ABSTRACT

This study investigates and compares the translation output quality of two statistical machine translation (SMT) systems – Google Translate and Bing Translator, by performing a human evaluation method called ‘linguistic evaluation’. The language pair in the translation tasks is Chinese – English (with English as the target language), and the domain is news articles. 50 Chinese sentences extracted from several lengthy Chinese news articles were automatically translated by Google Translate and Bing Translator into 50 sets of translations in English. Errors in the output of both systems were manually analysed and annotated based on the proposed error taxonomy, which allowed me to evaluate two MT systems at each linguistic level, namely the orthographical level, the morphological level, the semantic level, the lexical level, and the syntactic level.

A fine-grained taxonomy of linguistic errors is proposed and implemented in the study. Subcategories of errors at each linguistic level are tailored and defined for Chinese-English language pair (with English as the target language). The output sentences are analysed thoroughly, using a standardised form of ‘markup’ with an input-output mapping.

The results show that in the same quantity of Chinese-to-English translation tasks, Bing Translator, an SMT system which incorporates linguistic information, does outperform Google Translate, which is a pure SMT system that does not use linguistic rules to perform translation tasks. In general, Bing produces fewer linguistic errors, especially at syntactic level. The distribution of error types shows that syntactic and lexical errors are particularly problematic in both SMT systems, which suggests this is where developers should focus when attempting to improve the output quality of Chinese-English translation tasks.

Acknowledgements

This has been my most anticipated part. It means my project has finally come to a wonderful end. It also means I will soon proudly complete my master's study and fulfil one of my many dreams after over two years living and studying in Norway, an amazingly beautiful country. Looking back, my heart is now being filled with gratitude and appreciation.

I would like to express my heartfelt thanks to Professor Annjo K. Greenall, my research supervisor. Annjo has encouraged me to conduct this challenging project from the beginning and has supported me with lots of constructive comments and valuable opinions on my work during the whole process of my research. Without those extensive discussions with her, I would not have been able to complete this thesis.

I am especially and extremely grateful to Associate Professor Daniel Weston, for his tremendous help in my thesis. Daniel was not involved in this project from the start, but Annjo decided to get him in to offer advice on my study lately since September 2016. He is absolutely one of the smartest and nicest people I have ever seen in my life. His bright insights and generous comments brought a breath of fresh air into my project. Besides, his quick response and careful editing only made me more motivated to get things done as soon as possible. I will not forget the talk with him when I was once desperate and helpless because of some of my personal issues. His understanding and sensitivity made me feel safe to open up to him, and his listening and advice meant a lot to me. In addition to being a respectable teacher, he has become a life mentor.

Susan Seats is my best American friend. I met her about 6 years ago when she was teaching us Conversational English at Hubei University of Technology in Wuhan, China. Susan kept encouraging me to pursue a higher degree abroad because she knew this had been my dream for years. When I told her I was accepted into a master's program in Norway, she was thrilled and urged me to visit her in the US during my school break. I visited her twice in December 2015 and in June 2016, which were the best times and the most unforgettable experiences in my life. Susan was eagerly helping me with my thesis when I told her the topic of my project. She made a phone call directly to the Google Company trying to seek advice. Apart from that, she took me to community or university libraries in California and Arizona to look for any resources that might be useful to my research. She has also asked as many people as possible if any of them happened to know somebody who worked at Google or Microsoft. We did find someone and I reached out, but sadly I did not hear anything back. Nevertheless, I greatly appreciate what Susan has done in trying to help me with my project.

I met most of Susan's family and friends in California, Arizona and Idaho, and I am grateful to all of them for their help. Her nephew Brad Kula, who was living in a van and aiming to become a successful actor in Hollywood, took me to the Six Flags Magic Mountain for thrill rides when I told him I was stressed out because of my thesis. After my first-ever roller coaster experience, Brad said to me, 'Once you have ridden the real roller coasters, you can cope with all the roller coasters in your life.' Thanks to that, I became ready for more challenges in my life and felt more cheered up to complete my master's thesis as soon as possible.

Tom Kula and Penny Kula, Susan's older brother and his wife, hosted me when I was in Scottsdale, Arizona. I was writing my thesis at their place while I was enjoying the precious sunshine when it was dark and cold in Norway. Their generosity and hospitality made me feel home. A lot of work of my thesis was done when I was there. Penny is a dancer, painter and teacher. She shared her thoughts on my project and kindly collected a lot of newspaper clippings when she read any articles about Google translate or language technology. She took me to ballet and yoga classes, in order to reduce my stress level while I was working on my thesis. Tom, as an IT engineer, also did some editing and proofreading for my writings and offered bright advice.

I am also grateful to Joyce Buekers, who hosted me one night when I was in Phoenix, Arizona. She introduced me Dr. Robert W. Callan who is a senior systems management executive with more than 25 years of experience in Information Systems Development and Implementation. Joyce set up a Skype meeting for Robert and me, and he answered some of my research questions and we kept in touch via emails. His encouragement and wisdom was a great help to me.

Alec Yu, my best Chinese friend who currently lives in China, has been unconditionally supportive during my studies in Norway. Despite six to seven hours' time difference, he is always there, whenever I need him for a chat or discussion about my studies. He has also offered huge help with the formatting of my thesis and perfecting Excel tables and figures in my project, since he is a master of such work.

Thanks to Sigurd Gundersen, my yoga instructor for making me fall in love with yoga and keep going to yoga classes at least three times a week. Without the relaxation and meditation from yoga, I would not have been able to handle the enormous stress from three degrees that I am taking at the same time at NTNU and at UiO.

I am also very grateful to Håvard Fjær for his tremendous support and encouragement since December 2016. As a senior IT engineer, he has offered a lot of wise comments and advice on my study. He can always guide me through when I encountered difficulties in solving some logical or technical issues. I have to thank him for the 'endless' coffee that he has provided, giving me 'endless' energy to work harder so that I can complete my thesis just in time.

Of course, I want to thank Google and Bing, for creating such amazing products for us. Information technology has changed people's life, and I firmly believe that language technology in the field of computational linguistics will be a stronger area of international communication in the future, which will make the world truly connected and united.

My greatest gratitude goes to my family, my father Anhua Chen, my mother Baoyu Peng, my sister Lina Chen, and my dog Tiger. Without their support and love, I would not have come so far and made such an accomplishment in a foreign country far away from home. My thesis is dedicated to them. Lastly, a big hug for myself.

Ding Chen, March 2017

CONTENTS

List of Figures

List of Tables

List of Examples

1. Introduction

In March 2016, AlphaGo, a computer program developed by Google DeepMind, took on and defeated the legendary human player of the ancient Asian board game 'Go', marking a major milestone for artificial intelligence in human history. Ever since, increasing concerns have centred around the fact that technological developments will create even more unemployment in the next 20 years, because machines might eventually replace an incredibly large number of human professions, of which human translators might be one example. Whether or not machine translation (MT) will replace human translators in the near future is of course still disputable, but the popularity of using machine translation by different groups of people at present is undoubtedly increasing. In the past decades since the 1950s, this convergence between linguistics and computation has spread, sped up, blossomed and evolved. Machine translation has moved from the fringes of society where it was ignored and even stopped in the 1960s, to the centre stage of our modern digital information society today.

The usefulness of machine translation in a highly globalised, web-connected and multicultural world has been attested, and appreciated, by numerous people, especially Internet users. Nowadays, we can choose between many online machine translation services that can provide automated translation from a given source text, or even a whole webpage, into a target language in the blink of an eye. Businessmen use these translation services when they need to translate emails from a foreign language into their mother tongue; journalists use them when they need to understand news written in a foreign language; students use them when they are learning a second language, but need a convenient and efficient dictionary, or when they are studying in a foreign country only to find that information is written in a language they can barely pronounce. Machine translation has quietly become an almost indispensable utility for many of us in our daily lives.

Since the 1950s, the approach to machine translation has evolved from methods based on grammatical rules to methods based on corpus, and even some hybrid approaches that combine the best properties of the previous. One of the newest machine translation systems, known as statistical machine translation, is currently one of the best-favoured and most famous examples that provide free online automated translation services for Internet users. Such systems are not perfect, because they still make mistakes – sometimes minor, sometimes so serious. And language pairs like Chinese-English, which are typologically different, make accuracy of translation harder than pairs of related languages like German and English. Nevertheless, Google Translate, a popular online translation service and its free online competitors, such as Microsoft's Bing Translator, manage to offer a usable approximation. Years ago, many people complained that they were not even able to translate the simplest greetings correctly, but nowadays, more and more of us are finding that these services make it possible for people to understand even lengthy news reports.

There have been only a handful of studies which try to evaluate the output quality of statistical machine translation such as Google Translate or Bing Translator (Aiken and Balan, 2011, Balk et al., 2013, Li et al., 2014, Ghasemi and Hashemian, 2016), and the studied language pairs are quite limited. Plus, the comparison of these two translation services remains unexplored. This project is intended to investigate and compare Google Translate and Bing Translator, by performing a human evaluation method called 'linguistic evaluation'. Both translation systems are basically statistical machine translations, while Bing claims to be 'linguistically informed' which incorporates linguistic rules (I will come back to this in Chapter 2, Section 2.1.2). It is therefore expected that their performances would differ, where Bing should have a more linguistically well-formed output in general. This is the point of departure of my research, and my motivation is to explore how their performances would differ when they are performing the same quantity of Chinese-English translation tasks and what the causes are for their performance differences. I will be able to answer those questions in the end of my study.

The outputs of Chinese-English translation tasks (with English as the target language) are evaluated and compared, by detecting and annotating the errors that Google and Bing tend to make. All the errors are classified into five big categories at different linguistic levels, namely the orthographical level, the morphological level, the semantic level, the lexical level, and the syntactic level. The subcategories under each linguistic level for the Chinese-English translations are defined with specific examples. In the end, the total numbers of errors at each linguistic level generated by the two systems are counted and compared. A study such as this would be valuable to MT system developers, because the evaluation results can give an understanding of what types of errors are the most frequent in the outputs and what causes the errors, which will help them in building models to improve their translation systems.

The structure of the thesis is as follows: **Chapter 2** presents a literature review on machine translation (MT) and MT evaluation, followed by a brief description of research motivation and predictions. Comprehensive descriptions of the history of MT and the main approaches of MT (**rule-based MT**, **corpus-based MT**, and **hybrid MT**) are provided in section 2.1, and four major types of MT evaluation methods (the automatic evaluation, the adequacy and fluency judgements, the error analysis, and the linguistic evaluation) are introduced in section 2.2. The research questions and expectations are formulated in section 2.3. **Chapter 3** is the methodology part. It describes how the experiments are conducted, including how the corpora are chosen; how an output sentence is annotated and analysed; how the errors are classified in a taxonomy containing subcategories at different linguistic levels for the Chinese-English language pair. Then the error statistics and the interpretation of the statistics and the evaluation results are presented in **Chapter 4**, with tables, figures and charts. General observations extracted from the data are included. In **Chapter 5**, I briefly discuss the findings from the study in light of my research questions, including the discussion about the linguistic evaluation method that I refined and implemented, and the discussion about the evaluation results which indicate some critical, illustrative types of errors in the outputs. Conjectural sources of errors at each level and possible solutions are provided. In addition, the limitations and potential sources of problems of this study are also explored in this chapter. Finally, **Chapter 6** presents the conclusions along with suggestions for future work.

2. Theoretical Framework

2.1 Machine Translation

Machine translation refers to computerized systems that can be used to automatically translate texts or speech from one natural language into another with or without human assistance or intervention (Hutchins, 1995). The holy grail of the MT world is FAHQT, known as Fully Automatic High Quality Translation, where ‘high quality’ implies something approaching that of a good human translator for unrestricted input text (Bennett and Gerber, 2003). The primary objective of MT research during the early years was to develop MT systems that can produce high-quality translation without human translator’s assistance, before eventually replacing the human translators in the translation industry entirely, but no MT system has achieved the goal and no one considers MT a solved problem (Van et al, 2012).

Even though this objective has been disfavoured by most professional human translators and is obviously not possible at the moment or in the near future, the developments of MT through history have been strikingly dramatic and the objective has evolved over time. By now, MT has been subject to revived interest from an increasing number of researchers in the field of translation, multilingualism, computational linguistics, informatics, natural language processing, artificial intelligence, sociolinguistics, and so on.

2.1.1 A Brief History of Machine Translation

Research on MT began to emerge in the 1950s, soon after the computer was invented. In the past few decades MT has seen both marked advances and setbacks. Among a large number of researchers in the field of MT, John Hutchins has stood out as one of the most well-known MT researchers who has provided fruitful and salient studies of the history of MT in the open literature. I shall therefore refer mainly and particularly to Hutchins’ (1992, 1995, 2006 and 2010) descriptions of the history of MT in this section, to provide a concise picture of the history of MT from the 1950s till now.

In the 1950s, research on MT had already started at many universities in the US. This was because an American mathematics professor, Warren Weaver, had written a memorandum in July 1949 in which he put forward various proposals concerning MT, based on the achievements of code-breaking during the wartime, developments in information theory, and conjecture about universal principles underlying natural languages. His proposals led to and inspired the pioneering research on MT in the US, and the first public demonstration of the feasibility of MT was presented by the joint team of IBM and Georgetown University in 1954. This demonstration successfully stimulated a large amount of funding for research on MT in the US, despite its very limited capability for performing translation tasks. Ever since, MT research has spread enthusiastically to other parts of the world, especially the Soviet Union and Western Europe. The first decade in the 1950s was a decade of optimism for MT research, with researchers' predictions of impending breakthroughs. According to Slocum (1985), interest and support were mainly fuelled by visions of high-speed and high-quality translation of arbitrary texts at that time, especially in military organisations.

During the years 1954-1964, the second decade of MT encountered obvious linguistic problems that began to hinder further progress. Most of the major MT systems were based on bilingual dictionaries, where the entries of the source language had equivalences in the output, and simple grammatical rules were implemented in the models for analysing the linguistic information in the source language, and producing the correct word order in the target language. However, researchers were soon confronted with so-called 'semantic barriers', because they could not ignore that their systems lacked vital semantic information and syntactic disambiguation programs. Wilss (1982) notes that even though they had programmes that could provide grammatical analyses of sentences in the source language, they could not generate adequate output translations because of the large amount of remaining ambiguity.

By 1964, the funding sponsors of the US government had become more concerned about the progress of MT research and had set up the Automatic Language Processing Advisory Committee (ALPAC) in order to evaluate the work in MT research. With the disappointing results and slow progress, ALPAC published an 'infamous' report in 1966, indicating that MT was slower, less accurate and far more expensive than human translation and concluding that there was no immediate prospect of useful MT. This report stopped further investment and funding for the MT research and brought it into a remarkably dark period for nearly a decade both in the United States and elsewhere in the world. Paradoxically, an exception during that period was that one of the oldest machine translation companies – Systran – was founded in 1968 and was installed by the United States Air Force (USAF) in 1970. It was soon being employed by the Commission of the European Communities in 1976 as a tool for translating their increasing number of official documentations. The same year witnessed another system's installation in Canada for translating weather forecasts from English to French, known as the Météo system. These successful systems encouraged the MT research to continue.

From the 1980s till the early 1990s, MT underwent a revival in research, with the emergence of a diversity of MT systems, from various countries throughout the world. Famous systems such as Systran, Logos and the Metal were mainly built on 'mainframe' (large digital) computers. With the advent of microcomputers and personal computers, lower-end and cheaper MT systems were rapidly created in Europe, the US, and even Asia, including China, Japan and Korea. Apart from the PC versions of MT, some online translation services also began to appear.

In the late 1990s and the beginning of 2000s, with the explosion of the Internet, a higher efficiency and faster response performance from MT was urged, which triggered a growth of online automatic translation services in the market, such as Babelfish powered by Yahoo, Google Translate by Google and Bing Translator by Microsoft.

Until today, the use of online MT has expanded dramatically in a lot of areas including education, business, social media and more. These online translation services attract millions of users every year, and big IT companies such as Google and Microsoft have never stopped developing and improving their online translation products. In the past decade, they have made a lot of progress in performance, and user satisfaction has been steadily increasing. In part the continuous interest and motivation for MT research is because of more realistic expectations of what is possible in MT, specifically the fact that MT is acknowledged as useful if imperfect.

2.1.2 The Main Approaches of Machine Translation

The description above gives a historical perspective on how MT has developed in the past decades since 1950s until today. To better understand the differences among different MT systems and their applications, it is necessary to know some of the main approaches used in different phases of MT research. According to the core approaches, MT systems can be classified into three broad categories, which are **rule-based** machine translation (RBMT), **corpus-based** machine translation (also known as data-driven machine translation) and **hybrid** machine translation that combines the best properties of highly advanced pure rule-based or/and corpus-based approaches (Costa-jussà and Fonollosa, 2015).

2.1.2.1 Rule-based Machine Translation

From the 1950s to the early 1990s, RBMT constituted the main type of MT systems for research and commercial use. RBMT, just as its name implies, is a type of machine translation based on rules, namely linguistic rules. A rule-based translation consists of a process of analysing input sentences of source language and generating output sentences of target language by using dictionaries (lexicon) and/or grammar rules. The grammar rules basically involve syntax, semantics, morphology, part-of-speech tagging, orthographic features and so on. Three main approaches of RBMT are the **direct** approach, the **transfer** approach and the **interlingua** approach.

According to Jurafsky and Martin (2008: 867), the **direct** approach implements a large bilingual dictionary to facilitate word-by-word translations. In direct translations, the source language text is translated word-by-word using the bilingual dictionary. Basically, each entry in the dictionary can be seen as a small program which is responsible for translating a single word. Shallow morphological analysis and morphological generation can be applied before and after the words are translated. Besides, some very simple reordering rules or minor grammatical adjustment can be involved, for example, moving adjectives after nouns when translating from English to French. The direct approach is the most basic approach of MT systems. The process of a direct machine translation is illustrated in Figure 1.

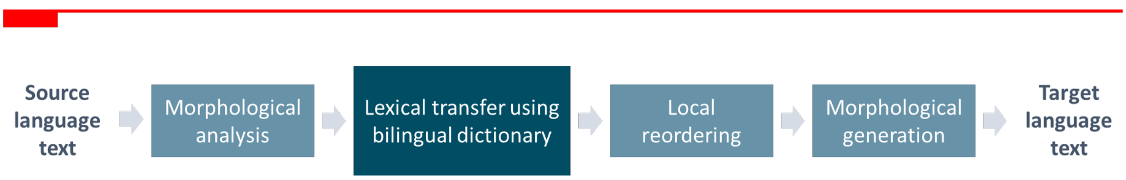


Figure 1. *Direct machine translation, modified from Jurafsky and Martin (2008)*

In the **transfer** approach, the source language input text is parsed and then linguistic rules for transfer (syntactic, semantic or lexical information) are applied to transform the source language parse into a target language parse. The output sentence is then generated from the parse tree (Nirenburg and Wilks, 2000). The transfer approach operates over three stages: analysis, transfer and generation. A simplified transfer model with two language pairs (French-English, English-French) is shown in Figure 2.

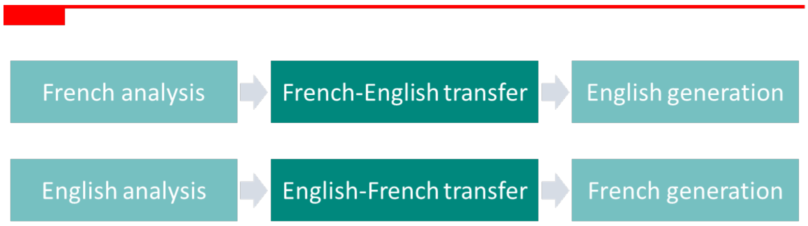


Figure 2. *The transfer model (analysis-transfer-generation), based on Hutchins (1992)*

Hutchins (1992:75) indicates that there are no language-independent representations in the transfer approach. The source language intermediate representation is specific to a particular language, as is the target language intermediate representation. Due to the language-dependent nature, different transfer rules for transformations in different language pairs are needed. For example, Yamada and Knight (2001) describe the syntactic transformations from English sentence structure to Japanese sentence structure as is illustrated in Figure 3.

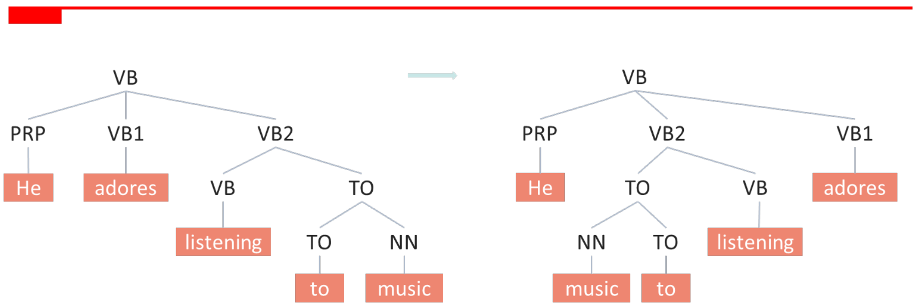


Figure 3. *Syntactic transformations from English order to Japanese order, based on Yamada and Knight (2001)*

For translating languages with SVO structure like English to languages with SOV structure like Japanese, specific syntactic transfer rules are required for moving the verb to the end after the NP and VP complements, changing prepositions to postpositions and so on. A big challenge with the transfer approach is that a distinct set of transfer rules will be required for every different pair of languages. The amount of knowledge that needs to be built for different pairs of languages is massive, which can take years to develop.

Jurafsky and Martin (2008) point out that the pure direct approach is no longer used, but the transformational intuition underlies most of the modern MT systems. The most obvious problem with the direct approach is that it is to a large extent focused on individual words and has no parsing component or any knowledge about grammatical structure in the source or target language. This can cause a lot of difficulties in handling long-distance reordering, phrases or larger structures. Therefore, phrasal and structural knowledge must be incorporated in the MT models in order to deal with real examples. The transfer approach is normally better able to cope with more complex source language phenomena than the direct approach, but it proves that some simple transfer rules such as SVO-SOV rules for translating from English to Japanese as mentioned before, are still not sufficient. Thus, in practice, more 'messy' rules that combine a large amount of lexical knowledge of both source language and target language with syntactic and semantic information are needed.

In fact, a lot of commercial MT systems tend to combine the direct and transfer approaches, using large bilingual dictionaries and also parsers/taggers. One of the famous RBMT systems, the Systran, founded in the 1960s as mentioned in the previous section, was a typical manifestation of the RBMT that combined these two approaches. Senellart et al. (2001) describe that the Systran system has three components, including a shallow analysis stage, a transfer stage and a synthesis stage. The analysis stage includes morphological analysis and part-of-speech tagging, chunking of NPs, PPs and larger phrases, and shallow dependency parsing (subjects, passives, head modifiers and so on). The transfer stage includes translation of idioms, word sense disambiguation, and assignment of prepositions according to governing verbs. The final synthesis stage includes lexical translation with a rich bilingual dictionary to do lexical translation, reordering, and morphological generation. Behaving like a direct system, Systran relies on the large bilingual dictionary for much of its processing. At the same time, it informs many of its steps by syntactic and semantic processing of the source language, like a transfer system.

In addition to the direct and the transfer approaches, another typical method used in the RBMT systems is the **interlingua** approach. This approach is based on the argument that MT must go beyond purely linguistic information and involve an 'understanding' of the content of texts (Hutchins and Somers, 1992). The interlingua idea arose in the hope of creating an abstract universal language-independent representation of meaning (the interlingua). According to Alansary (2011), the motivation behind the idea is that while languages can differ greatly in their 'surface structure', they all share a common 'deep structure'. It presupposes the existence of an interlingua that is able to represent all sentences that mean the same thing in the same way, no matter which language they are in. The interlingua approach regards translation as a process of extracting the meaning of the input text and then expressing that meaning in the target language. It translates texts by performing a deep semantic analysis (using semantic analyzer techniques) on the input from language X into the interlingua representation and generating from the interlingua to the output language Y. The architecture of interlingua systems is shown in Figure 4.

Figure 4. *Interlingua architecture*

As Peng (2013) puts it, interlingua compared with the other rule-based machine translation methods is the most attractive, better suitable approach for multilingual translation systems. However, despite its attractiveness and advantages for multilingual translations, this approach is too ideal. Hutchins (1992) points out that there are some major disadvantages of interlingua systems: it is extremely difficult to define an interlingua, even for closely related languages such as the Romance languages. A truly 'universal' and language-independent interlingua has defied the best efforts of linguists and philosophers from the seventeenth century onwards. Due to many complexities, only one interlingua MT system has ever been made in a commercial setting (Nyberg and Mitamura, 1992), and only a few have been taken beyond research prototype, including the Universal Translator, known as UNITRAN (Dorr, 1987), the Universal Networking Language, known as UNL (Uchida, 1996) and a few others.

In order to visualize these three approaches more clearly, it is common to use Vauquois Triangle, also known as the Machine Translation Pyramid (Vauquois, 1968) to illustrate the main differences of them, as is shown in Figure 5.

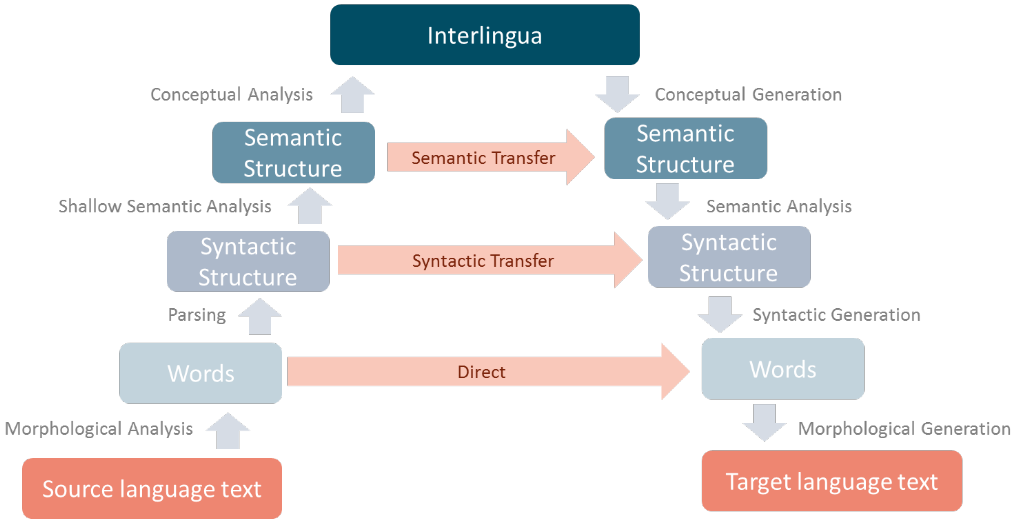


Figure 5. *The Vauquois Triangle (1968)*

First, the depth of analysis and generation required in the translation process in each of the three approaches is increasing. Moving from the bottom of the triangle to the top of the triangle, it goes from (shallow) morphological analysis to parsing, to shallow semantic analysis, then to conceptual analysis on the source-language side. On the target-language side, it goes from morphological generation, to syntactic generation, to semantic generation, then to conceptual generation. In other words, the interlingua approach does full analysis and generation, whereas the direct approach does a minimum of analysis and generation. The transfer approach is somewhere in between.

Second, the amount of transfer knowledge in different approaches is decreasing. At the direct level where a word-by-word approach is used, nearly all knowledge is ‘transfer knowledge’ (lexical transfer). As we move up the triangle, ‘transfer knowledge’ is only used for parse trees (syntactic transfer) and thematic roles (semantic transfer) in the transfer approach. On the top of the triangle, there is no specific transfer rule in the interlingua approach.

Since each of the approaches mentioned above has their pros and cons, the real systems of RBMT in the commercial settings tend to involve the combinations of the elements from these three approaches. A simplified model of the standard architecture of the common RBMT systems is presented in Figure 6 below.

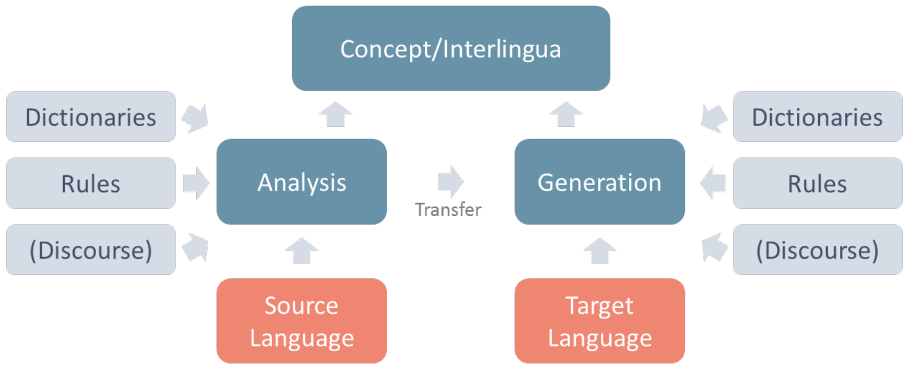


Figure 6. *The standard architecture of RBMT systems*

2.1.2.2 Corpus-based Machine Translation

The RBMT systems were mainly applied during the first four decades since the 1950s. In the early 1990s, corpus-based methods began to be experimented with. Since then, the research on exclusively rule-based methods has declined. According to Kaji (1988), the main reason is that the approaches of RBMT depend heavily on language theories, and the knowledge sources or grammatical rules formulated in the translation process have to be provided by linguistic experts, which requires huge effort in terms of human labour and large amount of money. In addition, the language pairs that can be applied in the machine translation systems are quite limited. Koehn (2010) also argues that language is so rich and complex and always ambiguous that it can never be fully analysed and distilled into a set of rules, which is the reason why MT has to take a new direction, from rule-based machine translation to corpus-based machine translation.

Two major approaches of corpus-based machine translation are: example-based machine translation (EBMT) and statistical machine translation (SMT). At present, it is obvious that the statistical method is the dominant framework in the field of MT research, as can be seen in the proceedings of the annual conference of the Association for Computational Linguistics (ACL). This study is mainly concerned with SMT, so EBMT will not be introduced.

Jurafsky and Martin (2008:875) explain that SMT uses a quite different way to approach the problem of translation compared to the rule-based approaches, because it focuses on the result rather than the process. In practice, the consensus of philosophers of translation seems to be that it is, strictly speaking, impossible for a sentence in one language to be a translation of a sentence in another. This is not only due to culture-specific problems, but also because of translation challenges whenever a language uses a metaphor, a word, a tense or a construction without an exact parallel in the other language. Therefore, we will have to compromise, in order to produce a translation, which is tolerably faithful to the source language and acceptably natural as an utterance in the target language. This perspective gives a hint for how to do MT – using a statistical method to find the 'most probable translation' of a sentence.

A typical SMT does not understand the languages or know any linguistic rules, but relies on a machine that discovers the rules of translations automatically from a large corpus of translated texts by pairing the input and output of the translation process and learning from the statistics over the data (Koehn, 2010). This method using probability to do translation tasks was fundamentally inspired by the statistical approach in the research of speech recognition, where a phrase like *going to go* would be assigned a higher probability of being uttered in speech than *going two go* or *going too go* (Sharman et al., 1990).

In the late 1980s, the very first statistical approaches to MT were pioneered by a group of researchers from IBM, known as IBM Model 1 and IBM Model 2 that are both **word-based** SMT (Brown et al., 1990). Since then, SMT has become overwhelmingly dominant in the field and has advanced from word-based to **phrase-based** models (Och and Ney, 2004). The intuition of phrase-based SMT is to use phrases (sequences of words) as well as single words as the fundamental units of translation. Xiong and Zhang (2015) illustrate a Chinese-to-English translation example that visualises the process of phrase-based SMT, as shown in Figure 7.

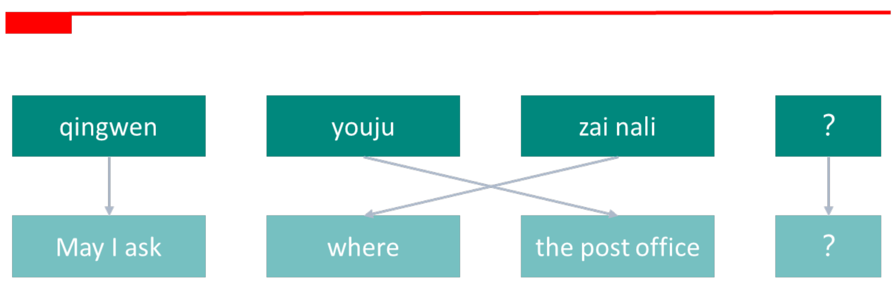


Figure 7. *A Chinese-to-English translation example by the phrase-based SMT*

According to Figure 7, there are three steps to translate a Chinese sentence into an English sentence by the phrase-based SMT: phrase segmentation, phrase translation and phrase reordering. The input is segmented into phrases in Chinese and then translated one-to-one into phrases in English, and finally reordered as a ‘comparatively fluent’ English sentence (the verb in the subordinate clause is missing). It is important to note that the phrase segmentation is not necessarily linguistically motivated, as most of the current phrase-based models are not rooted in any deep linguistic notion of the concept phrase and the process of segmentation is not modelled explicitly. That is to say, any segmentation in an input sentence is equally likely. In the translation step, the phrases or sequences of words are not translated by a bilingual dictionary as in the direct approach, but translated according to the probability in a phrase translation table (as in Figure 9). Besides, the system does not necessarily perform translation in the order from the leftmost phrase to the rightmost phrase of the input sentence. Additionally, in the reordering step, the output sentence is not reordered according to any grammatical rules as in the transfer approach of RBMT, but again, based on probability.

