School of Languages, Cultures and Societies

Faculty of Arts, Humanities and Cultures

UNIVERSITY OF LEEDS

Standard Cover Sheet for Assessed Coursework

From January 2017 all assessed coursework in LCS must be accompanied by this coversheet (*unless you have a specific learning disability, in which case you should use instead the Marking Awareness coversheet*).

This coversheet should be included as the first page of your submission. Please copy and paste it into the beginning of your assessment (Ctrl+A, then Ctrl+C to copy its entirety). The contents of the coversheet do not count towards your submission word count.

Please complete each of the following sections:-

Student ID:	201565992
Module Code:	MODL5007M
Module Tutor's Name:	Dr Serge Sharoff
Assessment description (eg essay 1):	Case Study
Word count	1,986 words

Please note that by submitting this piece of work you are agreeing to the University's Declaration of Academic Integrity. You can read the Declaration here.

NB Please ensure that you name your file with your SID, the module code, seminar/language tutor's name (if required) and a brief assessment description e.g. 200987654 MODL1234 seminar/language tutor Name (if required) Essay 1.pdf

Using comparable corpora to find translation equivalence in English and Chinese pair

Case study of basic, plain and simple vs 基本 *jīběn*, 普通 *pǔtōng* and 简单 *jiǎndān*

Introduction

Finding appropriate translation equivalence could be a difficult task for translators who are not experienced enough in a relevant subject field or are not native to one of their language pairs. Although bilingual dictionaries sometimes can offer translators some translation equivalents for a certain headword, those equivalents are word-for-word translations for a headword with the same part of speech (POS) (Sharoff, 2007, p.1). On top of that, since a translation unit that translators normally deal with can be more than a word unit, determining translation equivalence could be subject to its lexicogrammatical environment, semantic field and even cultural context which may differ significantly between different languages such as English and Chinese (Mikhailov and Cooper, 2016, p.150). On the other hand, a corpus, as a large body of naturally generated texts in a language, can help translators overcome the barriers mentioned above by offering annotated contexts and powerful query approaches (Newman and Cox, 2021). Corpora, therefore, become the ideal tools for translators to find contextually appropriate translation equivalents that are not listed in bilingual dictionaries.

This case study aimed to find the translation equivalents for the three English words and three Chinese words listed in the subtitle using two monolingual corpora in English and Chinese respectively as comparable corpora. Those terms belong to general lexicons meaning that they are normally polysemous and thus can be translated differently according to their semantic and pragmatic environment and the translation strategies applied. In the following sections, the research hypothesis, tools, procedures and data processing will be summarised and discussed in the methodology section, followed by a case study of the aforementioned English and Chinese lexicons, their translation equivalents and corresponding translation

conditions supported by corpus statistics. The main results and indications of the case study, potential further improvements and the implications of equipped corpus linguistic knowledge will be discussed in the final section.

Methodology

The research of this case study is formulated based on the following hypotheses:

Hypothesis 1: The corpora used in this case study should be large enough and well-sampled to represent the source language and target language (Rees, 2022).

Hypothesis 2: The corpora should also offer essential query options in their interfaces for translators to obtain concordances or keywords in contexts (KWIC) (Anthony, 2022).

Hypothesis 3: Translators should be able to read KWIC vertically and identify lexical and grammatical patterns (Hunston, 2022).

Hypothesis 4: Translation pairs that are semantically similar should share similar lexicogrammatical and semantic environments in KWIC of their own corpora (Manning and Schutze, 1999, p.295).

Hypothesis 5: Null hypothesis should be applied in the inferential statistics tests to provide evidence for the translation assumption in the case study (Zufferey, 2020).

According to the research hypotheses above, British National Corpus (BNC) and Modern Chinese Languages Corpus (MCLC) were chosen as the comparable corpora for the case study. Both corpora consist of around 100 million English words or Chinese characters and cover a wide range of similar topics and genres during a similar time period[1]. The size, composition and balanced sampling of the two general corpora justified the validity of using them as comparable corpora for corpus linguistics and translation study (Sharoff et al., 2013). BNC was mainly accessed through the Leeds Intellitext interface (http://corpus.leeds.ac.uk/itweb) and Leeds CQP interface

[1]See http://www.natcorp.ox.ac.uk/corpus/index.xml?ID=numbers and http://corpus.zhonghuayuwen.org/CorpusIntro.aspx?WebShieldDRSessionVerify=9Yg1DLG7kGdBv2kbs0ih for detailed information.

(http://corpus.leeds.ac.uk/protected/) (Wilson et al., 2010). MCLC has its own interface (in Chinese) with essential query functions.

The United Nations Parallel Corpus (UNPC) from Sketch Engine (Kilgarriff et al., 2014) and Kelly DB (http://kelly.sketchengine.co.uk/) were used to help identify the most common translation options for the selected words. Specialised web corpora in English and Chinese were built using web crawling to help further justify the translator's assumption of domain and genre by looking into the keyword list (Sharoff, 2021). It should be noted that although parallel corpora and specialised web corpora may have a limited volume of texts compared with large-scale general corpora, even a larger corpus should only be a part of the entire language set and offer a limited view of the language. Therefore, the point there is that those smaller corpora with their size can still help translators make good judgements on their assumptions in their translation decisions (Jones, 2022).

