Building a Parallel Corpus for Chinese Folk Songs Translation Studies: A Case Study of Northern Shaanxi and Hua'er Folk Songs

Article *in* Theory and Practice in Language Studies · February 2024

DOI: 10.17507/tpls.1402.17

Yan Lin
Universiti Putra Malaysia
4 PUBLICATIONS **12** CITATIONS

Hazlina Abdul Halim
Universiti Putra Malaysia
65 PUBLICATIONS **94** CITATIONS

Farhana Muslim Mohd Jalis
Universiti Putra Malaysia
13 PUBLICATIONS **7** CITATIONS

ISSN 1799-2591
Theory and Practice in Language Studies, Vol. 14, No. 2, pp. 454-468, February 2024
DOI: https://doi.org/10.17507/tpls.1402.17

Building a Parallel Corpus for Chinese Folk Songs Translation Studies: A Case Study of Northern Shaanxi and Hua'er Folk Songs

Yan Lin[*]
Department of Foreign Languages, Faculty of Modern Languages and Communication, Universiti Putra Malaysia,
43400 Serdang, Selangor, Malaysia

Hazlina Abdul Halim
Department of Foreign Languages, Faculty of Modern Languages and Communication, Universiti Putra Malaysia,
43400 Serdang, Selangor, Malaysia

Farhana Muslim Mohd Jalis
Department of Foreign Languages, Faculty of Modern Languages and Communication, Universiti Putra Malaysia,
43400 Serdang, Selangor, Malaysia

***Abstract*—Folk songs, collaboratively created by the public and transmitted orally, have gained widespread popularity. The translation of folk songs primarily centers on lyrics translation, a subset of literary translation. Recent advancements in corpus technology have highlighted the significance of corpus-based research approaches for the analysis of literary translation. The corpus method, now employed as a hybrid research approach, enables the generation of quantitative data for descriptive translation studies. Scholars are increasingly using parallel corpora containing both the source text (ST) and the target text (TT) to explore translation universals across diverse texts. Despite the growing body of literature on the translation of Chinese folk songs, most studies have involved straightforward analyses of a limited number of translated texts without the utilization of quantitative approaches. This article aims to bridge this gap by presenting a prospective study on the creation of the Chinese-English Parallel Corpus of Northern Shaanxi and Hua'er Folk Songs (CEPCNSHFS). The study covers essential aspects such as sampling, corpus structure, corpora selection, and corpora processing. Moreover, to assess the practical utility of the CEPCNSHFS, a pilot study was conducted. The primary contributions of this article reside in the potential of the CEPCNSHFS to support diverse research topics, including the exploration of translation language characteristics, styles, and methods employed in translating Northern Shaanxi and Hua'er folk songs, both of which hold significant positions within Chinese folk song traditions.**

***Index Terms*—parallel corpus, Northern Shaanxi and Hua'er folk songs, translation of Chinese folk songs into English, translation studies**

I. Introduction

The 1960s marked a significant development in the adoption of the corpus methodology, a research approach involving the compilation of extensive text collections based on predetermined selection criteria. The evolution of computer and network technologies ultimately resulted in the emergence of corpus linguistics as a distinct field of linguistic inquiry. A corpus-based approach offers the advantage of providing both quantitative data and qualitative insights for linguistic studies.

As of now, numerous well-established corpora are widely employed in linguistics research. These include the British National Corpus (BNC), Corpus of Contemporary American English (COCA), Corpus of Global Web-based English (GloWbE), iWeb Corpus, News on the Web (NOW) Corpus, Wikipedia Corpus, Hansard Corpus (representing the British Parliament's proceedings), and the Coronavirus Corpus.

In China, notable corpora comprise the Peking University Centre for Chinese Linguistics (CCL), the Tsinghua University THCHS-30 (a freely accessible database of Chinese speech invaluable for building comprehensive Chinese speech recognition systems), and the ICSA (Institute of Corpus Studies and Applications) Platform for Corpus Search and Applications, developed by Shanghai International Studies University (SISU).

In 1993, Mona Baker's seminal article titled "Corpus linguistics and Translation Studies: Implications and Applications" marked the beginning of corpus-based translation studies. The application of corpus methodology within translation studies has given rise to the field of corpus translatology and has generated a sustained demand for

[*] Corresponding Author. Email: linyan406@163.com

translation corpora (Hu & Tao, 2009). Translation corpora can be categorized into three main types: translation corpora, analogy corpora, and comparable or parallel corpora (Wang, 2004). Among these, a parallel corpus is the most commonly utilized, as it allows for the comparative analysis of original and translated texts through retrieval statistics.

In the realm of Chinese folk song research, scholars have undertaken efforts to establish monolingual corpora for various research purposes, including the study of musical communication, cultural introduction, and the preservation of this intangible cultural heritage. However, in the domain of Chinese folk song translation research, there is a notable gap that necessitates the creation of a parallel corpus to address this deficiency.

