Divergences in the Usage of Discourse Markers in English and Mandarin Chinese

David Steele and Lucia Specia

Department of Computer Science, The University of Sheffield, UK
dbsteele@sheffield.ac.uk l.specia@sheffield.ac.uk

Abstract. Statistical machine translation (SMT) has, in recent years, improved the accuracy of automated translations. However, SMT systems often fail to deliver human quality translations especially with complex sentences and distant language pairs. Current SMT systems often focus on translating single sentences with clauses being treated in isolation, leading to a loss of contextual information. Discourse markers (DMs) are vital contextual links between discourse segments and this paper examines the divergences in their usage across English and Mandarin Chinese. We highlight important structural differences in composite sentences extracted from a number of parallel corpora, and show examples of how these cases are dealt with by popular SMT systems. Numerous significant divergences, such as contextual omissions, were observed, which can lead to incoherent automatic translations. Our objective is to use these findings to guide a framework proposal to address divergences in DM usage in order to improve SMT output quality.

1 Introduction

In general "discourse" is used to signify an arbitrary length of coherent language-based communication consisting of either phrases, sentences or utterances [1]. With respect to natural language processing (NLP), and more specifically, Statistical Machine Translation (SMT) – our application of interest, discourse is mainly concerned with both written text and spoken dialogue consisting of some connected sequential units.

On a fundamental level discourse is linked in a meaningful way (lexical cohesion) by discourse markers (DMs, also known as discourse connectives), which separate the discourse into discourse segments or language structures, such as words, phrases, clauses or composite sentences [2], each of which contain a local coherence and context. However, DMs cover a range of connectives, conjunctions, conjunctives and other cue words and can be difficult to define precisely [3].

Despite the important role DMs have in terms of lexical cohesion, current SMT systems do not explicitly address DM constructions as such, and therefore translations can often lack the cohesive cues that DMs provide. Indeed, DMs are often translated into the target language in ways that differ from how they are used in the source language [4,5]. While recent developments in SMT potentially allow the modelling of discourse information across sentences [6], no efforts have been dedicated to address DMs in particular. Additionally it has been shown that single DMs can signal numerous

P. Sojka et al. (Eds.): TSD 2014, LNAI 8655, pp. 189–200, 2014.

discourse relations depending on where they occur and current SMT systems are unable to adequately distinguish between each of the relations during the translation process [7].

This paper examines the usage of a set of frequent DMs in Chinese[1] and English, highlighting some natural and common divergences observed in parallel corpora, and some of the problems that arise when the contextual information that surrounds them is not utilised by SMT systems. The focus is on Chinese into English translations. The results were produced from inspecting four corpora of various genres, domains and sizes, comparing given DMs in Chinese sentences against DMs in the English parallel human translation. Only DMs within compound sentences, rather than across discourse segments, were used for the analysis. The study shows that the parallelism in the usage of DMs in the two languages varies significantly across corpora. It also shows substantial divergences in the usage of DMs in a large proportion of the cases. This evidences the problem of using such parallel corpora as a source of information to build SMT systems without special treatment of DMs.

Popular online SMT systems were also used to translate the Chinese, with the resulting automated translation being compared to the given human translation, hence illustrating their limitations. The results show that these SMT systems are often unable to deal with the complex changes in word order and, because of DMs, struggle with contextual omissions, even across closely linked sentential clauses. As the sentences become more complex the problems are further compounded and more errors occur in the automated translations, ultimately suggesting that too much information is lost when the context carried by DMs is not utilised by SMT systems.

2 Discourse Markers in Chinese

Chinese and English stem from two very different language families (Sino-Tibetan and Indo-European respectively) which can be a chief cause of translation difficulty [9]. For example, Chinese is logographic and does not use inflection, relying on generating meaning through word order, which can often be quite flexible. Moreover, the positioning and order of connective markers is very fluid and syntactically they can take many positions including: “the initial position, the predicate-initial position, and the final position” [10]. English, on the other hand, has an alphabet and uses a degree of inflection with a relatively fixed word order where DMs can, for the main, only be placed in the initial position [10].

