A Comparative Study on Simplified-Traditional Chinese Translation

Xiaoheng Zhang

Department of Chinese and Bilingual Studies, Hong Kong Polytechnic University
ctxzhang@polyu.edu.hk

Abstract. Due to historical reasons, modern Chinese is written in traditional characters and simplified characters, which quite frequently renders text translation between the two scripting systems indispensable. Computer-based simplified-traditional Chinese conversion is available on MS Word, Google Translate and many language tools on the WWW. Their performance has reached very high precision. However, because of the existence of one-to-many relationships between simplified and traditional Chinese characters, there is considerable room for improvement. This paper presents a comparative study of simplified-traditional Chinese translation on MS Word, Google Translate and JFJ, followed by discussion on further development, including improvement of translation accuracy and support to human proofreading.

Keywords: Chinese characters, writing systems, simplified-traditional Chinese translation, one-to-many relationship, computer-aided proofreading

1 Introduction

Because of historical reasons [3], there are two Hanzi writing systems in China: the standard system of the Mainland is simplified Chinese, while in Taiwan, Hong Kong and Macau, traditional Chinese is the norm. In addition, these two writing systems are widely used internationally. For instance, the BBC Chinese website (http://www.bbc.co.uk/zhongwen/simp/) is presented in both simplified and traditional versions, as shown in Figure 1.

In addition to various documents, the user interfaces of popular software such as MS Windows, MS Office and Google also have their simplified and traditional Chinese versions. That means there is a tremendous demand for translation between the two Chinese writing systems.

Computer tools for simplified-traditional Chinese translation or conversion are widely available (For simplicity purpose, we will normally use the term “translation” in the following). Their performance has reached very high precision. However, because of the existence of one-to-many relationships between simplified and traditional Chinese characters there is no guarantee of 100% correct conversion [4]. For example, character 干 in simplified Chinese corresponds to 乾 (gan1), 幹 (gan4) and 干

(gan1) of different meanings in traditional Chinese, and traditional character 乾 corresponds to 乾 (qian2) and干 (gan1) of different meanings in simplified Chinese.

In the following sections, we will introduce an experiment of simplified-traditional Chinese translation on three representative tools, make a comparative analysis and provide some ideas for further improvement.

Fig. 1. BBC Chinese website (中文网) with simplified and traditional Chinese versions (简体版, 繁體版),

2 The Experiment

Our experiment was focused on simplified-to-traditional Chinese translation, which is more demanding than traditional-to-simplified translation and thus can better reflect the strength of modern technology in this area. A text in simplified Chinese was carefully selected, which was then translated into traditional Chinese on three representative tools.

2.1 Selection of the testing text

The testing text selected for the experiment is this year's Government Work Report by Premier Li Keqiang [5], which was presented to the annual meeting of the National People Congress of People's Republic of China in early March. Its original version is in simplified Chinese.

There are at least four reasons for our selection: (a) The government work report is an important document with attention of people all over the world; (b) Its contents cover every aspect of people's life in China; (c) It was a brand new document not likely to have been used to help improve any translation tool before our experiment. (d) The length of the document, 17,670 characters including spaces, is appropriate, because too short text may reveal few translation errors and too long text may require too much human proofreading.

2.2 Selection of translation tools

Three representative tools, including Microsoft Word, Google Translate and JFJ, were selected for our experiment on simplified-traditional Chinese translation.

MS Word has long been the most popular word processor in the world. And its Chinese Translation is probably the most popular off-lined tool for simplified-traditional Chinese translation. We used Word 2010, the newest version of the software available in our labs. When started on MS Word, Chinese Translation presents some options for the user to select, as shown in Figure 2.