The case study for the English and Chinese words followed a four-step methodology informed by the module MODL5007M and a series of seminal papers ((Sharoff, 2007; Philip, 2009; Bernardini, 2022).

A typical study begins by conducting a lemma query and studying the generated concordances with the aim to find the functional patterns of a word. Due to the time constraints and statistical distribution of major patterns of word use, 300 concordances for each word were used for pattern identification by copying them and implementing colour coding in a spreadsheet. Based on the first step, the context information of the word can then be generalised, and more importantly, neighbouring words or collocations that directly contribute to the meaning-making of the word can be spotted. These surrounding lexical components together with the word in the study should establish a solid basis for translation and serve as a constraint for translation equivalence screening in the target language corpus. To create a link between two comparable monolingual corpora, the translations of identified neighbouring words and collocations are generated assisted by either online bilingual dictionaries or parallel corpora. Finally, extensive queries using the translations of contextual information are conducted to find possible translation equivalence of the word in the study or to offer occurrences as evidence for the translator's justification of translation decisions.

In terms of the significance test, the log-likelihood (LL) score was used in this case study for calculating keyness and collocations because this widely applied, relatively reliable statistic allows a clear p-value threshold to set for the null hypothesis (Rayson and Garside, 2000; Pojanapunya and Watson Todd, 2018; Zufferey, 2020).

The details about the making of dictionary entries for the words discussed are mainly presented in the submitted dictionary entry XML file. In this case study, only selected senses and translation conditions were investigated in detail due to the word count limit.

Case study

English-Chinese translation conditions

Basic

In terms of POS, basic can be used as an adjective or a noun. The occurrences of the adjective basic take over 96% of the total hits in BNC, which indicates the term tends to be used as an adjective in most cases. However, the Chinese translation of basic can be tricky because after checking its collocation patterns, the translator found that the three most possible translation equivalents - 基本 (basic), 基础(basic) and 根本 (fundamental) – are interchangeable in the Chinese translations. Therefore, the translation conditions for these equivalents largely depend on the customary collocation of each sub-sense. This was addressed by several wildcard searches (基* and 根*) together with the translations of collocates with high LL scores ($p<0.001$) in the MCLC. By comparing the frequencies of different matches, the translator was able to assign each group of collocates to the Chinese equivalents.

Plain

Plain is most likely to function as an adjective (65%) and a noun (23%) based on POS-tag queries. Although there were over 400 hits for the adverb plain, a majority of wrong POS annotation was found in the concordance list. That reminded the translator of the importance of data pruning when processing the generated KWIC in a spreadsheet.

Table 1 Collocation list of JJ+N.* pattern

Collocation	Joint	Freq1	Freq2	LL score
plain flour	97		1071	351.78
plain clothe	68		7195	168.24
plain sailing	47		1360	146.92
plain chocolate	48		2230	138.53
plain fact	51		41485	74.43
plain English	44		22914	73.85
plain tile	21		1837	53.87
plain speaking	16		683	46.82
plain colour	24		15313	37.83
plain paper	26		21315	37.79
plain yoghurt	11		565	31.15

Plain has strong collocates with food-related nouns as shown in **Table 1**. This led to the further investigation of the translation conditions including domain, semantic prosody and genre for this sense. By crawling the BBC recipe website (https://www.bbc.co.uk/food/recipes) and a Chinese recipe website Xia Chu Fang (https://www.xiachufang.com/explore/classic/), both the collected web corpora were further used to identify emotion and genre labels[2]. The emotion and genre distribution in English and Chinese turned out to be overwhelmingly positive and promotional. By further calculating the keyness using the *word_keyness* script, the plain was finally justified as a positive keyword ($p<0.001$) compared with the BNC as a reference corpus. Therefore, based on the discussion above, plain, when collocating with food nouns, can be translated into "原味" (with original flavour) under the constraints of the aforementioned domain, emotion and genre.

Simple

Over 85% of the lemma simple hits were adjectives as this is its only POS. To describe tangible objects or abstract terms, simple presents two sub-senses although the corresponding Chinese translations are the same after doing similar wildcard queries in the Chinese corpus as in the Basic section. Both sub-senses show slightly positive semantic prosody as demonstrated in the highlighted terms in Table 2.

[2] For Chinese corpora, the Chinese RoBERTa-Base Model for Text Classification with Dianping data set was used in the *linguistic* script offered by the lecturer.