The points mentioned above indicate that most of the modern SMT models have essentially ignored linguistic aspects, unlike the traditional MT models that relied on various levels of linguistic analysis. This, in turn, implies that the output of SMT may be prone to errors on various linguistic levels (I will come back to this later in Chapter 3, Section 3.3).

To better understand how SMT system works, we need to learn three basic components required in SMT: a **translational model**, a **language model** and a **decoder**. For simplicity’s sake, the complex technical architectures and mathematical definitions or algorithms will not be introduced here in detail since they are not highly relevant to this study. One can, however, easily consult Jurafsky and Martin (2008: 877-859) to learn the workings of SMT. Here I present a simplified architecture of SMT as shown in Figure 8, to explain those three components.

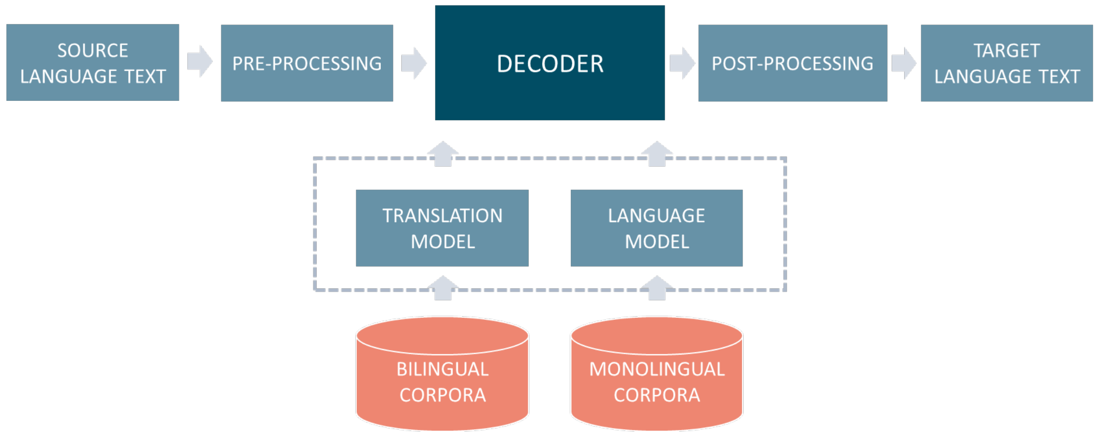


Figure 8. *The architecture of SMT*

A source language text is pre-processed (segmentation), and then a decoder searches the most probable words or phrases with the translation model and the language model. The found words or phrases are post-processed (word alignment) and possibly reordered, and at the end the target language text is produced.

A **translation model** is used to pick an input phrase and translate it to an output phrase. By consulting the phrase translation table, it looks up the highest probability for this pair of phrases. For example, a phrase translation table of English translations for the Chinese phrase *qingwen* may look like Figure 9. The translational model is trained with bilingual/ parallel corpora, which contains bodies of text that have been translated from one language to another. The parallel corpora are mainly sourced from international organisations such as UN and EU where there is sufficient documentation in multiple languages. Localised software manuals and translated literature are also important training data.

TRANSLATION	PROBABILITY
May I ask	0.5
Do you know	0.35
Please ask	0.15
Please, ask	0.05

Figure 9. *A translation table with p-value (probability)*

A **language model**, which is built on monolingual corpora, analyses large amount of text in a certain language and notes the frequency how certain words or phrases collocate. This model helps the system produce the most statistically correct word order in the target language output, and supports difficult decisions about choices of words when a source word has more than one translation. Formally, a language model is a function that takes an English sentence (English as the target language in the translation tasks) and computes the probability that it was produced by an English speaker. For example, a good language model would assign a higher probability to the sentence *the house is small* than the sentence *small the is house*. For another example, if a foreign word (such as *haus* in German) has multiple translations (*house, home, building*, …), the language model would give higher probability to a more natural word choice in a specific context as in *I am going home* rather than *I am going house* (Koehn, 2010:181).

A **decoder** functions as the statistical machine translator which uses the translation model and language model conjointly to produce the most probable translation output in a most probable word order.

According to Resnik and Smith (2003), massive numbers of parallel or monolingual corpora are required in order to improve the ability of the SMT systems to make decisions while performing translation tasks. Monolingual corpora are not too hard to build, but parallel human-translated corpora are harder to come by. Texts from bilingual/multilingual political documentation, manuals or literature are obviously not available in necessary quantity. Therefore, vast amounts of parallel or monolingual text have been hoovered up from the web. As a result, IT companies which boast powerful search engines and advanced computing technologies have more advantages while developing statistical machine translation systems. Although various free online statistical machine translation services have sprung up on the Internet, Google Translate developed by Google Company, stands out and enjoys considerable popularity.

Google Translate is a project that began in 2001 but was officially launched in 2006 in order to provide a free of charge online translation service for general Internet users. It used the Systran's engine (mentioned before, which was a rule-based machine translation system) until 2007 when Google developed its own proprietary, in-house phrase-based statistical machine translation system (Schwartz, 2007). The Economist (2017) reveals that in order to build the translation model for the system, Google trawled nearly a trillion web pages, searching any text that seemed to be a translation of another – for example, many pages are designed identically but have several versions with different languages.

Google Translate can now translate text, whole web pages, speech and even real-time video from one language into another at high speed. Compared with traditional RBMT systems, which takes linguists years to develop the rules for translating different language pairs, Google Translate, as an SMT service, does not apply any grammatical rules, which makes it possible to train the system on more data from more languages, more quickly. Google Translate serves more than 200 million people every single day and it translates over 100 billion words daily; more than the content of a million books (Shankland, 2013). According to its official website, it now supports 103 languages for 6 different features of translation tasks including TYPE, TALK, SNAP, SEE, WRITE and OFFLINE. TYPE is the most common and the most frequently used feature, as you can easily type or paste any text that you want to translate on Google's translation website (https://translate.google.com/) or on the smart phone application. TALK makes it possible for you to have a bilingual conversation with a person that speaks a different language, because it can translate your speech input instantly on the website or on the mobile application and read the output in the target language. SNAP can translate an image with text that you want to read in a different language. SEE is a feature that gets you to see real-time translations on your phone when you point your phone's camera at any text that appears on anything surrounding you in real life. WRITE allows you to draw letters or scribble characters with your finger as the input text. OFFLINE helps you get text translations without an Internet connection.

Currently, German, English, Chinese, Catalan, French and some others are supported for all 6 features, while some other languages are supported at various levels (at least TYPE works with all the supported languages). Besides, an increasing number of currently unsupported languages are also in development, which fits into Google's ultimate mission – to organise information worldwide and make it universally accessible (Shankland, 2013). Due to the nature of SMT, the translation ability to translate different languages is highly dependent on the language pairs, which means there is a higher possibility of producing better output for close language pairs (for example, Spanish and Portuguese) and languages for which larger amounts of parallel corpora are available (for example, English and other European languages, because of the prominence of human-translated EU documentation) (Barreiro et al., 2014). Worth mentioning here is that shortly after the translation service was launched, Google Translate won an international competition for English-Chinese machine translation. This may be attributed to the large number of parallel corpora available for English-Chinese language pair (Nielsen, 2011).

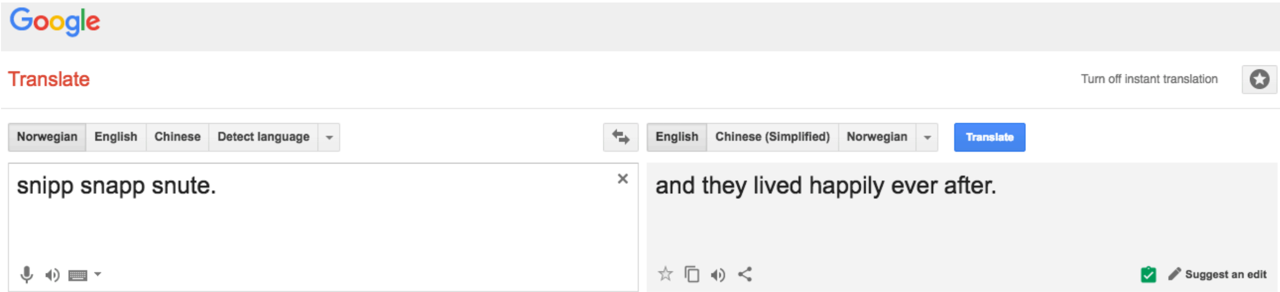


Figure 10*. Web interface of Google Translate*

Source language (Norwegian): snipp snapp snute.

Target language (English): and they lived happily ever after.

As is shown in Figure 10, the box on the left side is for the users to feed in text or website URL links as source language input. Before translating, one can easily choose any one of the supported languages by clicking the language option bars, or even just choose 'detect language' if one does not know what the source language is, so the system will automatically detect it. The box on the right side is for presenting the output text in the target language (which can also be chosen before translating). It takes just seconds for the system to complete the translation task, even from quite lengthy articles or a whole webpage, within a word limit of five thousand. One special feature is that anyone can click the right bottom icon 'Suggest an edit' if the output text is somehow unsatisfactory to the user. The edited data will be collected anonymously by Google for later research in order to improve their system (Babst, 2015).

By September 2016, Google announced a new breakthrough in their research on MT by implementing Google's Neural Machine Translation system, which has achieved more competitive translation results in terms of accuracy, speed and robustness (Wu et al., 2016). Their newest study can be seen as marking the state-of-the-art of MT research, and will undoubtedly stimulate more motivation for further studies as well as more optimistic perspectives for the future prospects of machine translation.

2.1.2.3 The Hybrid Approach

Since exclusively rule-based approaches were phased out, and the statistical method has become dominant, machine translation systems have made huge progress. Nevertheless, they continue to experience limitations, for example the linguistic shallowness of SMT limits its capabilities when applied to morphologically-rich languages and language pairs with highly divergent syntax (Lavie, 2011). Therefore, some hybrid approaches that integrate linguistic information with the statistics-based methods are at the forefront of MT research. The most common hybrids currently combine the SMT with syntactic modelling or morphological analysis in order to run the MT systems; the research at Microsoft is a prime example of this type of approach (Hutchins, 2010).

By employing language specific parsing, dependency and word alignment rules, MT at Microsoft is better able to generate more linguistically accurate translation output than the other conventional non-linguistic statistical machine translation systems (Dolan et al., 2002). The research at Microsoft has been focusing on developing new approaches that incorporate the power of phrase-based SMT with linguistic information. Quirk et al. (2005) at Microsoft describe a novel approach to SMT that combines syntactic information in the source language with the advances in their phrase-based translation system. Their system employs a source-language dependency parser and a target-language word segmentation component. To translate an input sentence, the system produces a dependency tree for the input sentence after parsing it, and then a decoder will find a combination of translation pairs that cover the source tree and have optimal probability according to a set of core models. Recall that in most SMT systems, the phrase segmentations are not linguistically motivated, this approach, however, gains phrases including combinations such as adjective-noun, article-noun, verb-object and so on. Another important advantage of Quirk et al.'s (2005) proposed approach is that they employ more powerful models that can incorporate information from the analysis of the source text. For instance, they may directly model the probability that an English pre-modifying adjective should be translated into a French post-modifier, or the probability that the translation of an object of a verb in English should precede the corresponding verb in Japanese.

Powered by Microsoft Company, Bing Translator is one of its online translation services, which is totally free of charge (they have their enterprise version, too). Its user interface and the way of using it are almost the same as Google Translate (see Figure 11 below), but without the 'editing' feature and with a far smaller language platform than Google Translate (50+ languages so far by October 2016, according to the official website of Microsoft).

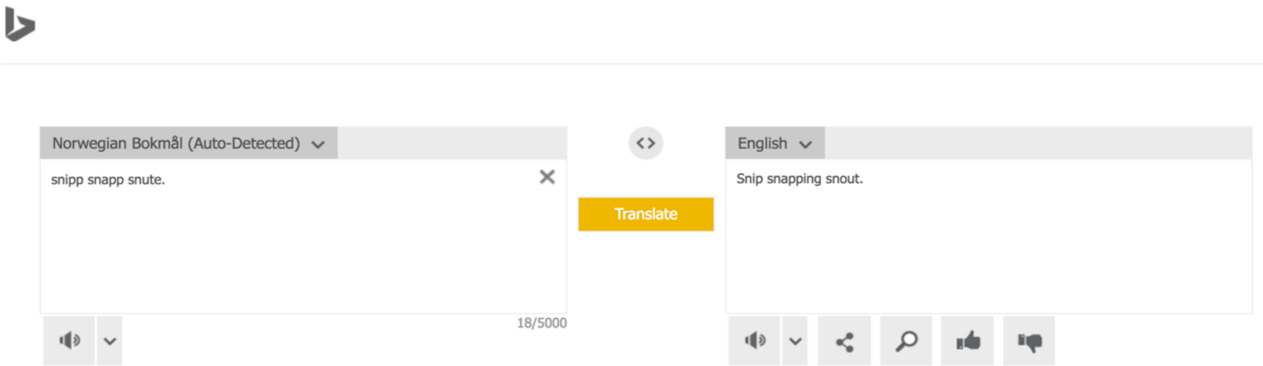


Figure 11. *Web interface of Bing Translator*

Source language (Norwegian): snipp snapp snute.

Target language (English): snip snapping snout.

Bing Translator was used internally for Microsoft groups since 2006 before becoming available as a SaaS (Software as a Service) web API (application programming interface) for customers in 2011. Similar to Google Translate, Bing has five main features – TEXT, SPEECH, PHOTO, CONVERSATION MODE and OFFLINE. Simply put, it can do all what Google Translate is able to do, except for the features of WRITE and SEE of Google as mentioned above. One important difference, however, is that it claims to be a **linguistically informed** phrasal statistical machine translation service, and thus typifies the hybrid approach of MT that uses statistical methods but with greater reliance on an incorporation with linguistic information.

2.2 Machine Translation Evaluation

White et al. (1994) notes that evaluation has always been central to the consciousness of those who are involved in the field of MT, though the role of MT itself has already evolved considerably since the 1950s, from a research subject to a practical utility in our daily life. Machine Translation evaluation (MT evaluation) is currently a very active field of research in the machine translation community (Koehn, 2004), but the evaluation of machine translation output is actually very challenging and controversial (Vilar et al., 2006). Historically, the evaluation of MT output has proven difficult, controversial, at times misleading (such as the notorious ALPAC report of 1966 mentioned before), but very often revealing and helpful (White et al., 1993). The quality of MT output is a major concern for both researchers and general users, because its evaluation will on one hand help boost the improvement of MT systems, while on the other benefitting the increasing number of general users in the market.

Evaluation of translation output can be carried out in various ways, yet there is no general accepted standard metric of doing it. As King (1997) explains, most frequently, an intuitive judgement is involved in an evaluation of translations, which may be based on knowledge of the languages in question or previous accumulated experience of translation. This sounds fine in terms of evaluation of human translations. However, when it comes to the evaluation of machine translation output, the task becomes more complex, simply because machine translation systems do not produce the same kind of translations as do human translators. Thus, the knowledge of the languages involved and the accumulated experience of translation are not quite reliable for making judgements on machine translation. One example described by White et al. (1994) is the first actual evaluation performed for machine translation in 1992, on behalf of Advanced Research Projects Agency (ARPA), a United States government agency tasked with funding research. A grading criterion which was normally used for evaluating human translations was given to a panel of translation experts to rate some MT output. First of all, they immediately found out that the grading criterion had to be modified to take the messy types of errors into account. Additionally, it seemed impossible for the panel of experts to reach a consensus of the grades to be assigned.

Though subject to considerable constraints and controversies, a wide variety of MT evaluation metrics have been proposed and used since the 1990s after the series of ARPA evaluations were carried out. MT evaluation can serve a purpose of judging whether an MT system adequately satisfies a set of specific needs (for example, aiding human translators), assessing whether a specific MT system has made progress in itself after development and adjustment, or diagnosing the system by determining and analysing what fails the system and why (King, 2007). Sometimes, MT evaluations are also used to compare several different types of MT systems, with the intention of comparing the performance of different systems, or/and diagnosing the source of problems of the systems.

Two major approaches to MT evaluation are: automatic evaluation and manual evaluation (human evaluation). Currently, the most popular automatic evaluation baseline metrics include, BLEU (bilingual evaluation understudy) (Papineni et al., 2002), an evaluation understudy developed by NIST (National Institute of Standards and Technology) (Doddington, 2002) and METEOR (Metrics for Evaluation of Translation with Explicit Ordering) (Lavie and Agarwal, 2007). A variety of human evaluation methods have been proposed and explored, and I will introduce three major types: Adequacy and Fluency Judgements (LDC, 2005), Error Analysis (Vilar et al., 2006) and Linguistic Evaluation (Farrús et al., 2010).

According to the official website of Microsoft, Bing Translator regularly evaluates the quality with BLUE standards and their own benchmarks (both automatic and human evaluations), constantly improving their machine learning engines and language models. However, what standard Google Translate uses to perform evaluation of their system remains unknown to the public (I have also sent inquiries about their evaluation method to Google which can be seen in Appendix A, but I have not received any answer by the time when I submit this thesis).

2.2.1 Automatic Evaluation of Machine Translation

The automatic evaluation method assesses the quality of machine translation output by a computer program based on human translations (also called reference translations). It compares the output of MT with the reference translations provided by experts, in terms of the statistics of short sequences of words, namely word N-grams (N-gram is a terminology in computational linguistics and probability, meaning a contiguous sequence of *n* items from a given sequence of text or speech and the items can be phonemes, syllables, letters or words). A word N-gram is a sequence of *n* words – for instance, a 2-gram is a two-word sequence of words and a 3-gram is a three-word sequence of words (Jurafsky and Martin, 2008:83). In the automatic evaluations, the more of the word N-grams that a translation output shares with the expert reference translations, the higher scoring is awarded to the translation (Doddington, 2002).

The intuition of such evaluation metrics derives from Miller and Beebe-Center (1958), who point out that good MT output is something that is very similar to a human translation. But the complication is that a source sentence could be legitimately translated in many possible ways and the human translations are not always able to cover all possibilities. In other words, a very good MT output sentence might resemble one human translation sentence, but could look very unlike another one. Even though multiple human translations of each sentence are normally provided for an automatic evaluation as references, this method is still quite problematic. It is inevitably subjective because it relies on the human translations while the quality of human translations varies. In addition, the program could wrongly decide that good output is in fact bad simply because it does not look like the human translation provided. More important, Vilar et al. (2006) and Farrús et al. (2010) point out that the identification of the most prominent sources of errors which is important for further development of a specific system, is impossible using those automatic metrics alone, because the results of the scoring only indicate which output words are correct. In other words, automatic evaluations can be useful for comparing several different MT systems or evaluating the progress of one MT system with the results of the scores, but cannot provide instructive information of the underlying problems of the systems. Nonetheless, automatic metrics still have been widely used in a number of MT systems for evaluation, especially for SMT, mainly because of its quick, inexpensive and language-independent features. Many of the automatic evaluation methods are particularly favoured by the developers of MT systems, because they need to monitor the effect of daily changes to their systems so that they can weed out bad ideas from good ideas in time (Papineni et al., 2002).

The technical architecture and mathematical algorithms of the automatic evaluation methods will not be introduced here due to limited space and its limited relevance, however, some of the most important manual/human evaluation methods, will be presented in detail in the following sections. As is acknowledged in the translation community, in spite of this being time consuming and expensive, human evaluation is still considered the best approach so far (Baisa, 2009).

2.2.2 Adequacy and Fluency Judgements

In order to evaluate the machine translation output, an obvious way is to look at the output and subjectively judge whether it is correct. Originally introduced by the Linguistic Data Consortium (LDC) for evaluation of MT, this type of human evaluation uses a straightforward numerical range and a coarse correctness standard to have evaluators (human annotators) make quality judgements of MT output (LDC, 2005). Highly proficient bilingual evaluators are generally best qualified to make these judgements because they have a good knowledge of both input and output languages. However, sometimes monolingual evaluators who only understand the output target language are used when such bilinguals are not available, in which case a reference translation (translated by expert human translators) is normally provided, so that they will only look at the output and directly make judgements based on the reference translations.

Two criteria are commonly used in such human perceptual evaluations: **adequacy** and **fluency** (Koehn, 2010: 218). There is by no means a clear definition of what a good-quality translation means, but it is quite natural to expect that a target sentence should preserve the meaning of the source sentence and that the target sentence is sufficiently comprehensible (Baisa, 2009). The former requirement is called *adequacy* (sometimes *accuracy*) and the latter one is called *fluency* (also known as *intelligibility*).

In the evaluation introduced by LDC (2005), **adequacy** is judged by asking if the output conveys the same meaning as the input sentence, or if part of the message is lost, added or distorted Options for evaluators include *all meaning, most meaning, much meaning, little meaning, none*. **Fluency** is judged by asking if the output sentence is good fluent language, which involves both grammatical correctness and idiomatic word choices. Options for evaluators include *flawless, good, non-native, disfluent,* and *incomprehensible*. When presented with a translation output, a source sentence and/or a reference sentence, evaluators need to score an output sentence generated by the MT, between 1-5 in adequacy and fluency (1 lowest, 5 highest). Figure 12 shows the score scale for the adequacy and fluency judgements (English as the target language).

ADEQUACY		FLUENCY	
1	All meaning	1	Flawless English
2	Most meaning	2	Good English
3	Much meaning	3	Non-native English
4	Little meaning	4	Disfluent English
5	None	5	Incomprehensible

Figure 12. *Adequacy and fluency scores for human evaluation, based on Koehn (2010)*

As is apparent, the standards and definitions for scoring are very vague and clearly too subjective, which easily makes the evaluation results unreliable. It is also exceedingly difficult for the human evaluators to be consistent in their judgement, especially when they fall on boundaries between the categories (Denkowski and Lavie, 2010). Therefore, this type of human evaluation method was not the most favoured one for MT output, and as a result of this, researchers began actively to seek novel ways to evaluate output, among which the *Error Analysis* developed by Vilar et al. (2006) became one of the most representative.

2.2.3 Error Analysis

Rather than evaluate an MT system by giving perceptual judgements or scoring based on harsh standard, it may be more convincing and less subjective simply to count how many errors the translation system produces. Vilar et al. (2006) propose an error analysis of SMT output based on a comparatively systematic classification of translation errors within a hierarchical structure. The classification is shown in Figure 13.

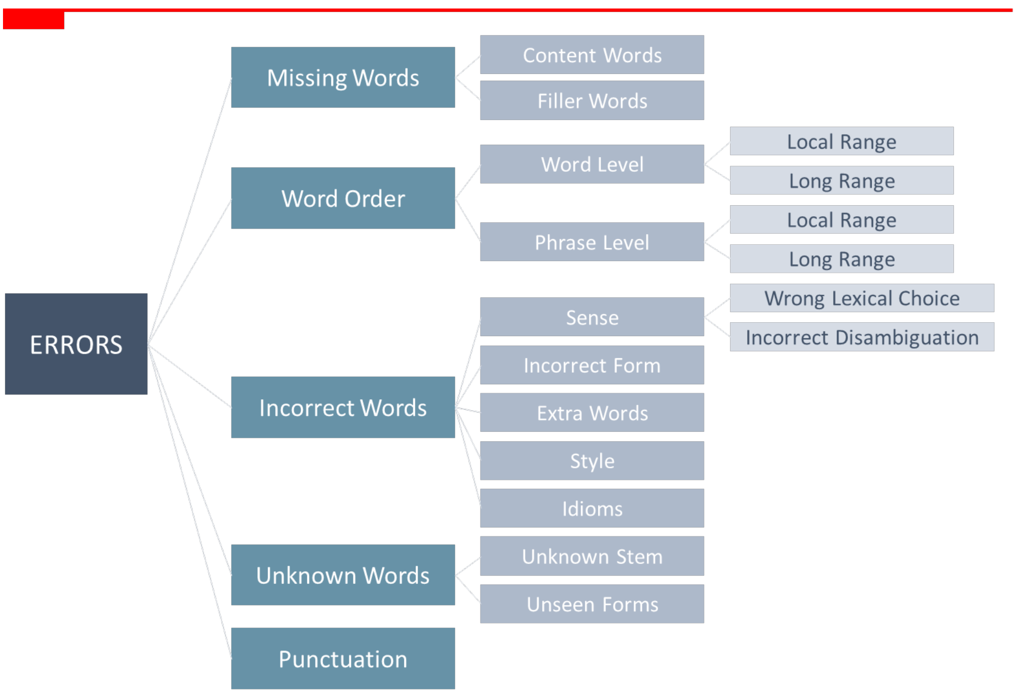


Figure 13. *Classification of translation errors — Vilar et al. (2006)*

In Vilar et al.'s (2006) model, all the errors in the MT output are classified into five big classes: **Missing Words, Word Order, Incorrect Words, Unknown Words and Punctuation Errors.**

- A **Missing Words** error is counted when a word in the generated output sentence is missing. The missing words may be essential for the meaning of the source sentence, or may be only necessary for forming a grammatically correct sentence. Vilar et al. (2006) define these two types as *content words* and *filler words*. They note that in most cases the first type of errors are caused by missing content words such as nouns or verbs, but sometimes the missing of a function word such as a preposition can change the meaning of the source sentence greatly. It is therefore, complex to sort all missing words into different subclasses, but as long as there are words found missing, it belongs to this type: a 'missing words' error.
- A **Word Order** error concerns the word order in the output sentence, either at *word level* or *phrase level*. Problems can be solved by moving words or phrases to another location within the sentence. Each level has two subclasses (i.e. local range and long range). As noted in their study, the distinction between local and long range is hard to define in absolute terms, but they intend to express the difference of reordering words/phrases in a local context (i.e. within the same syntactic chunk) or moving words/phrases into another chunk.
- **Incorrect Words** are identified when the translation wrongly translates some words. As classified in their taxonomy, they are defined in terms of *sense, incorrect form, extra words, style* and *idioms*. At the level of *sense*, errors fall into two subcategories – *wrong lexical choice* when the MT system chooses a wrong translation and *incorrect disambiguation* when the system fails to disambiguate the meaning of a source word (i.e. it chooses the wrong meaning of a polysemy word). *Incorrect form* is mainly concerned with morphological errors such as wrong verb or noun inflection. *Extra words* are defined as words produced by the system, which are redundant in the generated sentence. Here, Vilar at al. (2006) have not clarified if the *extra words* are seen as addition of meaning or redundant in the sense of grammar. The *style* and *idioms* errors are considered less important. The former concerns the influence of words on the style of the text, for example, a word is translated correctly but repeated too many times in a near context. The latter relates to errors when the system translates idiomatic expressions into a verbatim equivalence.
- **Unknown Words** are another source of errors of MT output when the MT system wrongly translates some unknown words. Unknown words literally refer to words that are not included in the training data. The MT system has never seen such words and as

a result can easily produce wrong translations. Proper names are a main source of such errors. Vilar et al. (2006) distinguish the unknown words with four categories for Chinese-English translations: *the person name*, *the location name*, *the organization name* and *other names*. Unlike translating languages that share the same alphabet (for example, many of the European languages), Chinese proper names cannot be translated by simply copying the input words to the generated sentence.

- **Punctuation** errors are considered minor but are nonetheless included in the translation errors classification.

Vilar et al. (2006) conclude that inevitably, there are some limitations in their proposed model. First, in order to find and count errors in a translation text, one or more human translations (reference translations) must be provided to contrast with the machine translation output. However, as is mentioned before, there are always multiple ways to correctly translate a given source text. Therefore, it is hard to keep consistent when identifying errors. For example, a translation may be diagnosed with an error of Missing Words based on one reference translation, but may be legitimately identified as an error of Incorrect Words based on another. The inconsistency poses problems particularly for comparison of several different MT systems. In addition, the classification of the errors of a machine translation system is by no means unambiguous; the error types, so defined, are not mutually exclusive. In a final analysis, it is quite common to see one kind of error causing another to occur. It is, therefore, hard to define which type one error belongs to. For instance, a missing word could also easily cause incorrect word order in the generated output sentences.

This method, although subject to limitations, can nevertheless, provide a more qualitative analysis of the errors and a slightly less subjective evaluation result, compared to automatic evaluation or human perceptual quality judgements. More important, it is, to some extent, helpful in identifying the most prominent source of errors in a given system which can then be revelatory for future research and development. In an error analysis carried out by Vilar er al. (2006), for evaluating an SMT system developed by RWTH Aachen University (a German research university), it is reported that the most important type of errors is language pair dependent. For example, word order errors in translation from Chinese into English, or the verb tense errors (i.e. *incorrect form* errors) in translation from English into Spanish are the most problematic in the translation output analysed. This conclusion can help the developers take account of word order more specifically when building models for Chinese-English pair, and spend more efforts in developing models with morphological analysis for English-Spanish pair.

Furthermore, this method sheds light on how to approach MT evaluation in a systematic manner of classification and even provides possibility of creating a set of automatic metrics which can perform automatic error classification for MT output (Popovic and Burchardt, 2011).

2.2.4 Linguistic Evaluation

Farrús et al. (2010) propose a novel way of identifying and classifying all the errors encountered in the MT translation output, using linguistic-based evaluation criteria. This method is called *linguistic evaluation*. Similar to the error analysis approach, a linguistic evaluation also presents a classification of all translation errors produced by MT systems. The major difference is that it takes specific linguistic information into account and classifies all errors into categories which are at five linguistic levels: **orthographical level**, **morphological level**, **semantic level**, **lexical level** and **syntactic level**. The guidelines for categorisation are summarised by Farrús et al. (2010) as follows:

- **Orthographical errors**: all the errors concerning the misuse of punctuation and misspelling of words in the target sentence, for example: accented vowels, apostrophes, letter capitalization and so on.

- **Morphological errors**: lack of gender and number concordance, errors in verbal inflection and lexical morphology (derivation and compounding), and morpho-syntactic changes due to changes in syntactic structures.

- **Lexical errors**: words that are incorrectly translated, words that are missing in the target language and words that are extra in the target language.

- **Semantic errors**: Wrong choice of meaning in the target language when a source word has multiple meanings (polysemy) and when a source word shares the same spelling and pronunciation but has different meanings (homonymy).

- **Syntactic errors**: errors in prepositions, relative clauses, verbal periphrasis, clitics, missing or spare articles in front of proper nouns, and syntactic element reordering.

Compared to the model developed by Vilar et al. (2006), this error specification is ambitious as it tries to include all types of possible errors that can occur in a translation from one language to another, and more importantly, it also offers more linguistic information about the type of errors which can be useful for developing or improving linguistically motivated MT systems.

The linguistic classification includes five linguistic categories, which is supposed to be generalizable to any other language in the translation output. However, their model has, as yet, only been applied to a Spanish-Catalan and English-Catalan language pairs (with Catalan as the target language). Moreover, the subcategories and annotation guidelines in these translation pairs are specifically related to Catalan, and sometimes related to both the source and target languages involved (Farrús et al., 2011, 2012). Thus, the five-category schema containing five linguistic levels is language pair independent, but the subcategories for a specific language pair are not.

Linguistic evaluation is one of the latest human evaluation approaches for MT output. Up to now, few linguistically-motivated evaluations of MT output have been done nor can they easily be found in the research literature. This method is very extremely time consuming as a single evaluator is required to finish the whole evaluating task. But it is less subjective than other human evaluation methods because very specific and strict evaluation guidelines have to be provided and this only one evaluator can be fairly consistent when performing the evaluation task. More importantly, the linguistic information obtained from the analysis of the linguistic evaluation results can be highly useful in the research and development of MT systems, especially SMT systems (Popovic and Ney, 2009).

2.3 Research Motivation and Predictions

Due to the fact that the linguistic evaluation approach remains largely unexplored but has promising future prospects, the motivation of this thesis is to explore it, by evaluating and comparing two different online MT services: Google Translate and Bing Translator. Since the Chinese-English language pair (with English as the target language), to date has not been studied in the linguistic evaluation approach, this study attempts to investigate and compare the translation output of two MT systems when translating from Chinese into English. Recall that Google Translate and Bing Translator both use phrase-based statistical approach when performing translation tasks, but the main difference is that Bing Translator incorporates linguistic information. Even though how much linguistic information Bing has employed in their translation system for Chinese-English language pair is unclear (to get such information from Microsoft is difficult), it is reasonable to presume that the outputs of Bing Translator are distinctive from that of Google Translate, and it will be interesting to see how the output quality differs.

