 是 这 5 小时 集中 在 同 一 天： 周六 或者 周日 。	View context
3	BMCS_CHN5.txt	3	一般 / 课程	在 米 盖尔·德·塞万提斯 图书馆 的 多媒体 教室 你 可以 免费 试用 AVE 各 类 课程 : 一般 课程： 包含 欧洲 共同 语言 参考 标准 制定 的 A1 , A2 , B1 , B2 及 C1 级别 课程 。	View context
4	ICP_CHN5.txt	4	普通 / 课程	北京 塞万提斯 学院 提供 丰富 多样 的 西班牙语 学习 课程 • 普通 课程 2600 元 • 紧凑 课程 2600 元 • 周末 课程 2600 元 (2015年 8月 更新 信息) 我们 所有 的 老师 都 是 以 西班牙语 为 母语， 受过 对外 西班牙语 教学 培训 并 且 拥有 对 外 西班牙语 教学 经验 的 老师 。	View context

Figure 9: Search result of the category adjective with the noun *kecheng* (课程)

The search tools and the POS information are important for Spanish-Chinese language learning but, on the other hand, discourse segmentation information is also relevant to support Spanish-Chinese language learning. The Spanish-Chinese language learners can compare the similarities and differences by using the segmentation of the parallel texts. Table 1 includes an example of discourse segmentation difference in our corpus.

Spanish	**Chinese**	**English translation of the Spanish text**
[La empresa española Aritex ha colaborado con la Corporación de Aeronaves Comerciales de China (COMAC) en la fabricación del C919, primer avión comercial diseñado y fabricado por China.] EDU_1	[西班牙 Aritex 公司与中国商飞 (COMAC) 合作，]EDU_1 [参与了中国首架国产 C 919 大型客机的制造过程。]EDU_2	[The Spanish company Aritex has collaborated with Commercial Aircraft Corporation of China (COMAC) in making the C919, the first commercial aircraft designed and manufactured by China.]EDU_1

Table 1: The segmentation difference between a parallel Spanish-Chinese

In this Spanish-Chinese parallel example, the whole Spanish sentence is an EDU, while the Chinese sentence is divided into two coordinated EDUs. This happens because of a translation strategy: in the Chinese translation, the Spanish phrase *en la fabricación* ('in the production') has been translated into *canyule* ('have participated in the production'), which has an elliptical subject "Aritex Company" and forms a coordinated part of EDU1 in Chinese passage.

Other useful information that foreign language learners can obtain from this parallel annotated-corpus is related with DMs. For example, they can compare the different DMs used in both languages. Taking the Chinese DMs *ruo* (若) ('if') and *ze* (则) ('then'), Table 2 shows two Spanish-Chinese parallel passages from the corpus.

Table 2 shows that in Spanish there are two sentences while in Chinese there is only 1 sentece. EDU1 and EDU2 in the Spanish passage correspond to EDU1 in the Chinese passage, and EDU3 in the Spanish passage corresponds to EDU2 in the Chinese passage. The number of different EDUs in Spanish and Chinese passages is due to the Spanish DM *para* ('for' or 'in order to') in the first sentence and the used translation strategy. The Spanish DM *para* is the signal of a PURPOSE relation. Therefore, the first sentence is segmented into two EDUs. The translation strategy causes the Chinese translation as one sentence. In the Spanish passage, there is no DM for holding a discourse relation between the two complete sentences. Instead, in the Chinese passage, there are two DMs at the beginning of each EDU, one DM is *ruo* (若), which means 'if' in English, and another DM is *ze* (则),

which means 'then'. The two DMs represent a CONDITION relation. In Chinese, it is necessary to use two DMs (*ruo* and *ze*) at the same time at the beginning of each EDU.