Table 2 Collocation list of RB+JJ pattern

Collocation	Joint	Freq1	Freq2	LL score
plain enough	18		31067	36.24
plain daft	7		624	24.42
plain stupid	6		3059	15.68
plain wrong	7		15756	13.12
plain silly	5		2619	13
plain English	7		22914	11.83

Chinese-English translation conditions

基本 jīběn

The Chinese term “基本” presented a similar issue as in the section of Basic; its two adjective senses, found in the Chinese concordances, have slightly different meanings due to customary usage of collocates but share the same translation in English. Therefore, the translation conditions for the adjective senses were quite loose although, in the source text, each sense is restricted by corresponding collocates obtained from LL score calculations.

普通 pǔtōng

The monosemy of “普通” in Chinese results in one single adjective sense when collocating with common terms like “工人” (worker). However, when “普通” forms a new noun with “话” or “法”, two frequent characters following after, the translation of “普通” differs drastically from its common translation due to the non-compositionality of the collocations.

简单 jiǎndān

The translation condition of the term “简单” presents a similar constraint imposed by non-compositionality. When collocating with a negative like “不” (No) in Chinese, the whole term functions as a strong compliment. This functional shift leads to a more semantically and functionally appropriate translation of “ordinary”.

Conclusion

In this case study, several new senses and translation equivalents for terms like plain were generated using comparable corpora. Due to the time and word count limits, the translator had to incorporate only part of the translation conditions in the case study and cannot present more exciting findings during queries. However, being equipped with essential knowledge and skills in corpus linguistics, the translator believes that this module serves as a stepping stone for future career development in the translation industry with the era of AI boom coming soon.

Word count: 1,986 words

Bibliography

Anthony, L. 2022. What can corpus software do? *In*: A. O'Keeffe and M. J. McCarthy, eds. *The Routledge Handbook of Corpus Linguistics*. London: Routledge, pp.103–125.

Bernardini, S. 2022. How to use corpora for translation *In*: A. O'Keeffe and M. J. McCarthy, eds. *The Routledge Handbook of Corpus Linguistics*. London: Routledge, pp.485–498.

Hunston, S. 2022. How can a corpus be used to explore patterns? *In*: A. O'Keeffe and M. J. McCarthy, eds. *The Routledge Handbook of Corpus Linguistics*. London: Routledge, pp.140–154.

Jones, C. 2022. What are the basics of analysing a corpus? *In*: A. O'Keeffe and M. J. McCarthy, eds. *The Routledge Handbook of Corpus Linguistics*. London: Routledge, pp.126–139.

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P. and Suchomel, V. 2014. The Sketch Engine: ten years on. *Lexicography*. **1**(1), pp.7–36.

Manning, C. and Schutze, H. 1999. *Foundations of Statistical Natural Language Processing*. MIT Press.

Mikhailov, M. and Cooper, R. 2016. 6 Applications of parallel corpora *In*: *Corpus linguistics for translation and contrastive studies: A guide for research*. London: Routledge.

Newman, J. and Cox, C. 2021. Chapter 2 Corpus Annotation *In*: M. Paquot and S. T. Gries, eds. *A practical handbook of corpus linguistics*. Springer Nature, pp.25–48.

Philip, G. 2009. Arriving at equivalence Making a case for comparable general reference corpora in translation studies *In*: *Corpus Use and Translating: Corpus Use for Learning to Translate and Learning Corpus Use to Translate*. John Benjamins Publishing, pp.59–73.

Pojanapunya, P. and Watson Todd, R. 2018. Log-likelihood and odds ratio: Keyness statistics for different purposes of keyword analysis. *Corpus Linguistics and Linguistic Theory*. **14**(1), pp.133–167.

Rayson, P. and Garside, R. 2000. Comparing corpora using frequency profiling *In*: *Proceedings of the workshop on Comparing corpora* - [Online]. Morristown, NJ, USA: Association for Computational Linguistics. [Accessed 25 January 2023]. Available from: http://dx.doi.org/10.3115/1117729.1117730.

Rees, G. 2022. Using corpora to write dictionaries *In*: A. O'Keeffe and M. J. McCarthy, eds. *The Routledge Handbook of Corpus Linguistics*. London: Routledge, pp.387–404.

Sharoff, S. 2021. Genre annotation for the Web. *Register Studies*. **3**(1), pp.1–32.

Sharoff, S. 2007. Harnessing the lawless: using comparable corpora to find translation equivalents. *Journal of Applied Linguistics*. **1**(3).

Sharoff, S., Rapp, R. and Zweigenbaum, P. 2013. Overviewing Important Aspects of the Last Twenty Years of Research *In*: S. Sharoff, R. Rapp, P. Zweigenbaum and P. Fung, eds. *Building and Using Comparable Corpora*. Berlin, Heidelberg: Springer, pp.1–20.

Wilson, J., Hartley, A., Sharoff, S. and Stephenson, P. 2010. Advanced corpus solutions for humanities researchers *In*: *Proceedings of the 24th Pacific Asia Conference on Language, Information and Computation*., pp.769–778.

Zufferey, S. 2020. How to Analyze Corpus Data *In*: *Introduction to corpus linguistics*. John Wiley & Sons, pp.195–226.