Two primary goals of this study are:

- To evaluate and compare the translation outputs of Google Translate and Bing Translator by the linguistic evaluation method
- To estimate the distribution of error types in order to identify which error types have most influence and which are particularly problematic for a given translation system.

Two sub-goals of this study are:

- To test the feasibility of the linguistic evaluation approach of Farrús et al. (2010), by exploring how time-consuming it is, how practical it is and how difficult it is. How much time does it take to perform the linguistic evaluation for my study? Is it possible to classify all errors produced by the MT systems into the categories at the five linguistic levels as proposed? What kinds of challenges and limitations have been encountered when carrying out the evaluation? All these questions are aimed to be answered in the end of my study.
- To propose and design a thorough and fine-grained taxonomy of linguistic evaluation for Chinese-English translation output, by presenting a list of subcategories and specific annotation guidelines that are tailored for the translation errors produced by these MT systems when translating Chinese into English.

Since Bing Translator claims to be a linguistically informed SMT service that incorporates various linguistic information, for example, syntactic analysis on the source side (Dolan et al., 2002 and Quirk et al., 2005), the first prediction for this study is:

- Bing Translator should outperform Google Translate with regard to the total number of errors when performing the same translation tasks. That is to say it will produce fewer errors in total, and have a better performance at least with regard to syntax.

In their linguistic evaluation experiments for English-to-Catalan and Spanish-to-Catalan translations by several different MT systems, Farrús et al. (2012), conclude that the most frequent errors in all systems are found at the syntactic and semantic levels, and that orthographical errors are the least frequent. Based on this conclusion, the second prediction for this study is:

- In the translation outputs of both Google Translate and Bing Translator, syntactic and semantic errors will have a larger distribution than the others, while there will be substantially fewer orthographical errors. The number of morphological and lexical errors will fall somewhere in between.

3. Experimental Setup

In this chapter I elaborate how the experiments are conducted, including: how the corpora are chosen, how the subcategories of errors at five different linguistic levels for the Chinese-English pair are determined, and the procedures by which an output sentence is annotated and analysed. The MT systems under comparison are Google Translate and Bing Translator, which have been described in Chapter 2.

3.1 Corpora

The initial research proposal for this project suggested evaluating the output of four language pairs, namely Chinese-English, English-Chinese, Norwegian-English and English-Norwegian (Note that *Chinese* in the whole study refers to the written mandarin Chinese, simplified). The source texts were intended to be extracted from five genres/domains, including texts from: a language phrasebook, online news, email commercials, abstracts from academic papers, and texts on new media such as Facebook or Twitter. The rationale for this was based on the following points:

1. Chinese and English are two big popular languages. Evaluation of the output for this language pair would therefore be of good value for the research on machine translation systems, especially for phrase-based statistical machine translation.
2. Norwegian-involved evaluations have rarely been carried out. Although Norwegian is a small language, the evaluation result was assumed to be useful for MT research where Scandinavian languages are involved.
3. These five specific text genres were identified because they constitute the most frequently-used source genres for online translation services. The diversity of source text was also intended to help provide a more comprehensive and representative evaluation of the translation services.

However, the actual process of the current research project showed that it was impossible for one evaluator to complete the necessary work within such a short period of time. Irrespective of the effort required for providing glosses or back translations for non-Chinese speakers, the error analysis also took too long time because the outputs were considerably messier than anticipated (some output sentences may contain 10 errors or even more, while a few are nearly flawless). Some of the errors were difficult to be classified into a certain category due to a lot of complexity (I will discuss this later), meaning that every single sentence needed to be considered with great care, hence again taking a lot of time. Furthermore, the task seemed to get increasingly intensive when too many language pairs were involved. The subcategories of errors for the output language was language dependent, which made it hard to generalise annotation guidelines for four languages pairs, and it will take much more time if I need to extend the taxonomy for each specific language pair.

The research scope was therefore finally confined to only one language pair Chinese-English (with English as target language). The domain of the source texts was also limited to online Chinese news for a number of reasons: the news text genre is, stylistically speaking, most formal, which helps to circumvent dealing with non-standard language; the accessibility of online news is fairly high, which made it easier for the evaluator (the author of this paper) to collect data; and the topics of news can cover most aspects of a person's life.

Totally 50 Chinese sentences of different lengths (about 1700 Chinese characters) were randomly chosen from 7 different online news texts from one of the most popular and authoritative news presses (Wang Yi News: http://news.163.com/), which covers most topics including culture, sports, technology, education and so on. The choice of the sentences was totally random in order to avoid any confounds in the data sampling. All the Chinese source sentences from the news texts can be found in Appendix B.

The same source sentence was fed into two systems at the same time, and the output translations were copied into a table for comparative analysis. An example of one source sentence and two target sentences generated by the two systems can be seen in Table1.

Note that Google launched their next-generation Neural MT approach to their translation system in September 2016, and the new technology has been applied to Chinese-English language pair (according to their official website). Wu et al. (2016), researchers at Google, have carried out extensive experiments on many Google-internal production data sets, and the results show that their new model reduces translation errors by nearly 60% for Chinese-English language pair (with English as the target language), compared to their phrase-based MT model. The data for my study (i.e. sentences translated by Google) were collected when Google was still using phrase-based statistical approach, before their new technology was launched.

3	丹麦去年排名第3位，落后于瑞士和冰岛。	Denmark last year ranked No. 3, behind Switzerland and Iceland.
		Denmark last year ranked 3rd, behind Switzerland and Iceland.

Table 1. *A translation example*

The Chinese source sentence is in blue, the English sentence generated by Google Translate is in red, and Bing's in green. The gloss for this Chinese sentence is: Denmark ranked No.3 last year, behind Switzerland and Iceland.

3.2 Procedures of Analysing an Output Sentence

As mentioned before, the five-category schema containing five linguistic levels is language pair independent, but the subcategories for a specific language pair are not. Assume that all the errors that I will encounter in my analysis of Chinese-English translation outputs are possible to be sorted into these five big categories (i.e. orthographical errors, morphological errors, semantic errors, lexical errors and syntactic errors), what I have to do is generalise and define the subcategories tailored for Chinese-English language pair (with English as the target language).

I will start with the first Chinese input sentence as shown in Appendix C, to elaborate the procedure for how a sentence analysis is actually performed. Because an output sentence may contain more than one error at different linguistic levels, it is not always possible to decide at a glance what types of errors this sentence has. It may, therefore, take several steps before the analysis can be completed, including segmenting the input and output sentences into phrases/sequences of words (may or may not be linguistically motivated); aligning and mapping words or phrases; detecting errors and categorising them. Explanations in detail are presented in Figure 14 - 22.

No.	Source Text	Target Text
1.	据BBC中文网3月16日报道,联合国可持续发展解决方案网络(SDSN)与哥伦比亚大学地球研究所周三(16日)共同发布的报告指出，丹麦是世界上最幸福的国家，而布隆迪是世界上最不幸福的国家。	According to the BBC Chinese network March 16th reported that the UN Sustainable Development Solutions Network (SDSN) and the Earth Institute at Columbia University on Wednesday (16th) jointly issued the report pointed out that Denmark is the happiest countries, Burundi is the world's most unhappy country.
		BBC Chinese website on March 16th, it was reported, solutions for the sustainable development of the United Nations network (SDSN) and the Earth Institute at Colombia University on Wednesday (16th) jointly issued the report points out that Denmark is the happiest country in the world, Burundi is the most happiest country in the world.

Figure 14. *An example of sentence analysis*

As the source sentence and target sentences are very long, I will discuss the analysis by dividing the Chinses sentence into four components:

“据 BBC 中文网 3 月 16 日报道,” as illustrated in Figure 15-16.

“联合国可持续发展解决方案网络（SDSN）与哥伦比亚大学地球研究所周三（16 日）共同发布的报告指出，” as in Figure 17-18.

“丹麦是世界上最幸福的国家，” as in Figure 19-20.

“而布隆迪是世界上最不幸福的国家。” as in Figure 21-22.

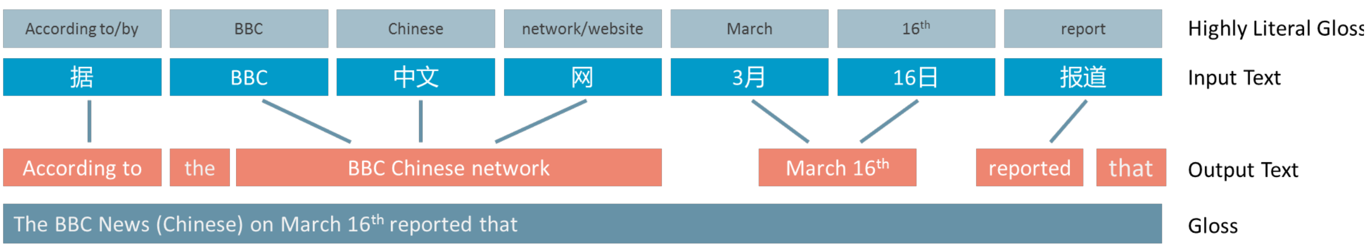


Figure 15. *Sentence component 1 (Google)*

The first line is the 'highly literal gloss' of the source text, which is provided here for readers who do not understand Chinese language. The second line is the input text segmented into phrases/sequences of words, in alignment with the third line, which is the output text produced by Google Translate. The last line is the English gloss for the input text. Remember that both systems are phrase-based, it is therefore necessary to segment the output text into 'phrases', and align them to the 'source phrases'. With such a standardised form of mapping, I will detect errors by going through the text phrase-by-phrase, checking if the 'phrases' have any errors at orthographical level, morphological level, semantic level, lexical error and syntactic level.

The procedure for how the outputs are analysed and compared, is presented in detail as follows:

1) Going through the output text of **Google Translate** from left to right, there are no orthographical errors found.
2) Going through it a second time, no morphological errors are found.
3) Going through the third time, no semantic errors are detected.
4) Going through the fourth time, *BBC Chinese network* is identified as a wrongly translated phrase. The Chinese source phrase is a proper noun meaning 'the BBC News (Chinese)'. I mark it as a **lexical error**, correct it as *BBC News (Chinese)*.
5) Going through the input text again, no more lexical errors are found. But *March 16th* is somehow problematic at syntactic level. In Chinese language, a preposition before a date is not necessary, but the lack of a preposition would render ill-formed English. Here I mark it as a **syntactic error**, correct it as *on March 16th*.
6) Likewise, going through the text one more time, a redundant preposition is identified: *according to*. Note that the Chinese word *据* in the source text, functions as a preposition meaning 'according to'. It is perfectly fine in Chinese syntax to use preposition *据* and verb *报道* jointly with an NP *BBC 中文网* in between, but not in English. Hence, I mark it as a **syntactic error**, correct it as ~~*according to*~~.
7) After 6 steps, the problematic input text was resolved as '~~According to~~ the BBC News (Chinese) on March 16th reported that'.

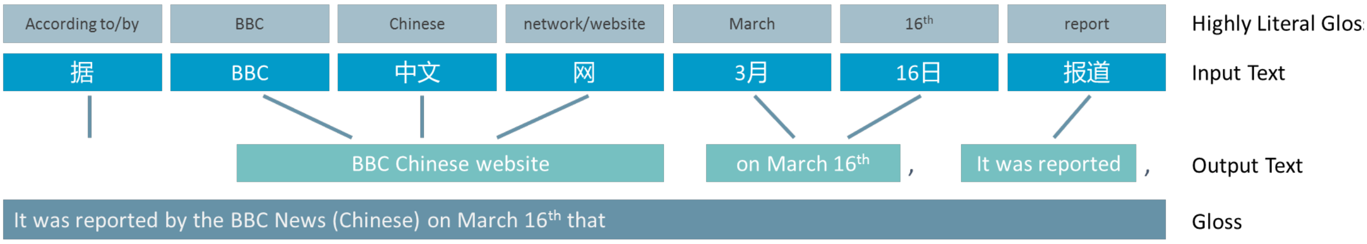


Figure 16. *Sentence component 1(Bing)*

8) Now I will analyse the translation output of **Bing Translator**. See Figure 16. Same as previous steps, no orthographical errors, morphological errors or semantic errors are found after going through the output text from left to right several times.

9) From Figure 16 we can see, there is no word/phrase mapped with the Chinese word *据*, hence a preposition is missing. It might be natural to see this as a syntactic error at first glance, since missing a preposition normally leads to an ungrammatical English sentence. While by adding the preposition to the text – 'by BBC Chinese website on March 16th, it was reported', the sentence after treatment is still syntactically ill-formed. Plus, in the source text the preposition *据* can actually be found. I define such errors as an omission of function word(s), namely a **lexical error**. Even though the omission causes syntactic problems, a treatment by adding the omitted word will not necessarily resolve a syntactically ill-formed sentence, but can render a correct phrase, in accordance with the source phrase. (I will discuss similar issues later in Section 3.3.4)

10) Proceeding the analysis, another **lexical error** is found – same as step 4), the proper noun was not correctly translated. I annotate it as *BBC News (Chinese)*.

11) In order to make the previous phrase *BBC News (Chinese)* completely correct, I need to add a definite article *the* in front of the proper noun. In Chinese language, no definite article is needed before a proper noun, while it is, normally, obligatory to have a definite article before certain proper nouns in English. For this title of organisation, adding a definite article *the* makes a correct English phrase. I mark it as lack of preposition, hence a **syntactic error**. Compared with step 9), an **omission** especially refers to an error caused by omitting word(s) that can be found originally in the source text, no matter whether it eventually leads to syntactic problems or not, it is defined as lexical error. However, a **lack** of syntactic element (e.g. a preposition in this case), which does not originally exist in the source text (mostly because of the difference of syntactic

structures/rules between Chinese and English), but necessary for grammaticality in the target text in English, is defined as a syntactic error.

12) After treatments from the previous steps, the sentence becomes ‘by the BBC News (Chinese) on March 16^{th}, it was reported’. Obviously, there is a **syntactic error** concerning word order. By moving *it was reported* to the front, the sentence is nearly resolved: ‘it was reported by the BBC News (Chinese)’.
13) Lastly, the lack of a syntactic element occurs, which makes another **syntactic error**. By adding a complementiser *that* at the end of the sentence, the text after correction becomes ‘it was reported by the BBC News (Chinese) on March 16^{th} that’.

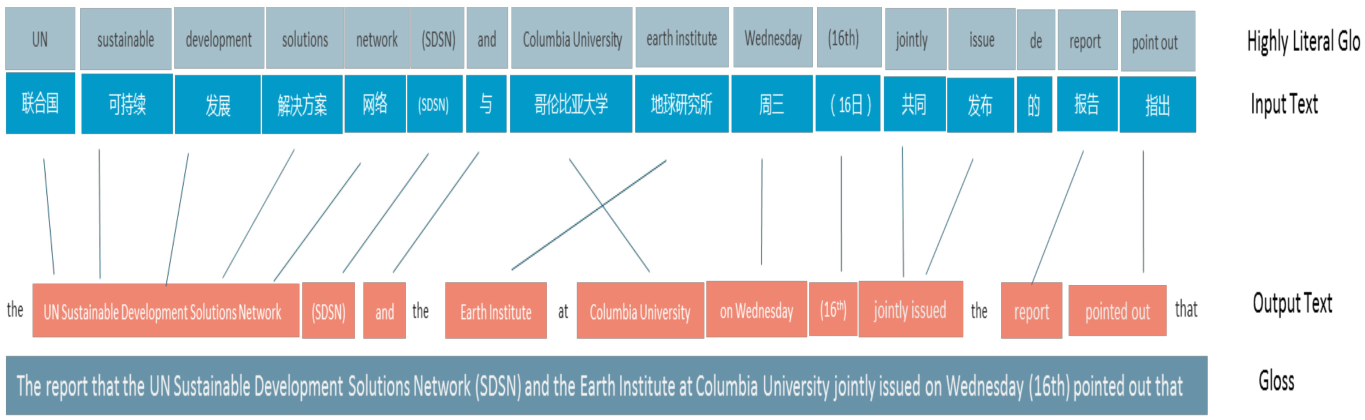


Figure 17. *Sentence component 2 (Google)*

14) Now I will continue to look at the second component of the translation by **Google**, see Figure 17. Going through similar process as before, no orthographical errors, morphological errors, semantic errors or lexical errors are detected in all the phrases in the output text.
15) From Figure 17 we can see that a Chinese word *的* (a structural particle *de*) is not aligned with any words in the output text. This does not mean that there is a lexical error of omission. It is actually an error of syntax. Normally, a relative clause in Chinese language, is similar to other adjectival phrases, since the clause proceeds the noun that it modifies, and ends with the particle *de*, as is in the input text. I note it as a **syntactic error**, and correct the sentence as ‘The report that the UN Sustainable Development Solutions Network (SDSN) and the Earth Institute at Columbia University jointly issued on Wednesday (16^{th}) pointed out that’.

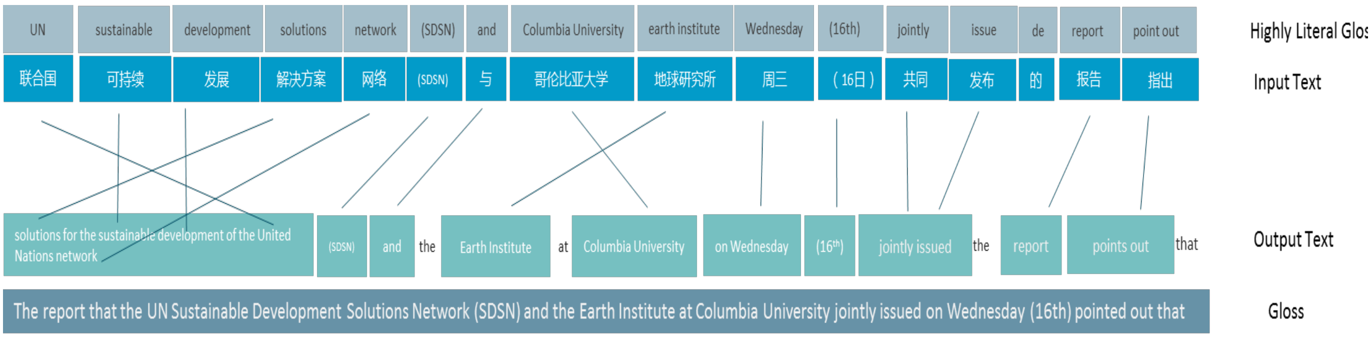


Figure 18. *Sentence component 2 (Bing)*

16) Figure 18 shows how **Bing Translator** translates the same input text. A **lexical error** is first detected in the first phrase box. The Chinese proper noun is not correctly translated, possibly because this proper noun is not trained in the bilingual translation model and the software cannot find the correct equivalence to it. I correct the phrase *solutions for the sustainable development of the United Nations network* as *the United Nations Sustainable Development Solutions Network*.
17) Same as step 15), Bing Translator fails to translate the Chinese relative clause with the particle *de*. A **syntactic error** is noted and the sentence after correction is 'The report that the UN Sustainable Development Solutions Network (SDSN) and the Earth Institute at Columbia University jointly issued on Wednesday (16^{th}) pointed out that'.

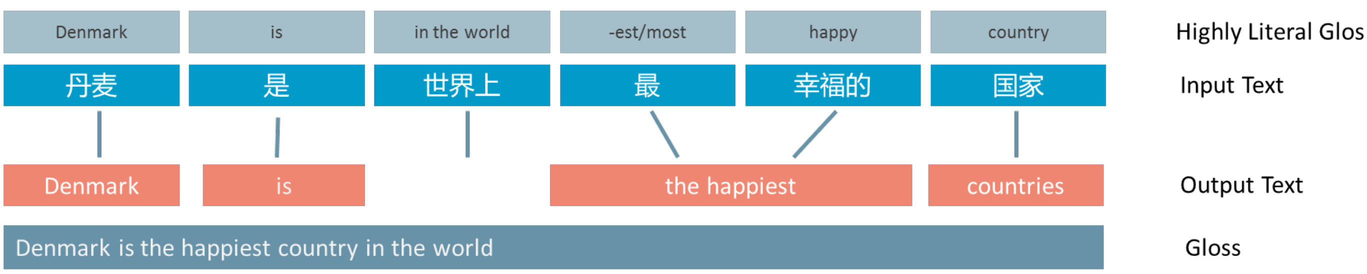


Figure 19. *Sentence component 3 (Google)*

18) Looking at the third component of the input text translated by **Google**, as shown in Figure 19, I do not find any orthographical errors, but an error at **morphological** level is found. The last NP in the text should not be in its plural form. By changing its

morphological form to singular (*countries-country*), the sentence is almost resolved. (It is marked as a **syntactic error** as well, which will be discussed later.)

19) Proceeding the analysis, no semantic errors are found, but an omission of words occurs. Though the output text after the previous treatment *Denmark is the happiest country* looks syntactically correct, the omission of *in the word* reduces the specificity of the output sentence. In order to be consistent and strict throughout the analysis in my study, such problems of omission are considered **lexical errors**. By adding *in the world* at the end of the input text, the analysis for this component is done: *Denmark is the happiest county in the world*.

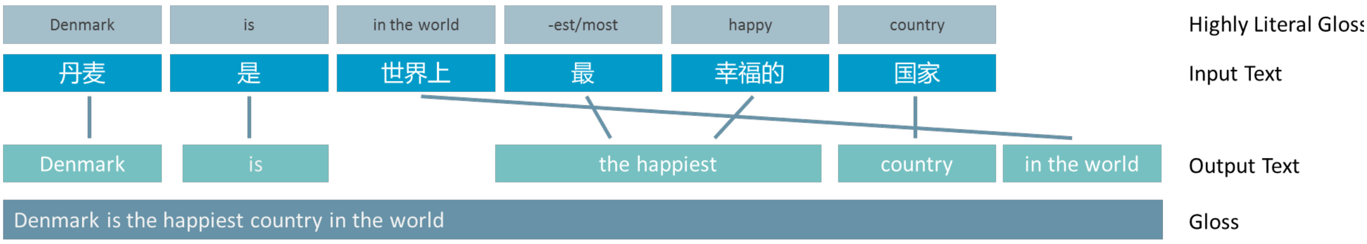


Figure 20. *Sentence component 3 (Bing)*

20) As shown in Figure 20, the third sentence component translated by **Bing Translator** is completely correct.

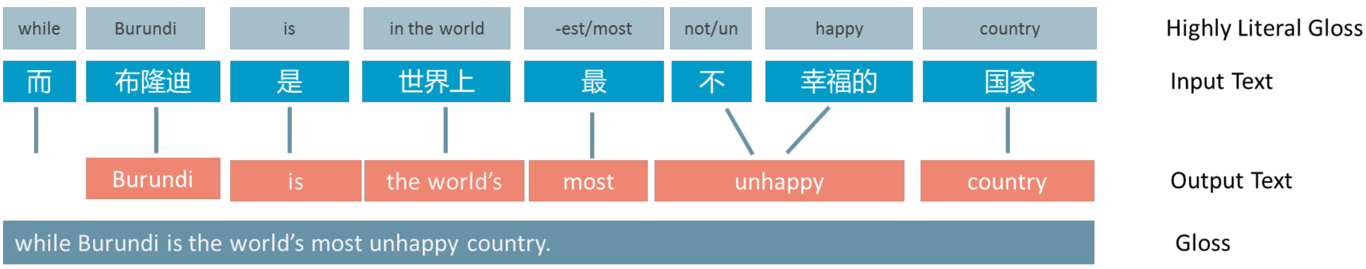


Figure 21. *Sentence component 4 (Google)*

21) Figure 21 illustrates the fourth component of the sentence, translated by **Google Translate**. Similar to step 19), the omission of the function word *while* is noted as a **lexical error**.

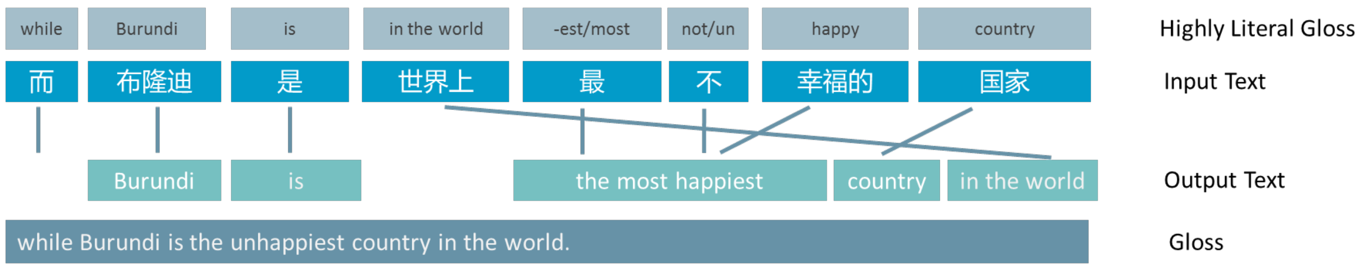


Figure 22. *Sentence component 4 (Bing)*

22) Figure 22 illustrates the fourth component of the sentence, translated by **Bing Translator**. No orthographical errors are found in this component, but the phrase *the most happiest* is not a correct adjectival inflection. I mark this as a **morphological error** and correct it to *the happiest*.
23) After referring to the source text, *the happiest* is the opposite meaning of the source phrase, here I define it as a **semantic error** of antonym. I change the phrase to *the unhappiest*.
24) Same as step 21), the omission of *while* is a **lexical error**.

Up to this point, the analysis of a set of output sentences translated from one input sentence by Google Translate and Bing Translator, has been completed. It takes 24 steps in total for this set, including identifying, annotating and correcting all the errors encountered during the process. 8 errors are produced by Google Translate, including 1 morphological error, 3 lexical errors and 4 syntactic errors. 10 errors are generated by Bing Translator, including 1 morphological error, 1 semantic error, 4 lexical errors and 4 syntactic errors. Table 2 shows how the sentences are analysed and annotated. Each error is marked in the table with a red circle (O) if it is produced by Google Translate, and a green cross (X) if it is produced by Bing Translator.

NO.	SOURCE TEXT	TARGET TEXT	DISSECTED COMPONENTS	ORTHOGRAPHICAL	MORPHOLOGICAL	SEMANTIC	LEXICAL	SYNTACTIC
1.	据BBC中文网3月16日报道,联合国可持续发展解决方案网络(SDSN)与哥伦比亚大学地球研究所周三(16日)共同发布的报告指出，丹麦是世界上最幸福的国家，而布隆迪是世界上最不幸福的国家。	**According to the BBC Chinese network March 16th reported that the UN Sustainable Development Solutions Network (SDSN) and the Earth Institute at Columbia University on Wednesday (16th) jointly issued the report pointed out that Denmark is the happiest countries, Burundi is the world's most unhappy country.**	*** According to the BBC Chinese network March 16th reported that** - According to the BBC News (Chinese) March 16th reported that				O	
			- -According to the BBC Chinese News (Chinese) on March 16th reported that					O
			- - - ~~According to~~ the BBC Chinese News (Chinese) on March 16th reported that					O
			***… jointly issued the report pointed out that** - the report that SDSN and … jointly issued					O
			*** Denmark is the happiest countries** - Denmark is the happiest country		O			O
			- - Denmark is the happiest country in the world				O	
			* Burundi is the world's most unhappy county. - while Burundi is the world's most unhappy country.				O	
		BBC Chinese website on March 16th, it was reported, solutions for the sustainable development of the United Nations network (SDSN) and the Earth Institute at Colombia University on Wednesday (16th) jointly issued the report points out that Denmark is the happiest country in the world, Burundi is the most happiest country in the world.	*** BBC Chinese website on March 16th, it was reported** - by BBC Chinese website on March 16, It was reported				X	
			- - by BBC News (Chinese) on March 16th, It was reported				X	
			- - - by the BBC News (Chinese) on March 16th, It was reported					X
			- - - - It was reported by the BBC News (Chinese) on March 16th					X
			- - - - - It was reported by the BBC News (Chinese) on March 16th that					X
			*** solutions for the sustainable development of the United Nations network (SDSN)** - the United Nations Sustainable Development Solutions Network (SDSN)				X	
			*** …jointly issued the report points out that** - the report that SDSN and … jointly issued					X
			*** Burundi is the most happiest country in the world.** - Burundi is the ~~most~~ happiest country in the world.		X			
			- - Burundi is the unhappiest country in the world.			X		
			- - - While Burundi is the unhappiest country in the world.				X	

***Table 2.** Analysis example.*

50 Chinese sentences have been translated into 50 sets of English sentences automatically by Google Translate and Bing Translator. The 100 output sentences are manually analysed by one evaluator (the author of this thesis) following in the previous procedures as illustrated above. Apparently, this testifies to the fact that the work of manual linguistic evaluation is extremely time-consuming and intensive.

3.3 Taxonomy of Linguistic Errors

Now that the procedure for how a sentence analysis is performed has been clarified, I will introduce my taxonomy of errors at five linguistic levels in the translation outputs of Chinese-English language pair (with English as target language). After all the 50 sets of sentences have been analysed and annotated, I am able to generalise and define the subcategories at each linguistic level. The taxonomy of my linguistic evaluation is presented in Figure 23, followed by specific examples. All the examples presented below are picked from the analyses of the 50 sets of sentences in my study, hence it can also be seen as an elaborate presentation of my data analysis.

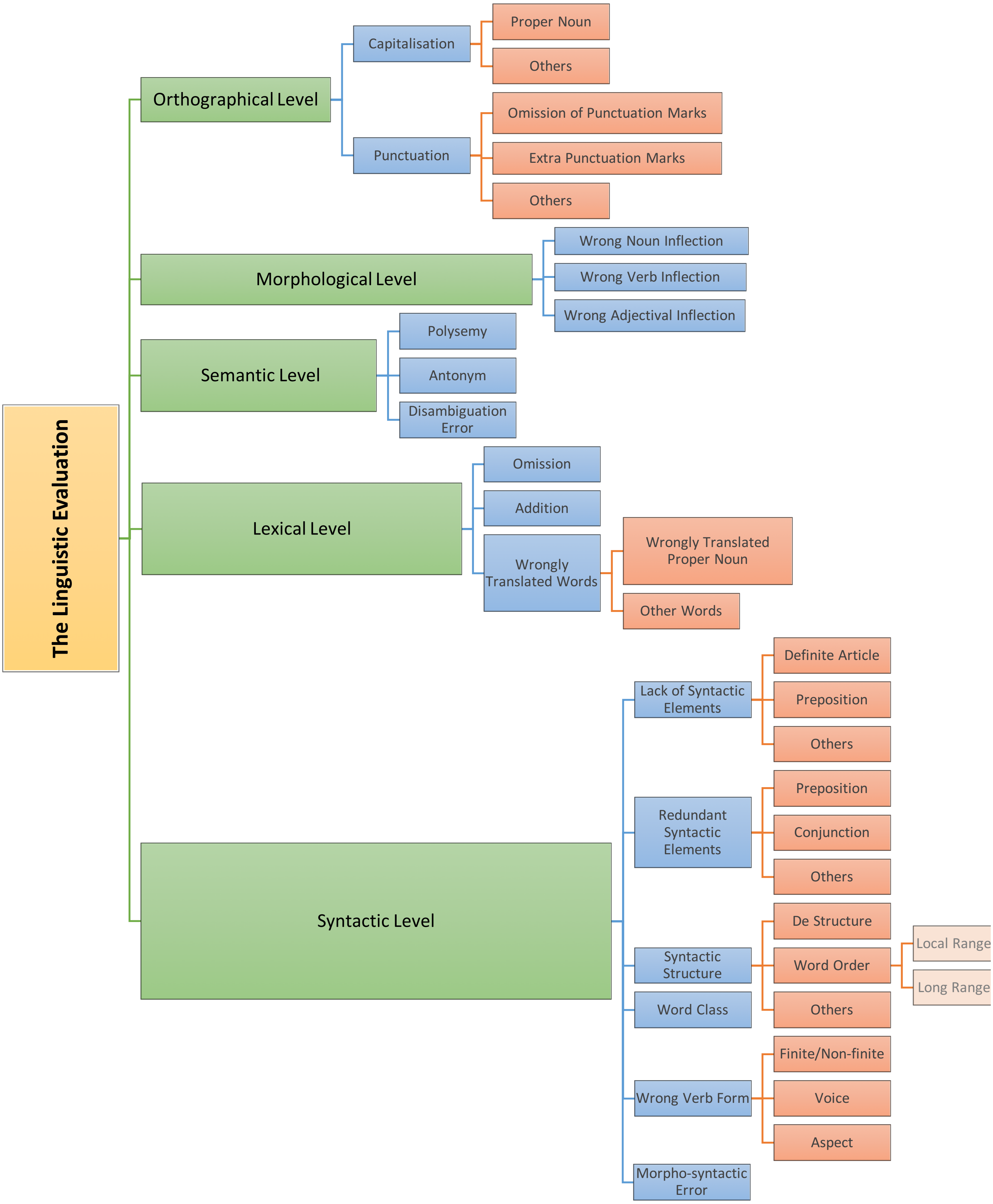


Figure 23. *Taxonomy of linguistic evaluation of Chinese-English translation output*

3.3.1 Orthographical Level

An orthography refers to a set of conventions for writing a language, including norms of spelling, capitalisation, word breaks, emphasis, and punctuation etc. The orthographical errors include all errors regarding violation of the norms. In the study of Farrús et al. (2012), a variety of orthographical errors are introduced, which are mostly related to the target language Catalan or the language pairs involved (Spanish-Catalan and English-Catalan). In their classification, errors include the misuse of punctuation marks (exclamation and interrogation marks, full stops, commas, colons, apostrophes and so on), wrong emphasis (especially related to accented vowels), capitalisation (lower case letters at the beginning of a sentence or in proper nouns and acronyms, capital letters in the output when they are lower case in the input or vice versa) and word breaks (extra space or lack of space between words).