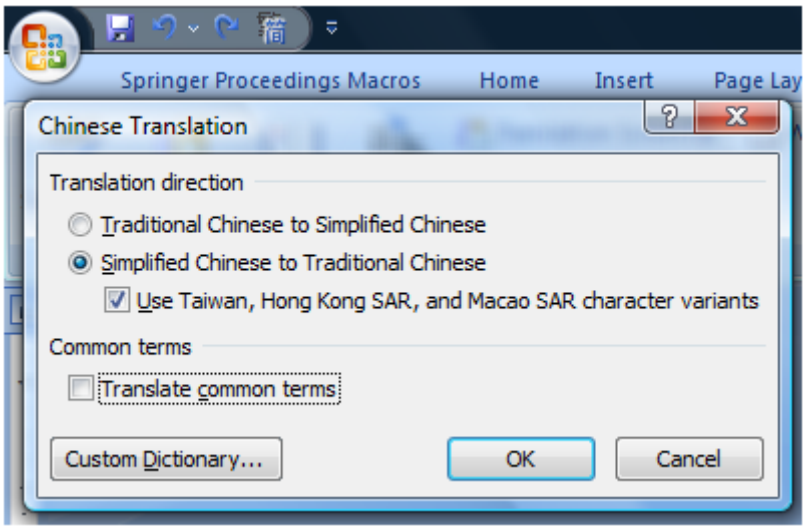

Fig. 2. Option setting on MS Word Chinese Translation

For our experiment, the Translation direction was set on "Simplified Chinese to Traditional Chinese". Normally, simplified-traditional Chinese conversion is performed in a character-to-character mode. For example, simplified 汉字信息处理 (Chinese characters information processing) is converted to漢字信息處理. MS Word can also translate common terms, in which case the previous translation would become 漢字資訊處理, because in Taiwan and Hong Kong, "information" is more often translated into 資訊 than 信息. Because neither Google Translate nor JFJ support terms translation, this option was set off for a fair comparison among the three tools. (As a matter of fact, MS Word sometimes translates common terms incorrectly. Examples will be presented in Section 4 of this paper.) When button OK is clicked on, MS Word will translate the whole document (or a selected part if it is highlighted beforehand).

Google Translate is an independent tool on the WWW (https://translate.google.com.hk/?hl=en). It is probably the most popular on-lined

translation tool. For our experiment, the source language was set to “Chinese (Simplified)” and the target language “Chinese (Traditional)”, as shown in Figure 3.

Fig. 3. Google Translate for simplified-traditional Chinese translation

JFJ (acronym of Pinyin Jian-Fan-Jian, or Simplified-Traditional-Simplified) is a small tool developed at Hong Kong Polytechnic University [15, 17]. The version used for the experiment is available on the Web at http://myweb.polyu.edu.hk/~ctxzhang/jfj30/ , as shown in Figure 4.

Fig. 4. JFJ simplified to high frequency traditional Chinese characters

The important features of JFJ include:

- Simplified-traditional Chinese bi-directional conversion with 4 options: (a) simplified to Hong Kong traditional Chinese, (b) simplified to Taiwan traditional Chinese, (c) simplified to high frequency traditional Chinese characters, and (d) traditional to simplified Chinese.

- Support for human proofreading by (a) high-lighting all characters with one-to-many relationship between simplified and traditional Chinese, (b) providing relevant dictionary information for reference, (c) correcting mistakes automatically by a single click.
- Employment of standard and frequently-used characters and punctuation marks in the target writing system.

2.3 The activities

The experiment was carried out in late March, 2014. We performed the following tasks on each of the three translation tools.

(a) Copy the text of the whole government report to the source text area of the translation tool;
(b) Translate the report from simplified Chinese to traditional Chinese with the tool (by character-to-character conversion, no translation of common terms);
(c) Proofread the translation manually and write down all the errors;
(d) Sort and categorize the errors on a table and calculate the error rates.

3 Experiment Results

The translation errors made by MS Word, Google Translate and JFJ are presented in Tables 1, 2 and 3. An error is an incorrect traditional Chinese translation of a simplified Chinese word. Word variants acceptable to the standards of Taiwan or Hong Kong are counted as correct. For example, either 裡面 or 裏面 is a correct translation of simplified Chinese word 里面, while 公里 can only be translated to 公里 (unchanged), because both 公裏 and 公裡 are not acceptable in traditional Chinese.