Known as “the king of songs” in the northwest region of China, Northern Shaanxi folk songs specifically pertain to those originating from Northern Shaanxi, as well as the broader regions of Yulin and Yan’an. In June 2008, the State Council of China officially designated Northern Shaanxi folk songs as a second-category national intangible cultural treasure. Among the diverse ethnic groups in China, Hua’er is a renowned folk song that resonates among the Hui, Tibetan, Yugu, Sala, Tu, Dongxiang, Han, Baoan, and Mongolian communities. Hua’er, often treated as “the soul of northwest China”, possesses distinctive regional and ethnic characteristics. It was among the pioneering initiatives recognized as part of China’s efforts to preserve intangible cultural assets. In 2009, Hua’er was also included in the list of intangible cultural heritages requiring international preservation efforts.

This article will focus on the construction of the Chinese-English Parallel Corpus of Northern Shaanxi and Hua’er Folk Songs (CEPCNSHFS) with the overarching aim of creating a substantial parallel corpus for translation studies. Such a corpus will serve as a valuable resource for investigating a wide array of topics, including but not limited to translation methods, translator profiles, translation universals, and more. The resulting findings are expected to foster the adoption of the parallel corpus approach within the field of Northern Shaanxi and Hua’er folk song translation research. Additionally, the corpus is poised to become a valuable reference tool for Chinese folk song translators and researchers, enriching their endeavors in this vibrant domain.

II. Related Work

The development and utilization of corpora, the majority of which encompass texts from diverse genres, often enriched with handwritten or manually validated annotations, stand as pivotal advancements in the realm of linguistic research methodologies and tools (Wang et al., 2021). Notably, the introduction of corpora such as the Brown Corpus and the Lancaster-Oslo-Bergen (LOB) Corpus has provided numerous corpus linguists with a straightforward means to delve into authentic texts through the use of corpus software. These accessible corpora and the methodologies grounded in corpus linguistics have undeniably yielded invaluable data for the field of translation studies.

The impact of corpus-based research became evident long before it significantly influenced the theoretical and descriptive aspects of translation studies, as emphasized by Baker (1993). A significant development in this context is the Translational English Corpus (TEC), located at the Centre for Translation and Intercultural Studies. TEC is a novel corpus designed specifically for translation studies, consisting of written texts translated into English from a wide range of source languages.

Baker (1995, p. 230) categorized three primary types of corpora within the field of translation studies: comparable, multilingual, and parallel corpora. These corpus types are classified as follows:

- Comparable Corpora: These are composed of two or more monolingual corpora that are not translations of each other and are not synchronized, but they share a common subject or topic. Comparable corpora allow for cross-linguistic comparison and analysis of texts on a similar theme or subject matter.
- Multilingual Corpora: These corpora consist of texts that have been translated into multiple languages and are structured similarly to parallel corpora. The key distinction is that multilingual corpora focus on texts translated into several languages, enabling the examination of translation variations across multiple target languages.
- Parallel Corpora: Parallel corpora encompass translated language equivalents of texts in the source language (language A) and their corresponding translations in the target language (language B). These corpora facilitate the direct alignment of original and translated texts, enabling detailed contrastive analysis.

Additionally, in the domain of translation research, bilingual parallel corpora are often developed or studied for various research objectives. These corpora contain both the source language texts and their translated counterparts, all centered around a common subject or theme. This configuration allows researchers to explore translation phenomena and language contrasts within a controlled and aligned context.

Koehn (2005) played a pivotal role in building a corpus of parallel texts sourced from European Parliament proceedings across 11 languages. This corpus was made accessible online and served as invaluable training material for the development of statistical machine translation (SMT) systems. Lapshinova-Koltunski et al. (2018) undertook the creation of a parallel corpus that featured comprehensive coreference chain annotations. This innovative approach addressed a significant challenge associated with coreference translation across languages, a problem that has implications for machine translation and other multilingual natural language processing (NLP) technologies. Yuan (2020) made significant contributions by developing a parallel corpus containing Russian and Chinese languages, comprising 27,664 sentence pairs and a staggering 2,010,632 words. This corpus was instrumental in the extraction of phrases relevant to the context of political diplomacy. In a similar vein, Wang (2021) capitalized on machine translation software to establish a parallel corpus tailored for teaching English translation, facilitating language learning and translation instruction. Guo and Zhou (2019) played a crucial role in the creation of a substantial English-Chinese parallel corpus, which was specifically geared towards popular science content. This corpus spans both sides of the Taiwan Strait and serves as a valuable resource for linguistic and cross-cultural studies in the realm of popular science communication.

The predominant corpus resources employed for the investigation of Chinese folk songs have primarily originated from music research (Guo, 2020), literary research (Zhou, 2017), and retrieval research (Zhang, 2010), and are mostly monolingual. One notable example is the ACCESS database, recognized as the first Chinese folk song corpus. Comprising over 2,000 samples of Northern Shaanxi folk songs, ACCESS serves as a valuable resource of extensive firsthand information pertaining to Northern Shaanxi folk songs. However, the availability of untagged raw texts and the fact that it is monolingual restrict its applicability in the context of translation research.