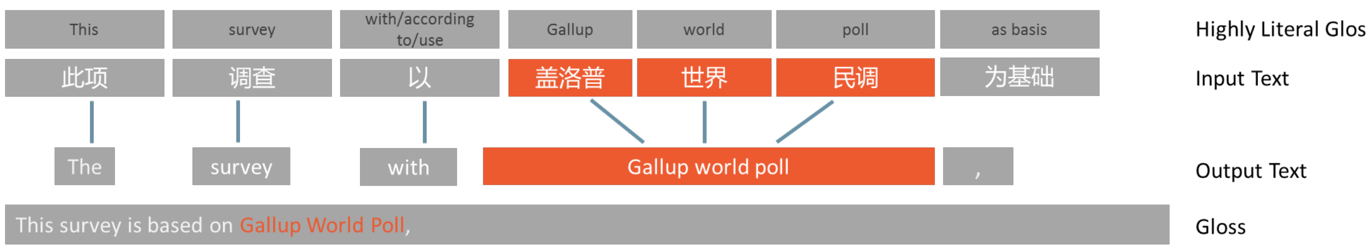
Most of the statistical machine translation systems are well trained with the language model of large English corpora. Therefore, in this study for Chinese-English translation output, few spelling errors are found. Plus, unlike the majority of European languages or some languages that share the same alphabet, Chinese and English are typologically different languages. Errors concerning word breaks or emphasis are quite rare. Several errors of capitalisation and punctuation occurred in the analyses. Examples are presented below.

Example 1: Capitalisation (Proper Noun)

CN: 此项调查以<u>盖洛普世界民调</u>为基础,

Gloss: This survey is based on <u>Gallup **W**orld **P**oll</u>,

EN by Google Translate: The survey with <u>Gallup world poll</u>,



Example 1. *Capitalisation (Proper noun)*

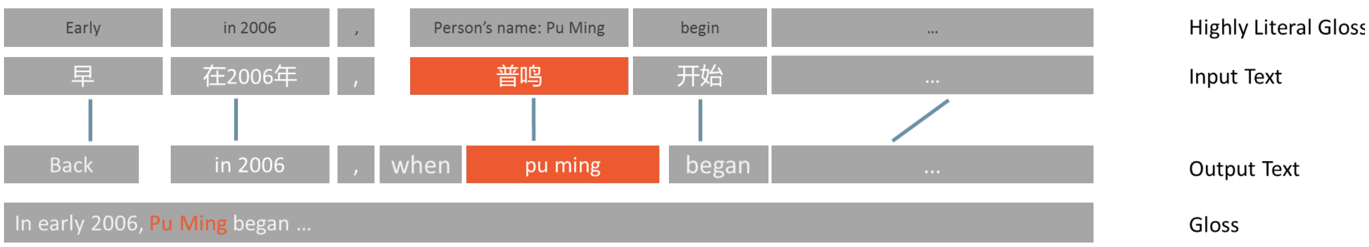
In example 1, the proper noun Gallup World Poll is not correctly presented in the English output sentence with all three initial capital letters. In Chinese writings, the characters do not change orthography no matter whether it is a proper noun or not. From the example, we can see that each word (*Gallup*, *world*, *poll)* is correctly selected by the MT system as translation output, but not properly capitalised. One possible reason of such errors is that during the automatic translation process, the source phrase is segmented into three independent words, and the system does not recognise it as a proper noun because it is not included in the training data.

Example 2: Capitalisation

CN: 早在 2006 年，普明开始…

Gloss: In early 2006, Pu Ming began to …

EN by Bing Translator: Back in 2006, when pu ming began to …



Example 2. *Capitalisation (Person's name)*

Person's name is another source of orthographical errors in the Chinese-English translation outputs. Such kind of translation task is difficult for MT systems, because unlike translating languages that share the same alphabet, Chinese names cannot be translated by simply copying the input words to the generated sentence. Especially for person's name, if the system has never seen this name in the training, it tends to translate the words according to the pinyin (pronunciation). In this case, the pinyin is chosen correctly, but not capitalised.

Example 3: Punctuation (Omission of punctuation marks)

CN: …加强了他们的科研，在国际期刊上更多地发表论文…

Gloss: … strengthened their scientific research, and published more papers in international journals…

EN by Google Translate: … strengthen their research in international journals published more papers…

Example 3. *Punctuation (Omission of commas)*

In example 3, a necessary comma is omitted which represents a punctuation error. This error occurs perhaps because with the language model in the statistical machine translation system, a string ‘strengthen their research in international journals’ is assigned a higher probability than ‘strengthen their research, in international journals’.

3.3.2 Morphological Level

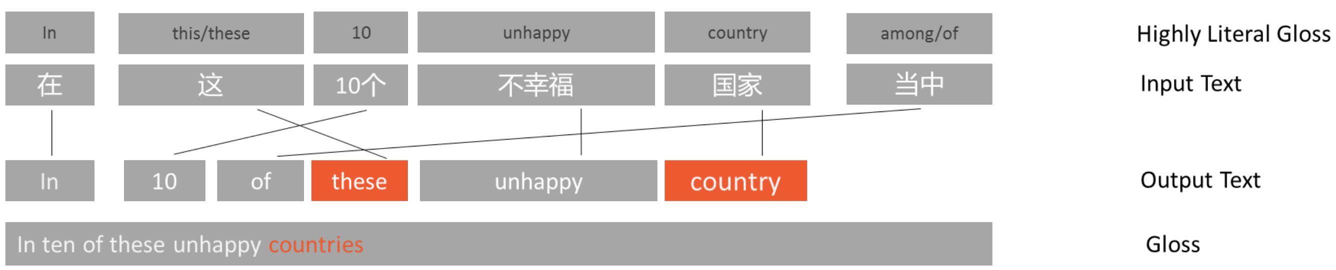
Errors related to the morphology of words are put under this category, including **wrong noun inflection**, **wrong verb inflection** and **wrong adjectival inflection**. A tricky problem here is that Chinese language, in general, does not possess overt inflectional morphology (Norman, 1988); in fact, Chinese is such an isolating language that it brings into question what we mean by morphology (Packard, 2000). This makes the translation analysis difficult for the reason that although English has a relatively poor inflectional system, it still has more overt inflectional morphology than Chinese, as can be seen in the existence of noun inflection, verbal inflection and adjectival inflection. In most cases during the analyses of this study, the Chinese words in the input texts do not have any morphological changes with regard to comparative/superlative form of adjectives, singular/plural form of nouns or verbal inflections for tense, number and person. How can we define whether a target word is inflected correctly or not, if the source word does not have any inflection at all? Example 4 and 6 represent how morphological errors are found by only looking at the output English sentences without referring to the source text. Example 5 shows how a wrong verb inflection in the output is identified, by referring to an aspectual marker in the source text. Example 7 is the only case of wrong adjectival inflection found in the analyses of all the 50 sets of sentences in my study.

Example 4: Wrong noun inflection

CN: 在这 10 个不幸福国家当中 …

Gloss: In ten of these unhappy countries …

EN by Bing Translator: In 10 of these unhappy country…



Example 4. *Wrong noun inflection*

Example 4 represents the major type of morphological errors – wrong noun inflection. Without referring to the source text, it is already clear that *country* should be in its plural form in order to be grammatically correct. Note again that the source word *国家* (meaning 'country') does not show any morphological changes in its plural form. I conjecture that such errors occur because the SMT systems translate segmented phrases/words independently (e.g. *国家*→ *country*) in the decoding module, and do not take account of other words near them (e.g. *these*) after reordering in the post-processing module.

In the translation analysis for Chinese-English translation outputs, this type of error is identified when a noun in the target language has been incorrectly rendered into its plural form, or vice versa, as can be seen in Table 3.

*many profession
*one of the main concern
*a batch of new car
*this boys
*Denmark is the happiest countries.

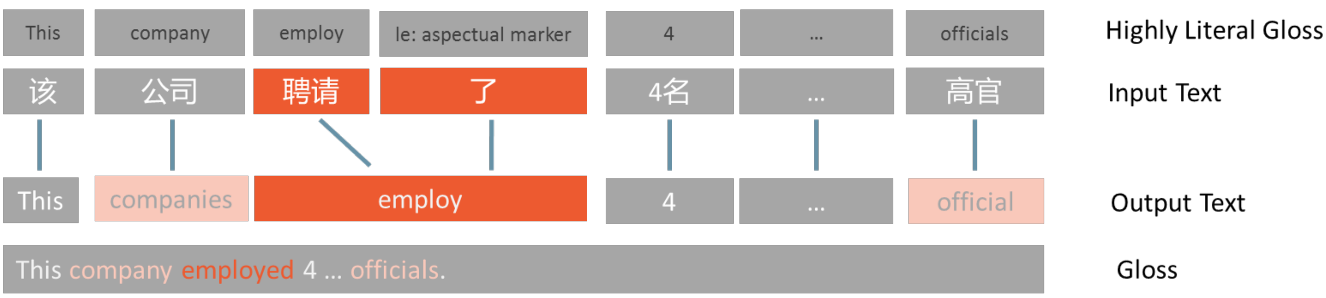
Table 3. *More examples of wrong noun inflection*

Example 5: Wrong verb inflection

CN: 该公司聘请了 4 名… 高官,

Gloss: This company employed 4 … officials,

EN by Bing Translator: This companies employ 4 … official,



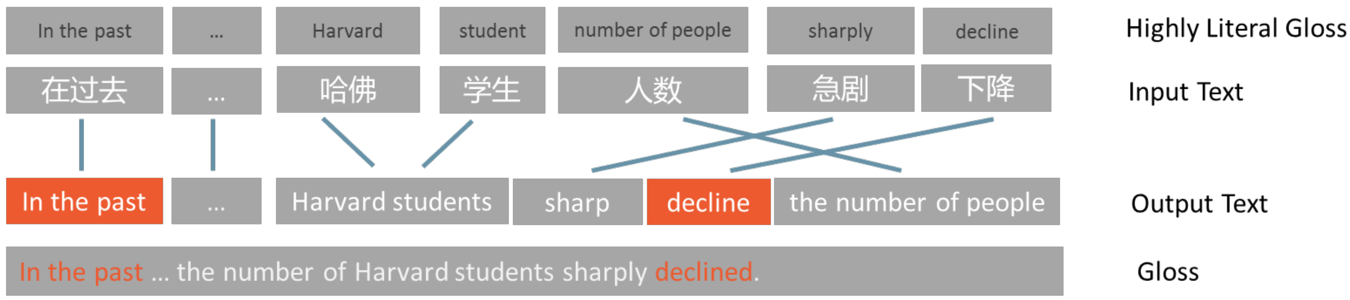
Example 5. *Wrong verb inflection*

Example 6: Wrong verb inflection

CN: 在过去… … 哈佛学生人数急剧下降。

Gloss: In the past … … the number of Harvard students declined sharply.

EN by Google Translate: In the past … … Harvard students sharp decline the number of people.



Example 6. *Wrong verb inflection*

Example 5 and 6 are two cases of wrong verb inflection in the translation outputs. They also represent the difference in the marking of tense and aspect of Chinese and English. Chinese language does not mark tense morphologically, but does mark aspect (Duff and Li, 2002). Smith (1997) points out that past tense aspectual markers in Chinese are oriented towards discourse and pragmatic factors that, in some cases, make the use of aspectual markers syntactically optional.

The aspectual marker *le* 了 in the input text of Example 5 is used behind the verb to indicate a perfective aspect, which gives information about the temporal flow of an event which took place in the past. While more commonly in Chinese language, the time at which an action is conceived as taking place is indicated by expressions of time (e.g. *yesterday*, *now*, *tomorrow*, *last year* and so on) or simply inferred from the context. As we can see from those two examples above, there is an aspectual marker in Example 5, while there is not in Example 6 when an expression of time (i.e. *in the past*) is used in the sentence.

In example 5, two errors of noun inflection (i.e. *companies*, *official*) can also be found, which will be ignored here. The verb *employ* was not correctly inflected according to the past tense denoted by the perfective aspect marker *le* in the source text. Without referring to the source text, we can identify a morphological error in the output text in Example 6. *In the past* indicates the time at which the action *decline* took place. Therefore, the verb *decline* should be inflected as *declined*.

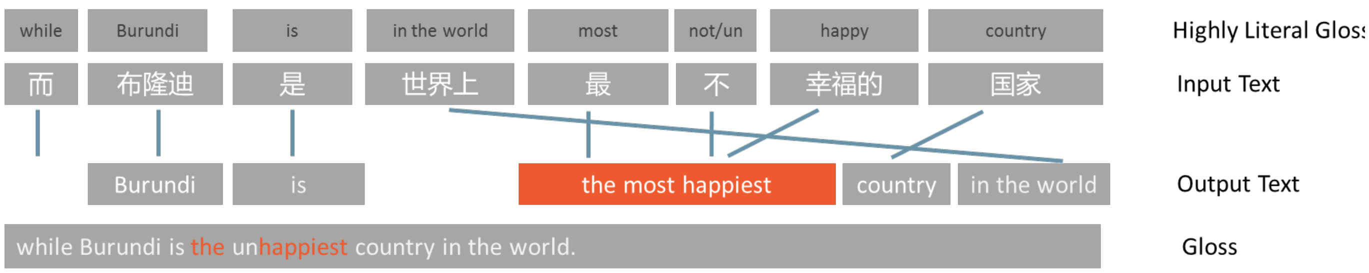
From the examples discussed above, we can see that in order to identify either an error of noun inflection or an error of verb inflection, syntactic/grammatical rules are more or less taken into consideration (e.g. the errors of noun inflection are related to syntactic agreement and concord, the errors of verb inflection are related to tense, aspect etc.). In fact, it is not always easy to perform analysis at morphological level without considering syntactic features. Therefore, such ‘morpho-syntactic’ errors found in my analyses are usually marked as a morphological error as well as a syntactic error.

Example 7: Wrong adjectival inflection

CN: 布隆迪是世界上<u>最不幸福的</u>国家。

Gloss: Burundi is the <u>unhappiest</u> country in the world.

EN by Bing Translator: Burundi is the <u>most happiest</u> country in the world.



***Example 7.** Wrong adjectival inflection*

In Example 7, it is apparent that the superlative form of the adjective *happy* was wrongly inflected as *the most happiest*, hence a wrong adjectival inflection error. (There is obviously a semantic error in this sentence, which will be discussed later.) The cause of such errors is mysterious, and similar cases are rarely found in this study, which means this error might be quite random. Due to curiosity, I feed several similar Chinese sentences into Bing Translator, with different names of nations. It generates correct outputs when the subject is *Finland*, *Norway*, *Ethiopia*, *Denmark*, *India* and many more, while produces *the most happiest* when the subject is *Burundi* and *Sweden*. See Figure 24.

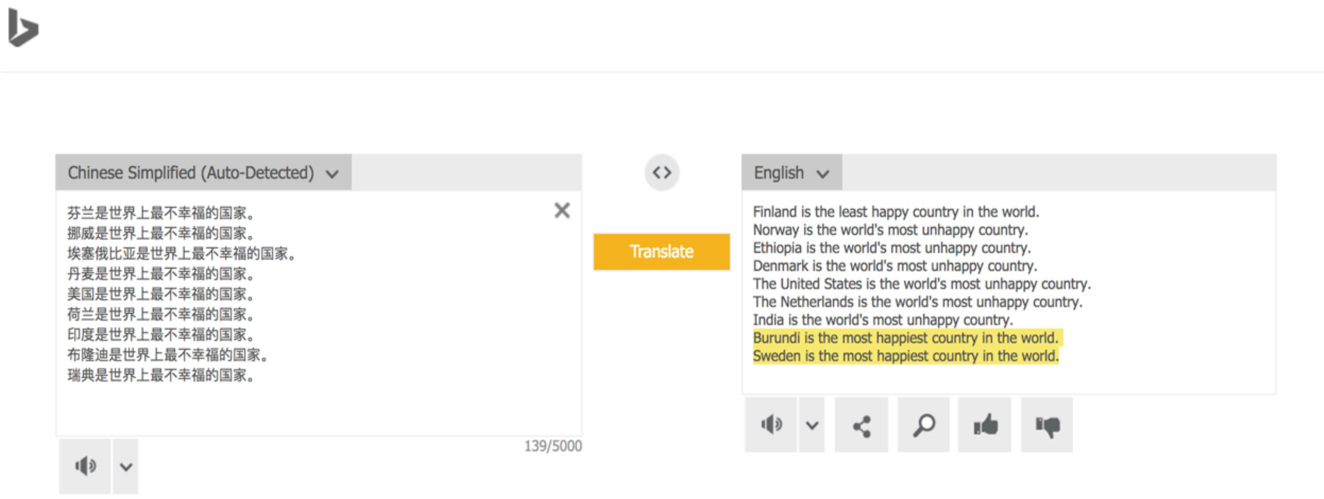


Figure 24. *Translation of superlative form of adjectives by Bing Translator*

This is an interesting phenomenon that developers should pay attention to, when attempting to improve their translation system. It also implies that specific morphological information for translating Chinese superlative forms of adjectives might be helpful to increase accuracy of the MT output. For example, if there is a morphological analysis in the pre-processing module on the source side to detect and tag the word *最* (meaning 'most') before a Chinese adjective, and there is a morphological generation in the post-processing module on the target side to assure the superlative form of the adjective in the target language is correct, such errors might be reduced.

3.3.3 Semantic Level

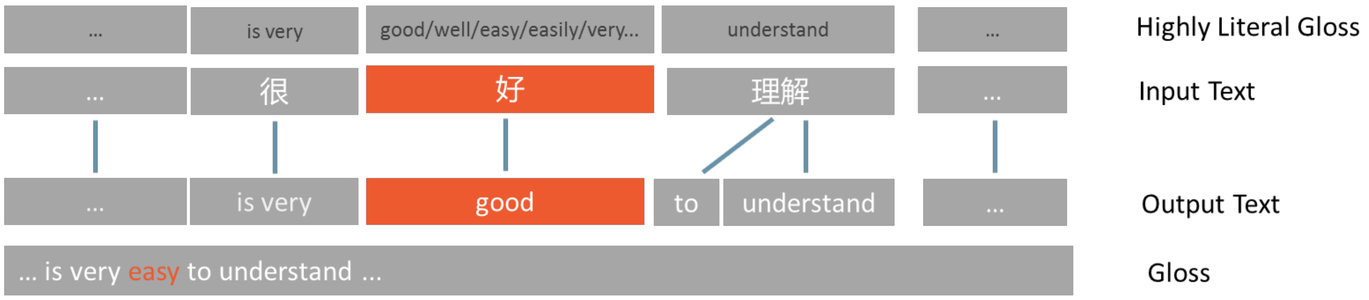
In the Chinese-English translation output analysis, three major types of semantic errors are found: polysemy errors, antonym errors and disambiguation errors. **Polysemy errors** concern the cases when a word with multiple meanings in the source language, is translated into one of the possible meanings, but turns out to be a wrong choice in that certain context. **Antonym errors** occur when a source word is translated into its opposite meaning. **Disambiguation errors** refer to cases where the systems are not able to disambiguate some words that have similar meanings, pronunciations or characters in Chinese. See examples:

Example 8: Polysemy errors

CN: …很<u>好</u>理解…

Gloss: … is very <u>easy</u> to understand…

EN by Google Translate: … is very <u>good</u> to understand…



Example 8. Polysemy errors

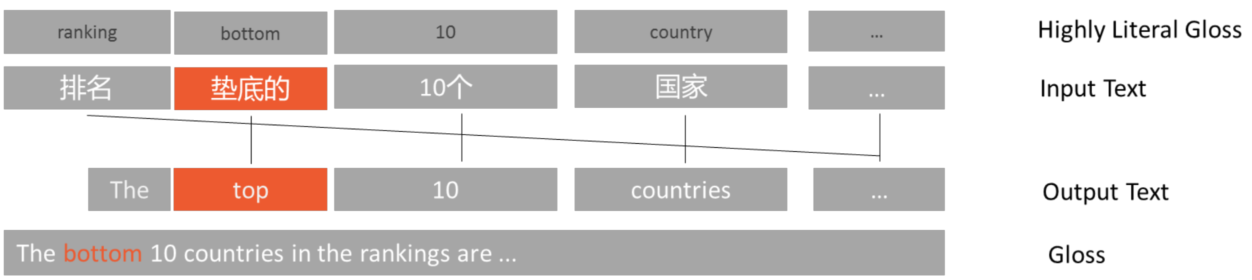
In Chinese, *好* (hao) is a polyseme that has a variety of meanings under different circumstances, including 'easy', 'good', 'well', 'very' and so on. In example 8, *good* was chosen as the translation of *好*, while *easy* is the correct translation in that context, hence a polysemy error.

Example 9: Antonym errors

CN: 排名<u>垫底</u>的 10 个国家…

Gloss: The <u>bottom</u> 10 countries in the rankings…

EN by Google Translate: The <u>top</u> 10 countries…



Example 9. Antonym errors

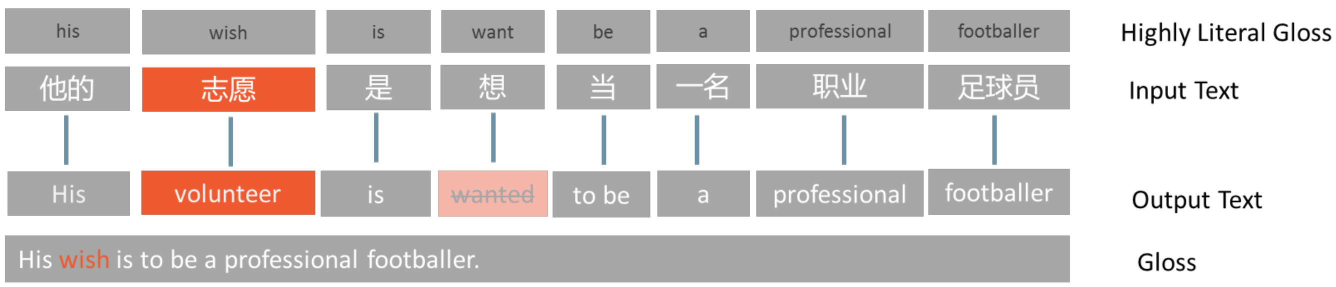
In this case, the opposite meaning of the source word was chosen (*bottom-top*), which mangles the original meaning of the whole sentence. Antonym errors do not occur very often in the analysis. A tentative explanation for such errors is that a string such as *top 10* is assigned a much higher probability than *bottom 10*, simply because the system has seen *top 10* appear more often in a sentence in the language model (the monolingual corpora).

Example 10: Disambiguation errors

CN: 他的志愿是想当一名职业足球员。

Gloss: His wish is to be a professional footballer.

EN by Bing Translator: His volunteer is wanted to be a professional footballer.



Example 10. *Disambiguation errors*

Example 10 represents another type of semantic error. A Chinese noun *志愿* (meaning 'wish') is translated into *volunteer*. This is a disambiguation error because the MT system fails to disambiguate two similar Chinese words properly. Chinese words *志愿* and *自愿* (meaning 'volunteer') in some way have some semantic convergence (but strictly speaking they are not synonyms or near-synonyms), just as the English word pair *wish* and *volunteer*. Plus, these two Chinese words have some phonetic and orthographical similarities (as is presented in Figure 25). Thus, it poses difficulty for the MT system to perform disambiguation tasks. Interestingly, both Google Translate and Bing Translator wrongly choose *volunteer* instead of *wish* as the translation of *志愿* in this translation task.

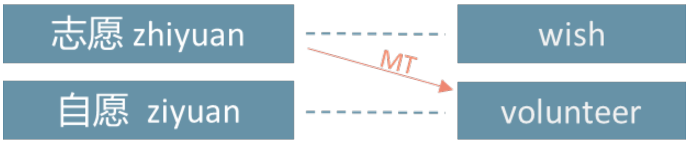


Figure 25. *Disambiguation error*

3.3.4 Lexical Level

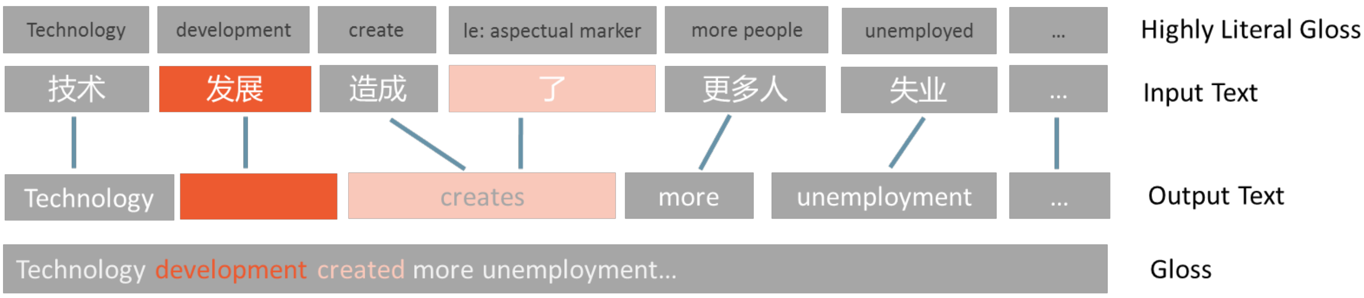
Errors at the lexical level are divided into: **omission**, **addition** and **wrongly translated words**.

Example 11: Omission

CN: …技术 <u>发展</u> 造成了 更多人 失业， …

Gloss: … technology <u>development</u> created more unemployment, …

EN by Google Translate: … technology __________ creates more unemployment, …



Example 11. *Lexical error: omission*

In Example 11, the word *development* in the output sentence is missing, which is seen as a lexical error (i.e. omission). As discussed before in Section 3.2, similar cases can be seen in Figure 19, 21 and 22. The omission of words usually reduces the specificity of the output sentence, though it does not have a dissimilar meaning to the source sentence, as in Example 11. However, an omission of some words could greatly change the meaning in the target sentence, for example, an omission of a negation word. In those cases, the omission does not necessarily make the target sentence syntactically ill-formed, while at other times, omission of words can easily cause ungrammaticality in the output sentences, for example an omission of a content word (a verb) or an omission of a functional word (a pronoun).

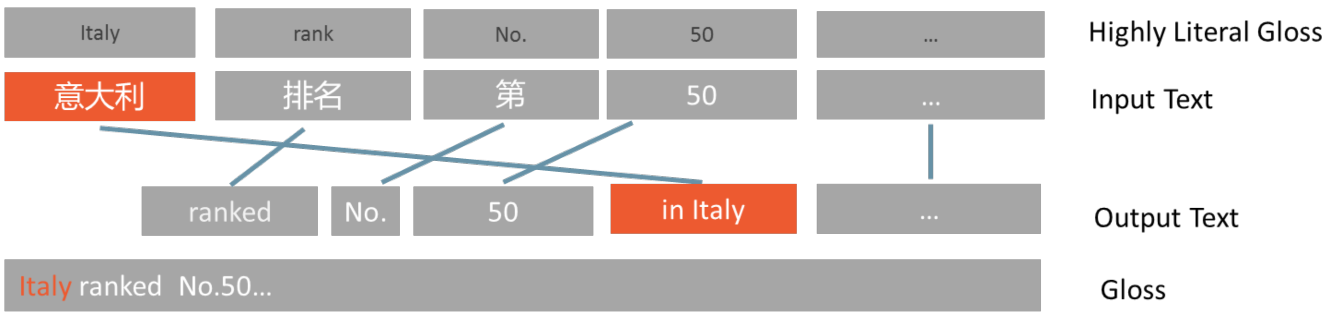
This causes difficulty for sentence analysis in my study – an omission sometimes changes the meaning of the sentence, sometimes does not; it sometimes leads to syntactically ill-formed sentences, while sometimes not. In order to be consistent and strict throughout the analysis in my study, such problems of omission are all considered **lexical errors,** as long as there are words omitted which can be found originally in the source sentence. It is also a solution to avoiding derivative errors (i.e. a lexical error causes a syntactic error).

Example 12: Addition

CN: 意大利排名第 50…

Gloss: Italy ranked No.50 …

EN by Google Translate: ranked No.50 <u>in</u> Italy …



Example 12. *Lexical error: addition*

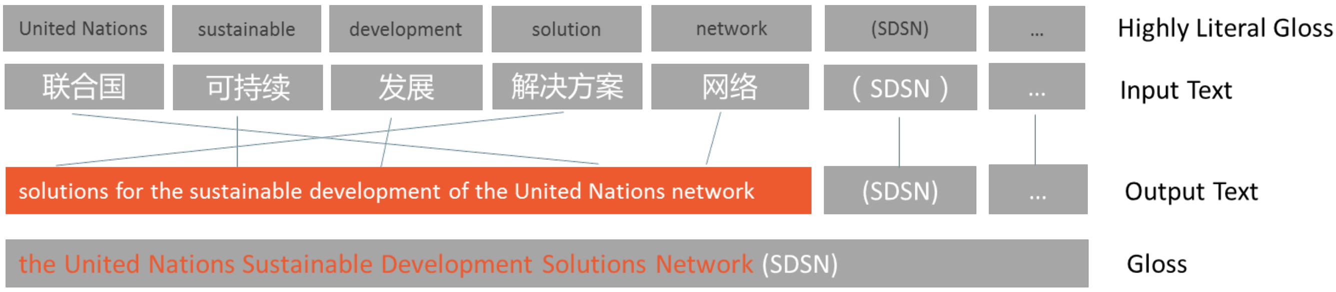
In example 12, a preposition *in* is added to the output text, while there is not an equivalent preposition found in the source text (the word order error is ignored here). Similarly, an addition of word(s) may or may not change the meaning/cause ungrammaticality in the target language. I will stick to the same principle that applies to omission as described above, as long as there are words added to the output text, which cannot be found originally in the source text, it is defined as an error at lexical level, namely addition.

Example 13: Wrongly translated words (proper nouns)

CN: 联合国可持续发展解决方案网络 （SDSN）…

Gloss: The United Nations Sustainable Development Solutions Network (SDSN)…

EN by Bing Translator: solutions for the sustainable development of the United Nations network (SDSN)…



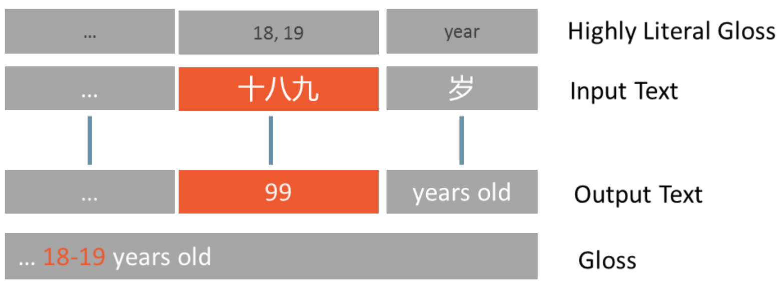
__Example 13.__ Wrongly translated words (proper noun)

Example 14: Wrongly translated words (other words)

CN: … 十八九岁

Gloss: … 18-19 years old

EN by Bing Translator: … 99 years old



__Example 14.__ Wrongly translated words (other words)

Example 13 and 14 present the last subcategory of lexical errors – wrongly translated words. Two major types of wrongly translated words are found, namely proper nouns as shown in Example 13, and other words as shown in Example 14.

When a proper noun is written in Chinese language (no capitalisation, no separators between words etc.), it is not easy for the MT system to determine how to make a correct translation. Example 13 indicates that the system is confused by the long complex phrase and does not even recognise it as a proper noun. Recall that I introduced an error related to proper nouns at orthographical level in section 3.3.1, as discussed in Example 1. The distinction between these two types of error is, when the MT system chooses correct words as the translation of the source proper noun but not properly capitalised, it is an orthographical error (e.g. Example 1); when the MT system does not produce a correct translation of the source proper noun (e.g. Example 13), it is a lexical error.

Example 14 shows how MT systems wrongly translated other words. In such cases, the MT system chooses a completely wrong translation of the source word. The most possible reason for such errors is that some Chinese words/characters are not included in the translation model (the bilingual corpora) of the MT system, but a target word is picked anyway from the Translation Table simply because of highest probability computed by the system.