Defining DMs is not necessarily a trivial task. Chinese uses a rich array of DMs to link parts of speech in both simple and complex sentences [2]. Chinese conjunctions appear in two main types: those linking words or short phrases (simple conjunctions) such as: 和 (hé – and), 跟 (gēn – and/with), 或 (huò – or) as in 刀和叉 (dāo hé chā – knife and fork), and those that link clauses (composite conjunctions). Conjunctives are also used often appearing in the main (usually second) clause of a sentence and link back to the previous clause [11]. Additionally, there are instances where clauses may be linked in a sentence without the use of any DM (zero connective structures). In

[1] For the purposes of this paper the term Chinese is used to mean Mandarin or Mandarin Chinese, considered to be the main standardised language of China [8].

these cases the meaning or context is strongly inferred across the clauses, leading to the creation of sentences that have natural omissions, which can cause problems for current SMT systems.

3 Settings: Corpora and SMT Systems

We used four well known corpora for gathering the data necessary for observing DM frequency and pertinent translations:

• Basic Travel Expression Corpus (BTEC): This corpus is primarily made up of short simple phrases and utterances that occur in travel conversations. For this study, 44016 sentences in each language were processed with over 250000 Chinese characters and over 300000 English words [12].

• Foreign Broadcast Information Service (FBIS) corpus: This corpus uses a variety of news stories and radio podcasts in Chinese. For this study 302996 parallel sentences were used containing 215 million Chinese characters and over 237 million English words.

• Ted Talks corpus (TED): This corpus is made up of approved translations of the live Ted Talks presentations[2]. This corpus contains over 300.000 Chinese characters and over 2 million English words [13] spread across 156805 parallel sentences.

• Multi-UN corpus (UN): This is a parallel corpus (for 6 languages) using data extracted from the United Nations Website. It includes over 220 million words in English and over 629 million Chinese characters in 8.8 million parallel sentences [14].

The SMT systems used to produce the automatic translations are Google Translate[3] and Bing Translator[4]. Whilst these are specific commercial translation tools and they may not represent the best quality translation systems for Chinese-English, they are good representatives of statistical translation approaches, known to use state of the art techniques and achieve reasonable translation quality. In addition they are freely available, making it possible to reproduce and expand the analysis presented here.

4 Analysis of Chinese Discourse Markers

In this section we examine the main types of Chinese DMs, including conjunctions for composite sentences, sequential paired conjunctions and zero connectives [11,15,16,17,18,19]. Our first step was a simple quantitative analysis to identify the most commonly used DMs in our corpora, so that we could select a few cases of interest to analyse in more detail. Table 1 shows the proportion of sentences containing the ten most frequent disyllabic DMs in the four different corpora. It also shows one or more frequent English translations for each DM, but we note that variants of these translations are possible.

While the percentages of sentences containing specific DMs in Table 1 may seem small at first, overall DMs are present in a significant proportion of sentences. The

[2] http://www.ted.com

[3] http://translate.google.com

[4] http://www.bing.com/translator

Table 1. Ten most frequently occurring DMs in the four corpora.

TED	UN
因为 (4.72%) : because	因此 (1.70%) : so/therefore
如果 (4.32%) : if	以便 (1.42%) : so that
所以 (4.05%) : so/therefore	因为 (1.24%) : because
但是 (3.58%) : but	由于 (1.22%) : due to/as a result of
或者 (1.68%) : or	如果 (1.05%) : if
还有 (1.59%) : furthermore	而且 (1.04%) : moreover
那么 (1.59%) : then/in that case	为了 (0.88%) : in order to
而且 (1.47%) : moreover	但是 (0.81%) : but
并且 (1.34%) : and also	并且 (0.73%) : and also
因此 (1.24%) : so/therefore	虽然 (0.62%) : although
FBIS	**BTEC**
因为 (1.39%) : because	如果 (1.18%) : if
如果 (1.30%) : if	但是 (1.10%) : but
因此 (1.19%) : so/therefore	那么 (0.44%) : then/in that case
为了 (1.13%) : in order to	还是 (0.39%) : or
由于 (1.10%) : due to/as a result of	所以 (0.29%) : so/therefore
但是 (1.01%) : but	因为 (0.25%) : because
而且 (0.85%) : moreover	或者 (0.23%) : or
虽然 (0.80%) : although	并且 (0.17%) : and also
然而 (0.79%) : however/but	只有 (0.17%) : only
甚至 (0.72%) : even	而且 (0.13%) : moreover

frequency analysis highlights certain trends, for instance 如果 (rúguǒ – if) and 因为 (yīnwèi – because) have a relatively high frequency in all four corpora. 因为 (yīnwèi) is classed as one of the high frequency (causal) connectives [20] and is considered to have a strong correlation in usage with ‘because’. In what follows we pinpoint some of the divergences in the use of these markers through examples of constructions, and connect these divergences to the behaviour of SMT systems when faced with such constructions.