Beyond its designation as a category in Chinese literature, Northern Shaanxi and Hua’er folk songs occupy a special role as means of disseminating Chinese culture. Given the wide-ranging significance, it becomes essential to create a Chinese-English parallel corpus that includes both the source texts (ST) and target texts (TT) of Northern Shaanxi and Hua’er folk songs. Such a corpus has the potential to yield illustrative and quantitative data crucial for the advancement of translation studies within this domain.

III. Corpus Construction Procedure

Corpus construction represents the initial and key phase in any corpus investigation, as the components of the corpus and the selection procedures employed have far-reaching implications for all subsequent research processes (Sinclair, 1999). It is imperative to ascertain whether existing corpora are available for your research, as building a corpus demands significant time and effort. Throughout this research, it has been revealed that there is currently no existing parallel corpus specifically dedicated to Northern Shaanxi and Hua’er folk songs for the purpose of translation studies.

Furthermore, it is essential to construct a corpus with a clear understanding of its intended purpose, construction methodology, and theoretical as well as practical relevance (Zhang, 2020). These considerations directly influence whether the corpus can effectively address the research objectives. This article undertakes the creation of the Chinese-English Parallel Corpus of Northern Shaanxi and Hua’er Folk Songs (CEPCNSHFS) with the specific aim of providing valuable corpus data for the study of Northern Shaanxi and Hua’er folk songs translation.

The CEPCNSHFS will take the form of a parallel corpus covering both the Chinese texts of Northern Shaanxi and Hua’er folk songs and their corresponding English translations. In this section, we will explore the detailed construction process of the CEPCNSHFS.

A. Sampling & Corpus Structure

According to Baker (2010), sampling, balancing, and representativeness are the three key theoretical principles that support corpus linguistics. Since it is difficult to determine the makeup of the complete linguistic community from a sample of isolated linguistic units, conventional statistical sampling procedures are inadequate for building a language corpus (Atkins et al., 1992). However, the corpus methodology offers the flexibility to sample specific words, sentences, paragraphs, or even entire texts, depending on the research objectives. The goals of the study must be taken into consideration while choosing the sampling scope, which may include words, phrases, paragraphs, or entire texts. Additionally, in instances where the text’s length exceeds the requirements of the research goals, it may become necessary to extract text fragments from the entirety of the text during the corpus selection process.

Due to the fact that Northern Shaanxi and Hua’er folk songs frequently follow a concise and clear pattern resembling Chinese ancient poetry (Wang, 2014), which is distinguished by brevity and few lines, the CEPCNSHFS will adopt a sampling approach that requires choosing the full length of each song’s text as a sample. This method aligns with the inherent structure of these folk songs, ensuring comprehensive coverage in the corpus for this study.

To maintain balance and prevent any individual texts from disproportionately influencing the corpus as a whole, the selection of texts for inclusion in a corpus must be conducted with careful consideration. A corpus should aim to be representative of a certain language, language variety, or topic, therefore maintaining a balance is crucial. In the case of Chinese folk songs, which can be categorized into different types based on geographical regions, it is important to mention that Northern Shaanxi and Hua’er folk songs, chosen as the focus for the CEPCNSHFS, represent the musical traditions of northwest China. Candidate texts for the CEPCNSHFS are carefully chosen to cover a diverse range of subjects and to maintain an equitable distribution across the same time span. This method prevents any undue skewing of the corpus’s overall composition and guarantees that the corpus remains representative of the vast and varied terrain of Northern Shaanxi and Hua’er folk songs. As a result, once the sampling is complete, the CEPCNSHFS structure, which is indicated in Figure 1, will consist of two parts.

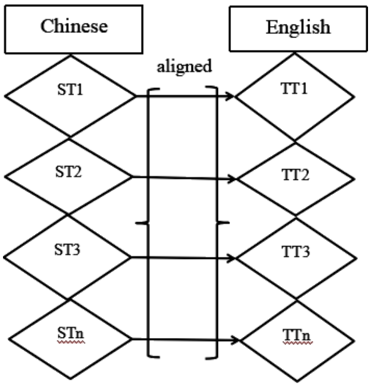


Figure 1. The Mapping Structure of Texts in the CEPCNSHFS

The CEPCNSHFS comprises a significant body of texts (ST1~STn), each thoughtfully matched with its corresponding English translations (TT1~TTn). This alignment occurs at the sentence level, a structural choice aligned with the specific research objectives.

B. Corpora Selection

Texts, comprising coherent sentences and paragraphs, serve as the fundamental building blocks of corpora in the realm of translation studies. Hence, the selection of suitable texts from printed books holds significant value. These texts, which consist of an essential component of the CEPCNSHFS, were carefully selected from four published books. Table 1 provides a comprehensive overview of book titles, the names of translators, publication years, publisher, and the total count of songs featured in these books.