3.3.5 Syntactic Level

This level presents the most various types of errors, which are classified into the subcategories as follows: **lack of syntactic elements**, **redundant syntactic elements**, **syntactic structure errors**, **word class errors**, **wrong verb forms**.

In a whole output sentence, syntactic errors are identified in the last step, after all the errors are analysed and treated at the other four levels, namely orthographical errors, morphological errors, semantic errors and lexical errors in turn. This is to avoid any situation where a 'non-syntactic error' causes a syntactic error, namely a derivative error.

Example 15: Lack of syntactic elements (definite article)

CN: 《纽约时报》不久前报道，…

Gloss: The New York Times recently reported that …

EN by Google Translate: ___ New York Times recently reported that …

Example 15. *Lack of syntactic elements (definite article)*

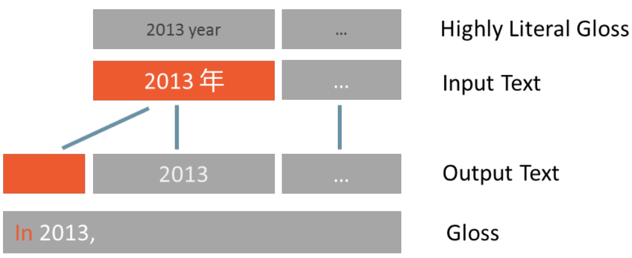
The definite article *the* is one of the most common words in English. It is used under many different circumstances, for example, before some proper nouns including names of geographical areas, rivers, oceans, names of organisations, and names of newspapers etc. However, there are not articles (definite/indefinite) in Chinese at all and no words need to come at the beginning of the proper nouns. Example 15 represents the lack of a definite article in the English output sentence. Usually, such errors are caused by the difference of the syntactic structures between these two languages. Some words are required in English to form a well-formed sentence, but are not necessary in Chinese. For another example, it is quite common to express time (day, month, and year) without preposition in a Chinese sentence. But the lack of a preposition before the figure of a year in English leads to ungrammaticality, as can be seen in Example 16.

Example 16: Lack of syntactic elements (preposition)

CN: 2013 年，…

Gloss: In 2013, …

EN by Google Translate: ___ 2013, …



Example 16. *Lack of syntactic elements (preposition)*

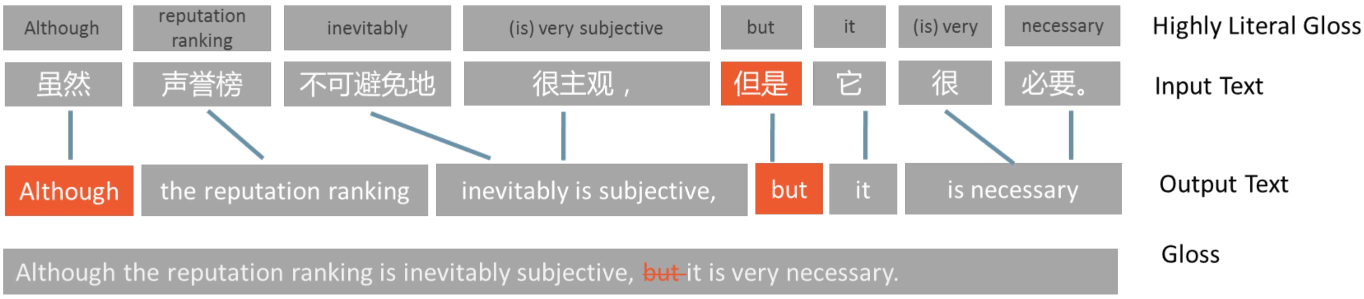
Example 17: Redundant syntactic elements

CN: <u>虽然</u>声誉榜不可避免地很主观，<u>但是</u>它很必要。

Gloss: Although the reputation ranking is inevitably subjective, it is very necessary.

The reputation ranking is inevitably subjective, but it is very necessary.

EN by Google Translate: <u>Although</u> the reputation ranking is inevitably subjective, <u>but</u> it is very necessary.



Example 17. Redundant syntactic elements

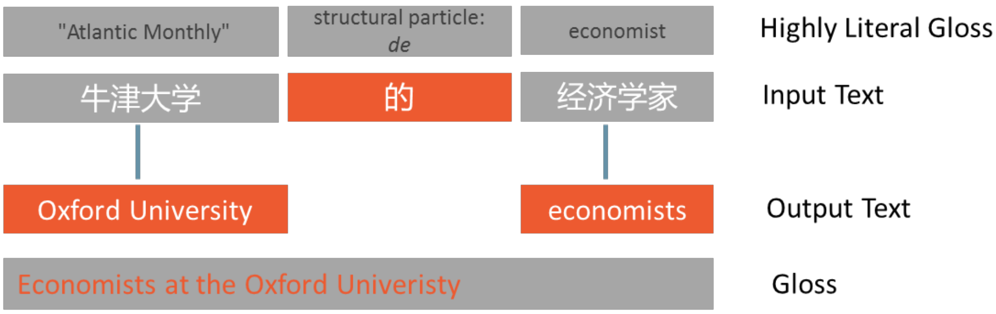
In Chinese, '虽然……但是……' is one of the most used patterns (literally meaning 'although..., but...') and two words are required to be used together. But the subordinating conjunction *although* and the coordinating conjunction *but* cannot appear together in one English sentence. Example 17 represents a syntactic error, called redundant syntactic element(s). Such errors are identified in the analyses when some syntactic elements are needed in the source language but are actually redundant in the target language.

Example 18: Syntactic structure errors (De structure)

CN: 牛津大学的经济学家

Gloss: Economists at the Oxford University

EN by Google Translate: Oxford University economists



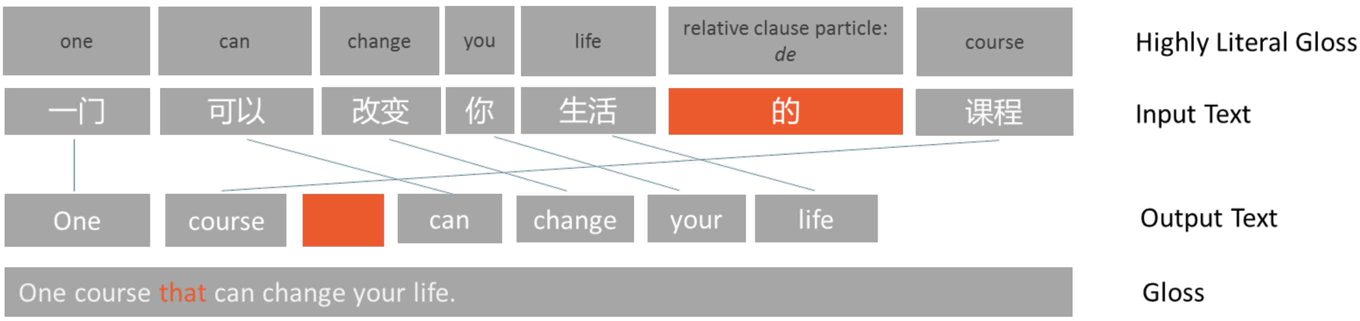
Example 18. *Syntactic structure errors: De structure (noun phrase)*

Example 19: Syntactic structure errors (De structure)

CN: 一门可以改变你生活的课程。

Gloss: One course that can change your life.

EN by Bing Translator: One course can change your life.



Example 19. *Syntactic structure errors: De structure (relative clause)*

De structure is another important grammatical pattern in Chinese language that the MT systems have difficulty dealing with. Example 18 and 19 represent how the MT systems wrongly translate Chinese De structure phrases.

The particle *de* in Chinese is usually used to mark possession or modification for noun phrases. A De phrase is somewhat similar to other attributive adjectival phrases in English, since the De phrase proceeds the noun that it modifies, as is shown in Figure 26.

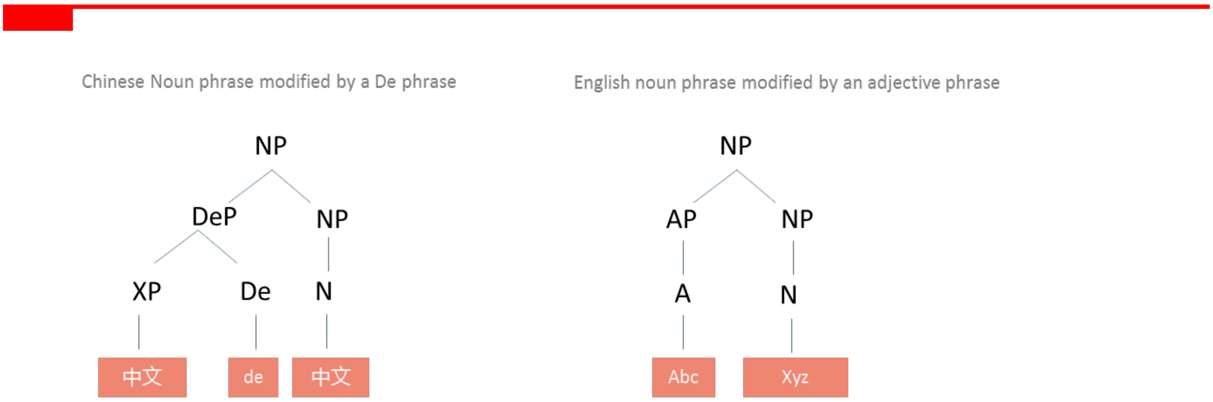


Figure 26. *De structure in Chinese*

One possible way for non-Chinese speakers to understand the De phrase is that it works as *'s* in English. For example, the noun phrase in Example 18 may literally mean *Oxford University's economists*. Unlike English, noun phrases can be modified by either an adjective phrase at the beginning of the noun or an attributive prepositional phrase/a relative clause that comes after the noun, the Chinese De phrase always proceeds the noun phrase that it modifies. The tree diagrams of the noun phrases in Example 18 and 19 are illustrated in Figure 27.

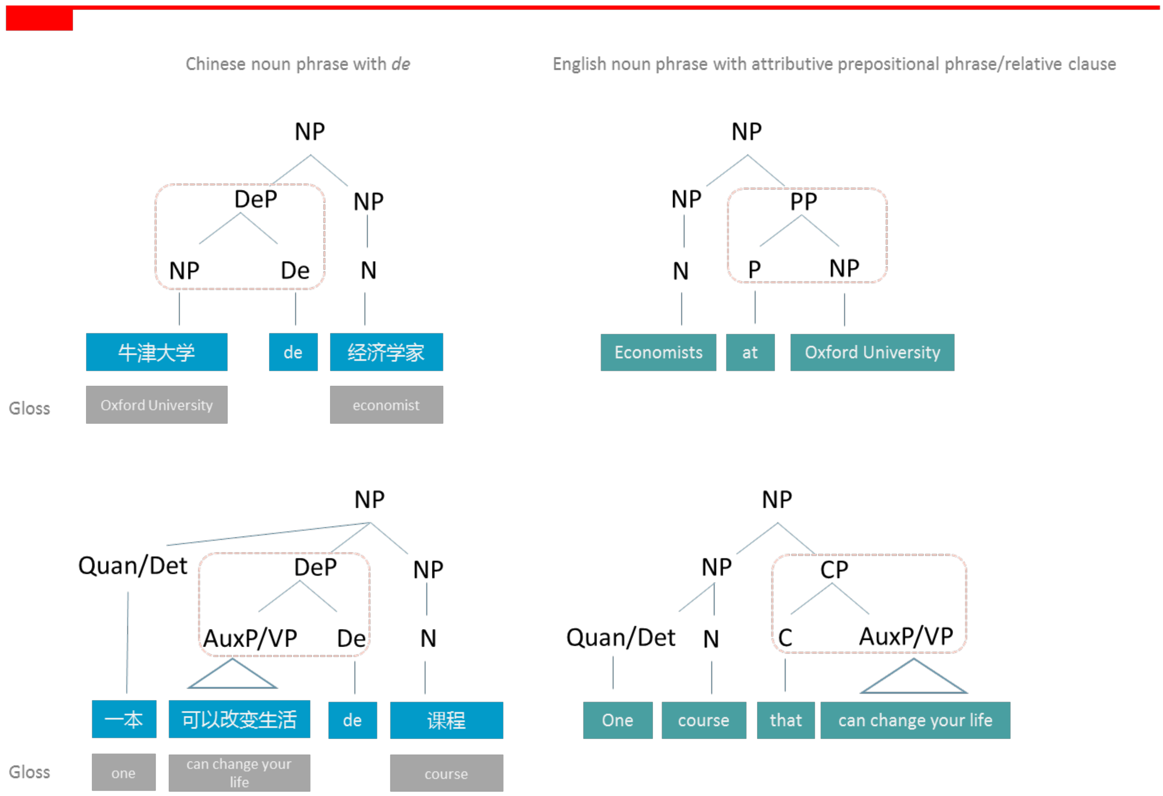


Figure 27. *Tree diagrams of Chinese and English noun phrases*

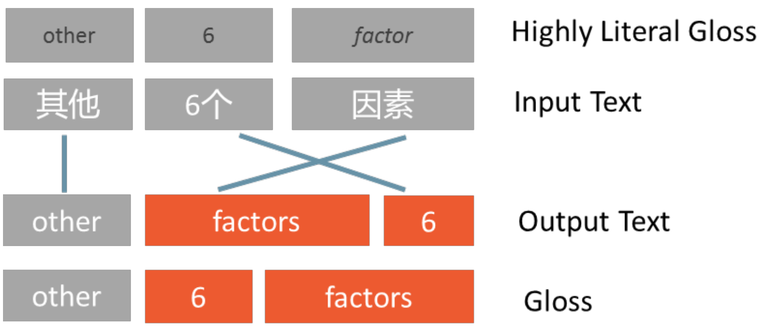
This again, proves that the difference of syntactic structures can easily lead to syntactic errors in MT translations. Because the MT systems are phrase-based, they tend to translate words/phrases independently with the Translation Model and then run reordering with the Language Model to produce a 'most probable' fluent English in the output, which can easily ignore the Chinese particle *de* and leave it out in the output (since it is literally untranslatable). This tendency can be seen in Example 19 – *one*, *can*, *change*, *your life*, *course* are picked by the Translation Model as translations of each phrase/word, and then the Language Model tries to produce a fluent English output, which became 'One course can change your life'.

Example 20: Syntactic structure errors (word order: local range)

CN: 其他 6 个因素…

Gloss: other 6 factors …

EN by Bing Translator: other factors 6 …



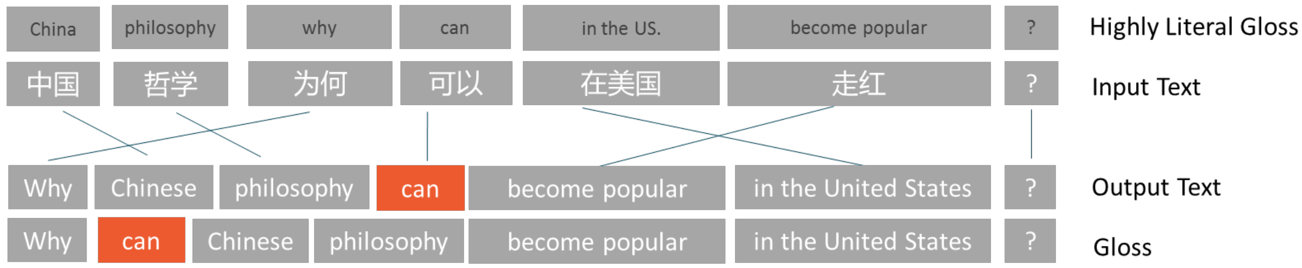
Example 20. *Syntactic structure error: word error (local range)*

Example 21: Syntactic structure errors (word order: long range)

CN: 中国哲学为何可以在美国走红？

Gloss: Why can Chinese philosophy become popular in the United States?

EN by Google Translate: Why Chinese philosophy can become popular in the United States?



Example 21. *Syntactic structure error: word error (long range)*

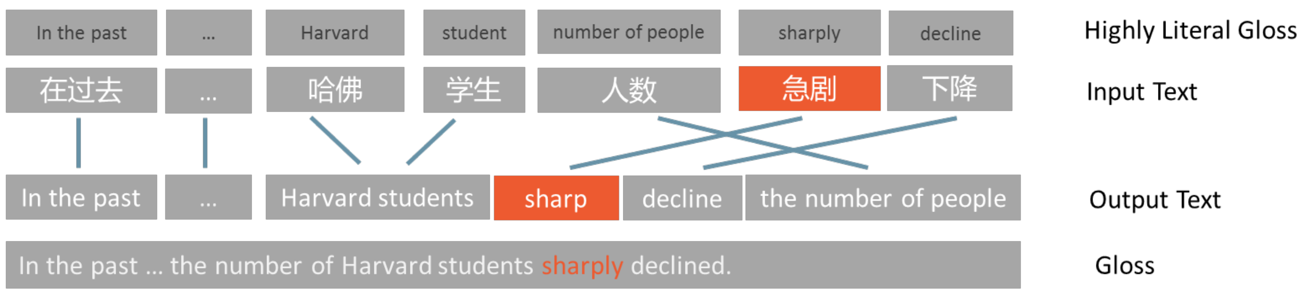
Incorrect word order is another common type of errors that SMT systems generate. Just as mentioned before, the Language Model takes the main responsibility for reordering. It uses probability to produce a 'most probable' word order, but cannot make sure it is a correct output every time. It may prefer *other factors* to be reordered in sequence rather than *other 6* in Example 20, which leads to an ill-formed output. Similarly, *Chinese philosophy can* ... may have a higher probability than *can Chinese philosophy* ... in Example 21. The former represents an error that occurs in a local range, the latter takes place in a long range. Both types are very common in the analyses of my study.

Example 22: word class

CN: 在过去 …… 哈佛学生人数<u>急剧</u>下降。

Gloss: In the past … the number of Harvard students <u>sharply</u> declined.

EN by Google Translate: In the past … Harvard students <u>sharp</u> decline the number of people.



Example 22. *Word class*

Another feature of Chinese language that makes MT Chinese-to-English translation difficult is that it is hard to tell the grammatical category of Chinese words when they do not exhibit morphological changes to show different grammatical functions. Generally, the grammatical category can only be distinguished from its place or the word order in a Chinese sentence. For some reasons, the grammatical category of a word is not even indicated in a Chinese dictionary. In many cases, a same Chinese word can convey the meaning of different grammatical functions. The MT system, therefore, may easily produce a word that is not in accordance with the grammatical category of the source word, and lead to an ill-formed output sentence, as in Example 22. More examples of word class error are presented in Table 4.

Gloss	Input Word	Output Word	Grammatical Category
sharply	急剧 jiju	sharp	adverb-adjective
analyse	分析 fenxi	analysis	verb-noun
liberate	解放 jiefang	liberation	verb-noun
controversy	争议 zhengyi	controversial	noun-adjective

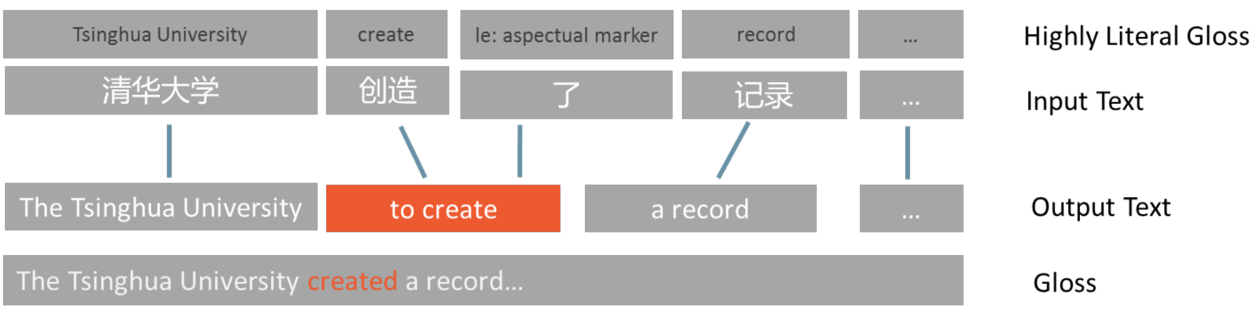
Table 4. *More examples of wrong word class*

Example 23: wrong verb form (finite/non-finite, voice, aspect)

CN: 清华大学创造了记录……

Gloss: The Tsinghua University created a record...

EN by Google Translate: The Tsinghua University to create a record …



Example 23. *Wrong verb form (finite/non-finite)*

Again, due to the fact that Chinese words do not show morphological changes, the MT system tends to make mistakes with the verb form in the English output quite commonly. In Example 23, the non-finite form of the verb *create* is chosen, which causes a syntactically ill-formed sentence. More examples of wrong verb form are listed in Table 5, with regard to voice, aspect, and type (finite/non-finite).

Gloss	Input	Output	Error Type
results show	结果显示	results are shown	voice
has come	已经来临	is coming	aspect
use	使用	using	finite/non-finite

Table 5. *More examples of wrong verb form*

3.4 Further Glimpses into the Analyses

One of the primary goals of this study is to evaluate and compare the outputs of Google Translate and Bing Translator, in order to investigate the difference of the performances of these two SMT systems when translating Chinese into English. The linguistic evaluation method and the taxonomy of the linguistic errors at different levels introduced above allow me to analyse the outputs in a systematic manner and even conjecture some of the possible reasons underlying the linguistic errors produced by these two systems, which may provide potential solutions to some of the challenges that the SMT systems are facing.

During the process of my analyses, it is hard to make a straightforward comparison of these two translation systems by considering individual sentences, because the performance of each system appears to vary considerably at translating one identical source sentence and one set of output sentences usually contain various types of errors at different linguistic levels. For example, Table 6 presents the analysis of outputs of the first source sentence, in which Google generates 8 errors and Bing produces 10. That is a case where neither Google nor Bing performs satisfactorily. Table 7 shows the analyses of source sentence No. 28 and 29, in which both systems produce non-identical outputs without any error. Moreover, Table 8 is an example where Google generates much more errors at different levels (8 in total, 5 syntactic errors, 2 lexical errors and 1 morphological error) than Bing does (1 syntactic error and 1 lexical error) when translating source sentence No.18. On the contrary, Table 9 exhibits how Bing performs much worse than Google at translating sentence No.38, in which the output of Bing contains 7 errors (6 morpho-syntactic errors and 1 lexical error) but the output of Google has only 1 syntactic error. Lastly, Table 10 shows Google and Bing produces same number and type of errors, with two non-identical output sentences when translating the same De structure noun phrase in Chinese.

NO.	SOURCE TEXT	TARGET TEXT	DISSECTED COMPONENTS	ERROR TYPE	ORTHOGRAPHICAL	MORPHOLOGICAL	SEMANTIC	LEXICAL	SYNTACTIC
1.	据BBC中文网3月16日报道,联合国可持续发展解决方案网络(SDSN)与哥伦比亚大学地球研究所周三(16日)共同发布的报告指出，丹麦是世界上最幸福的国家，而布隆迪是世界上最不幸福的国家。	**According to the BBC Chinese network March 16th reported that the UN Sustainable Development Solutions Network (SDSN) and the Earth Institute at Columbia University on Wednesday (16th) jointly issued the report pointed out that Denmark is the happiest countries, Burundi is the world's most unhappy country.**	*** According to the BBC Chinese network March 16th reported that** - According to the BBC News (Chinese) March 16th reported that	*proper noun*				o	
			- -According to the BBC Chinese News (Chinese) on March 16th reported that	*lack of preposition*					o
			- - - ~~According to~~ the BBC Chinese News (Chinese) on March 16th reported that	*redundant preposition*					o
			***… jointly issued the report pointed out that** - the report that SDSN and … jointly issued	*syntax: relative clause*					o
			*** Denmark is the happiest countries** - Denmark is the happiest country	*wrong noun inflection*		o			o
			- - Denmark is the happiest country in the world	*omission*				o	
			* Burundi is the world's most unhappy county. - while Burundi is the world's most unhappy country.	*omission*				o	
		BBC Chinese website on March 16th, it was reported, solutions for the sustainable development of the United Nations network (SDSN) and the Earth Institute at Colombia University on Wednesday (16th) jointly issued the report points out that Denmark is the happiest country in the world, Burundi is the most happiest country in the world.	*** BBC Chinese website on March 16th, it was reported** - by BBC Chinese website on March 16, It was reported	*omission*				x	
			- - by BBC News (Chinese) on March 16th, It was reported	*proper noun*				x	
			- - - by the BBC News (Chinese) on March 16th, It was reported	*lack of definite article*					x
			- - - - It was reported by the BBC News (Chinese) on March 16th	*word order*					x
			- - - - - It was reported by the BBC News (Chinese) on March 16th that	*lack of complementiser*					x
			*** solutions for the sustainable development of the United Nations network (SDSN)** - the United Nations Sustainable Development Solutions Network (SDSN)	*proper noun*				x	
			*** …jointly issued the report points out that** - the report that SDSN and … jointly issued	*syntax: relative clause*					x
			*** Burundi is the most happiest country in the world.** - Burundi is the ~~most~~ happiest country in the world.	*wrong adjectival inflection*		x			
			- - Burundi is the unhappiest country in the world.	*antonym*			x		
			- - - While Burundi is the unhappiest country in the world.	*omission*				x	

Table 6. Error analysis No1.

NO.	SOURCE TEXT	TARGET TEXT	DISSECTED COMPONENTS	ERROR TYPE	ORTHOGRAPHICAL	MORPHOLOGICAL	SEMANTIC	LEXICAL	SYNTACTIC
28.	我们很乐意看到越来越多的中国高校出现在榜单之中。	We are happy to see more and more Chinese universities appear on the list.	√						
		We would be happy to see more and more Chinese universities appear on the list.	√						
29.	2016年，他们共收到来自133个国家的10323份有效回复。	In 2016, they received a total of 10,323 valid responses from 133 countries.	√						
		In 2016, they received a total of 10,323 valid replies from 133 countries.	√						

Table 7. *Error analysis No.28-29*

NO.	SOURCE TEXT	TARGET TEXT	DISSECTED COMPONENTS	ERROR TYPE	ORTHOGRAPHICAL	MORPHOLOGICAL	SEMANTIC	LEXICAL	SYNTACTIC
18	事实上，自19世纪工业革命以来，就不断有争议指出技术发展造成了更多人失业，最终会让机器取代人类。	In fact, since the 19th century Industrial Revolution, we have continued to point out the controversial technology creates more unemployment, the machines will eventually replace humans.	*** since the 19th century Industrial Revolution,** - since the Industrial Revolution in the 19th century,	*syntax: NP*					O
			*** we have continued to point out the controversial technology creates more unemployment,** * ~~we~~	*addition*				O	
			*** controversial** - controversies	*word class*					O
			- - we controversies have continued to point out the technology creates more unemployment,	*word order*					O
			- - - controversies have continued to point out that the technology creates more unemployment,	*lack of syntactic elements*					O
			- - - - controversies have continued to point out that the technology development creates more unemployment,	*omission*				O	
			- - - - - controversies have continued to point out that the technology development created more unemployment,	*wrong verb inflection*		O			O
		In fact, ever since the industrial revolution in the 19th century, there have been controversies pointed out that technological development has created more unemployment, and will eventually supersede humanity.	*** there have been controversies pointed out that technological development has created more unemployment,** - there have been controversies pointing out that technological development has created more unemployment,	*wrong verb form-voice*					X
			*** and will eventually supersede humanity.**	*omission*				X	

Table 8. *Error analysis No.18*

NO.	SOURCE TEXT	TARGET TEXT	DISSECTED COMPONENTS	ERROR TYPE	ORTHOGRAPHICAL	MORPHOLOGICAL	SEMANTIC	LEXICAL	SYNTACTIC
38.	该公司聘请了4名前美国国家高速公路交通安全管理局高官，他们帮助谷歌游说美国政府接受谷歌的自动驾驶汽车技术。	**The company hired four former US National Highway Traffic Safety Administration officials who helped Google lobbying the US government to accept Google's autonomous vehicles technology.**	*** The company hired four former US National Highway Traffic Safety Administration officials who helped Google lobbying the US government to accept Google's autonomous vehicles technology.** - The company hired four former US National Highway Traffic Safety Administration officials who helped Google lobby the US government to accept Google's autonomous vehicles technology.	*wrong verb form-finite/non-finite*					o
		The companies employ 4 former United States National Highway Traffic Safety Administration official, they help Google's lobbying the United States Government to accept Google's self-driving car technology.	*** The companies employ 4 former United States National Highway Traffic Safety Administration official,** - This companies employ 4 former United States National Highway Traffic Safety Administration official,	*wrongly translated words*				x	
			- - This company employ 4 former United States National Highway Traffic Safety Administration official,	*wrong noun inflection*		x			x
			- - - This company employed 4 former United States National Highway Traffic Safety Administration official,	*wrong verb inflection*		x			x
			- - - - This company employed 4 former United States National Highway Traffic Safety Administration officials,	*wrong noun inflection*		x			x

Table 9. *Error analysis No.38*

NO.	SOURCE TEXT	TARGET TEXT	DISSECTED COMPONENTS	ERROR TYPE	ORTHOGRAPHICAL	MORPHOLOGICAL	SEMANTIC	LEXICAL	SYNTACTIC
41.	一门可以改变你生活的课程。	One can change your course of life.	*** One can change your course of life.** -One course **that** can change your life.	*syntax: relative clause*					o
		A course can change your life.	*** A course can change your life.** - A course **that** can change your life.	*syntax: relative clause*					x

Table 10. *Error analysis No.41*

All the errors identified in the output sentences are put under different subcategories in the tables, marked in yellow. The detailed error annotations are also included in the tables, so I will not elaborate the descriptions of the procedures for all the sentence analyses due to limited room. From the examples mentioned above, we can see how considerably different the performances of two systems are when they are translating one identical source sentence. It is, therefore, necessary to count the total number of each error type that each system produces, and then compare their performance by looking at the error statistics.

4. Results

4.1 Error Statistics

All 50 sets of translations produced by Google Translate and Bing Translator were analysed manually based on the guidelines described above. When the analyses were completed, the numbers of errors were counted, as is shown in Table 11:

TRANSLATOR	ERRORS IN TOTAL	ORTHOG.	MORPHO.	SEMANT.	LEXICA.	SYNTAC.	SENTENCES WITHOUT ERRORS
Google	170	4	11	10	42	103	6
Bing	153	5	15	8	49	76	5

Table 11. *Number of errors detected in Google and Bing*

In the same quantity of Chinese-English translation tasks (with English as the target language), the outputs of Google contain 170 linguistic errors in total, and 4 orthographical errors, 11 morphological errors, 10 semantic errors, 42 lexical errors and 103 syntactic errors respectively. Google generates 6 sentences without errors out of 50, which can be found in Appendix C (No.3, 8, 28, 29, 42 and 49). By comparison with Google Translate, Bing produces 153 linguistic errors totally, in which there are 5 orthographical errors, 15 morphological errors, 8 semantic errors, 49 lexical errors and 76 syntactic errors. There are 5 sentences translated by Bing that are flawless, which also can be seen in Appendix C (No.3, 28, 29, 42 and 49).

The first impression from the statistics obtained is that Bing Translator has a slightly better performance than Google Translate at the same quantity of translation tasks, because Bing produces less linguistic errors in total, plus the gap at syntactic level is especially noticeable. In order to compare these two SMT systems in more depth, I will look into the statistics of the subcategories at each linguistic level, presented in Table 12, 13, 14, 15 and 16.