Ex (1) shows the 因为 (yīnwèi) DM being used in a relatively short causal sentence, and it is clear that the SMT system has problems with the DM, dropping it completely from its position before the comma.

Ex (1)[5] 他因为病了，没来上课。
he because ill, not come class.
Because he was sick, he didn’t come to class. [18]
He is ill, absent. (Bing)

[5] Each example in this paper has the following format: Line 1 is the correct Chinese in characters; line 2 is a literal word-for-word translation; line 3 is the given translation and line 4 is (usually) the best translation returned by the SMT system. In some cases more than one automated translation is given for comparison purposes.

In Ex (1) the two parts of the sentence appear to have a very weak link in the translation as the DM is simply not used at all in the automated translation. The information after the comma (in the Chinese sentence) is correct and as Chinese does not use inflection, a sentence segment similar to ‘did not come to class’ should appear in the translation rather than simply having ’absent’. In Ex (2) the problem seems to be the reverse. The 因为 (yīnwèi – because) being present in the Chinese sentence causes problems for the SMT system as it tries to force ‘because’ into the translation (rather than omitting it) and by doing so significantly alters the meaning.

Ex (2) 你因为这个在吃什么药吗?
you because this (be) eat what medicine [MA]
Have you been taking anything for this? (BTEC)
What are you eating because of this medicine? (Google)

The automated translation gives the impression that the person has changed their diet due to having medicine, rather than their being required to take medicine for an ailment.

4.1 Sequential Constructions: Paired Conjunctions/Conjunctives

Paired DMs are frequently used in Chinese [21] and feature in many translations of complex sentences. Some paired constructions are formed using two conjunctions, but other formations are also possible such as: ‘conjunction …conjunctive’. Typical conjunctives include: 才 (cái – only/only if/ not unless), 就 (jiù – then/that), 却 (què – but/yet/while) and are often treated as connecting referential adverbs [11] . Conjunctions tend to appear in both clauses, while conjunctives frequently appear in just the second clause. They represent even more challenging problems for both human and machine translation.

Table 2 shows (for each corpus) the proportion of sentences that contain at least one occurrence of the given paired marker patterns. The main outcome of this frequency analysis is that for each corpus the …一…就…(…yī…jiù…) pattern appears with the highest frequency. However, manual inspection of a random sample of sentences showed that the …一…就…(…yī…jiù…) was only being used as a sequential paired marker construction in around one quarter of the cases.

Chinese does not have a specific word which maps one-to-one exactly with ‘then’ and so 就 (jiù) and 那么 (nàme – so) are often utilised to perform a similar function [18] . It is difficult to categorise 就 (jiù) on its own as it serves numerous functions. Many other characters such as 来 (lái) and 的 (de) can also be difficult to categorise for a similar reason, but perhaps none more so than the character ‘一’ (yī – one/single/ whole/same…) which covers six pages in the Oxford Chinese dictionary. By themselves 一 (yī) and 就 (jiù) can be ambiguous, but as a sequential construction they work together as a pair in a specific pattern with a relatively fixed meaning. Ex (3) shows a short five-character sentence that uses the …一…就…(…yī…jiù…) pattern as a sequential paired construction to mean: ‘…no sooner…than…’; ‘the moment…’; ‘as soon as…’; ‘once…’

Ex (3) 他一学就会。

Table 2. Ten most frequently occurring paired DMs in the four corpora.