TABLE 1
CORPORA RESOURCE FOR THE CEPCNSHFS

Folk song	Book Title	Translator	Songs	Publishing Year	Publisher
Northern Shaanxi folk song	Voice from the Northwest-Folk Songs of Northern Shaanxi	Wang Hongyin	105	2009	Culture and Art Publishing House
	Lyrics of Northern Shaanxi Folk Songs:Translation and Exegesis	Wang Zhanbin	88	2021	Nankai University Press
Hua’er	Hua’er-Folk Songs from the Silk Road	Yang Xiaoli & Caroline Elizabeth Kano	88	2016	The Commercial Press
	Voice from the Silk Road-Hua’er of Northwest China	Yang Xiaoli et al.	105	2022	The Commercial Press

The CEPCNSHFS covers two sub-corpora: Northern Shaanxi folk songs and Hua’er folk songs, each totaling 193 songs, respectively. As shown in Table 1, “Voice from the Northwest-Folk Songs of Northern Shaanxi”, consisting of 105 folk songs, was published by the Culture and Art Publishing House in 2009. This collection represents a nine-year translation endeavor by Wang Hongyin, a professor at Nankai University (Li, 2009). Wang Zhanbin, a Professor of Commerce at Tianjin University, translated 90 songs featured in the book “Lyrics of Northern Shaanxi Folk Songs: Translation and Exegesis”, which was released in 2021 by Nankai University Press.

A condensed translation of “Comprehensive Discussion on Chinese Hua’er Songs” by Wu (2008) resulted in the book “Hua’er-Folk Songs from the Silk Road”, published by the Commercial Press in 2016. This publication contains approximately 240 songs and provides an overview of Hua’er, a distinctive type of folk song found exclusively in northwest China. Following six years of preparation, in June 2022, the Commercial Press released “Voice from the Silk Road-Hua’er of Northwest China”, a book containing 105 songs. This book serves as a companion to “Hua’er-Folk Songs from the Silk Road”. To better align the CEPCNSHFS with our research objectives, we chose to select only 88 songs from two books: “Lyrics of Northern Shaanxi Folk Songs: Translation and Exegesis” and “Hua’er-Folk Songs from the Silk Road”, in the corpus.

The translators of these four books are all Chinese university teachers with extensive expertise in both translation practice and academic research. This background assures the quality of their translation versions. Additionally, their English translations are currently the most widely recognized and representative in the field of translating Northern Shaanxi and Hua’er folk songs.

C. Corpora Processing

With the growth of the Internet and personal computer usage, an increasing number of corpora are becoming accessible in the form of digitized texts that are machine-readable. Corpora processing indeed involves the conversion of candidate corpora into machine-readable files. The processing of corpora for the CEPCNSHFS was conducted through a five-stage approach, which included raw data acquisition, noise filtering, metadata markup, segmentation & POS tagging, and parallel alignment.

(a). Raw Data Acquisition

Raw data can be acquired through various means, including scanning paper books, gathering information online, or manual typing. One of the most prevalent methods for digitizing content is scanning, which transforms physical books into editable digital texts suitable for computer use. Books can undergo scanning processes and be converted into editable file formats like Word, Excel, or plain text with the assistance of optical character recognition (OCR) software or programs such as OneNote, ABBYY FineReader, and others.

When scanning paper volumes to obtain raw texts for the CEPCNSHFS, it is essential to remove the table of contents, prologue, images, footnotes, references, and any other materials not required for the corpus. After this editing process to ensure accuracy, the raw corpora are saved as plain text files in UTF-8 format for the subsequent stage. Each Chinese file may be labeled with the title “cn_00x” (x=1~n), which corresponds to “en_00x” (x=1~n) for the English version. Each file should go through a comprehensive review and revision process to ensure the utmost accuracy of the texts.

(b). Noise Filtering

Standardizing the text format is crucial to make corpora more readable and compatible with computers. Once the key texts have been extracted from the original paper books or Portable Document Format (PDF) files, the raw data should be improved and cleaned through noise filtering. Noise filtering entails removing extra spaces from the obtained texts. Consequently, it’s important to avoid using white spaces and tab spaces for Chinese characters, whereas for English and closely related languages, only one space should be used between two words. In this study, unnecessary spaces can be eliminated using text editors like EditPlus or the “search-replace” feature in a WORD document. For instance, if the command “^p” is used to replace “^p^p”, it will reduce the text to only one space instead of two. In addition, spaces between paragraphs and at the beginning of each paragraph are also removed in this stage.

(c). Metadata Markup

Annotation involves the use of tags to mark various text features in a corpus to serve research objectives. Common types of annotation include voice and speech mistake tagging, POS tagging, syntactic tagging, semantic tagging, pragmatic tagging, and metadata markup. In the context of this work, metadata markup involves applying a standardized set of readable labels or tags to annotate detailed information in the text.

For the CEPCNSHFS, metadata annotations were performed using the Corpus of Contemporary American English (COCA) annotation system. This system is characterized by enclosing annotation symbols and contents within “< >” brackets. As a result, the “<p>” and “</p>” tags are used to indicate the start and end of each paragraph, respectively. “<S>” and “</S>” are employed as sentence annotators, while “<H>” and “</H>” signify header information for each text. Given the structural similarity between Northern Shaanxi and Hua’er folk songs and Chinese ancient poetry, the annotation framework for the CEPCNSHFS follows the structure depicted in Figure 2.