Spanish	Chinese	English translation of the Spanish text
[Los resultados que se obtienen no son aún los que se precisarían]EDU_1 [para efectuar un vaciado absolutamente automático.]EDU_2 [Se ha de encontrar el equilibrio entre la cobertura (recall) y la precisión (precision).]EDU_3	[若上述过程中获得的结果仍无法完全自动构建一个精确的术语条目，]EDU_1 [则必须在覆盖度（召回率）和精确度（精确性）之间达到平衡。]EDU_2	[The obtained results are still cannot be completely required to make a precisely term automatically.]EDU_1 [It must find a balance between coverage (recall) and accuracy (precision).]EDU_2

Table 2: The difference of DMs between a Spanish-Chinese parallel sentence

The Spanish-Chinese language learners can consult any segmentation case in the corpus by using the "Bilingual EDUs" column, which is manually aligned. The different search functions are adequate for different learning tasks carried out by Spanish-Chinese language learners.

Besides of the discourse segmentation information, in the future we will annotate and align the discourse structure for the whole corpus. Spanish and Chinese learners will obtain aligned relational discourse information for language learning related to the following aspects: nuclearity, discourse relation, discourse structure and central discourse unit.

5 Conclusion and Future Work

As a complementary methodology, the use of corpora is a very adequate and useful strategy for language learning in comparison with the traditional methods (Baker, 2007). In this work, we introduce the first online POS-tagged, discourse-based segmented and manually aligned Spanish-Chinese parallel corpus for foreign language learning purposes between this language pair. This corpus offers several search possibilities for different Spanish-Chinese language learning needs. For Spanish L2 learners and Chinese L2 learners, their level must be intermediate or advanced to use the research corpus.

In the future, we will annotate the discourse structure of the whole corpus under RST. This parallel Spanish-Chinese discourse treebank will be available online, together with the search tool. It will be possible to search for parallel passages including a specific RST relation.

References

Albert-Ludwigs Christian Mair. 2007. *The FLOB Corpus* (online). http://www.helsinki.fi/varieng/CoRD/corpora/FLOB/index.html [Last consulted: 27 of July of 2016].

Asher Nicholas and Alex Lascarides. 2003. *Logics of conversation*. Cambridge: Cambridge University Press.

Baker Mona. 2007. Corpus-Based Translation Studies in the Academy. *Journal of Foreign Languages*, 171:50.

Cao Shuyuan, da Cunha Iria, and Iruskieta Mikel (in press). Toward the Elaboration of a Spanish-Chinese Parallel Annotated Corpus. *The EPiC Series in Language and Linguistics*, 2, ISSBN 2398-5283.

Cao Shuyuan, da Cunha Iria, and Bel Nuria. 2016. An analysis of the Concession relation based on the discourse marker aunque in a Spanish-Chinese parallel corpus. *Procesamiento del Lenguaje Natural*, 56: 81-88.

Carreras Xavier, Chao Isacc, Padró Lluís, and Padró Muntsa. 2004. FreeLing: An Open-Source Suite of Language Analyzers. In *Proceedings of the 4th International Conference on Language Resources and Evaluation* (*LREC' 2004*), 239-242.

Chien Yi-Shan. 2012. *Análisis contrastivo de los marcadores condicionales del español y del chino*. PhD thesis. Salamanca: Universidad de Salamanca.

Eisele Andreas, and Chen Yu. 2010. A Multilingual Corpus form United Nations Documents. In *Proceedings of Language Resource and Evaluation Conference* (*LREC 2010*), 2868-2872.

Iruskieta Mikel, da Cunha Iria, and Taboada Maite. 2015. A Qualitative Comparison Method for Rhetorical Structures: Identifying different discourse structures in multilingual corpora. *Language Resources and Evaluation*, 49*: 263-309.*

Johns Tim. 2002. Data-Driven learning: The perpetual challenge. *Language and Computers*, 1: 107-117.

Lavid Julia, Arús Jorge, and Zamorano Juan Rafael. 2010. Desiging and exploiting a small online English-Spanish parallel corpus for language teaching purposes. Corpus-Based Approach to English Language Teaching, 138-148.