Translator	Orthographical Errors in Total	Capitalisation	Punctuation
Google	4	1	3
Bing	5	3	2

Table 12. *Numbers of orthographical errors*

Translator	Morphological Errors in Total	Wrong Noun Inflection	Wrong Verb Inflection	Wrong Adjective Inflection
Google	11	8	3	0
Bing	15	10	4	1

Table 13. *Numbers of morphological errors*

Translator	Semantic Errors in Total	Polysemy	Antonym	Disambiguation Error
Google	10	7	1	2
Bing	8	5	1	2

Table 14. *Numbers of semantic errors*

Translator	Lexical Errors in Total	Omission	Addition	Proper Nouns	Wrongly Translated Words
Google	42	19	9	7	6
Bing	49	28	8	9	5

Table 15. *Numbers of lexical errors*

Translator	Syntactic errors in total	Lack of syntactic elements	Redundant syntactic elements	De structure	Word order	Word class	Wrong verb form	Morpho-syntactic errors
Google	103	19	7	12	31	4	12	11
Bing	76	27	4	7	22	1	8	14

Table 16. *Numbers of syntactic errors*

4.2 Interpretation of the Statistics and the Evaluation Results

To better visualise the differences between the performances of these two SMT systems, Figure 28 is presented below:

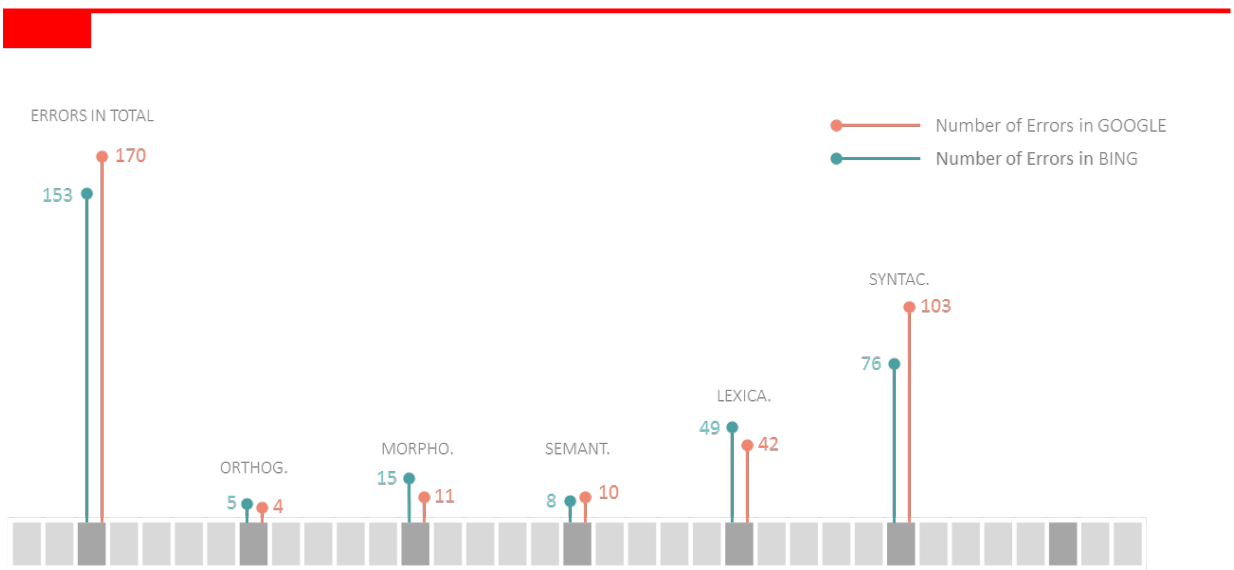


Figure 28. *Numbers of errors in Google and Bing*

Based on the comparison of the numbers of errors, my first expectation for this study is met. Since Bing Translator claims to be a linguistically informed SMT service that incorporates various linguistic information, for example, syntactic analysis on the source side (Dolan et al., 2002 and Quirk et al., 2005), my first prediction was:

- Bing Translator should outperform Google Translate with regard to the total number of errors when performing the same translation tasks. That is to say it will produce fewer errors in total, and have a better performance at least with regard to syntax.

As mentioned in Chapter 2, Section 2.3, the evaluation study done by Farrús et al. (2012), shows that the most frequent errors in all studied systems are found at the syntactic and semantic levels, and that orthographical errors are the least frequent, in the translation outputs of English-Catalan and Spanish-Catalan language pairs. On the basis of this, my second prediction for my study was:

- In the translation outputs of both Google Translate and Bing Translator, syntactic and semantic errors will have a larger distribution than the others, while there will be substantially fewer orthographical errors. The number of morphological and lexical errors will fall somewhere in between.

This expectation is partly met according to the statistics obtained. Indeed, the syntactic errors have a larger distribution than the others and the orthographical errors have the least percentage. But semantic errors actually do not show significant influence on the output quality, and the numbers are slightly larger than orthographical errors. Lexical errors, however, takes up the second largest proportion of the result, following the syntactic errors. The number of morphological errors stands in the middle.

One interesting result that pertains to both systems is that the distribution of errors at five linguistic levels shows the same pattern, which can be seen in Figure 29 and 30:

- Syntactic > Lexical > Morphological > Semantic > Orthographical

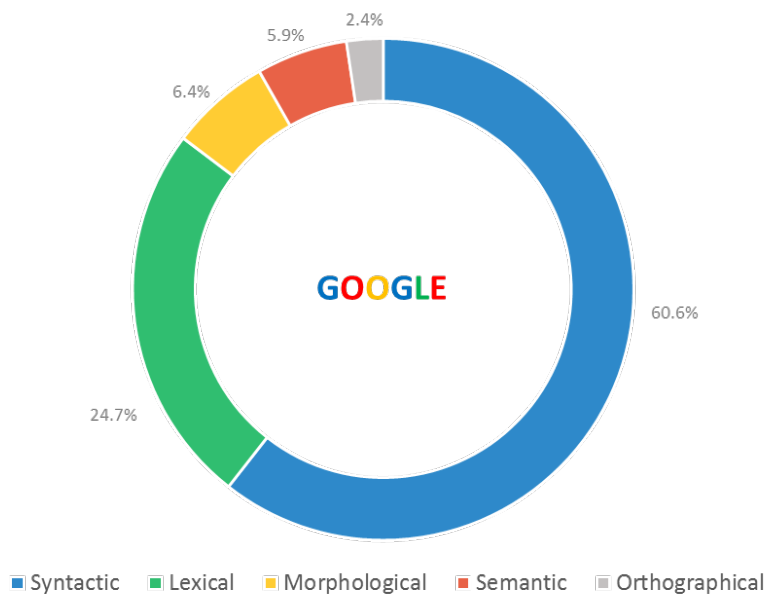


Figure 29. *Distribution of Errors at each level in **Google***

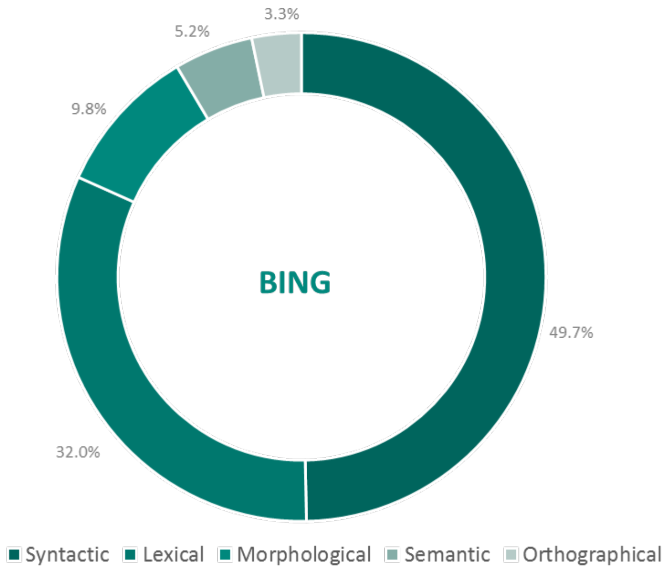


Figure 30. *Distribution of errors at each level in* ***Bing***

This result provides evidence that errors at syntactic and lexical level are the most frequent ones which will more or less affect the translation output quality of both of the two SMT systems studied. More specifically, the comparisons of syntactic and lexical errors at sublevels are exhibited below, in Figure 31 and 32.

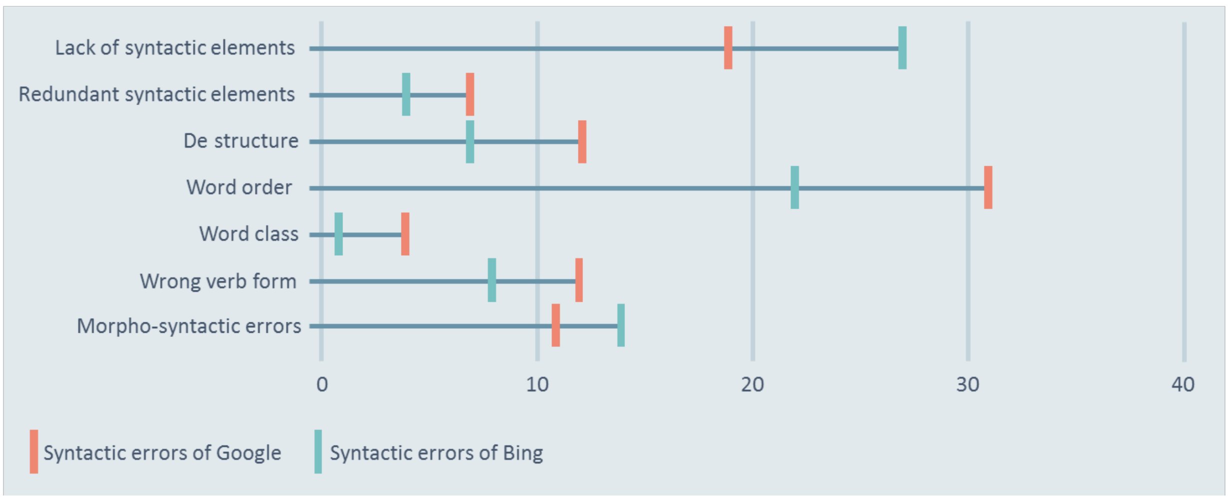


Figure 31. *Numbers of syntactic errors at sublevels in Google and Bing*

According to Figure 31, Bing generally performs better at all sublevels except for morpho-syntactic errors and the lack of syntactic elements, producing less syntactic errors than Google in regard to redundant syntactic elements, de structure, word order, word class and verb form. **Lack of syntactic elements** seems to be the most discernible syntactic errors in Bing, while errors of **word order** comprise the largest subcategory of syntactic errors in Google.

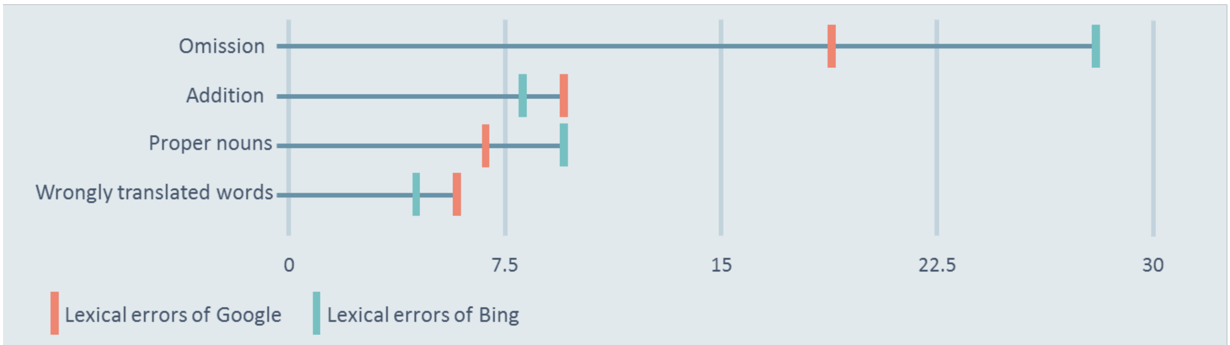


Figure 32. *Numbers of lexical errors at sublevels in Google and Bing*

Figure 32 shows that at the lexical level, **omission** is strikingly frequent in both systems, with Bing producing more errors of omission than Google. The numbers of addition, wrongly translated proper nouns and wrongly translated other words are quite close, in the range of 5 – 9.

Three general observations can be extracted from the data:

i. In the same quantity of Chinese-to-English translation tasks, Bing Translator, an SMT system which incorporates linguistic information, does outperform Google Translate, which is a pure SMT system that does not use linguistic rules. In general, Bing produces fewer linguistic errors, especially at syntactic level.

ii. When performing Chinese-to-English translation tasks, **syntactic errors** and **lexical errors** appear to be the types that have the most influence on translation quality, in both Google Translate and Bing Translator. The influence of orthographical errors is nuanced.

iii. **Word order** is Google's biggest weakness at syntactic level, and the **lack of syntactic elements** is Bing's most frequent syntactic problem. At lexical level, **omission** is the main source of errors in both systems.

5. Discussion

In this chapter I will address the findings from the study in light of my research goals and questions. First, I will discuss the linguistic evaluation method that I adopted and to some extent refined, based on the study of Farrús et al. (2010, 2011 and 2012). Second, I will discuss the results of this linguistic evaluation and some thoughts on the value of the results for MT system developers or for further academic research. Lastly, the limitations and potential sources of problems of this study will be explored.

5.1 Discussion about the Linguistic Evaluation Method

I have introduced four types of MT evaluation methods in the beginning, including automatic evaluation (Papineni et al., 2002, Doddington, 2002, Lavie and Agarwal, 2007), adequacy and fluency judgements (LDC, 2005), error analysis (Vilar et al., 2006) and linguistic evaluation (Farrús et al., 2010, 2011 and 2012). The **automatic evaluation** is considered problematic because it uses human translations as references which could easily cause subjectivity, and the reliability is shady because it could wrongly decide that a good output is bad just because it does not look like the human translation reference. Moreover, the results of automatic evaluation cannot always identify the most frequent types of errors in a specific system, because it is just a scoring tool in essence. In addition, the **adequacy and fluency judgements** by human evaluators are more subjective and inconsistent, based on their coarse judgement standards and vague scoring definitions. The **error analysis** method provides a new perspective to do MT evaluation, namely identifying and counting errors that a specific MT system produces. But, still, they need human translations as references which are sources of being subjective. Besides, the classification of the errors is, admitted by Vilar et al. (2006) themselves, by no means unambiguous. The errors defined are not mutually exclusive and it is common to see one type of error causing another to occur, namely derivative errors. Lastly, the **linguistic evaluation** that my current study is inspired by and based on, uses linguistic-based evaluation criteria to define and classify errors at five linguistic levels, which is less subjective and more systematic, compared to the others. One problem of their method is that the guidelines for categorisation are sometimes unclear when it comes to the exclusiveness of errors at different levels. For example, in their guidelines, 'morpho-syntactic changes due to changes in syntactic structures' and 'lack of gender and number concordance' are defined as morphological errors, ignoring the syntactic aspects. In addition, 'errors in prepositions' and 'missing or spare articles' are categorised as syntactic errors, whereas they may happen at lexical errors in a certain situation.

The linguistic evaluation method that I have experimented on my study, is an updated and expanded version of the previous one proposed by Farrús et al. (2010, 2011 and 2012). It is a more fine-grained linguistic evaluation method which is more objective, consistent, exclusive and systematic, despite being time-consuming as any other type of human evaluation method.

First, it avoids subjectivity as much as possible. In the other MT evaluation methods, two elements can result in subjectivity, including human's perceptual judgement as in the method of adequacy and fluency judgements and human reference translation as in many of the other methods. In my study, more scientific and objective 'linguistic judgements' are made on the outputs, rather than perceptual judgements on whether the translation is 'good' or 'bad'. More important, no reference translation is provided nor required in this approach. In the analyses, I did not decide how a source sentence should be translated and then compare it with the output of an MT system. The output and input conjointly, in fact, function as a 'reference' themselves. As explained before, strictly speaking, there is no true translation for a foreign input sentence, that is why researchers try to use statistical methods to do translation. The output of the SMT systems, is already 'the best' translation that the machines could produce, since the system decides it is the most probable one. The input, at the same time, is always a true translation if we think backwards, pretending the actual output (target sentence) as an 'input', and the actual input (source sentence) as an 'output'. That is to say, when I am evaluating one output of an SMT system, I do not compare it with any other references which are very likely to cause subjectivity, but I segment the output sentences into phrases, align them backwards with the source, then directly look at the 'most probable' output and the 'always true' input conjointly, to find any divergences that occur, which namely are the errors identified at five linguistic levels.

Second, this type of evaluation is consistent in terms of evaluation criteria. There is only one evaluator required to evaluate the outputs, with the strict error classification and annotation guidelines that the evaluator sticks to. While in other types of human evaluation methods when several evaluators are involved, it is always hard to reach absolute agreement or consensus of some cases, hence creating inconsistency in the evaluation results.

Thirdly, the linguistic errors in the taxonomy I presented are more mutually exclusive, which gets rid of some difficulties in identifying errors and avoids derivative errors during analyses.

Indeed, in the actual study, there are occasionally complex cases where I find difficult to distinguish error types that seem to fall on boundaries. One output sentence may, at first glance, look 'superficially' syntactically ill-formed. We cannot at once decide that this is a syntactic error or suchlike. Both SMT systems are phrase-based, and we need to dissect the output into a 'corrupted' sentence and then inspect each component independently. The important distinctions between lexical and syntactic errors, morphological and syntactic errors and orthographical and lexical errors are illustrated in Figure 33, 34 and 35, which Farrús et al.'s studies did not clarify thoroughly.

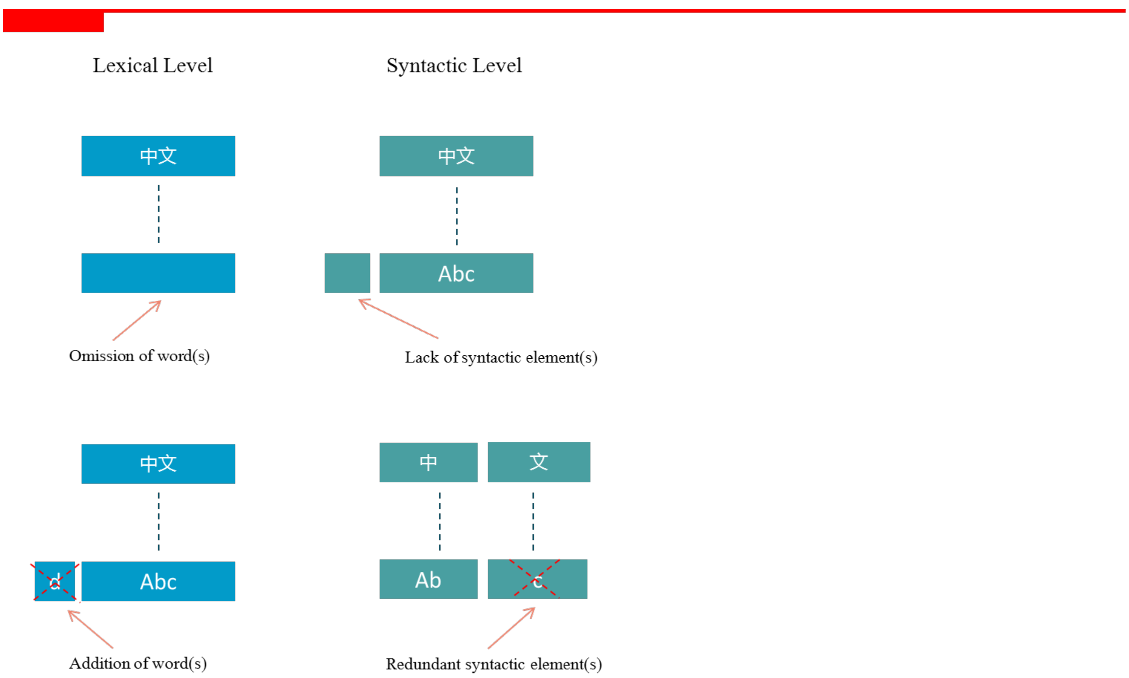
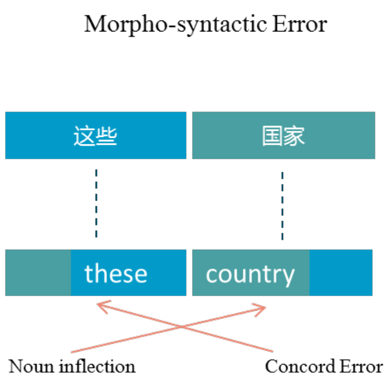


Figure 33. *Distinctions between lexical and syntactic errors*



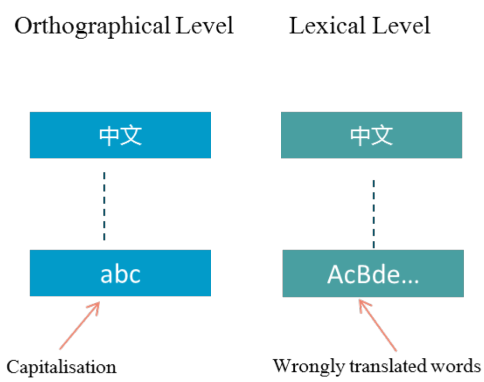


Figure 34. *Morpho-syntactic errors*

Figure 35. *Distinctions between ortho.and lexic. errors*

Lastly, the evaluation has been manually performed in a systematic way which can be potentially incorporated with automatic methods, so some of the work could be done by machines and the evaluation time may be largely reduced. Indeed, with some of the existing purely automatic evaluation methods, the evaluation time can be considerably shorter, which is very important for the developers to supervise and monitor the progress of their systems. But Koehn (2010) notes that compared to automatic evaluation approaches, it is more reasonable that we tend to put more trust into the judgement of humans rather than machines, because only human evaluators can examine the output sentence by sentence, assess each sentence with context, and conclude with an overall performance of each translation system. My current study attempts to, ideally speaking, present some possibilities of combining human evaluation methods with automatic methods. In computer programming, it is common to run a program with a 'loop' to do some repetitive and monotonous tasks. When I was introducing the procedures of analysing an output sentence, I intentionally went through the dissected components of the output again and again, inspecting them repetitively from orthographical level, through morphological and semantic levels, to lexical and syntactic levels, which was actually a program-like 'loop'. Some studies have already made efforts in developing the hybrids of human evaluation and automatic metrics, for example, Popovic and Burchardt (2011) propose an automatic classification for MT output based on human error analysis.

5.2 Discussion about the Evaluation Results

The results of the study indicate that the performances of Google Translate and Bing Translator do differ, when they are translating the same quantity of translation tasks. The difference is clearer at syntactic level. This proves that incorporating linguistic information in the SMT systems may, to some extent, improve the output quality. The syntactically-informed approach in Microsoft's product shows its robustness and reliability in producing sentences at the target side, compared to the purely statistical method in Google's. Nevertheless, this does not mean that Bing manages to always produce 'better' outputs than Google, since we can clearly see from the results obtained, both systems made a certain number of errors at different linguistic levels, with only 5-6 flawless output sentences out of 50. It implies that SMT systems are still prone to linguistic errors, due to the differences between languages. Dorr (1994) refers to the differences between languages as translation divergences and points out that an understanding of what causes the divergences will help us in building models that overcome them. Based on the analyses of this study, we can see that some translation divergences, including lexical divergences, typological differences and other structural divergences, do affect the translation output quality when the systems are translating the Chinese-English language pair.

5.2.1 Syntactic Errors

Various types of errors are present at syntactic level, including the lack of syntactic elements, redundant syntactic elements, de structure errors, word order errors, word class errors and wrong verb forms.

Structural divergences between Chinese and English are the main sources of errors such as the **lack of syntactic elements**, **redundant syntactic elements**, **de structure** and **word order**. For example, the lack of definite article as in Example 15 occurs quite frequently because there are not definite or indefinite articles in Chinese. The redundant syntactic elements are found because words such as *although* and *but* can appear together in one Chinese sentence but not in English. In addition, the de structure in noun phrases or relative clauses in Chinese is very common, which makes it hard to translate long Chinese sentences with a de structure.

To reduce such errors, researchers should develop better Language Model that can takes account of syntactic rules. For example, add definite article *the* in front of certain proper nouns, and prohibit *although* and *but* from being present together in one sentence. To avoid word errors, better reordering models should be explored to produce more fluent English output. For example, the model should be able to detect that *other factors 6* is less probable than *other 6 factors*, and auxiliary verbs such as *can* should proceed the noun phrase in an English interrogative sentence (see Example 21). A more advanced Chinese-English Translation Model should also be developed, which is better trained with Chinese de structures so it can correctly translate long Chinese sentences with relative clauses or noun phrases with de structure.

As explained in Example 22 and 23 in Chapter 3, errors regarding **word class** and **verb forms** are mainly due to the typological differences between Chinese and English. Chinese words do not show morphological changes, so the same words may have different grammatical categories, and the verbs always remain unchanged. A possible solution to decreasing word class errors might be to implement 'word class tagging' or 'part-of-speech tagging' technology to perform an analysis on the source target and mark up a word as corresponding to a particular part of speech. Researchers should also consider to build a 'penalty model' on the target side to prevent words that have different word class as in the source text. A remedy for the wrong verb form is challenging to find, because the Chinese verbs are invariable under all conditions. But the researchers may consider taking context or words that denote syntactic information into consideration. For example, use the aspectual markers in Chinese to produce a correct aspect or tense in English.

5.2.2 Lexical Errors

Lexical errors comprised the second largest category in both systems, with **omission** being the most common ones. The systems leave out a source word if the word has never been seen during training; such words are typically obscure or low-frequency in the source language. Nevertheless, certain source words were omitted in the output sentences in this study despite their being quite common in Chinese language. This phenomenon can, at least, provide a possible avenue of investigation for the SMT system developers. In Google's newest approach, namely GNMT (Google's Neural Machine Translation system), Wu et al. (2016) introduce that their beam search technique employs a length-normalisation procedure and uses a coverage penalty, which encourages generation of an output sentence that is most likely to cover all the words in the source sentence. This method may be optimal for reducing omission errors in the outputs. Likewise, a similar penalty model could be used to 'punish' the system for adding words that are not found in the source sentence, namely errors of **addition**. In Example 12, we have an output sentence *in Italy ranked No.50*, the preposition *in* should not be added when *Italy* was supposed to be the subject of the sentence. Even though *in Italy* may be assigned by the computer a higher probability than *Italy*, the output should receive a penalty because the preposition *in* does not exist in the source at all when the source word *Italy* is the subject of the sentence. Lastly, **wrong translated words** are mainly because of lack of training, and such problems can be improved by enlarging the bilingual corpora of the Translation Model.

Omission and addition are very likely to cause ungrammaticality in the output sentences as discussed before, which may crucially affect the overall accuracy of the MT systems. Further researches should make more efforts to focus on how to tackle such problems.

5.2.3 Morphological and Semantic and Orthographical Errors

Compared to syntactic and lexical errors, the other three categories are much less problematic in both Google and Bing, as seen in the error distribution. But special attention should be paid to the **morphological errors**. Even though this type of error does not show significant influence on the general performance of both systems in this study, the differences of morphological features between Chinese and English could be a potentially important source of errors when Chinese is the source language and English is the target. Human translators can easily overcome this challenge because they can use the context to decide whether there should be a noun inflection or verbal inflection in English. An SMT system does not understand context, thus it cannot always correctly decide to inflect a word or not. The system developers should not neglect this factor if they want to improve the quality of Chinese-English translation tasks.

These two SMT systems had an impressive performance on selecting an appropriate meaning where a given source word had multiple meanings; both systems produced only a fraction of **semantic errors**. This shows that the 'semantic barriers' encountered in the rule-based machine translation systems since the 1950s have been greatly overcome in the SMT systems. Besides, it was anticipated that orthographical errors would only have a minor impact on the translation quality of both systems. **Orthographical errors** tend to be more frequent in human translations, but a machine system is usually trained with texts that do not have many such errors.

5.3 Limitations of This Study

The inevitable limitation of this study is that online translation services update their systems very quickly. Given they constantly update their language models and translation models (including monolingual and bilingual corpora), the output sentences will also change on a regular basis. That is to say, with the same source sentences, the target language sentences these systems generate will now be somewhat different from the time when I collected my data (April 2016 – July 2016). The difference in Google is enormous since they launched their newest Google's Neural Machine Translation (GNMT) in late September of this year (2016). When I feed in some of the same source sentences into Google Translate, the output sentences are quite different from those in this study, and surprisingly similar to the sentences after corrections. According to Wu et al. (2016), using a human evaluation on a set of isolated simple sentences, the newer system reduces translation errors by an average of 60% compared to Google's previous phrase-based production system. I have picked several source texts which are translated by GNMT, the comparison of the outputs is presented in Table 17.

Source Text	Target Text of Google (SMT)	Target Text of Google (GNMT)
一门可以改变你生活的课程。	One can change your course of life.	A course that can change your life.
《纽约时报》不久前报道	"New York Times" recently reported that	"The New York Times" recently reported
中国哲学为何可以在美国走红？	Why Chinese philosophy can be became popular in the United States?	Why can Chinese philosophy be popular in the United States?
事实上，自19世纪工业革命以来，就不断有争议指出技术发展造成了更多人失业，最终会让机器取代人类。	In fact, since the 19th century Industrial Revolution, we have continued to point out the controversial technology creates more unemployment, the machines will eventually replace humans.	In fact, since the 19th century industrial revolution, there has been controversy that technological development has led to more people unemployed, and ultimately let the machine replace mankind.

Table 17. *Comparison of the outputs of Google (SMT) and Google (GNMT)*

That the output is changing and updating on a regular basis is a big challenge for all types of MT evaluations, especially human evaluations. It is difficult to ensure that the pace of evaluation keeps up with the speed of system development. The performance results of the two systems in this study can nonetheless be internally consistent and meaningful for this period of time (April 2016 – July 2016). In addition, the evaluation method including the taxonomy of linguistic errors and the annotation guidelines proposed in this study can be used for further studies that aim to identify dominant error types in a specific MT system. Last but not the least, the data collected from the current study could be used to compare with the outputs produced by the newer system, in order to investigate how much the new system has improved its coverage, so that the developers would know something more about their system, in particular, that it is better, or worse than it once was.

6. Conclusions and Future Work

A central goal of this thesis was to evaluate and compare the translation quality of Google Translate and Bing Translator, by using the linguistic evaluation method. A detailed taxonomy of errors at 5 linguistic levels was developed and applied to Chinese-English translation tasks (with English as the target language). 50 Chinese sentences selected from news articles were automatically translated into English by both Google Translate and Bing Translator. Errors in the output sentences were manually analysed and annotated based on the proposed error taxonomy, which allowed me to evaluate two MT systems at each linguistic level, namely the orthographical level, the morphological level, the semantic level, the lexical level and the syntactic level. The results show that in the same quantity of Chinese-to-English translation tasks, Bing Translator, an SMT system which incorporates linguistic information, does outperform Google Translate, which is a pure SMT system that does not use linguistic rules. In general, Bing produces fewer linguistic errors, especially at syntactic level. The distribution of error types shows that syntactic and lexical errors are particularly problematic in both SMT systems, which suggests this is where SMT developers should focus when attempting to improve the translation quality of Chinese-English translation tasks.

Although this thesis has been subject to considerable time and other constraints, it has nonetheless demonstrated where future research efforts might be mostly fruitfully directed. This might involve, for example: increasing the size and domain of the corpora so that the results would be more accurate and more statistically significant; performing an automatic evaluation and a human perceptual evaluation of the same corpora and then studying the correlations of automatic, human perceptual judgements and linguistic evaluation; conducting an English-Chinese translation tasks (with Chinese as the target language) to extend the taxonomy that has been proposed in this study; getting more language pairs involved to test and develop the evaluation method which is used in this study. Furthermore, it would be interesting to compare the outputs of Google's newest Neural Machine Translation with the previous generation of Google Translate, or to compare the newer Google Translate with Bing Translator, to get an understanding of the improvement that Google has achieved. Lastly, an automatic linguistic method could be explored, ideally speaking, by using statistical deep learning approaches to detect some types of errors automatically in order to make linguistic error analysis faster.

REFERENCES

Aiken, M. and Balan, S. (2011). An analysis of Google Translate Accuracy. In *Translators and Computers*, *Translation Journal*, 16(2).

Alansary, S. (2011). Interlingua-based Machine Translation Systems: UNL versus Other Interlinguas. In *11th International Conference on Language Engineering*, Ain Shams University, Cairo, Egypt.

Babst, A. (2015). Google Translate Keeps Getting Better with Your Help. Retrieved September 20, 2016, from https://translate.googleblog.com/2015/06/google-translate-keeps-getting-better.html.

Baisa, V. (2009). Problems of Machine Translation Evaluation. In *Proceedings of Recent Advances in Slovanic Natural Language Processing*, pp. 21.

Balk, E., Chung, M., Chen, M., et al. (2013). Assessing the accuracy of Google Translate to Allow Data Extraction from Trials Published in Non-English Languages. Rockville, MD: Agency for Health Care Research and Quality.

Barreiro, A., Monti, J., Orliac, B., et al. (2014). Linguistic Evaluation of Support Verb Construction by OpenLogos and Google Translate. In *Proceedings of Language Resources and Evaluation Conference*, pp.35-40.

Bennett, S. and Gerber, L. (2003). Inside Commercial Machine Translation. In *Computers and Translation*, pp. 175-190.

Brown, P.F., Cocke, J., Della-Pietra, S.A., et al. (1990). A Statistical Approach to Machine Translation. *Computational Linguistics*, 16(2):76-85.

Costa-jussà, M.R. and Fonollosa, J.A.R. (2015). Latest Trends in Hybrid Machine Translation and its Application. *Computer Speech and Language*, 32(1):3-10.

Denkowski, M. and Lavie, A. (2010). Choosing the Right Evaluation for Machine Translation: An Examination of Annotator and Automatic Metric Performance on Human Judgement Task. In *Proceedings of the Association for Machine Translation in the Americas.*

Doddington, G. (2002). Automatic Evaluation of Machine Translation Quality Using N-gram Cooccurrence Statistics. In *Proceedings of the HLT- NAACL*, pp. 138–145, San Diego.

Dolan, B., Pinkham, J. and Richardson, S.D. (2002). MSR-MT: the Microsoft Research Machine Translation System. In: *AMTA*, pp. 237-239.