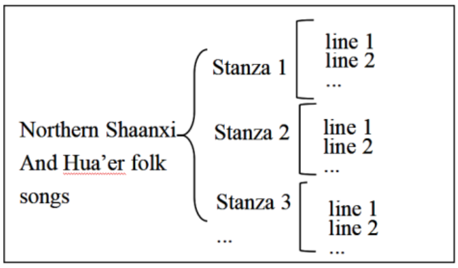
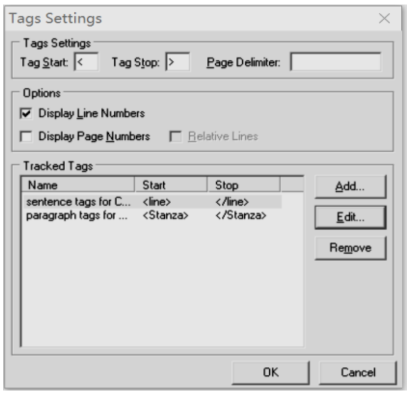


Figure 2. Annotation Framework of Northern Shaanxi and Hua’er Folk Songs

Just like Chinese ancient poetry, Northern Shaanxi and Hua’er folk songs typically consist of multiple stanzas, with each stanza containing several lines. Consequently, it is essential to configure the software tags in preparation for alignment. Figure 3 illustrates the tag settings for ParaConc.



Tags Settings

Tag Start: <   Tag Stop: >   Page Delimiter:

Options

☑ Display Line Numbers

□ Display Page Numbers   □ Relative Lines

Tracked Tags

Name	Start	Stop
sentence tags for C...	<line>	</line>
paragraph tags for ...	<Stanza>	</Stanza>

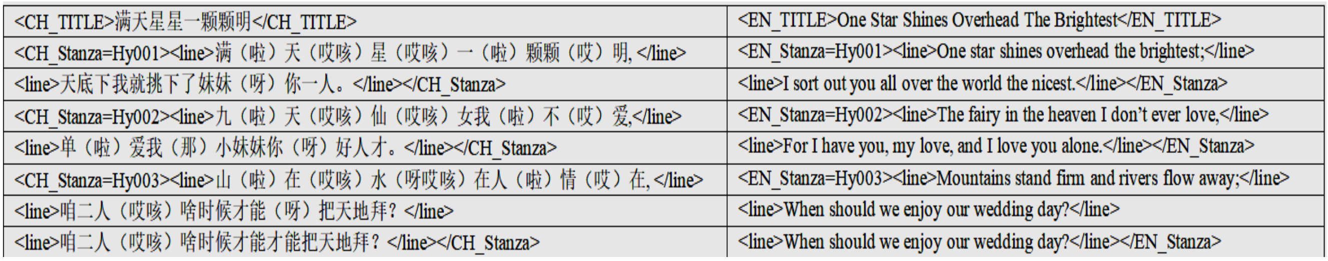
Figure 3. ParaConc Tag Settings

In ParaConc, sentence tags for Northern Shaanxi and Hua'er folk songs are set as <line>...</line>, while paragraphs are represented as <stanza>...</stanza>. Table 2 provides a list of the additional tags used for the CEPCNSHFS, apart from sentences and paragraphs.

TABLE 2
THE METADATA FOR THE CEPCNSHFS

Content	Tag	Content	Tag
Chinese title	<CH_TITLE>...</CH_TITLE>	gender	<GENDER>...</GENDER>
English title	<EN_TITLE>...</EN_TITLE>	publication date	<PUB_DATE>...</PUB_DATE>
subtitle	<SUBTITLE>...</SUBTITLE>	publication place	<PUB_PLACE>...</PUB_PLACE>
translator's name	<TRANSLATOR>...</TRANSLATOR>	publisher	<PUBLISHER>...</PUBLISHER>
author's name	<AUTHOR>...</AUTHOR>	words	<WORDS>...</WORDS>

These tags are employed to annotate crucial information in the corpus, including details such as the title, subtitle, translator's name, author's name, gender, publication information, publication location, publisher, terms, and more. For example,