TED	UN
…一…就…(3.67%) : once/as soon as, (then)	…一…就…(0.92%) : once/as soon as, (then)
…如果…就…(1.33%) : if,(then)	…越…越…(0.30%) : more, more
…如果…那…(0.95%) : if, (then)	…由于…因…(0.24%) : due to, because
…也…也…(0.49%) : also, and	…如果…就…(0.22%) : if,(then)
…越…越…(0.49%) : more, more	…不仅…而且…(0.21%) : not only, but also
…从…开始…(0.48%) : starting from…	…从…起…(0.17%) : starting from…
…是…还是…(0.48%) : [be], or	…从…开始…(0.14%) : starting from…
…如果…那么…(0.34%) : if, (then)	…是…还是…(0.14%) : [be], or
…不是…而是…(0.29%) : not, but(is)	…虽然…但是…(0.12%) : although, but
…从…起…(0.27%) : starting from…	…也…也…(0.11%) : also, and
FBIS	**BTEC**
…一…就…(2.20%) : once/as soon as, (then)	…一…就…(0.28%) : once/as soon as, (then)
…越…越…(0.63%) : more, more	…如果…就…(0.22%) : if, (then)
…也…也…(0.40%) : also, and	…从…开始…(0.15%) : starting from…
…从…起…(0.38%) : starting from…	…如果…那…(0.10%) : if, (then)
…如果…就…(0.36%) : if,(then)	…从…起…(0.09%) : starting from…
…从…开始…(0.35%) : starting from…	…是…还是…(0.06%) : [be], or
…不仅…而且…(0.30%) : not only, but also	…只要…就…(0.06%) : as long as, (then)
…是…还是…(0.27%) : [be], or	…又…又…(0.05%) : both, and
…既…又…(0.25%) : both, also	…越…越…(0.03%) : more, more
…既…也…(0.24%) : both, also	…的话…就…(0.03%) : …if, (then)

he as soon as study then can.
He learned it (the trick) in a jiffy. [22]
He learn. (Google)

In Ex (3) it is clear that very little concrete information can be extracted from the five characters alone, and there is a lot of inference such as the speed in which the person learned to do something (in this case – a trick). To identify both the ‘trick’ and ‘speed’ would require additional contextual information. The overarching pattern for the …一…就…(…yī…jiù…) construct is fairly simple: …一 VP^{a} 就 VP^{b}

The 一 (…yī…) should come immediately before the prepositional phrase and/or verb or verb phrase [18], although it can have some subject information that precedes it. In the case of Ex (3) a pronoun is used for the subject.

It is possible that by itself the sentence in Ex (3), while grammatically correct, has too much inference for an SMT system to manage and sentences that contain more information may produce better translations. The actions in the structure do not have to be related and the subjects in each clause do not have to be the same, but it is often the case that the second action is as a direct result of the first.

Ex (4) 一有空位我们就给你打电话。
As soon as have space we then give you make phone.

We'll call you as soon as there is an opening. (BTEC)
A space that we have to give you a call. (Google)

In Ex (4) the SMT system tries to remain closer to the actual order of the given sentence, but once again misses the 'as soon as'. If the word order is to be kept close to the original then a sentence similar to 'as soon as we have a vacancy (then) we will give you a call' could be used.

4.2 Linking Clauses Without Discourse Markers (Zero Connectives)

The zero connective [11] is often used to link closely set clauses where the meaning of the second clause is contextually implied by the meaning of the first clause. This can be done through repetition, answering or qualifying conditions as in Ex (5) or for rhythmic balance [17].

Ex (5) 东西太贵, 我不买。
things too expensive, I not buy
If things are too expensive, I won't buy them. [17]
Too expensive, I do not buy it. (Google)

In this case, the SMT system, appears to translate the Chinese word for word, but loses some meaning. The gist of the condition is evident, but the translation is not adequate. Manual insertion of two standard DMs into the sentence is actually required for the SMT system to produce a better output.

Ex (6) 如果东西太贵，我就不买 (了)。
If things too expensive, I then not buy(le).
If something is too expensive, I do not buy it. (Google)

5 Analysis of Chinese and English Discourse Markers in Parallel Corpora

In this Section we perform a quantitative analysis on the usage of DMs in both Chinese and English (human translation). SMT systems learn translation models primarily from parallel corpora with examples of translations aligned at the sentence level. The goal of this analysis is to study whether Chinese markers and their corresponding English markers appear in sentences that are aligned in parallel corpora. For a given DM, a high percentage of aligned sentences containing the marker in both Chinese and English could be an indication that learning the translation of such a marker from the corpus is potentially feasible. On the other hand, a low percentage of aligned sentences containing both Chinese and English markers could be an indication that the markers might be dropped or translated using different linguistic constructs, making the learning of SMT models a more difficult task.

Given that we start the analysis with Chinese DMs, a question that arises is how to find their corresponding English DMs. Each of the given DMs (Tables 1 and 2) are

Table 3. Frequencies of six Chinese DMs and their corresponding translations in parallel corpora.

