The Construction and Application of an English-Chinese Parallel Corpus of Hydrogen Energy

Zheng Li*
Foreign Languages School, Shanghai Dianji University, Shanghai, China
**Corresponding Author.*

Abstract

This paper describes the construction and application of the first English-Chinese Parallel Corpus of Hydrogen Energy (PCOHE) in China, including design framework, collecting and processing of corpus materials, and its applications.Thepaper adopts statistical analysis with the help of corpussoftware to portray the linguistic characteristics of the hydrogen energy texts and study the translation strategy of the translated texts, which is expected to improve the efficiency and accuracy of translation. Although PCOHE is intended for translation studies of hydrogen energy texts, it can also be applied to such areas as the learning of hydrogen energy and machine translation.

Keywords*: Parallel Corpus, Hydrogen Energy, Construction and Application*

I. Introduction

With the development of energy revolution strategy, hydrogen energy, as a clean energy, plays a supporting role in coping with climate change and protecting the environment. At present, countries all over the world have promoted the development of hydrogen industry to the height of national energy strategy. China has followed the international trend of hydrogen energy development, and the development of hydrogen industry has stepped into the fast lane. With the increasingly frequent international exchanges in this field, the demand for professional translation and translation skills in hydrogen energy industry is increasing. Lack of parallel corpora in this field makes the related translation tricky, treacherous and time-consuming.Parallel corpora have proved to be a valuable resource in translation studies [1].Therefore the construction and study of English-Chinese Parallel Corpus of Hydrogen Energyhave graduallybecome an urgent issue and the focus study of this field.

Corpus linguistics takes a large number of carefully collected authentic texts as research materials, and draws conclusions mainly through the method of probability and statistics, which integrates language analysis, language education, statistics, information technology and other disciplines[2]. According to Sammut C. (2011), a parallel corpus is a document collection composed of two or more disjoint subsets, each written in a different language, such that documents in each subset are translations of documents in each other subset [3]. The contents of the parallel corpus are aligned side-by-side in order to be used for comparison purpose. The parallel corpus of specialized scientific fields provides a useful tool to extract specialized knowledge and its pairs in other languages, which enables to achieve the comparability of bidirectional texts and paves the new way for constructing domain ontology and translation study. The specialized parallel corpora not only provide abundant resources for researchers in this field but also benefit the research on translation, text features and the term data mining in professional field. Few specialized parallel corpora in energy domain have been provided in China, for example, the industrial Chineseparallel corpus based on equipment manufacturing industry is developed for the research of bilingual terminology of equipment manufacturing [4]. However, to the best of our knowledge, the English-Chinese Parallel Corpus of Hydrogen Energy has not been built yet. Therefore, it is of practical value to construct the Parallel Corpus of Hydrogen Energy.

This paper describes the construction and application of the first English-Chinese Parallel Corpus of Hydrogen Energy (PCOHE) in China, including design framework, collecting and processing of corpus materials, and its applications. The paper then adopts statistical analysis with the help of corpus software to portray the linguistic characteristics of the hydrogen energy texts and study the translation strategy of the translated texts, which is

expected to improve the efficiency and accuracy of translation. Although PCOHE is intended for translation studies of hydrogen energy texts, it can also be applied to such areas asthe learning of hydrogen energy and machine translation.

II. Framework Design

Bilingual parallel corpora provide a solid empirical foundation for generalpurpose language tools and descriptions, and the construction of bilingual parallel corpora is an interdisciplinary technical work.[5] It requires that the researchers can skillfully master the related corpus software and python language, meanwhile they have a good command of language proficiency and specialized knowledge.This paper designs and constructs the English-Chinese Parallel Corpus of Hydrogen Energy based on Python language and related corpus software. The design framework is shown in figure 1. The availability of design benefits not only the process of corpus compilation, but also the tasks of accurately validating and evaluating the results obtained through the corpus by its users.