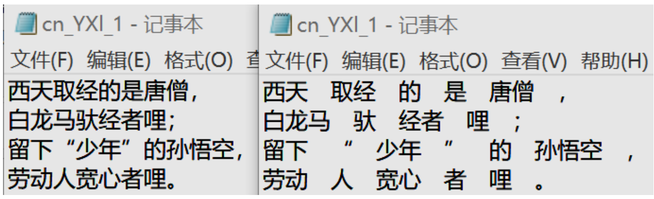
<CH_TITLE>满天星星一颗颗明</CH_TITLE>	<EN_TITLE>One Star Shines Overhead The Brightest</EN_TITLE>
<CH_Stanza=Hy001><line>满（啦）天（哎咳）星（哎咳）一（啦）颗颗（哎）明, </line>	<EN_Stanza=Hy001><line>One star shines overhead the brightest;</line>
<line>天底下我就挑下了妹妹（呀）你一人。</line></CH_Stanza>	<line>I sort out you all over the world the nicest.</line></EN_Stanza>
<CH_Stanza=Hy002><line>九（啦）天（哎咳）仙（哎咳）女我（啦）不（哎）爱,</line>	<EN_Stanza=Hy002><line>The fairy in the heaven I don't ever love,</line>
<line>单（啦）爱我（那）小妹妹你（呀）好人才。</line></CH_Stanza>	<line>For I have you, my love, and I love you alone.</line></EN_Stanza>
<CH_Stanza=Hy003><line>山（啦）在（哎咳）水（呀哎咳）在人（啦）情（哎）在, </line>	<EN_Stanza=Hy003><line>Mountains stand firm and rivers flow away;</line>
<line>咱二人（哎咳）啥时候才能（呀）把天地拜? </line>	<line>When should we enjoy our wedding day?</line>
<line>咱二人（哎咳）啥时候才能才能把天地拜? </line></CH_Stanza>	<line>When should we enjoy our wedding day?</line></EN_Stanza>

Figure 4. An Example of Metadata Annotation

Figure 4 illustrates the metadata annotation for a folk song from Northern Shaanxi in the source text (ST) and its English translation (TT). This annotation serves to facilitate information retrieval within the CEPCNSHFS through corpus software. And the extent of annotation may vary depending on the study's objectives.

(d). Segmentation & POS Tagging

Word segmentation is most frequently applied to the source texts (ST) of Northern Shaanxi and Hua'er folk songs. This process involves dividing a sequence of characters into distinct and recognizable components. Since English words are separated by spaces, it is essential to segment Chinese texts into individual words or insert spaces between each Chinese character to facilitate retrieval using corpus software. Various methods can be employed for segmentation, such as manually creating gaps between Chinese characters using the "search-replace" tool in Word documents or employing corpus software for automatic segmentation. Figure 5 provides an example of how the text in the ST of Northern Shaanxi and Hua'er folk songs was segmented using SegmentAnt in this study.


cn_YXl_1 - 记事本
文件(F) 编辑(E) 格式(O) 查
cn_YXl_1 - 记事本
文件(F) 编辑(E) 格式(O) 查看(V) 帮助(H)

西天取经的是唐僧，
白龙马驮经者哩；
留下“少年”的孙悟空，
劳动人宽心者哩。

西天 取经 的 是 唐僧 ，
白龙马 驮 经者 哩 ；
留下 “ 少年 ” 的 孙悟空 ，
劳动 人 宽心 者 哩 。

Figure 5. Segmentation in the ST of Northern Shaanxi and Hua'er Folk Songs

Figure 5 indicates the format of the text before segmentation on the left side, where Chinese characters are not spaced, and the format after segmentation on the right side. Following segmentation, each line of this Hua'er folk song was divided into several units, typically about four or five pieces. For instance, based on the basic usage or function of certain words or phrases, the first line "西天取经的是唐僧" was segmented into "西天", "取经", "的", "是", and "唐僧" with spaces.

Part of speech (POS) tagging involves categorizing words in a sentence or group of words in a paragraph into specific grammatical categories such as verbs, adjectives, adverbs, nouns, etc., based on their contextual usage (Chiche & Yitagesu, 2022). This process helps to understand the grammatical functions of words within a corpus. In the ST of Northern Shaanxi and Hua'er folk songs, POS tagging was carried out using CorpusWordParser, a tool owned by the Chinese Ministry of Education Institute of Language Application. For the TT of Northern Shaanxi and Hua'er folk songs, POS tagging was performed using TreeTagger, a tool developed by the Institute for Computational Linguistics at the University of Stuttgart. Figure 6 shows an example of POS tagging in the ST of a Hua'er folk song, generated using CorpusWordParser.

```
日头/n
月亮/n 盘场/ns 是/vl 刮风/v 哩/u，/w
日头/n 盘场/ns 是/vl 下/v 哩/u。/w
女婿/n 知道/v 是/vl 活杀/v 哩/u，/w
公婆/n 们/k 知道/v 是/vl 骂/v 哩/u！/w
```

Figure 6. POS Tagging by CorpusWordParser in the ST of a Hua'er Folk Song

In this example, characters like "日头", "月亮", "女婿", and "公婆" are marked with the "/n" symbol, indicating that they are "nouns" in the ST of this Hua'er folk song. Other tags include "/ns" for place names[1], "/vl" for linking verbs, "/u" for auxiliary words, "/k" for proclitics, and "/w" for punctuation. The information in Table 3 provides a guide for understanding and exploring these tags that have been assigned by CorpusWordParser.

TABLE 3
CORPUSWORDPARSER TAGSETS

Tag	Meaning	Tag	Meaning	Tag	Meaning
/n	noun	/nt	time noun	/d	adverb
/nh	human name	/p	preposition	/u	particle
/ni	institution name	/o	onomatopoeia	/c	conjunction
/vl	link verb	/nd	location noun	/ns	place name
/m	numeral	/v	verb	/e	interjection
/r	pronoun	/a	adjective	/w	punctuation

Indeed, every software tool has its own tagset, comprising a range of tags. Table 4 shows the tags established by TreeTagger for reference.