Chinese Marker	Occurrence rate in Chinese (%)				Occurrence rate in human translation (%)				Appear in both the Chinese and English translation (%)			
	BTE	FBIS	UN	TED	BTE	FBIS	UN	TED	BTE	FBIS	UN	TED
因为 (because)	0.25	1.39	1.24	4.72	0.20	1.01	0.48	3.92	80	73	39	83
如果 (if)	1.18	1.30	1.05	4.32	1.15	1.09	0.76	3.84	89	84	72	89
因此 (consequently)	0.02	1.19	1.70	1.24	0.02	0.83	1.09	1.07	100	70	64	86
但是 (but)	1.10	1.01	0.81	3.58	1.07	0.89	0.54	3.19	97	88	67	89
而且 (moreover)	0.13	0.85	1.04	1.47	0.13	0.59	0.69	1.15	100	69	66	78
虽然 (although)	0.02	0.80	0.16	0.36	0.02	0.65	0.15	0.15	100	81	94	42

relatively common, but can naturally have variance in the associated translations. For example, a strong link has already been suggested between 因为 (yīnwèi) and ‘because’, but there are numerous comparable ways of uttering or writing ‘because’ such as: ‘in light of’, ‘for this reason’, ‘as a result of’ [23,24]. For this paper, interchangeable values are classed as variance rather than ambiguity. Ambiguity is taken to mean a word that has numerous different functions as per the individual characters ‘一’ yī and ‘就’ jiù discussed in Section 4.1.

Table 3 shows the occurrence percentages of six frequently used Chinese DMs in the four corpora. The first column shows the Chinese DM with its commonly associated English equivalent. Column two shows the occurrence rate of the Chinese marker in sentences across the corpora. Column three shows the occurrence rate where a directly equivalent English DM (with variance included) is used in the parallel translations (e.g. 因为 = ‘because’ or a variant of ‘because’); that is, for each set of sentences with a given Chinese DM, a subset is formed from the parallel translations of the sentences. The percentages in column three show the size of the resulting subsets compared to the size of the whole corpus. The final column shows the percentages of sentences that contain, within a set, both the Chinese DM along with the equivalent usage of an English DM in the translation. The percentages in the fourth column can be used as general measure of the strength of correlation.[6]

We note that the source language of our corpora is not always Chinese. For TED it is English, while for UN it could be any of the six languages. BTEC and FBIS however consist of segments originally in Chinese, and their translations into English. Therefore the implications of the numbers in Table 3 will be different for different corpora.

Overall, the numbers show that in short everyday sentences (BTEC) the main DMs are used as expected (e.g. 因为 maps closely to ‘because’ or a strong variant of ‘because’). As the sentences become more complex and are used at a higher level (FBIS and TED), then the way DMs are used becomes more fluid. The markers appear to be

[6] It must be noted that whilst the percentages show trends, there is still a small degree of error where less common variant phrases may have possibly been used in the parallel translation (e.g. because = this is down to). Detailed discussion of further variance is beyond the scope of this paper and can be considered in future work. The given percentages are considered to offer a close enough approximation for the related discussion.

increasingly omitted or absorbed into the general meaning of a clause rather than translated directly. As expected with the UN corpus, where complex language is used and discourse is divided into subsections, addenda and annexes, there is even less need for certain markers and there are inevitably fewer occurrences of items such as ‘if’ and ‘but’.

Ex (7) 这将是一次规模盛大, 而且受到广泛国际关注的聚会.
This will be one scale grand, moreover receive wide international attention [DE] meeting.
This will be a grand gathering with wide international concern. (FBIS)
This will be a grand scale, but widespread international concern gatherings. (Google)

In Ex (7), the 而且 (érqiě – moreover) is serving as a link that brings together the qualities of the meeting; that is, it will be on a ‘grand scale’ and will receive ‘wide international attention’. Clearly the human translation is very succinct and does away with the need for the ‘moreover’ or ‘furthermore’ type link.

For an SMT system to reach a similar translation it would need to be aware of when to drop the marker, and how to reorder the sentence accordingly. Additionally the [DE] adds complication as grammatically it implies that the described qualities (scale and attention) belong to the meeting, which is not necessarily an easy connection to automatically recognise.

6 Related Work

In the last few years much work has gone into improving machine translation from Chinese into English, including major efforts as part of the DARPA GALE program [25]. A number of useful parallel corpora and wordlists have been developed. Additionally, due to the contextual information connected to DMs, there has been a shift to working with them to improve MT. Initially projects such as the Penn Discourse (Chinese) Treebank (PDTB) [26] started identifying DMs according to type for parts of speech (POS) tagging. The Chinese Discourse Treebank (CDTB) [21] was designed to add a discourse annotation layer to the PDTB.