Dorr, B. (1987). UNITRAN: An Interlingua Approach to Machine Translation. In *Proceedings of the 6th Conference of the American Association of Artificial Intelligence,* Seattle, Washington, USA.

Dorr, B. (1994). Machine Translation Divergences: A Formal Description and Proposed Solution. *Computational Linguistics*, 20(4), pp. 597-633.

Duff, P. and Li, D. (2002). The Acquisition and Use of Perfective Aspect in Mandarin. In Salaberry, R. and Shirai, Y. (Eds.), *The L2 Acquisition of Tense-Aspect Morphology*, pp. 417-453. Philadelphia, PA: John Benjamins.

Farrús, M., Costa-jussà, M.R., Mariño, J. and Fonollosa, J.A.R. (2010). Linguistic-based Evaluation Criteria to Identify Statistical Machine Translation Errors. In *Proceedings of the 14th Annual Meeting of the European Association for Machine Translation*, pp. 167–173. Saint-Raphaël, France.

Farrús, M., Costa-jussà, M.R., Mariño, J., Poch, M., Hernández, A., Henríquez, C. and Fonollosa, J.A.R. (2011). Overcoming Statistical Machine Translation Limitations: Error Analysis and Proposed Solutions for the Catalan-Spanish Language Pair. *Language Resources and Evaluation*, 45:181.

Farrús, M., Costa-jussà, M.R., Popovic, M. (2012). Study and Correlation Analysis of Linguistic, Perceptual, and Automatic Machine Translation Evaluations. *Journal of the American Society for Information Science and Technology*, 63(1):174-184.

Ghasemi, H. and Hashemian, M. (2016). A Comparative Study of Google Translate Translations: An Error Analysis of English-to-Persian and Persian-to-English Translations. *English Language Teaching*, 9(3): 13-17.

Hutchins, W.J. and Somers, H.L. (1992). *An Introduction to Machine Translation,* pp*. 75, 362.* London: Academic Press.

Hutchins, W.J. (1995). Machine translation: A Brief History. In Koerner EFK & Asher RE (Eds), *Concise History of the Language Sciences: From the Sumerians to the Cognitivists*, pp. 431–445. Oxford: Pergamon Press.

Hutchins, W.J. (2006). Machine translation: History of Research and Use. In K. Brown (Eds.), *Encyclopaedia of languages and linguistics 2nd Edition*, Vol. 7, pp. 375-383. Oxford: Elsevier.

Hutchins, W.J. (2010). Machine Translation: A Concise History. University of East Anglia, United Kingdom, *Journal of Translation Studies, 13* (1 & 2):29–70.

Jurafsky, D. and Martin, H.J. (2008). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, 2nd Edition. Upper Saddle River, NJ: Prentice-Hall.

Kaji, H. (1988). An Efficient Execution Method for Rule-Based Machine Translation. In *Proceedings of the 12th Conference on Computational Linguistics Association for Computational Linguistics*, 2:824-829.

King, M. (1997). Evaluating Translation. In *Christa Hauenchild and Susanne Heizmann, eds., Machine Translation and Translation Theory*. Berlin: Mouton de Gruyter, pp. 251-263.

Koehn, P. (2004). Statistical Significance Tests for Machine Translation Evaluation. In *EMNLP*, pp. 388-395.

Koehn, P. (2010). *Statistical Machine Translation*. Cambridge: Cambridge University Press.

Lavie, A. and Agarwal, A. (2007). METEOR: An Automatic Metric for MT Evaluation with High Levels of Correlation with Human Judgements. In *Proceedings of Workshop on Machine Translation 07*, pp. 228-231.

Lavie, A. (2011). Statistical MT with Syntax and Morphology: Challenges and Some Solutions. Presentation at *LTI Colloquium*. Barcelona, Spain.

LDC. (2005). Linguistic Data Annotation Specification: Assessment of Fluency and Adequacy in Translations. *Linguistic Data Consortium, Revision* 1.5.

Li, H.Y., Graesser, A.C. and Cai, Z.Q. (2014). Comparison of Google Translation with Human Translation. In *Proceedings of the Twenty-seventh International Florida Artificial Intelligence Research Society Conference*, pp. 190-195

Miller, G.A. and Beebe-Center, J.G. (1958). Some Psychological Methods for Evaluating the Quality of Translations. *Mechanical Translation*, 3: 73–80.

Norman, J. (1988). *Chinese*. Cambridge: Cambridge University Press.

Nielsen, M. (2011). *Reinventing Discovery: The New Era of Networked Science.* Princeton, NJ: Princeton University Press.

Nirenburg, S. and Wilks, Y. (2000). Machine Translation. *Advances in Computer*, 52:160-189.

Nyberg, E.H. and Mitamura, T. (1992). The KANT System: Fast, Accurate, High-Quality Translation in Practical Domains. In *Proceedings of the International Conference on Computer Linguistics*. Nantes, France.

Och, F.J. and Ney, H. (2004). The Alignment Template Approach to Statistical Machine Translation. *Computational Linguistics*, 30(4):417-449.

Packard, J.L. (2000). *The Morphology of Chinese: A linguistic and cognitive approach*. Cambridge: Cambridge University Press, pp xvi, 335.

Papineni, K., Salim R., Todd W., and Wei, J.Z. (2002). BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 311–318.

Peng, L. (2013). A Survey of Machine Translation Methods. *TELKOMNIKA Indonesian Journal of Electrical Engineering*, 11(12): 7125-7130

Popovic, M. and Ney, H. (2004). Towards the Use of Word Stems and Suffixes for Statistical Machine Translation. In *Proceedings of International Conference on Language Resources and Evaluation*, Lisbon, Portugal.

Popovic, M. and Ney, H. (2006). POS-based Word Reorderings for Statistical Machine Translation. In *Proceedings of International Conference on Language Resources and Evaluation*, Genoa, Italy.

Popovic, M. and Ney, H. (2009). Syntax-oriented Evaluation Measures for Machine Translation Output. In *Proceedings of the Fourth Workshop on Statistical Machine Translation,* pp. 29–32. Stroudsburg, PA: Association for Computational Linguistics.

Popovic M. and Burchardt A. (2011). From Human to Automatic Classification for Machine Translation Output. In *Proceedings of European Association for Machine Translation 2011,* pp. 265-272.

Quirk, C., Menezes, A., and Cherry, C. (2005). Dependency Tree Translation: Syntactically Informed Phrasal SMT. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, pp. 271–279.

Resnik, P. and Smith, N.A. (2003). The Web as a Parallel Corpus. *Computational Linguistics*, 29(3): 349-380.

Senellart, J., Dienes, P. and Váradi, T. (2001). New Generation SYSTRAN Translation System. In *MT Summit 8.*

Schwartz, B. (2007). Google Translate Drops Systran for Home-Brewed Translation. Retrieved October 3, 2016, from http://searchengineland.com/google-translate-drops-systran-for-home-brewed-translation-12502.

Sharman, R.A., Jelinek, F. and Mercer, R. (1990). Generating a Grammar for Statistical Training. In *Proceedings of Work on Speech and Natural Language*, pp. 267-274.

Shankland, S. (2013). Google Translate Now Serves 200 Million People Daily. Retrieved September 15, 2016, from https://www.cnet.com/news/google-translate-now-serves-200-million-people-daily/.

Slocum, J. (1985). A Survey of Machine Translation, *Computational Linguistics*, 11(1).

Smith, C.S. (1997). *The Parameter of Aspect. Studies in Linguistics and Philosophy*. Springer.

The Economist. (2017). Finding a Voice. *Technology Quarterly*, Retrieved January 05, 2017, from http://www.economist.com/technology-quarterly/2017-05-01/language.

Uchida, H. (1996). Universal Networking Language – An Electronic Language for Communication, Understanding and Collaboration, UNU/IAS/UNL Centre, Tokyo, Japan.

Van, R.A., Snyman, C. and Lotz, S. (2012). Applying Google Translate in a Higher Education Environment: Translation Products Assessed. *Southern African Linguistics and Applied Language Studies*, 3(4), 511-524.

Vauquois, B. (1968). A Survey of Formal Grammars and Algorithms for Recognition and Transformation in Machine Translation. In *IFIP Congress*, pp. 254–260, Edinburgh, UK

Vilar, D., Xu, J., D'Haro, L.F. and Ney, H. (2006). Error Analysis of Machine Translation Output. In *Proceedings of the 5th Internation Conference on Language Resources and Evaluation (LREC'06)*, pp. 697–702, Genoa, Italy.

White, J.S., O'Connell, T.A. and Carlson, L.M. (1993). Evaluation of Machine Translation. In *Proceedings of the Workshop on Human Language Technology*, Stroudsburg, PA, USA.

White, J.S., O'Connell, T.A. and O'Mara, F. (1994). The ARPA MT Evaluation Methodologies: Evolution, Lessons, and Future Approaches. In *Proceedings of the Association for Machine Translation in the Americas Conference*, Columbia, Maryland, USA.

Wilss, W. (1982). *The Science of Translation. Problems and Methods.* Tübingen: G. Narr.

Wu, Y. H., Schuster, M., Chen, Z.F., et al. (2016). Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. *Technical Report*, Google.

Xiong, D. and Zhang M. (2015). *Linguistically Motivated Statistical Machine Translation*. Springer-Verlag.

Yamada, K. and Knight, K. (2001). A Syntax-based Statistical Translation Model. In *Proceedings of Association for Computational Linguistics-01*, Toulouse, France.

APPENDICES

Appendix A: An Inquiry Letter to Google

20.12.2015

To Google,

My name is Ding Chen and I am a student currently studying linguistics at the Norwegian University of Science and Technology (NTNU), in Trondheim, Norway.

I am currently working on my master's thesis for a Master's degree in English Linguistics. To that end, I am asking for your assistance in answering some questions.

My master's thesis will explore the subject of machine translation. The main objective is to compare the translation outputs of Google Translate and Bing Translator using English-Chinese-Norwegian language pairs in order to evaluate the quality of the translation outputs.

This evaluation will be performed by using a newly proposed evaluation method (i.e., a linguistic evaluation which was first proposed by Marta. R et al. 2010) rather than the commonly used automatic evaluation methods: BLEU, NIST, etc., or the human perceptual evaluation method. By using this new linguistic evaluation method, errors will be detected and then sorted into different linguistic categories, such as orthographic, lexical, morphological, semantic, and syntactic errors.

From my research, I found that Google Translate and Bing Translator are basically using statistical methods. However, the Bing Translator claims to have united the power of statistical methods with linguistic information, as indicated on their official website. Microsoft calls this unified method as 'a linguistically informed statistical machine translation service.'

To better understand the translation methodologies, I am asking for your assistance with the following questions. Any thoughts or ideas on any or all of the questions would be greatly appreciated!

1. Is Google Translate currently using a purely statistical method without any linguistic information/knowledge to perform translation works?

2. How does Google Translate measure its translation quality? Microsoft uses mainly the BLEU (bilingual evaluation understudy) standard and also their own benchmarks (including automatic and human evaluations). Does Google have similar benchmarks?

3. How satisfied is Google with its own translation quality? Are they aiming for a ‘high-quality/linguistic-error free’ online translation service in the future? And how much do the developers care about the ‘Linguistic errors’ generated by Google Translate? Would developers consider a linguistic evaluation as meaningful or helpful in their further research and development?

4. Does Google Translate have any plan of improving its translation quality in the future by developing a more hybrid engine which combines the statistical method with linguistic rules/information?

5. How do you see the differences between Google Translate and Bing Translator? Have you ever compared the translation output of these two?

6. Have you used different training methods to train the engine for translating different languages, like Asian languages, and European/ Germanic languages? Do you think it’s smarter or more accurate in translating European languages (for example, English-Norwegian; Norwegian-English), than in translating Asian-European pairs (for example, English-Chinese; Chinese- English; Norwegian-Chinese; Chinese-Norwegian)? If there are indeed differences between these different language pairs, how do you account for it?

Thank you so much in advance if you can help me with any of the above questions.
Your replies will be kept in confidence and will be used only in the development of my thesis.

Best regards,

Ding Chen
Master student, Department of Language and Literature, Norwegian University of Science and Technology, 7491 Trondheim, Norway

Appendix B: 50 Chinese Source Sentences from the News Texts

1	据 BBC 中文网 3 月 16 日报道,联合国可持续发展解决方案网络(SDSN)与哥伦比亚大学地球研究所周三(16 日)共同发布的报告指出，丹麦是世界上最幸福的国家，而布隆迪是世界上最不幸福的国家。
2	调查显示，今年跻身全球 10 大最幸福国家依次为丹麦、瑞士、冰岛、挪威、芬兰、加拿大、荷兰、新西兰、澳大利亚和瑞典。
3	丹麦去年排名第 3 位，落后于瑞士和冰岛。
4	排名垫底的 10 个国家分别为马达加斯加、坦桑尼亚、利比里亚、几内亚、卢旺达、贝宁、阿富汗、多哥、叙利亚和布隆迪。
5	在这 10 个不幸福国家当中，撒哈拉沙漠以南非洲国家就占了 8 个。
6	在西方强国方面，美国排名第 13，英国排名第 23，法国排名第 32，意大利排名第 50。
7	美国这次名列第 13 位，有所上升。
8	中国在这次评比中位列第 83 位，在菲律宾之后。
9	台湾排名第 35名，提前了 3 位。
10	此项调查以盖洛普世界民调为基础，分析人均 GDP、健康、预期寿命等 6 个因素。
11	今年的评比，首次使用幸福感差距来代替收入差距。
12	除经济因素外，还包括了自然环境等可持续发展的因素。
13	据报道，这名智商超过天才科学家爱因斯坦和物理学家霍金的男童，上星期接获门萨信件，称他是全球 1%最聪明的人。
14	但是当大家都以为他会成为科学家时，他的志愿却是想当一名职业足球员。
15	据悉，哈默尔目前是英国足球队高云地利 12 岁以下球员。
16	而在机器人将取代人类的争议中，大面积失业是主要担忧之一，然而 20 世纪以来的科技发展的确让很多职业消失，但这只是部分事实。
17	2013 年，牛津大学的经济学家们对美国就业市场上现有的 702 种职业进行了量化评估，结果显示在未来 20 年，有 46%的职业可能被机器替代 。
18	事实上，自 19 世纪工业革命以来，就不断有争议指出技术发展造成了更多人失业，最终会让机器取代人类。
19	让人类从枯燥重复的纯体力劳作中解放出来更是科技的一大贡献。
20	《泰晤士报高等教育专刊》(Times Higher Education，下文简称 THE)公布 2016 年全球大学声誉排行榜（World Reputation Rankings ）。
21	榜单显示，哈佛大学连续六年蝉联第一。
22	麻省理工、斯坦福大学比去年上升两位。
23	此次，中国内地高校表现不俗 。
24	其中，清华大学在泰晤士报全球大学声誉榜上创造中国高校历年最高记录，位居 18。
25	其它三校均首次入围，分别是复旦大学、上海交通大学和浙江大学。
26	亚洲高校在庞大资金的基础上，加强了他们的科研，在国际期刊上更多地发表论文。

27	伦敦大学国王学院政策研究所教授 Paul Blackmore 也表示，高等教育和科研的实力的平衡已经开始改变 。
28	我们也很乐意看到越来越多的中国高校出现在榜单之中。
29	2016 年，他们共收到来自 133 个国家的 10323 份有效回复。
30	Phil Baty 表示，虽然声誉排行榜不可避免地带有主观色彩，但是这个排行榜很有存在的必要。
31	谷歌公司和菲亚特克莱斯勒汽车公司 3 日宣布，两家公司将合作生产 100 辆自动驾驶汽车，但这些车辆仅供谷歌测试无人驾驶技术。
32	这是谷歌第一次和传统汽车制造商合作研发自动驾驶汽车。
33	谷歌表示目前没有将其无人驾驶汽车技术授权或转让给任何汽车公司。
34	谷歌目前正在美国 4 个城市的道路上测试无人驾驶技术。
35	新一批自动驾驶车将首先在谷歌自己的测试场地内测试，然后投入道路测试。
36	去年，谷歌开始在得克萨斯州奥斯汀市测试无人驾驶车。
37	今年测试城市增加到 4 个。
38	该公司聘请了 4 名前美国国家高速公路交通安全管理局高官，他们帮助谷歌游说美国政府接受谷歌的自动驾驶汽车技术。
39	美国国家高速公路交通安全管理局会在 7 月份之前发布无人驾驶汽车的准则。
40	美国交通部长安东尼·福克斯对路透社表示：“这项技术已经来临，不管我们有没有准备好，它已经来临。”
41	一门可以改变生活的课程。
42	如今这已成为不少哈佛学生的共识。
43	《纽约时报》不久前报道，在哈佛，中国哲学仅次于计算机和经济学，排在最受欢迎课程的前三名。
44	中国哲学为何可以在美国走红？
45	一些中国哲学著作被美国人奉为“生活哲学”，因此成为流行读物。
46	早在 2006 年，普鸣开始向本科生传授中国哲学概论时，该课程就受到了大量哈佛学子的欢迎。
47	美国媒体的相关报道显示，在过去相当长的时间内，哈佛学生选择人文科学的人数急剧下降。
48	并且这样的趋势在美国其他文科学院中也同样存在。
49	这样的背景令中国哲学课程更加走红。
50	《大西洋月刊》的报道称，这些中国思想能够帮助那些十八九岁的年轻人思考如何成为一个好人，如何创造一个良好的社会。

Appendix C: Error Analysis and Comparison

NO.	SOURCE TEXT	TARGET TEXT	DISSECTED COMPONENTS	ERROR TYPE	ORTHOGRAPHICAL	MORPHOLOGICAL	SEMANTIC	LEXICAL	SYNTACTIC
1.	据BBC中文网3月16日报道,联合国可持续发展解决方案网络(SDSN)与哥伦比亚大学地球研究所周三(16日)共同发布的报告指出，丹麦是世界上最幸福的国家，而布隆迪是世界上最不幸福的国家。	**According to the BBC Chinese network March 16th reported that the UN Sustainable Development Solutions Network (SDSN) and the Earth Institute at Columbia University on Wednesday (16th) jointly issued the report pointed out that Denmark is the happiest countries, Burundi is the world's most unhappy country.**	*** According to the BBC Chinese network March 16th reported that** - According to the BBC News (Chinese) March 16th reported that	*proper noun*				O	
			- -According to the BBC Chinese News (Chinese) on March 16th reported that	*lack of preposition*					O
			- - - ~~According to~~ the BBC Chinese News (Chinese) on March 16th reported that	*redundant preposition*					O
			***… jointly issued the report pointed out that** - the report that SDSN and … jointly issued	*syntax: relative clause*					O
			*** Denmark is the happiest countries** - Denmark is the happiest country	*wrong noun inflection*		O			O
			- - Denmark is the happiest country in the world	*omission*				O	
			* Burundi is the world's most unhappy county. - while Burundi is the world's most unhappy country.	*omission*				O	
		BBC Chinese website on March 16th, it was reported, solutions for the sustainable development of the United Nations network (SDSN) and the Earth Institute at Colombia University on Wednesday (16th) jointly issued the report points out that Denmark is the happiest country in the world, Burundi is the most happiest country in the world.	*** BBC Chinese website on March 16th, it was reported** - by BBC Chinese website on March 16, It was reported	*omission*				X	
			- - by BBC News (Chinese) on March 16th, It was reported	*proper noun*				X	
			- - - by the BBC News (Chinese) on March 16th, It was reported	*lack of definite article*					X
			- - - - It was reported by the BBC News (Chinese) on March 16th	*word order*					X
			- - - - - It was reported by the BBC News (Chinese) on March 16th that	*lack of complementizer*					X
			*** solutions for the sustainable development of the United Nations network (SDSN)** - the United Nations **S**ustainable **D**evelopment **S**olutions **N**etwork (SDSN)	*proper noun*				X	
			*** …jointly issued the report points out that** - the report that SDSN and .. jointly issued	*syntax: relative clause*					X
			*** Burundi is the most happiest country in the world.** - Burundi is the ~~most~~ happiest country in the world.	*wrong adjectival inflection*		X			
			- - Burundi is the unhappiest country in the world.	*antonym*			X		
			- - - While Burundi is the unhappiest country in the world.	*omission*				X	

NO.	SOURCE TEXT	TARGET TEXT	DISSECTED COMPONENTS	ERROR TYPE	ORTHOGRAPHICAL	MORPHOLOGICAL	SEMANTIC	LEXICAL	SYNTACTIC
2	调查显示，今年跻身全球10大最幸福国家依次为丹麦、瑞士、冰岛、挪威、芬兰、加拿大、荷兰、新西兰、澳大利亚和瑞典。	**Surveys show that this year among the world's 10 happiest countries were Denmark, Switzerland, Iceland, Norway, Finland, Canada, the Netherlands, New Zealand, Australia and Sweden.**	*** this year among the world's 10 happiest countries were** - this year among the world's top 10 happiest countries were	*omission*				O	
			- - this year ~~among~~ the world's 10 happiest countries were	*redundant preposition*					O
		Investigations revealed that this year rank among the 10 happiest countries in the world followed by Denmark, and Switzerland, and Iceland, and Norway, and Finland, and Canada and the Netherlands, and New Zealand, and Australia and Sweden.	*** this year rank among the 10 happiest countries in the world followed by** - this year ~~rank~~ among the 10 happiest countries in the world followed by	*redundant verb*					X
			- - this year the top 10 happiest countries in the world followed by	*omission*				X	
			- - - this year the top 10 happiest countries in the world are	*wrongly translated words*				X	
			*** and Switzerland, and Iceland, and Norway, and Finland, and Canada and the Netherlands, and New Zealand, and Australia and Sweden.** - Switzerland, Iceland, Norway, Finland, Canada, the Netherlands, New Zealand, Australia and Sweden.	*addition*				X	
3.	丹麦去年排名第3位，落后于瑞士和冰岛。	**Denmark last year ranked No. 3, behind Switzerland and Iceland.**	√						
		Denmark last year ranked 3rd, behind Switzerland and Iceland.	√						
4.	排名垫底的10个国家分别为马达加斯加、坦桑尼亚、利比里亚、几内亚、卢旺达、贝宁、阿富汗、多哥、叙利亚和布隆迪。	**The top 10 countries are Madagascar, Tanzania, Liberia, Guinea, Rwanda, Benin, Afghanistan, Togo, Syria and Burundi.**	*** The top 10 countries are Madagascar, Tanzania, Liberia, Guinea, Rwanda, Benin, Afghanistan, Togo, Syria and Burundi.** - The bottom 10 countries are Madagascar, Tanzania, Liberia, Guinea, Rwanda, Benin, Afghanistan, Togo, Syria and Burundi.	*antonym*			O		
		The bottom 10 countries namely Madagascar, Tanzania, Liberia, Guinea, Rwanda, Benin, Afghanistan, Togo, Syria and Burundi.	*** namely Madagascar, Tanzania, Liberia, Guinea, Rwanda, Benin, Afghanistan, Togo, Syria and Burundi.** - are namely Madagascar, Tanzania, Liberia, Guinea, Rwanda, Benin, Afghanistan, Togo, Syria and Burundi.	*omission*				X	
5.	在这10个不幸福国家当中，撒哈拉沙漠以南非洲国家就占了8个。	**In this 10 unhappy countries, the sub-Saharan African countries accounted for eight.**	*** In this 10 unhappy countries,** - In these 10 unhappy countries,	*wrong noun inflection*		O			O
		In 10 of these unhappy country, sub-Saharan African countries accounted for 8.	*** In 10 of these unhappy country,** - In 10 of these unhappy countries,	*wrong noun inflection*		X			X

NO.	SOURCE TEXT	TARGET TEXT	DISSECTED COMPONENTS	ERROR TYPE	ORTHOGRAPHICAL	MORPHOLOGICAL	SEMANTIC	LEXICAL	SYNTACTIC
6.	在西方强国方面，美国排名第13，英国排名第23，法国排名第32，意大利排名第50。	**In terms of Western powers, the United States ranked No. 13, ranked No. 23 United Kingdom, France ranked No. 32, ranked No. 50 in Italy.**	*** ranked No. 23 United Kingdom,** - ranked No.23 the United Kingdom,	*lack of definite article*					O
			- United Kingdom ranked No.23,	*word order*					O
			*** ranked No. 50 in Italy** - ranked No.50 ~~In~~ Italy	*addition*				O	
			- - Italy ranked No.50	*word order*					O
		In terms of Western powers, United States ranked 13th, United Kingdom ranked 23rd, France ranked 32nd, Italy ranked 50th.	*** United States ranked 13th,** - The United States ranked 13th,	*lack of definite article*					X
			*** United Kingdom ranked 23rd,** - The United Kingdom ranked 23rd,	*lack of definite article*					X
7.	美国这次名列第13位，有所上升。	**The United States ranked 13th, has increased.**	*** has increased.** - has moved up/risen.	*polysemy*			O		
			- - has moved up/risen moderately.	*omission*				O	
			- - - having moved up/risen moderately.	*wrong verb form-finite/non-finite*					O
		United States ranked 13th, has increased.	*** United States ranked 13th,** - The United States ranked 13th,	*lack of definite article*					X
			*** has increased.** - has moved up/risen.	*polysemy*			X		
			- - has moved up/risen moderately.	*omission*				X	
			- - - having moved up/risen moderately.	*wrong verb form-finite/non-finite*					X
8.	中国在这次评比中位列第83位，在菲律宾之后。	**China ranked No. 83 in the rankings, after the Philippines.**	√						
		In this competition, China ranked 83rd, after the Philippines.	*** In this competition,** - In this ranking ,	*disambiguation error*			X		
9.	台湾排名第35名，提前了3位。	**Taiwan ranked No.35, ahead of the three.**	*** ahead of the three** - climbed the three	*polysemy*			O		
			- - climbed ~~the~~ three	*addition*				O	
			- - - climbed three places	*omission*				O	
			- - - - which climbed three places	*lack of pronoun*					O
		Taiwan ranked No.35, ahead of 3 bits.	*** ahead of 3 bits** - climbed 3 bits	*polysemy*			X		
			- - climbed 3 places	*polysemy*			X		
			- - - which climbed 3 places	*lack of pronoun*					X

NO.	SOURCE TEXT	TARGET TEXT	DISSECTED COMPONENTS	ERROR TYPE	ORTHOGRAPHICAL	MORPHOLOGICAL	SEMANTIC	LEXICAL	SYNTACTIC
10.	此项调查以盖洛普世界民调为基础，分析人均GDP、健康、预期寿命等6个因素。	**The survey with Gallup world poll, based on analysis of per capita GDP, health, life expectancy and other factors 6.**	*** Gallup world poll** - Gallup World Poll	*capitalisation*	O				
			* and other - and so on	*wrongly translated words*				O	
			* per capital GDP - GDP per capital	*word order*					O
			* factors 6 - 6 factors	*word order*					O
			* The survey with Gallup World Poll, based on … - The survey based on with Gallup World Poll,	*word order*					O
			* with - ~~with~~	*redundant preposition*					O
			* analysis of … 6 factors - analysis of 6 factors GDP per capita, health, life expectancy and so on	*word order*					O
			* analysis of 6 factors GDP per capita, … - analysis of 6 factors including GDP per capita, …	*lack of preposition*					O
			* analysis of 6 factors including GDP per capita, … - analysed 6 factors including GDP….	*word class*					O
		The survey is based on Gallup World poll, analysis of per capita GDP, health, life expectancy and other factors 6.	*** Gallup World poll** - Gallup World Poll	*capitalisation*	X				
			* and other - and so on	*wrongly translated words*				X	
			* per capital GDP - GDP per capital	*word order*					X
			* factors 6 - 6 factors	*word order*					X
			* analysis of … 6 factors - analysis of 6 factors GDP per capita, health, life expectancy and so on	*word order*					X
			* analysis of 6 factors GDP per capita, … - analysis of 6 factors including GDP per capita, …	*lack of preposition*					X
			* analysis of 6 factors including GDP per capita, … - analysing 6 factors including GDP….	*wrong verb form-non-finite*					X

NO.	SOURCE TEXT	TARGET TEXT	DISSECTED COMPONENTS	ERROR TYPE	ORTHOGRAPHICAL	MORPHOLOGICAL	SEMANTIC	LEXICAL	SYNTACTIC
11	今年的评比, 首次使用幸福感差距来代替收入差距。	**This year's competitions, the first use of happiness gap instead of the income gap.**	*** This year's competitions, the first use of happiness gap instead of the income gap.** - This year's competitions, make the first use of happiness gap instead of the income gap.	*lack of syntactic elements*					O
		This year's rankings, for the first time using the happiness gap instead of the income gap.	*** This year's rankings, for the first time using the happiness gap instead of the income gap.** - This year's rankings, using the happiness gap instead of income gap for the first time.	*word order*					X
			- - This year's rankings, use the happiness gap instead of income gap for the first time.	*wrong verb form-finite/non-finite*					X
12	除经济因素外，还包括了自然环境等可持续发展的因素。	**In addition to economic factors, but also includes elements of sustainable development of natural environment.**	*** but also includes elements of sustainable development of natural environment.** - but also includes factors of sustainable development of natural environment.	*polysemy*			O		
			- - ~~but~~ also includes factors of sustainable development of natural environment.	addition				O	
			- - - also includes factors of sustainable development like/such as natural environment.	*wrongly translated words*				O	
			- - - - it also includes factors of sustainable development like natural environment.	*lack of pronoun*					O
		In addition to economic factors, sustainable development, including factors of environment.	*** sustainable development** **- also sustainable development**	*omission*				X	
			- - also includes sustainable development, including …	*omission*				X	
			- - - natural environment	*omission*				X	
			- - - - also includes factors of sustainable development…	*word order*					X
			- - - - - it also includes ….	*lack of pronoun*					X

NO.	SOURCE TEXT	TARGET TEXT	DISSECTED COMPONENTS	ERROR TYPE	ORTHOGRAPHICAL	MORPHOLOGICAL	SEMANTIC	LEXICAL	SYNTACTIC
13	据报道，这名智商超过科学家爱因斯坦和物理学家霍金的男童，上星期接获门萨信件，称他是全球1%最聪明的人。	**According to reports, this IQ than Einstein and physicist Hawking boy Mensah received a letter last week, said he was the most intelligent 1% of the world's people.**	*** Einstein** - scientist Einstein	*omission*				O	
			* than scientist Einstein - higher than scientist Einstein	*omission*				O	
			* this IQ higher than scientist Einstein and physicist Hawking boy - this boy IQ higher than scientist …	*word order*					O
			- - this boy whose IQ is higher than …	*syntax: relative clause*					O
			* Mensah received a letter last week - received a letter Mensah last week	*word order*					O
			- - received a letter from Mensah last week	*lack of preposition*					O
			*** said he was the most intelligent 1% of the world's people.** - saying he was the most intelligent 1% of the world's people.	*wrong verb form - non-finite*					O
			- - saying he was 1% of the world's most intelligent people.	*word order*					O
		It was reported that this IQ more than boys, scientist Albert Einstein and hawking, Mensa received letters last week, says he is 1% the smartest people in the world.	*** this IQ more than boys, scientist Albert Einstein and hawking,** -this IQ more than boys scientist Albert Einstein and hawking,	*punctuation*	X				
			- - this …. Hawking	*capitalisation*	X				
			- - - this IQ higher than boys…	*polysemy*			X		
			- - - - this boy IQ higher than ….	*wrong noun inflection*		X			X
			- - - - - this boy whose IQ is higher than …	*syntax: relative clause*					X
			*** received Mensa letters last week,** - received letters Mensa last week	*word order*					X
			- - received letters from Mensa last week	*lack of preposition*					X
			*** says he is 1% the smartest people in the world.** - saying he is 1% the smartest people in the world.	*non-finite*					X