Table 1. Errors and corrections of MS Word's translation

Original text in Simplified Chinese	Translation error in Traditional Chinese	Correction	Occurrences
精准	精**准**	精準	3
免征	免**征**	免徵	1
深松整地	深**松**整地	深鬆整地	1
松了绑	**松**了綁	鬆了綁	1
页岩气	**菜**岩氣	頁岩氣	1
这只…的手	這**只**…的手	這隻…的手	2
征地	**征**地	徵地	1
制售	**制**售	製售	1
Total:			11
Error rate against all characters in the source text = (11/17670)*100%=0.062%			
Error rate against one-to-many characters = (11/912)*100%=1.206%			

Table 2. Errors and corrections of Google Translate's translation

Original text in Simplified Chinese	Translation error in Traditional Chinese	Correction	Occurrences
遨游太空	遨**游**太空	遨遊太空	1
备案制	備案**製**	備案制	1
复制	複**制**	複製	1
赶超	**赶**超	趕超	1
公有制	公有**製**	公有制	3
骨干水源	骨**乾**水源	骨幹水源	1
旅游近亿人	旅**游**近億人	旅遊近億人	1
牵一发而动全身	牽一**發**而動全身	牽一髮而動全身	1
深松整地	深**松**整地	深鬆整地	1
示范建设	示**范**建設	示範建設	1
所有制	所有**製**	所有制	2
体系不健全	體**係**不健全	體系不健全	1
维护发展中国家共同利益	維護**髮**展中國家共同利益	維護發展中國家共同利益	1
行政复议	行政**復**議	行政複議/覆議	1
这只…的手	這**只**…的手	這隻…的手	2
制度	**製**度	制度	13
Total			**32**
Error rate against all characters in the source text = (32/17670)*100% = 0.181%			
Error rate against one-to-many characters = (32/912)*100%=3.509%			

Table 3. Errors and corrections of JFJ's translation

Original text in Simplified Chinese	Translation error in Traditional Chinese	Correction	Occurrences
牵一发而动全身	牽一**發**而動全身	牽一髮而動全身	1
松花江	**鬆**花江	松花江	1
台风	**台**風	颱風	1
维护发展中国家共同利益	維護**髮**展中國家共同利益	維護發展中國家共同利益	1
行政复议	行政**復**議	行政複議/覆議	1
制售假冒伪劣	**制**售假冒偽劣	製售假冒偽劣	1
Total:			6
Error rate against all characters in the source text = (6/17670)*100%=0.034%			
Error rate against one-to-many characters =(6/912)*100%=0.658%			

The three tables above contained the combined results of 45 individual experiments made by the author and a class of Chinese native students according to the requirements introduced in Section 2. In other words, the experiment was performed

45 times by different people and the results were then checked and integrated into three tables. That means the data reported here should be very comprehensive and reliable, with all the translation errors made by each tool counted.

The first column of the tables presents all the words and phrases in the original text which have been incorrectly translated by MS Word, Google Translate or JFJ. The second column presents the error translations. Each error is caused by an incorrectly converted character, which is presented with sufficient context to show its incorrect usage and to cause the mistake. For instance, Google can translate 骨干 correctly into 骨幹, but would translate 骨干水源 incorrectly into骨乾水源. Hence 骨干水源 and 骨乾水源 appear in the table. The third column presents the corrected expressions in traditional Chinese. And the final column shows the number of times each error occurs in the entire translation text. The rows are sorted by the first column in Putonghua Pinyin order. At the end of the table, there are two error rates: the error rate against all characters in the source text is calculated by dividing the "total number of incorrectly-translated characters" with "total number of characters in the text" before converting into a percentage, the error rate against one-to-many characters is calculated by dividing the "total number of incorrectly-translated characters" with "total number of characters of one-to-many relationship". According to JFJ's report at the end of its translation, there are totally 17,670 characters (including spaces) in the source text, among which 912 have multiple counterparts in traditional Chinese.

The references we used in human proofreading include

- Dictionary of Commonly Used Words in Mainland and Taiwan (Mainland version) [7]
- Dictionary of Commonly Used Words in Mainland and Taiwan (Taiwan version) [8]
- Lexical Items for Fundamental Chinese Learning in Hong Kong [2]
- Revised Edition of the Dictionary of the Chinese National Language [9]
- Academia Sinica Balanced Corpus of Modern Chinese [1]
- Corpus of Modern Chinese by the National Language Commission of China [10]
- Decoding the *Standard List of Commonly-used Chinese Characters* [13].