TABLE 4
TREETAGGER TAGSETS

Tag	Meaning
DT	Article and determiner
IN	Preposition or subordinating conjunction
MD	Modal verb
NN	Common noun, singular or mass
NNS	Common noun, plural
PP	Personal pronoun
PP$	Possessive pronoun
VV	Lexical verb, base form (e.g. live)
VVD	Past tense verb of lexical verb (e.g. lived)
VVP	Present tense (other than third-person singular) of lexical verb (live)
VVZ	Present tense (third-person singular) of lexical verb (lives)

These tags serve as a kind of roadmap for researchers, assisting in data analysis and interpretation. Figure 7 displays the result of POS tagging in the TT of the Hua'er folk song from Figure 6, achieved through the use of TreeTagger.

```
The_DT halo_NN around_IN the_DT moon_NN forecasts__VVZ wind_VVP,_,
The_DT halo_NN around_IN the _DT sun_NN foretells_VVZ rain_NN._SENT
If_IN my_PP$ husband_NN knew_VVD,_, he_PP would_MD blame_VV me_PP,_,
If_IN my_PP$ parents-in-law_NN heard_VVD,_, they_PP would_MD scold_VV me_PP._SENT
```

Figure 7. POS Tagging by TreeTagger in the TT of a Hua'er Folk Song

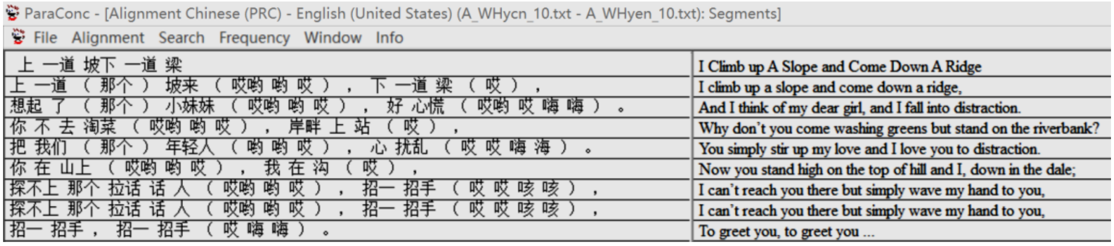
The TT of this Hua'er folk song can be understood and analyzed based on the tags provided in Table 4. To make the underlying meaning of the ST more visible and comprehensible to the TT reader, the translator rendered the last two lines, "女婿知道是活杀哩,/公婆知道是骂哩" into two hypothetical if-clauses: "If my husband knew, he would blame me,/If my parents-in-law heard, they would scold me". These four verbs- "刮风" (forecasts wind), "下" (foretells rain), "活杀" (blame), and "骂" (scold)-were faithfully translated in the TT, appearing in each line of the ST, respectively. It has been observed that POS tagging the corpora will facilitate linguistic comparison between the ST and TT of Northern Shaanxi and Hua'er folk songs. In addition to tagging words or sentences for the study of lexical and syntactic features in the ST and TT, other data such as rhetorical devices, translation strategies and methods, translation universals, and more can also be tagged and explored.

(e). Parallel Alignment

The term "alignment" refers to the process of matching up the translated and original texts. Given that there are not many relevant factors requiring further investigation, text or paragraph alignment may be insufficient for translation studies. Moreover, determining word alignment is challenging due to significant differences between Chinese and English. Hence, sentence alignment is recommended for research on Northern Shaanxi and Hua'er folk songs

[1] The Chinese character "盘场" in this song does not refer to a place name. Instead, it is a dialect term from northwestern China that alludes to the natural phenomenon of a halo, a circle that forms around the moon in the previous evening, indicating that it will be windy the following day.

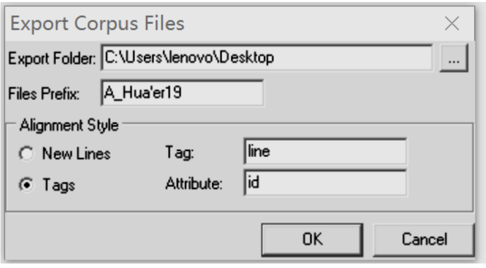
translation. In practice, there are numerous approaches to parallel alignment, which can be carried out manually or automatically using computer programs such as Trados, SISU Aligner, ParaConc, and others. In this study, sentence alignment for Northern Shaanxi and Hua’er folk songs is conducted using ParaConc, and the result is displayed in Figure 8.