Levy Roger and Manning Christopher. 2003. Is it harder to parse Chinese, or the Chinese Treebank? In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics* (*ACL' 2003*), 439-446.

Mann William C. and Thompson Sandra A. 1988. Rhetorical Structure Theory: Toward a functional theory of text organization. *Text&Talk*, 8(3): 243-281.

Marcu Daniel. 2000. The rhetorical parsing of unrestricted texts: A surface-based approach. *Computational Linguistics*, 26(3): 395-448.

McEnery Tony, and Xiao Richard. 2004. *The Lancaster Corpus of Mandarin Chinese* [online]. http://www.lancaster.ac.uk/fass/projects/corpus/LCMC/ [Last consulted: 27 of July of 2016].

Neff-van Aertselaer JoAnne. 2015. *Learner Corpora and Discourse*. Cambridge: Cambridge University Press.

O'Donnell Michael. 2000. RSTTool 2.4 – A Markup Tool For Rhetorical Structure Theory. In *Proceedings of First International Conference on Natural Language Generation*, 253-256.

Pórtoles José. 2001. *Marcadores del discursivo*. 4th edition. Barcelona: Ariel.

Rafalovitch Alexandre, and Dale Robert. 2009. United Nations general assembly resolutions: A six-languages parallel corpus, In *Proceedings of Machine Translation Summit XII*, 292-299.

Resnik Philip, Olsen Mari Broman, and Diab Mona. 1999. The Bible as a Parallel Corpus: Annotating the 'Book of 2000 Tongues'. *Computers and the Humanities*, 33(1-2): 129-153.

Qian Zhiying. 2005. *Yinghan/Hanying pingxingfanyi yuliaoku de sheji jiqi zai fanyi zhong de yingyong* (汉英/汉英平行翻译语料库的设计及其在翻译中的应用 *[*The Design of Chinese-English/English-Chinese Parallel Translation Corpus and its Application in Translation Studies*]*). Master thesis. Shanghai: East China Normal University.

Vyatkina Nina. 2016. What can multilingual discourse-annotated corpora do for language learning and teaching? In *Proceedings of TextLink – Structure Discourse in Multilingual Europe Second Action Conference*, 21-24.

Wang Ling, Guang Xiang, Dyer Chris, Black Alan, and Trancoso Isabel. 2013. Mircoblogs as Parallel Corpora. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics* (*ACL' 2013*), 176-186.

Wang Yi-Chen. 2013. *Los marcadores conversacionales en el subtitulado del español al chino: análisis de La mala educación y Volver de Pedro Almodóvar*. PhD thesis. Barcelona: Universitat Autònoma de Barcelona.

Wu Shangyi. 2014. On Application of computer-based corpora in translation. In *Proceedings of 2nd International Conference on Computer, Electrical, and Systems Sciences, and Engineering* (*CESSE' 2014*), 173-178.

Xiao Richard, and McEnery Tony. 2010. *Corpus-Based Contrastive Studies of English and Chinese*. New York: Routledge

Yang Yunmei. 2008. *Hanxi yanyu duibi yanjiu---Yi Tangjikede weili* (汉西谚语对比研究---以《唐吉坷德》为例 *[*Comparative study of Spanish and Chinese proverbs --- case study of *Don Quijote]*). Master thesis. Shandong: Shandong University.

Yao Junming. 2008. *Estudio comparativo de los marcadores del discurso en español y en chino a través de diálogos cinematográficos*. PhD thesis. Valladolid: Universidad de Valladolid.

Yu Shiwen, Duan Huiming, and Zhu Xuefeng. 2005. *Ciyujianlei ji dongci xiang mingci piaoyi xianxiang de jiliang fenxi* (词语兼类暨动词向名词飘移现象的计量分析 *[*A Quantitative Analysis on Multi-class Words and Shift from Verbs to Nouns in Chinese*]*). *Natural language understanding and large-scale computing content* (自然语言理解与大规模内容计算), 70-76.