Efforts have also been made to improve identification of Chinese DMs through applying machine learning [2] and indeed categorising them in terms of relationship (e.g. causal and conditional). Additional work has gone into identifying the meaning of DMs [15] to ascertain their type (e.g. concession or contrast) and improve classification techniques. More recently, word reordering of grammatical structures around DMs, a known translation difficulty, is also being explored [27] and tools such as the Stanford Parser have been built.

Further work has gone into cross-lingual identification of DMs and disambiguation [28], which builds on information from bi-lingual dictionaries, the PDTB and parallel corpora. There is now a fresh trend with a focus on lexical and grammatical cohesion as well as the disambiguation of connectives [29,30,31] and recognising the variety of discourse relations attached to DMs [7].

7 Conclusions and Future Work

Chinese and English both belong to very different language families leading to numerous structural differences between the two languages including differing word order and the use of DMs. DMs in particular provide a level of lexical cohesion between phrases and clauses, but are not always utilised during the MT process. This means that sentential positioning is often incorrect and words are frequently omitted leading to unclear translations with a loss of context and information.

In many cases Chinese discourse has significant subject inference carried across clauses and sentences leading to contextual omission of many items (often pronouns) within a sentence. Ex (9) shows a modified version of Ex (2) where the pronoun and second marker have been manually inserted into the Chinese sentence. With the extra information Bing returns a better translation, highlighting the importance of preserving DM/contextual information.

Ex (8) 他因为病了, 所以他没来上课。(modified version of Ex (2))
he because ill, so he not come class. (extra *he* and *so* in the 2nd clause)
Because he was sick, he didn't come to class.
He is ill, so he did not come to class. (Bing)

In the case of paired DMs, especially with the 一 (yī) and 就 (jiu) structure, the SMT systems struggled with inference and disambiguation, often failing to spot the 'as soon as' relation. The main focus of this paper has been on Chinese to English translation. A positive next step would be to analyse DM usage and translation patterns for English to Chinese translation, which would enable comparisons of DMs in both directions. A detailed analysis of the comparisons would look at the relative sentential positioning of DMs and examine where direct equivalents do and do not exist. Additionally, where DMs are used, it will be important to examine the changes in word order that are required to accommodate the respective DMs in the target language. Part of this will include an in depth analysis of contextual omissions (e.g. dropped pronouns) within a sentence, but also will examine the distance that context can be carried in discourse through larger discourse segments that go beyond sentence level.

It is also expected that analysis of translations in both directions will produce more data detailing the variance that can be applied to individual DMs and hence work can go into developing improved recognition of such variance. The results of this further analysis of the corpora will provide insights to help develop a framework to model discourse markers in SMT between Chinese and English.

References

1. Zuffery, S., Degand, L.: Annotating the Meaning of Discourse Connectives in Multilingual Corpora. Corpus Linguistics and Linguistic Theory. Volume 0, Issue 0, pp. 1–24. (2013)
2. Tsou, B., Gao,W., Lai, T., Chan, S.: Applying Machine Learning to Identify Chinese Discourse Markers. In: International Conference on Information, Intelligence and Systems, Chania Crete, Greece. (1999)