NO.	SOURCE TEXT	TARGET TEXT	DISSECTED COMPONENTS	ERROR TYPE	ORTHOGRAPHICAL	MORPHOLOGICAL	SEMANTIC	LEXICAL	SYNTACTIC
14.	但是当大家都以为他会成为科学家时，他的志愿却是想当一名职业足球员。	**But when everyone thought he would become a scientist, he wanted to volunteer is a professional football player.**	*** he wanted to volunteer** - he wanted to wish	*disambiguation error*			O		
			- - he ~~wanted to~~ wish	*redundant verb*					O
			- - - his wish	*disambiguation error*			O		
			* is a professional football player - is to become a professional football player.	*omission*				O	
		But when everyone thought he was going to be a scientist, and his volunteers are wanted to be a professional footballer.	*** and his volunteers are** - and his volunteer are	*wrong noun inflection*		X			X
			- and his volunteer is	*wrong verb inflection*		X			X
			- - and his wish is	*disambiguation error*			X		
			- - - ~~and~~ his wish is	*addition*				X	
			- - - - his wish is ~~wanted~~ to be a …	*redundant verb*					X
15.	据悉，哈默尔目前是英国足球队高云地利12岁以下球员。	**It is reported that, Hamel is currently 12 years old British soccer team Gaoyundeli players.**	*** players** - a player	*wrong noun inflection*		O			O
			*** Gaoyundeli** - Coventry City	*proper noun*				O	
			*** 12 years old** - under 12 years old	*omission*				O	
			* British soccer team - in British soccer team	*lack of preposition*					O
			- - in the British soccer team	*lack of definite article*					O
			* Hamel is currently under 12 years old in the British soccer team Coventry City a player. - Hamel is currently a player under 12 years old in the British soccer team Coventry City.	*word order*					O
		It is reported that Hamel is currently the United Kingdom football team Coventry City player under age 12.	*** the United Kingdom football team** - in the United Kingdom football team	*lack of preposition*					X
			* player - a player	*lack of indefinite article*					X
			* Hamel is currently in the United Kingdom football team Coventry City a player under age 12. - Hamel is currently a payer under age 12 in the United Kingdom football team Coventry City.	*word order*					X
16.	而在机器人将取代人类的争议中，大面积失业是主要担忧之一，然而20世纪以来的科技发展的确让很多职业消失，但这只是部分事实。	**The robots will replace humans in dispute, a large area of unemployment is one of the main concern, however, technological development since the 20th century did make many profession disappear, but this is only partially true.**	*** The robots will replace humans in dispute,** - In dispute that the robots will replace humans,	*syntax: relative clause*					O
			- - In the dispute that the robots will replace humans,	*lack of definite article*					O
			*** is one of the main concern** - one of the main concerns	*wrong noun inflection*		O			O
			*** many profession** - many professions	*wrong noun inflection*		O			O
		Robots will replace humans in the controversy, widespread unemployment is one of the main concerns, but since the 20th century, technological developments make many disappear, but this is only part of the truth.	*** Robots will replace humans in the controversy,** - In the controversy that robots will replace humans,	*syntax: relative clause*					X
			* make many disappear, - make many professions disappear,	*omission*				X	

NO.	SOURCE TEXT	TARGET TEXT	DISSECTED COMPONENTS	ERROR TYPE	ORTHOGRAPHICAL	MORPHOLOGICAL	SEMANTIC	LEXICAL	SYNTACTIC
17	2013年，牛津大学的经济学家们对美国就业市场上现有的702种职业进行了量化评估，结果显示在未来20年，有46%的职业可能被机器替代。	**2013, Oxford University economists for the US job market, the existing 702 kinds of occupations quantitative assessment results are shown in the next 20 years, 46% of the profession are replaced by machines.**	*** 2013,** - In 2013,	*lack of preposition*					O
			*** Oxford University economists ....results are shown....** - Oxford University economists for the US job market the existing 702 kinds of occupations quantitative evaluation, results are shown …	*punctuation*	O				
			- - Oxford University economists for the US job market the existing 702 kinds of occupations conducted quantitative assessment,	*omission*				O	
			- - - economists at Oxford University ….	*syntax: NP*					O
			- - - - economists at the Oxford University …	*lack of definite article*					O
			- - - - - … in the US job market ..	*lack of preposition*					O
			- - - - - - economists at the Oxford University conducted a quantitative evaluation for …	*word order*					O
			- - - - - - - … for the existing 702 kinds of occupations in the US job market,	*word order*					O
			*** results are shown …** - results showed	*wrong verb form-voice*					O
			* are replaced by machines - could be replaced by machines	*wrong verb form*					O
		In 2013, economists at the University of Oxford US 702 jobs were available in the job market assessment, results show that over the next 20 years, 46% the career may be replaced by machines.	*** economists at the University of Oxford US 702 jobs were available in the job market assessment,** - ~~were~~ available	*addition*				X	
			- - for 702 jobs ..	*omission*				X	
			- - - conducted quantitative assessment	*omission*					
			- - - - in the US job market	*word order*					
			- - - - - … economists … conducted quantitative assessment for 702 jobs ….	*word order*					
			* 46% the career - 46% the careers	*wrong noun inflection*		X			X
			- - 46% of the careers	*lack of preposition*					X

NO.	SOURCE TEXT	TARGET TEXT	DISSECTED COMPONENTS	ERROR TYPE	ORTHOGRAPHICAL	MORPHOLOGICAL	SEMANTIC	LEXICAL	SYNTACTIC
18	事实上，自19世纪工业革命以来，就不断有争议指出技术发展造成了更多人失业，最终会让机器取代人类。	In fact, since the 19th century Industrial Revolution, we have continued to point out the controversial technology creates more unemployment, the machines will eventually replace humans.	* since the 19th century Industrial Revolution, - since the Industrial Revolution in the 19th century,	*syntax: NP*					O
			* we have continued to point out the controversial technology creates more unemployment, * we	*addition*				O	
			* controversial - controversies	*word class*					O
			- - we controversies have continued to point out the technology creates more unemployment,	*word order*					O
			- - - controversies have continued to point out that the technology creates more unemployment,	*lack of syntactic elements*					O
			- - - - controversies have continued to point out that the technology development creates more unemployment,	*omission*				O	
			- - - - - controversies have continued to point out that the technology development created more unemployment,	*wrong verb inflection*		O			O
		In fact, ever since the industrial revolution in the 19th century, there have been controversies pointed out that technological development has created more unemployment, and will eventually supersede humanity.	* there have been controversies pointed out that technological development has created more unemployment, - there have been controversies pointing out that technological development has created more unemployment,	*wrong verb form-voice*					X
			* and will eventually supersede humanity. - and machines will eventually supersede humanity.	*omission*				X	

NO.	SOURCE TEXT	TARGET TEXT	DISSECTED COMPONENTS	ERROR TYPE	ORTHOGRAPHICAL	MORPHOLOGICAL	SEMANTIC	LEXICAL	SYNTACTIC
19	让人类从枯燥重复的纯体力劳作中解放出来更是科技的一大贡献。	**Let mankind from boring repetitive physical labor pure liberation is a major contribution to science and technology.**	*** Let mankind from boring repetitive physical labor pure liberation is a major contribution to science and technology.** - Let mankind from boring repetitive pure physical labor liberation is a major contribution to science and technology.	*word order*					O
			- - Let mankind from boring repetitive pure physical labor liberating is a major contribution to science and technology.	*word class*					O
			- - - Let liberating mankind from boring repetitive pure physical labor is a major contribution to science and technology.	*word order*					O
			- - - - ~~Let~~ liberating mankind from boring repetitive pure physical labor is a major contribution to science and technology.	*redundant verb*					O
			- - - - - Liberating mankind from boring … contribution of science and technology.	*syntax: NP*					O
		Humans freed from boring repetition of pure physical labor was a great contribution of science and technology.	*** Humans freed from boring repetition of pure physical labor was a great contribution of science and technology.** - freed humans from boring repetition of pure physical labor was a great contribution of science and technology.	*word order*					X
			- - freeing humans from boring repetition of pure physical labor was a great contribution of science and technology.	*wrong verb form-voice*					X
20.	《泰晤士报高等教育专刊》(Times Higher Education，下文简称THE)公布2016年全球大学声誉排行榜（World Reputation Rankings）。	**"The Times Higher Education Supplement" (Times Higher Education, hereinafter referred to as THE) announced the 2016 Global University reputation rankings (World Reputation Rankings).**	- - The Times Higher Education	*proper noun*				O	
			* 2016 Global University reputation rankings - World Reputation Rankings 2016	*proper noun*				O	
		Special issue of the times higher education (the Times Higher Education, hereinafter THE) world in 2016 College reputation rankings (World Reputation Rankings).	*** Special issue of the times higher education (the Times Higher Education, hereinafter THE)** - The Times Higher Education (the Times Higher Education, hereinafter THE)	*proper noun*				X	
			*** world in 2016 College reputation rankings** - World Reputation Rankings 2016	*proper noun*				X	
			- - announced world college reputation rankings in 2016	*omission*				X	

NO.	SOURCE TEXT	TARGET TEXT	DISSECTED COMPONENTS	ERROR TYPE	ORTHOGRAPHICAL	MORPHOLOGICAL	SEMANTIC	LEXICAL	SYNTACTIC
21	榜单显示，哈佛大学连续六年蝉联第一。	List shows, Harvard University, for six consecutive years won the first place.	* List shows - The list shows	*lack of definite article*					O
			* for six consecutive years won the first place. - for six consecutive years won ~~the~~ first place.	*addition*				O	
		List, for six consecutive years ranked first at Harvard University.	* List, - List shows,	*omission*				X	
			- - The list shows	*lack of definite article*					X
			* for six consecutive years ranked first at Harvard University. - at Harvard University ranked first for six consecutive years.	*word order*					X
			- - ~~at~~ Harvard University ranked first for six consecutive years.	*addition*				X	
22.	麻省理工、斯坦福大学比去年上升两位。	MIT, Stanford University last year, which rose two points.	* MIT, Stanford University last year, which rose two points. - compared to last year	*omission*				O	
			- - climded	*polysemy*			O		
			- - - two places	*polysemy*			O		
			- - - - ~~which~~	*addtion*				O	
			- - - MIT, Stanford University climbed two places compared to last year.	*word order*					O
		MIT, Stanford University, up two from last year.	* MIT, Stanford University, up two from last year. - climbed	*word class*					X
			- - two places	*omission*				X	
23.	此次，中国内地的高校表现不俗。	This time, the Chinese mainland universities performed well.	* the Chinese mainland universities performed well. - the universities in the Chinese mainland	*syntax: NP*					O
		This time, performed well in colleges and universities in mainland China.	* performed well in colleges and universities in mainland China. - ~~in~~ colleges and universities	*addition*				X	
			- - colleges and universities in mainland China performed well	*word order*					X

NO.	SOURCE TEXT	TARGET TEXT	DISSECTED COMPONENTS	ERROR TYPE	ORTHOGRAPHICAL	MORPHOLOGICAL	SEMANTIC	LEXICAL	SYNTACTIC
24.	其中，清华大学在泰晤士报全球大学声誉榜上创造中国高校历年最高记录，位居18。	**Among them, the Tsinghua University to create a global reputation in the Times top Chinese Universities over the years, ranked 18.**	* **to create** - created	*wrong verb form-finite/non-finite*					O
			* global reputation in the Times - the Times Global University Reputation Rankings	*proper noun*				O	
			* top - top record of Chinese Universities over the years	*omission*				O	
			* created a in the Times Global University Reputation Rankings top record of Chinese Universities over the years - created a top record of Chinese Universities over the years in the Times Global University Reputation Rankings	*word order*					O
			* ranked 18 - ranking 18	*wrong verb form-finite/non-finite*					O
		Among them, the Tsinghua University in the times of global University reputation creating Chinese colleges and universities the highest on the list, 18.	* **the Tsinghua University in the times of global University reputation creating Chinese colleges and universities the highest on the list ,** - the Tsinghua University in the Times Global University Reputation Rankings creating Chinese colleges and universities the highest on the list,	*proper noun*				X	
			- - the Tsinghua University in the Times Global University Reputation Rankings created Chinese colleges and universities the highest on the list,	*wrong verb form-finite/non-finite*					X
			- - - the Tsinghua University in the Times Global University Reputation Rankings created the highest record of Chinese colleges and universities on the list,	*omission*				X	
			- - - - the Tsinghua University created the highest record of Chinese colleges and universities on the list in the Times Global University Reputation Rankings	*word order*					X
			* **18,** - ranking 18,	*omission*				X	

NO.	SOURCE TEXT	TARGET TEXT	DISSECTED COMPONENTS	ERROR TYPE	ORTHOGRAPHICAL	MORPHOLOGICAL	SEMANTIC	LEXICAL	SYNTACTIC
25.	其它三校均首次入围，分别是复旦大学、上海交通大学和浙江大学。	**The other three schools are the first finalists, namely Fudan University, Shanghai Jiaotong University and Zhejiang University.**	*** The other three schools are the first finalists,** - for the first time	*wrongly translated words*				O	
			- - shortlisted	*wrongly translated words*				O	
		Three other schools were nominated for the first time, Fudan University, Shanghai Jiaotong	*** Three other schools** **- other three schools**	*word order*					X
			- - The other three schools	*lack of definite article*					X
26.	亚洲高校在庞大资金的基础上，加强了他们的科研，在国际期刊上更多地发表论文。	**Universities in Asia on the basis of huge capital, strengthen their research in international journals published more papers.**	*** strengthen** - strengthened	*wrong verb inflection*		O			O
			* their research - their scientific research	*omission*				O	
			- - their scientific research,	*punctuation*	O				
			* in international journals published more papers. - published more papers in international journals.	*word order*					O
		Asia on the basis of colleges and universities in money, enhance their scientific research, more papers published in international journals.	*** Asia on the basis of colleges and universities in money,** - colleges and universities in Asia on the basis of money,	*word order*					X
			- - colleges and universities in Asia on the basis of huge amount of money,	*omission*				X	
			*** enhance their scientific research,** - enhanced their	*wrong verb inflection*		X			X
			* more papers published in international journals. -published more papers in international journals.	*word order*					X
27.	伦敦大学国王学院政策研究所教授Paul Blackmore也表示，高等教育和科研的实力的平衡已经开始改变。	**Professor King's College, University of London Institute for Policy Paul Blackmore also said that the balance of the strength of the higher education and scientific research has begun to change.**	*** Professor King's College, University of London Institute for Policy Paul Blackmore** - Policy Institute at King's College London	*proper noun*				O	
			- - the Policy Institute at King's College London	*lack of definite article*					O
			- - - Professor Paul Blackmore from the Policy Institute, King's College London	*syntax: NP*					O
		Policy Institute at King's College, University of London Professor Paul Blackmore said, higher education and scientific research of the balance of power has begun to change.	*** Policy Institute at King's College, University of London Professor Paul Blackmore said,** - King's College London	*proper noun*				X	
			- - the Policy Institute at King's College London	*lack of definite article*					X
			- - Professor Paul Blackmore from the Policy….	*syntax: NP*					X
			*** higher education and scientific research of the balance of power has begun to change.** - the balance of power of higher education and scientific research has begun to change.	*word order*					X

NO.	SOURCE TEXT	TARGET TEXT	DISSECTED COMPONENTS	ERROR TYPE	ORTHOGRAPHICAL	MORPHOLOGICAL	SEMANTIC	LEXICAL	SYNTACTIC
28.	我们很乐意看到越来越多的中国高校出现在榜单之中。	We are happy to see more and more Chinese universities appear on the list.	√						
		We would be happy to see more and more Chinese universities appear on the list.	√						
29.	2016年，他们共收到来自133个国家的10323份有效回复。	In 2016, they received a total of 10,323 valid responses from 133 countries.	√						
		In 2016, they received a total of 10,323 valid replies from 133 countries.	√						
30.	Phil Baty表示，虽然声誉排行榜不可避免地带有主观色彩，但是这个排行榜很有存在的必要。	Phil Baty said that although the reputation rankings inevitably have a subjective color, but the very existence of this list is necessary.	* have a subjective color - are subjective	*wrongly translated words*				O	
			* but the very existence of this list is necessary. - ~~but~~ …	*redundant conjunction*					O
		Phil Baty said, although the reputation rankings are inevitably subjective, but this list is needed.	* but this list is needed. - ~~but~~	*redundant conjunction*					X
31.	谷歌公司和菲亚特克莱斯勒汽车公司3日宣布，两家公司将合作生产100辆自动驾驶汽车，但这些车辆仅供谷歌测试无人驾驶技术。	Google Fiat and Chrysler announced Wednesday, the two companies will co-produce 100 autonomous vehicles, but these vehicles are for Google test driverless technology.	* Google Fiat and Chrysler announced Wednesday, - Google and Fiat Chrysler	*word order*					O
			- - announced on 3rd	*wrongly translated words*				O	
			* but these vehicles are for Google test driverless technology. - are only for Google	*omission*				O	
			- but these vehicles are only for Google to test driverless technology.	*wrong verb form - non-finite*					O
		Google, Fiat and Chrysler announced on 3rd, the two companies will cooperate in the production of 100 vehicles self-driving cars, these vehicles are intended for Google to test driverless technology.	* Google, Fiat and Chrysler announced on 3rd, - Google and Fiat Chrysler	*word order*					X
			* the two companies will cooperate in the production of 100 vehicles self-driving cars, - ~~vehicles~~	*addition*				X	
			* are intended for Google - are only intended for Google ..	*omission*				X	
32.	这是谷歌第一次和传统汽车制造商合作研发自动驾驶汽车。	This is the first time Google and the traditional car manufacturers to develop autonomous vehicles.	* Google and the traditional .. - Google with ….	*polysemy*			O		
			- - Google has worked with …	*omission*				O	
		This is the first time Google and traditional automotive manufacturers to develop autonomous vehicles.	* Google and the traditional .. - Google with ….	*polysemy*			X		
			- - Google has worked wth	*omission*				X	

NO.	SOURCE TEXT	TARGET TEXT	DISSECTED COMPONENTS	ERROR TYPE	ORTHOGRAPHICAL	MORPHOLOGICAL	SEMANTIC	LEXICAL	SYNTACTIC
33.	谷歌表示目前没有将其无人驾驶汽车技术授权或转让给任何汽车公司。	**Google currently does not represent its driverless car technology licensed or transferred to any car company.**	*** Google currently does not represent its driverless car technology licensed or transferred to any car company.** - Google currently does not says its driverless car technology licensed or transferred to any car company.	*polysemy*			O		
			- - Google says currently its driverless car technology does not licensed or transferred to any car company.	*word order*					O
			- - - Google says currently its driverless car technology is not licensed or transferred to any car company.	*wrong verb form-voice*					O
		Google said its driverless car technology is not authorized or to any car company.	*** Google said its driverless car technology is not authorized or to any car company.** - Google said currently its driverless car technology is not authorized or to any car company.	*omission*				X	
			- - Google said currently its driverless car technology is not authorized or transferred to any car company.	*omission*				X	
34.	谷歌目前正在美国4个城市的道路上测试无人驾驶技术。	**Google is currently testing driverless technology is on the road four US cities.**	*** Google is currently testing driverless technology is on the road four US cities.** - Google is currently testing driverless technology ~~is~~ on the road four US cities.	*addition*				O	
			- - on the road of four cities in US	*syntax: NP*					O
			- - - the US	*lack of definite article*					O
			- - - - the roads of	*wrong noun inflection*		O			O
		Google is currently United States 4 cities on the road to test unmanned technologies.	*** Google is currently United States 4 cities on the road to test unmanned technologies.** - Google is currently to test unmanned technologies on the road 4 United States cities.	*word order*					X
			- - Google is currently testing unmanned technologies on the road 4 United States cities.	*wrong verb form - non-finite*					X
			- - - Google is currently testing unmanned driving technologies on the road 4 United States cities.	*omission*				X	
			- - - - Google is currently testing unmanned driving technologies on the roads 4 United States cities.	*wrong noun inflection*		X			X
			- - - - - the roads of 4 cities in United States	*syntax: NP*					X
			- - - - - the United States	*lack of definite article*					X

NO.	SOURCE TEXT	TARGET TEXT	DISSECTED COMPONENTS	ERROR TYPE	ORTHOGRAPHICAL	MORPHOLOGICAL	SEMANTIC	LEXICAL	SYNTACTIC
35.	新一批自动驾驶车将首先在谷歌自己的测试场地内测试，然后投入道路测试。	The new car will be the first batch of automatic driving test within Google's own test site, then put the road test.	*** The new car will be the first batch of automatic driving test within Google's own test site, then put the road test.** - The new batch of automatic driving car will be test the first within Google's own test site, then put the road test.	*word order*					O
			- - The new batch of automatic driving cars will be test the first within Google's own test site, then put the road test.	*wrong noun inflection*		O			O
			- - - The new batch of automatic driving cars will be tested the first within Google's own test site, then put to the road test.	*lack of preposition*					O
		A new batch of self-driving car will be tested first in Google's own testing grounds, and then into road test.	*** A new batch of self-driving car will be tested first in Google's own testing grounds, and then into road test.** - A new batch of self-driving cars will be tested first in Google's own testing grounds, and then into road test.	*wrong noun inflection*		X			X
			- - A new batch of self-driving cars will be tested first in Google's own testing grounds, and then put into road test.	*omission*				X	
36.	去年，谷歌开始在得克萨斯州奥斯汀市测试无人驾驶车。	Last year, Google began Austin, Texas test driverless cars.	*** Austin, Texas** - in Austin Texas	*lack of preposition*					O
			- - testing	*wrong verb form - non-finite*					O
			- - - Google began testing driverless cars in Austin, Texas	*word order*					O
		Last year, Google started in Austin, Texas to test driverless cars.	*** Last year, Google started in Austin, Texas to test driverless cars.** - Google started to test driverless cars in Austin, Texas.	*word order*					X
37.	今年测试城市增加到4个。	Test this year to four cities.	*** Test this year to four cities.** - Test cites this year to four.	*word order*					O
			- - Test cities this year increased to four.	*omission*				O	
		Test cities this year to 4.	*** Test cities this year to 4.** - Test cities this year increased to 4.	*omission*				X	

NO.	SOURCE TEXT	TARGET TEXT	DISSECTED COMPONENTS	ERROR TYPE	ORTHOGRAPHICAL	MORPHOLOGICAL	SEMANTIC	LEXICAL	SYNTACTIC
38.	该公司聘请了4名前美国国家高速公路交通安全管理局高官，他们帮助谷歌游说美国政府接受谷歌的自动驾驶汽车技术。	**The company hired four former US National Highway Traffic Safety Administration officials who helped Google lobbying the US government to accept Google's autonomous vehicles technology.**	*** The company hired four former US National Highway Traffic Safety Administration officials who helped Google lobbying the US government to accept Google's autonomous vehicles technology.** - The company hired four former US National Highway Traffic Safety Administration officials who helped Google lobby the US government to accept Google's autonomous vehicles technology.	*wrong verb form- finite/non-finite*					o
		The companies employ 4 former United States National Highway Traffic Safety Administration official, they help Google's lobbying the United States Government to accept Google's self-driving car technology.	*** The companies employ 4 former United States National Highway Traffic Safety Administration official,** - This companies employ 4 former United States National Highway Traffic Safety Administration official,	*wrongly translated words*				x	
			- This company employ 4 former United States National Highway Traffic Safety Administration official,	*wrong noun inflection*		x			x
			- - This company employed 4 former United States National Highway Traffic Safety Administration official,	*wrong verb inflection*		x			x
			- - - This company employed 4 former United States National Highway Traffic Safety Administration officials,	*wrong noun inflection*		x			x
39.	美国国家高速公路交通安全管理局会在7月份之前发布无人驾驶汽车的准则。	**US National Highway Traffic Safety Administration guidelines issue before July driverless car.**	*** US National Highway Traffic Safety Administration guidelines issue before July driverless car.** - The ..	*lack of definite article*					o
			- - will issue	*omission*				o	
			- - - The US National Highway Traffic Safety Administration will issue driverless car guidelines before July.	*word order*					o
		The US National Highway Traffic Safety Administration will release guidelines for driverless cars before the July.	*** the July.** - ~~the~~ July	*addition*				x	

NO.	SOURCE TEXT	TARGET TEXT	DISSECTED COMPONENTS	ERROR TYPE	ORTHOGRAPHICAL	MORPHOLOGICAL	SEMANTIC	LEXICAL	SYNTACTIC
40.	美国交通部长安东尼·福克斯对路透社表示：“这项技术已经来临，不管我们有没有准备好，它已经来临。“	United States Transportation Secretary Anthony Fox told Reuters: "The technology has come, whether we have are not ready, it has arrived."	* US Transportation Secretary Anthony Fox - United Sates Secretary of Transportation	*proper noun*				O	
			- - the United Sates Secretary of Transportation	*lack of definite article*					O
			* whether we have are not ready, - ~~have~~	*addition*				O	
			- - whether we are ready not	*word order*					O
			- - - whether we are ready or not	*lack of conjunction*					O
		United States Transportation Secretary andongniđukesi told Reuters: "this technology is coming, whether or not we are ready, it is already here. “	* United States Transportation Secretary andongniđukesi - United Sates Secretary of Transportation	*proper noun*				X	
			- - the United Sates Secretary of Transportation	*lack of definite article*					X
			- - - Anthony Fox	*proper noun*				X	
			* this technology is coming - this technology has come	*wrong verb form-aspect*					X
41.	一门可以改变你生活的课程。	One can change your course of life.	* One can change your course of life. -One course that can change your life.	*syntax: relative clause*					O
		A course can change your life.	* A course can change your life. - A course that can change your life.	*syntax: relative clause*					X
42.	如今这已成为不少哈佛学生的共识。	Today, this has become the consensus of many Harvard students.	√						
		Now it has become the consensus of many Harvard students.	√						

NO.	SOURCE TEXT	TARGET TEXT	DISSECTED COMPONENTS	ERROR TYPE	ORTHOGRAPHICAL	MORPHOLOGICAL	SEMANTIC	LEXICAL	SYNTACTIC
43.	《纽约时报》不久前报道，在哈佛，中国哲学仅次于计算机和经济学，排在最受欢迎课程的前三名。	**"New York Times" recently reported that, at Harvard, Chinese philosophy and economics behind the computer, ranked in the top three of the most popular courses.**	*** "New York Times" recently reported that, at Harvard, Chinese philosophy and economics behind the computer, ranked in the top three of the most popular courses.** - The New York Times	*lack of definite article*					o
			- - The New York Times recently reported that, at Harvard, behind the computer and economics, Chinese philosophy, ranked in the top three of the most popular courses.	*word order*					o
		The New York Times recently reported that at Harvard, second only to computer science and Economics in Chinese philosophy, ranked in the top three of the most popular courses.	*** The New York Times recently reported that at Harvard, second only to computer science and Economics in Chinese philosophy, ranked in the top three of the most popular courses.** - The New York Times recently reported that at Harvard, second only to computer science and Economics, Chinese philosophy, ranked in the top three of the most popular courses.	*punctuation*	x				
			- -The New York Times recently reported that at Harvard, second only to computer science and Economics ~~in~~ Chinese philosophy, ranked in the top three of the most popular courses.	*addition*				x	
44.	中国哲学为何可以在美国走红？	**Why Chinese philosophy can be became popular in the United States?**	*** Why Chinese philosophy can be became popular in the United States?** - ~~became~~	*redundant verb*					o
			- Why can Chinese philosophy be popular in the United States?	*word order*					o
		Why Chinese philosophy became popular in the United States?	*** Why Chinese philosophy became popular in the United States?** - Why can Chinese philosophy	*omission*				x	
			- -become	*wrong verb form*					x

NO.	SOURCE TEXT	TARGET TEXT	DISSECTED COMPONENTS	ERROR TYPE	ORTHOGRAPHICAL	MORPHOLOGICAL	SEMANTIC	LEXICAL	SYNTACTIC
45.	一些中国哲学著作被美国人奉为"生活哲学“，因此成为流行读物。	**Some Chinese philosophical writings by the Americans regarded as a "philosophy of life" and therefore become popular books.**	*** Some Chinese philosophical writings by the Americans regarded as a "philosophy of life" and therefore become popular books.** - Some Chinese philosophical writings regarded as a "philosophy of life" by the Americans and therefore become popular books.	*word order*					O
			- -Some Chinese philosophical writings are regarded as a "philosophy of life" by the Americans and therefore become popular books.	*lack of verb*					O
		Some Chinese philosophical works by Americans as the "philosophy of life" became popular reading.	*** Some Chinese philosophical works by Americans as the "philosophy of life" became popular reading.** - Some Chinese philosophical works as the "philosophy of life" by Americans became popular reading.	*word order*					X
			- -Some Chinese philosophical works are regarded as the "philosophy of life" by Americans became popular reading.	*omission*				X	
			- - -Some Chinese philosophical works are regarded as the "philosophy of life" by Americans and therefore became popular reading.	*omission*				X	
			- - - -Some Chinese philosophical works are regarded as the "philosophy of life" by Americans and therefore became popular reading**s**.	*wrong noun inflection*		X			X

NO.	SOURCE TEXT	TARGET TEXT	DISSECTED COMPONENTS	ERROR TYPE	ORTHOGRAPHICAL	MORPHOLOGICAL	SEMANTIC	LEXICAL	SYNTACTIC
46.	早在2006年，普鸣开始向本科生传授中国哲学概论时，该课程就受到了大量哈佛学子的欢迎。	**As early as 2006, Pu Ming began to teach undergraduate Introduction to Chinese philosophy, the course will receive a large number of Harvard students welcome.**	*** As early as 2006, Pu Ming began to teach undergraduate Introduction to Chinese philosophy,** - As early as 2006, When Pu Ming began to teach undergraduate Introduction to Chinese philosophy,	*lack of conjunction*					O
			- -As early as 2006, When Pu Ming began to teach undergraduates Introduction to Chinese philosophy,	*wrong noun inflection*		O			O
			*** the course will receive a large number of Harvard students welcome.** - the course ~~will~~ receive a large number of Harvard students welcome.	*addition*				O	
			- - the course was welcomed by a large number of Harvard students.	*wrong verb form - passive voice*					O
		Back in 2006, when pu ming began to teach undergraduate introduction to Chinese philosophy, which were welcomed by a large number of Harvard students.	*** Back in 2006, when pu ming began to teach undergraduate introduction to Chinese philosophy, which were welcomed by a large number of Harvard students.** - Back in 2006, when Puming began to teach undergraduate introduction to Chinese philosophy, which were welcomed by a large number of Harvard students.	*capitalisation*	X				
			- -Back in 2006, when Puming began to teach undergraduate introduction to Chinese philosophy, which was welcomed by a large number of Harvard students.	*wrong verb inflection*		X			X
			- - -Back in 2006, ~~when~~ Puming began to teach undergraduate introduction to Chinese philosophy, which was welcomed by a large number of Harvard students.	*redundant complementizer*					X
			- - - -Back in 2006, Puming began to teach undergraduates introduction to Chinese philosophy, the course was welcomed by a large number of Harvard students.	*wrong noun inflection*		X			X

NO.	SOURCE TEXT	TARGET TEXT	DISSECTED COMPONENTS	ERROR TYPE	ORTHOGRAPHICAL	MORPHOLOGICAL	SEMANTIC	LEXICAL	SYNTACTIC
47.	美国媒体的相关报道显示，在过去相当长的时间内，哈佛学生选择人文科学的人数急剧下降。	US media reports show that in the past for a long time, Harvard students choose humanities sharp decline the number of people.	* US media reports show that in the past for a long time, Harvard students choose humanities sharp decline the number of people. - The US media reports	lack of definite article					O
			- - ~~people~~	addition				O	
			- - - sharply	word class					O
			- - - - declined	wrong verb inflection		O			O
			- - - - - the number of Harvard students choose humanities declined sharply.	word order					O
			- - - - - - the number of Harvard students who choose humanities declined sharply.	syntax: relative clause					O
		United States media reports showed that in the past for a long time, a sharp decline in the number of Harvard students choose humanities.	* United States media reports showed that - The United States media reports	lack of definite article					X
			* a sharp decline in the number of Harvard students choose humanities. - a sharp decline in the number of Harvard students who choose humanities.	syntax: relative clause					X
			- - there was a sharp decline	lack of syntactic elements					X
48.	并且这样的趋势在美国其他文科学院中也同样存在。	And such a trend in the US also exists in the other liberal arts.	* And such a trend in the US also exists in the other liberal arts. - the other liberal arts colleges	omission				O	
			- - And such a trend also exists in the other liberal arts colleges in the US.	word order					O
		And this trend also exists in other arts institution in the United	* And this trend also exists in other arts institution in	wrongly translated words				X	
49.	这样的背景令中国哲学课程更加走红。	This background makes Chinese philosophy more popular.	√						
		Chinese philosophy courses are more popular with that background.	√						

NO.	SOURCE TEXT	TARGET TEXT	DISSECTED COMPONENTS	ERROR TYPE	ORTHOGRAPHICAL	MORPHOLOGICAL	SEMANTIC	LEXICAL	SYNTACTIC
50.	《大西洋月刊》的报道称，这些中国思想能够帮助那些十八九岁的年轻人思考如何成为一个好人，如何创造一个良好的社会。	"Atlantic Monthly," the report said, these thoughts can help Chinese young people think about how to become a good person, how to create a good society.	*** "Atlantic Monthly," the report said** - "Atlantic Monthly" the report said	*punctuation*	O				
			- the -Atlantic Monthly“	*lack of definite article*					O
			- - the report of the Atlantic Monthly said,	syntax: NP					O
			*** these thoughts can help Chinese young teens to think about how to become a good person, how to create a good society.** - these Chinese thoughts can help young people think about how to become a good person, how to create a good society.	*word order*					O
		The Atlantic reports, these ideas can help those who are 99-year-olds thinking about how to be a good person, and how to create a good society.	*** The Atlantic reports, these ideas can help those who are 99-year-olds thinking about how to be a good person, and how to create a good society.** - The Atlantic Monthly	*proper noun*				X	
			- - who are teenagers	*wrongly translated words*				X	
			- - - think about	*wrong verb form*					X