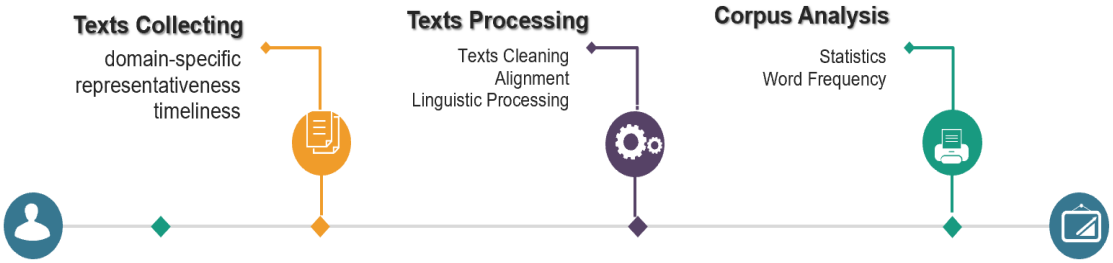

Fig 1: Framework design for the construction of PCOHE

III. Workflow Construction

Two essential pre-processing steps are used toobtain a high quality of English-Chinese Parallel Corpus of Hydrogen Energy and make the corpus standardization,unity and practicability, which are texts collecting and texts processing.

3.1Texts Collecting

The English-Chinese Parallel Corpus of Hydrogen Energy (PCOHE) is designed and built to satisfy the research of hydrogen energy and the translation of hydrogen energy texts, therefore the domain-specific, representativeness and timeliness are the essential issues that should be taken into consideration when selecting the texts. Hence, the following guidelines summarize the criteria used to collect the raw materials, which are interrelated and complement one another to guarantee the high quality of the corpus.

3.1.1Text Domains

All the corpus texts should be in hydrogen domain, which mainly involves the production, storage, transportation and application of hydrogen as well as the Hydrogen safety.These are the core topics in the field of hydrogen energy.

3.1.2 Representativeness

Biber (2010) argues that “a corpus is not simply a collection of texts. Rather, a corpus seeks to represent a language or some part of language. The appropriate design for a corpus therefore depends upon what it is meant to represent. The representativeness of the corpus, in turn, determines the kinds of research questions that can be addressed and the generalizability of the results of the research.” [6] In this sense, texts were obtained from different sources such as the academic monographs, related textbooks, journal articles, web pages, contracts, agreement and other legal documents concerning hydrogen energy, which guarantee the representativeness of PCOHE.

3.1.3Timeliness
All the corpus texts should reflect the current frontier progress or important research results oriented hydrogen energy research, and can meet the current social needs. In doing so, the foundations for the future diachronic studies of hydrogen energy are to be laid. Texts obtained were during the latest ten years.

3.2Texts Processing

The English-Chinese Parallel Corpus of Hydrogen Energy (PCOHE) is composed of the original English text sub-corpus and its corresponding Chinese translated text sub-corpus. Real-world data is generally full of noise and irrelevant information. These raw texts require a time-consuming three-step data pre-processing to make the data clean and accurate: texts cleaning, alignment and linguistic processing.

3.2.1Texts Cleaning
In general, texts cleaning process consists of digitalization with OCR, material input, manual correction (error codes or spelling mistakes etc.), data denoising and the removing of irrelevant materials (titles, web navigation controls, copyrights and dates, tables, formulas, references and acknowledgements etc.). Tools like python or EmEditor can perform a series of preprocessing on the corpus, including data cleaning, noise removal, feature extraction, data reduction, and data personalization, which make the data more accurate.

3.2.2Alignment
The utmost purpose of compiling parallel corpus is to align bilingual texts. Sentence-level alignment and paragraph-level alignment can be achieved by Trados or Tmxmall, thus facilitating the language research and apply it to specific hydrogen energy industry.

3.2.3 Linguistic Processing
LancsBox can help perform the following series of automatic linguistic processing: POS-tagging, tokenization and lemmatization, word frequency statistic and concordance which facilitate the corpus retrieval and analysis. The value of those corpus analysis lies in the visualization and induction of parallel corpus with language pairs, so as to reduce the cognitive burden of researchers or translators.

IV. Statistical Analysis of PCOHE and Its Applications

A detailed statistical analysis of the English-Chinese Parallel Corpus of Hydrogen Energy is carried out in attempts to explore the linguistic and discourse feature of bilingual hydrogen energy texts. The written part of this bilingual parallel corpus covers 582528 tokens and 25662 types, which includes hydrogen energy texts between 2010 to 2020.The English-Chinese Parallel Corpus of Hydrogen Energy (PCOHE) is comprised of an original English text sub-corpus and its corresponding Chinese translated texts. The translation texts are done by professional translators. Texts compiled are in different forms and published between 2010 and 2020. It was automatically aligned at sentence level, as well as manually post-corrected. Lancbox and Python both can be used for English tokenization. The corpus size and token distribution of corpora are shown in table 1.