3. Hussein, M.: Two Accounts of Discourse Markers in English. University of Damascus, Syria. (2002)
4. Hardmeier, C.: Discourse in Statistical Machine Translation: A Survey and a Case Study. In: Discours – Revue de linguistique, psycholinguistique et informatique, Caen, Presses Universitaires de Caen. (2012)
5. Meyer, T., Webber, B.: Implicitation of Discourse Connectives in (Machine) Translation. In: Workshop on Discourse in Machine Translation (DiscoMT), pp. 19–26. Sofia, Bulgaria. (2013)
6. Hardmeier, C., Stymne, S., Tiedemann, J,. Nivre, J.: Docent: A Document-Level Decoder for Phrase-Based Statistical Machine Translation. In: 51st Annual Meeting of the ACL. Sofia, Bulgaria, pp. 193–198. (2013)
7. Hajlaoni, N., Popsecu-Belis, A.: Translating English Discourse Connectives into Arabic: a Corpus-based analysis and an Evaluation Metric. In: CAASL4 Workshop at AMTA (Fourth Workshop on Computational Approaches to Arabic Script-based Languages), San Diego, CA, pp. 1–8. (2013)
8. Swan, M., Smith, B.: Learner English (2nd Edition) Cambridge University Press, Cambridge, UK. (2004)
9. Chang, P., Jurafsky, D., Manning, C.: Disambiguating "DE" for Chinese-English Machine Translation. In: 4th Workshop on SMT, pp. 215–223, Athens, Greece. (2009a)
10. Li, Y.: Sensitive Positions and Chinese Complex Sentences: A Comparative Perspective. Journal of Chinese Language and Computing, 18(2): pp. 47–59. (2008)
11. Po-Ching, Y., Rimmington, D.: A Comprehensive Grammar. Routledge, London. (2004)
12. Takezawa, T., Sumita, E., Sugaya, F., Yamamoto, H., Yamamoto, S.: Toward a broad-coverage bilingual corpus for speech translation of travel conversations in the real world. In: LREC, pp. 147–152. Las Palmas, Spain. (2002)
13. Cettolo, M., Girardi, C., Federico, M.: WIT3: Web Inventory of Transcribed and Translated Talks. In: EAMT, pp. 261–268, Trento, Italy. (2012).
14. Eisele, A.,Chen, Y.: MultiUN: A Multilingual Corpus from United Nation Documents. In: 7th conference on International Language Resources and Evaluation, Pages 2868-2872, La Valletta, Malta. (2010)
15. Hutchinson, B.: Acquiring the Meaning of Discourse Markers. In: 42nd meeting of ACL, Main Volume, pp. 684–691. Barcelona, Spain. (2004)
16. Po-Ching, Y., Rimmington, D.: Chinese: Intermediate Chinese, A Grammar and Workbook. Routledge, London. (1998)
17. Po-Ching, Y., Rimmington, D.: Chinese: An Essential Grammar (2nd Edition). Routledge, London. (2010)
18. Ross, C., Sheng Ma, J.: Modern Mandarin Chinese Grammar. Routledge, London. (2006)
19. The Conjunction 2010, http://www.chineseteachers.com/blog/resource_content.jsp?id=142
20. Wang, C., Huang, L.: Grammaticalisation of Connectives in Mandarin Chinese: A Corpus-Based Study. Language and Linguistics, Volume 7, No. 4, pp. 991–1016. (2006)
21. Xue, N.: Annotating Discourse Connectives in the Chinese Treebank. In: ACL Workshop on Frontiers in Corpus Annotation 2: Pie in the Sky. (2005)
22. Oxford Chinese Dictionary: English-Chinese Chinese-English. Oxford University Press, UK. (2009)
23. Macmillan Publishers Limited 2009–2014. http://www.macmillandictionary.com/thesaurus-category/british/
24. Thesauraus.com. Roget's 21st Century Thesaurus, Third Edition, http://thesaurus.com/
25. Olive, J., Christianson, C., McCary, J.: Handbook of Natural Language Processing and Machine Translation. Springer, New York. (2011)

26. Xia, F.: The Part-Of-Speech Tagging Guidelines for the Penn Chinese Treebank (3.0). In Technical Reports, IRCS Report 00-07. Pennsylvania. (2000)
27. Chang, P., Tseng, H., Jurafsky, D., Manning, C.: Discriminative Reordering with Chinese Grammatical Relations Features. In: 3rd Workshop on Syntax and Structure in Statistical Translation at NACCL HTL, Boulder, Colorado. (2009b)
28. Zhou, L., Gao, W., Li, B., Wei, Z., Wong, K.: Cross-lingual Identification of Ambiguous Discourse Connectives for Resource-Poor Language. In: 24th International Conference on Computational Linguistics (COLING). Mumba, India. (2012)
29. Tu, M., Zhou, Y., Zong, C.: Enhancing Grammatical Cohesion: Generating Transitional Expressions for SMT. In: 52nd annual meeting of the ACL, June 23–25, Baltimore, USA. (2014)
30. Guilou, L.: Analysing Lexical Consistency in Translation. In: Workshop on Discourse in Machine Translation (DiscoMT), pp. 10–18. Sofia, Bulgaria. (2013)
31. Wong, B., Kit, C.: Extending machine translation Evaluation Metrics with Lexical Cohesion to Document Level. In: 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pp. 1060–1068, Jeju Island, Korea. (2012)