4 Error Analysis and Solution

Generally speaking, the three tools all performed very well in translating the government work report from simplified Chinese into traditional Chinese. The accuracy of Google Translate is 1-0.181%=99.819%. MS Word performed even better with an accuracy of 1-0.062%=99.938%. And JFJ was the champion with an accuracy of 1-0.034%=99.966%.

4.1 Error analysis

The distribution of different errors is shown in Table 4.

Table 4. Distribution of different errors made by Word, Google and JFJ

	Word & Google	Word & JFJ	Google & JFJ	Word	Google	JFJ
	深**松**整地	**制**售	**復**議	精**准**	遨**游**	**鬆**花江
	這**只**...的手		牽一**發**	免**征**	備案**製**	**台**風
			維護**髮**展	**松**了綁	複**制**	
				葉岩氣	**赶**超	
				征地	公有**製**	
					骨**乾**	
					旅**游**	
					示**范**	
					所有**製**	
					體**係**	
					製度	
Total	2	1	3	**5**	11	2

Totally the three tools made 24 different errors in their translations, none of which is shared by all the tools. 6 errors were made by two tools, including 3 by Google and JFJ, 2 by Word and Google and 1 by Word and JFJ. 19 errors were made by a single tool, including 11 by Google, 5 by Word and 2 by JFJ. The wide distribution of errors means that the three tools employ quite different approaches in translation, and hence are good representatives of this area of language computing.

The errors listed in the tables are mostly self-evident. We will discuss a few cases which are more interesting or informative. Errors in simplified-traditional Chinese translation normally happen in characters corresponding to more than one counterpart in the target writing system. For example, character 松 in simplified Chinese corresponds to 松 (pine tree) and 鬆 (loose, slack) in traditional Chinese. In the original phrases of 松了绑 and 深松整地, 松 means "to set the hands loose (untied)" and "make the land loose (for farming)", hence should be converted to 鬆 in traditional Chinese. On the other hand, 松花江 is a river's name relevant to the pine tree, hence its correct translation is 松花江.

Simplified Chinese character 只 corresponds to 只 (zhi3) and 隻 (zhi1) in traditional Chinese. When used as a measure word, it should be converted into 隻. So the correct translation of 这只…的手 in traditional Chinese is 這隻…的手, not 這只…的手.

Character 制 also has two counterparts in traditional Chinese, 製 (to make, manufacture) and 制 (system). Word 制售 is the short form of 制造销售, or "to manufacture and sell (products)". 复制 means to make a copy. Hence the character 制 in both cases should be converted to 製. On the other hand, words 备案制, 公有制, 所有制 and 制度 all refer to some systems. Hence their embodied character 制 should remain unchanged in traditional Chinese.

It is surprising that MS Word translates 页岩气 to 葉岩氣, even when the option of "Translate common terms" is turned off. According to Dictionary of Commonly Used Words in Mainland and Taiwan (Taiwan version) [8], the standard translation is

頁岩氣. In fact, none of the many dictionaries we have consulted tells us that the traditional form of simplified 页 can be 葉.

复议 is a less frequently-used word, and few Mainlanders can be confident about its correct translation in traditional Chinese. According to Dictionary of Commonly Used Words in Mainland and Taiwan (Taiwan version) again, the standard counterparts of simplified Chinese 复议 are synonyms 複議 and 覆議, not 復議. And 行政複議 appears in an example sentence of the book.

As for 维护发展中国家共同利益, the computer mistakenly considered 护发 as a word and translated it into 護髮. However the correct word segmentation is 维护/发展/中/国家/共同/利益, where 发展 is a word and should be converted to 發展. 發 and 髮 are different characters in traditional Chinese.

4.2 Error solution

In this section, we will discuss improvement of the three tools based on further analysis of the errors they made in the experiment.

In the case of MS Word, the function of "Convert common terms" should not be allowed to interfere with translation when the user selects not to use it, such as incorrectly converting 页岩气 to 葉岩氣. On the other hand, the function should be improved. The facts that MS Word would translate 主席缺位期间由副主席代理 to 主席缺位元期間由副主席代理，文件资料 to 檔資料，and 循环经济 to 迴圈經濟 shows that MS Word is not good enough at Chinese words segmentation and is quite confined to the domain of computer science. Word 缺位 in the first example means 职位空缺 or "seat/post vacant". MS Word mistakenly considered character 位 as a word with the meaning of "bit" in information technology, and converted it into 位元. 檔資料 is another strange expression in Chinese. And translating 循环经济 into 迴圈經濟 is not acceptable either, though in computer programming 循环 often means "loop" and can be safely converted to 迴圈 in a Taiwan style. Some people even consider term conversion not worthwhile because of two reasons: loss of original "flavor" of the source text, and introduction of new mistakes [12].