Table 1 Texts and Token distribution of the PCOHE

Cprpus/subcorpora	Token	Type	Lemmas
Original English Texts	557994	15578	12932
Chinese translated texts	24534	10863	10564
PCOHE	582528	25662	23014

The sentence length,lexical diversity, lexical densityand lexical sophistication of original English texts and Chinese translated texts can be indicated respectively through the data analysis. It provides convenience for empirical research, comparative linguistics and translation strategy and theory.These in-depth exploration and applicationscan ensure good translation.

Python including ‘jieba’ can be used to obtain wordfrequency statistics. Higher word frequency indicatesmore importance. The functional words like pronouns, numerals, adverbs, adjectives, articles, prepositions, conjunctions

were deleted and only meaningful nominal nouns were saved. The top 20 word frequency among the whole corpus can be seen in Table 2.

Table 2Frequency of Notional Words

No.	Type	Frequency:01-Freq
1	hydrogen	7842
2	fuel	3676
3	energy	3076
4	cell	2506
5	gas	2482
6	storage	1888
7	used	1796
8	system	1726
9	power	1636
10	production	1382
11	temperature	1232
12	electricity	1156
13	pressure	1154
14	water	890
15	sample	854
16	materials	832
17	measurement	784
18	adsorption	784
19	efficiency	778
20	vehicles	750

The frequency of words can be applicable to the compilationthe technical term dictionary within the specialized hydrogen energy domain. PCOHE presents parallel pairs which can help the learners understand the exact meaning of certain words.Similarly, the collocation and association of words and the proper useof specialized vocabulary can be easily shown in the statistical study of the PCOHE, which is conducive to the learning and teaching hydrogen energy, soas to ensure the sustainable development of hydrogen energy industry.

Machine translation is translating the text into one language usingthe software in another language by combining computational and linguistic knowledge.[7] Machine translation requires a large scale bilingual parallel corpusbecause statistical machine translation and neural machine translation systems arebased on probabilistic models, which are created using features extracted from theparallel corpus. PCOHE, as the first bilingual parallel corpus for hydrogen energy industry in China, will contribute to machine translation.

V. Conclusion

This paper aims to build the English-Chinese Parallel Corpus of Hydrogen Energy. The design framework, collecting and processing of corpus materials, and its applications are analyzed. As the first hydrogen energy parallel corpus, it could be applied to many application scenarios for various purpose and can benefit researchers, language learners, as well as dictionaries and textbooks developers in several ways. First, the original English texts provides abundant authentic sources, which assist them with the acquisition of hydrogen energy field knowledge. Second, as it can be seen in Table 2, English-Chinese Parallel Corpus of Hydrogen Energy have a great deal to offer bilingual lexicography and word frequency in hydrogen energy domain, which can be employed by materials developers to compile and develop dictionaries and textbooks related to the specialized field. Most importantly, English to Chinese translation pairs can be extracted in the PCOHE, which can contribute to the translation research and practice. Although intended for translation studies of hydrogen energy texts, PCOHE have provided a solid foundation for machine translation and the teaching of hydrogen energy.

Acknowledgements

This research was supported by Cooperative Education Program of Higher Education Department of Ministry of Education (Grant No. 202002242018).

References

[1] W. Xie, X. Wang, "Building a Parallel Corpus for English Translation Teaching Based on Computer-Aided Translation Software," Journal of Computer-Aided Design & Applications, 18(S3), pp. 12-22, 2021.

[2] M.C. Liang, W. Z.Li, J.J. Li, "Using Corpora: A Practical Coursebook," Foreign Language Teaching and Research Press, 2020.

[3] C.Sammut, G.I. Webb, "Parallel Corpus" In: Encyclopedia of Machine Learning,"Springer, Boston, MA. 2011.

[4] D.Liu, J.Liu, K.Lin, et al., "Research on the Construction and Related Problems of Industrial ChineseParallel Corpus Based on Equipment Manufacturing Industry," Journal of software, vol. 42, no.1, pp. 8-11, 2021.

[5] D. Liu et al.,"Establishment of Parallel Text Corpus of Equipment ManufacturingIndustry Based on Data Mining Technology," In: Proceedings of the 2nd International Conference on Computing and Data Science. 2021.

[6] Biber, D.& B.Gray. "Nominalizing the verb phrase in academic science writing," In B. Aarts, J. Close, G.Leech&S. Wallis(eds.). The Verb Phrase in English: Investigating Recent Language Change with Corpora. Cambridge University Press.pp.99-132,2010.

[7] S.Zhao, "English corpus translation system based on FPGA and machine learning," Journal of Microprocessors and Microsystems, vol. 15, no. 5, pp. 594-600, 2020.