Google Translate made 32 errors in the experiment, as shown in Table 2. Word 制度 appears 33 times in the source document, among which 13 were incorrectly converted to 製度, more than the total number of errors made by MS Word or JFJ. That means there is something seriously wrong. A possible reason is that Google over relies on corpora and statistics MT, and paid insufficient attention to deep and comprehensive linguistic analysis [6]. In the n-gram algorithm, we pre-suppose the appearance of a word to be independent of other words in the context, which is not true in real-world natural languages. Another shortcoming of n-gram is in its implementations, where n is often set to 2 or 3 due to limitation of computing resources. In natural languages, however, the correct analysis of a word often involves a more remotely-located word. For instance, in 放開市場這只“看不見的手”，用好政府這只“看得見的手”, the correct translation of 只 is dependent on its modifying relation

with noun 手, which is 5 characters away. It seems more reasonable to employ a flexible context length in language analysis, not a fixed bigram or trigram for all cases.

It is surprising that JFJ clearly outperforms the two IT giants in simplified-traditional Chinese translation. And the six errors made by JFJ can be eliminated by better use of contextual information. Let's take 牵一发而动全身 as an example. Simplified Chinese 发 has two counterparts in traditional Chinese: 發 (fa1: develop, grow) and 髮 (fa4: hair). And 一发 is still ambiguous, considering 一發不可收拾, 一發子彈 and 千鈞一髮. However, putting the lengthy 牵一发而动全身 and its correct translation into the conversion dictionary is not cost-effective. And if we do, its variant expressions would not be dealt with as well, including 牽一發動全身 (without 而) and "牽一發, 動全身", etc. We turned to the modern Chinese corpus of the National Language Commission of China [10], and found that 牵一发 appears 12 times, in 牵一发而动全身, 牵一发动全身, "牵一发，动全身" and 牵一发而动全局. In all of these cases, 牵一发 can be safely converted to 牽一髮. That means 牵一发 is a minimum translation segment to unambiguously convert 发 into 髮. It is more cost-effective to put "牵一发-牽一髮" in the computer's conversion dictionary. And the possible appearances of 牵一发而动全国, 牵一发而动全市, 牵一发而动全校 and so on will be covered as well.

5 Conclusion

Simplified-traditional Chinese translation achieved an accuracy of 99% more than a decade ago [11, 14]. The percentage has now been raised to 99.9% and beyond, as shown by our experiments on Google Translate, MS Word and JFJ. Notwithstanding, all of the three tools made more than 5 errors. That means there is substantial space for further progress, as discussed in the previous section.

In another aspect, because of the one-to-many relationships between simplified and traditional Chinese characters and the great complexities of natural languages, there is no guarantee of 100% correct translation in the foreseeable future. As we know, linguistically, every rule leaks; statistically, even the most probable event may not happen. That means human proofreading is needed for machine translation, especially when high quality text output is required [15, 16]. Another lovely feature of JFJ is that it goes on to support human proofreading after automatically translating a text from simplified Chinese into traditional Chinese or vice versa. And this supporting function can be further improved as well. A newly revised version of JFJ taking into account of feedback from our experiment is available for testing on the Web at http://www.acad.polyu.edu.hk/~ctxzhang/jfj/ .

Last but not least, the experiment reported in this article was performed on the single document of Premier Li's government report. To be better representative of modern Chinese language, the testing text will be greatly enriched in our further research, both quantitatively and typologically.

Acknowledgements

The author would like to thank his MA postgraduate students of subject "Modern Chinese Characters and Information Technology" at Hong Kong Polytechnic University for their support in the experiment. He is also very grateful to the three anonymous reviewers, whose informative comments and constructive suggestions played an important role in the revision of the paper.

References

1. Academia Sinica (1997). Academia Sinica Balanced Corpus of Modern Chinese (中央研究院現代漢語語料庫). http://app.sinica.edu.tw/cgi-bin/kiwi/mkiwi/kiwi.sh?language=1.
2. Education Bureau of Hong Kong (2009). *Lexical Items for Fundamental Chinese Learning in Hong Kong* (香港學校中文學習基礎字詞, http://www.edbchinese.hk/lexlist/).
3. Fu, Y. (傅永和, 2005). Fifty Years of Chinese Characters Simplification (汉字简化五十年回顾). *Languages of China* (中国语文), No. 6, 2005.
4. Halpern, J. and Kerman, J. (1999). The Pitfalls and Complexities of Chinese to Chinese Conversion. Fourteenth International Unicode Conference, Boston, 1999.
5. Li, K. (李克强, 2014). Government Work Report 2014 (政府工作报告 2014). http://www.gov.cn/guowuyuan/2014-03/05/content_2629550.htm.
6. Li, M. Wu, Y. Zeng, Y. Yang, P. and Ku, T. (2010). Chinese Characters Conversion System Based on Lookup Table and Language Model. *Computational Linguistics and Chinese Language Processing*. Vol. 15, No. 1, pp. 19-36 19.
7. Li, X. (李行健, 2014, editor). *Dictionary of Commonly Used Words in Mainland and Taiwan* (两岸常用词典, Mainland version). http://www.zhonghuayuwen.org/.
8. Liu, Z. (劉兆玄, 2014, editor). *Dictionary of Commonly Used Words in Mainland and Taiwan* (兩岸常用詞典, Taiwan version). http://chinese-linguipedia.org/clk/index.php.
9. Ministry of Education (1994). *Revised Edition of Dictionary of the Chinese National Language* (重編國語辭典修訂本). Taipei: Ministry of Education.
10. National Language Commission of China (2010). Corpus of Modern Chinese (现代汉语语料库). (http://www.cncorpus.org/CCindex.aspx, Consulted on May 31, 2014.)
11. Shen, D and Sun M. (沈达阳，孙茂松, 1996). An Intelligent Simplified-Traditional Chinese Conversion System (汉字简繁体智能化转换系统). *Chinese Information* (中文信息), No. 6, 1996.
12. Wang, L. Wang, X, and Wu, J. (王立军，王晓明 and 吴健, 2013). The Correspondence Simplified Characters and Traditional Characters and the Mutual Conversion (简繁对应关系与简繁转换). *Journal of Chinese Information Processing* (中文信息学报). Vol. 27, No. 4.
13. Wang, N. (王宁, 2013). Decoding the *Standard List of Commonly-used Chinese Characters* (《通用规范汉字表》解读). Beijing: The Commercial Press. (Chapter 3: Relationship between Simplified and Traditional Chinese Characters 简繁关系).
14. Xin, C. and Sun, Y. (辛春生, 孙玉芳, 2000). Design and Implementation of Simplified-Traditional Chinese Conversion System (简繁汉字转换系统的设计与实现). *Journal of Software* (软件学报). No. 11, 2000.

15. Zhang, X. (2011). A Simplified-Traditional Chinese Conversion Tool with a Supporting Environment for Human Proofreading (一个支持人工校对的中文简繁体转换工具). In Sun, M. and Chen, Q. (孙茂松, 陈群秀编, eds.), *Advances of Computational Linguistics in China* 2009-2011 (中国计算语言学研究前沿进展2009-2011). Beijing: Tsinghua University Press (清华大学出版社), pp. 569-575.
16. Zhang, X. (2012). Existing Space for Improvement in Simplified-Traditional Chinese Character Conversion (计算机汉字简繁体转换有待解决的问题). In Li, X., Zhang, J. and Xu, J. eds. (李晓琪，张建民，徐娟主编), *Digital Teaching of Chinese Language 2012* (数字化汉语教学 2012). Beijing: Tsinghua University Press. (清华大学出版社), pp 219-226.
17. Zhang, X. (2013). Simplified-Traditional Chinese Conversion with Assistance to Human Proofreading. Invited short paper for *Newsletter of the Chinese Language Teachers Association*, Volume 37, January 2013, Number 1, *p 30*.