ParaConc - [Alignment Chinese (PRC) - English (United States) (A_WHycn_10.txt - A_WHyen_10.txt): Segments]
File Alignment Search Frequency Window Info

上 一道 坡下 一道 梁	I Climb up A Slope and Come Down A Ridge
上 一道 （ 那个 ） 坡来 （ 哎哟 哟 哎 ） ， 下 一道 梁 （ 哎 ） ，	I climb up a slope and come down a ridge,
想起 了 （ 那个 ） 小妹妹 （ 哎哟 哟 哎 ） ， 好 心慌 （ 哎哟 哎 嗨 嗨 ） 。	And I think of my dear girl, and I fall into distraction.
你 不 去 淘菜 （ 哎哟 哟 哎 ） ， 岸畔 上 站 （ 哎 ） ，	Why don’t you come washing greens but stand on the riverbank?
把 我们 （ 那个 ） 年轻人 （ 哟 哟 哎 ） ， 心 扰乱 （ 哎 哎 嗨 海 ） 。	You simply stir up my love and I love you to distraction.
你 在 山上 （ 哎哟 哟 哎 ） ， 我 在 沟 （ 哎 ） ，	Now you stand high on the top of hill and I, down in the dale;
探不上 那个 拉话 话 人 （ 哎哟 哟 哎 ） ， 招一 招手 （ 哎 哎 咳 咳 ） ，	I can’t reach you there but simply wave my hand to you,
探不上 那个 拉话 话 人 （ 哎哟 哟 哎 ） ， 招一 招手 （ 哎 哎 咳 咳 ） ，	I can’t reach you there but simply wave my hand to you,
招一 招手 ， 招一 招手 （ 哎 嗨 嗨 ） 。	To greet you, to greet you ...

Figure 8. Sentence Alignment by ParaConc

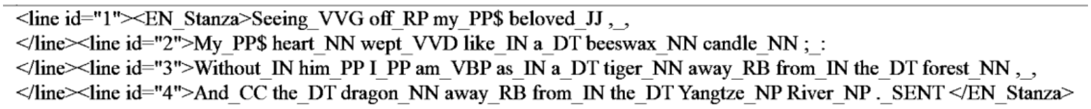
In the example provided, each sentence of the Northern Shaanxi folk song is aligned separately. The user interface showcases how ParaConc arranges the source text (ST) on the left and the translation (TT) on the right, facilitating the comparison between the ST and TT for researchers, allowing them to track specific words or phrases. Once completed, the parallel alignment files can be saved in either workspace or “txt” formats. However, it is essential to set the tags before saving, as demonstrated in Figure 9.



Export Folder: C:\Users\lenovo\Desktop
Files Prefix: A_Hua'er19
Alignment Style
○ New Lines Tag: line
● Tags Attribute: id

Figure 9. Tag Setting Before Exporting the Alignment Files in ParaConc

As a result, when the “seg” tag is changed to “line” and the attribute “id” is set, each line of a folk song will be automatically numbered from <line id= “1”> to <line id= “n”>. Figure 10 offers an example of the completed alignment file for the translation text (TT) of a Hua’er folk song.



```
<line id="1"><EN_Stanza>Seeing_VVG off_RP my_PP$ beloved_JJ ,_,
</line><line id="2">My_PP$ heart_NN wept_VVD like_IN a_DT beeswax_NN candle_NN ;_:
</line><line id="3">Without_IN him_PP I_PP am_VBP as_IN a_DT tiger_NN away_RB from_IN the_DT forest_NN ,_,
</line><line id="4">And_CC the_DT dragon_NN away_RB from_IN the_DT Yangtze_NP River_NP ._SENT </EN_Stanza>
```

Figure 10. The Alignment File for the TT of a Hua’er Folk Song

Therefore, as seen in Figure 10, the completed alignment file can serve as a valuable resource for research on Northern Shaanxi and Hua’er folk songs translations, as it contains both the Chinese lyrics and their English translations along with any segmentation, POS tagging, or other annotations. Since each song has very few sentences, the alignment quality is quite high. At this stage, the corpus processing is complete, and the exported files are ready for research use. It’s worth noting that postscripts, translators’ assessments, and book introductions can be stored in separate files for future reference.

IV. Result and Evaluation

This section provides essential details about the CEPCNSHFS and assesses its applicability in research related to folk song translation. It includes preliminary studies that utilize various analytical techniques, such as Word Frequency analysis, Concordance analysis, Cluster analysis, and N-grams analysis, to demonstrate the corpus’s potential utility.

The CEPCNSHFS consists of 118,497 characters for the ST with 15,491 tokens and 3,677 word types. For the TT, it contains 27,719 words with 3,240 word types. In the CEPCNSHFS, two bilingual sub-corpora are established: one for Northern Shaanxi folk songs (Sub_PC 1) and the other for Hua’er folk songs (Sub_PC 2). Sub_PC 1 contains 26,576 word tokens and 5,721 word types, while Sub_PC 2 contains 9,184 word tokens and 2,818 word types.

Indeed, the Type-Token Ratio (TTR) and word density are important metrics in corpus linguistics and text analysis. The TTR is calculated as (number of types/number of tokens) * 100%. Word density is measured as “number of words/total text length”. A high TTR might indicate a rich and varied vocabulary in the songs, while word density can provide insights into the songs’ efficiency in conveying cultural and lyrical content. Researchers can use these indicators to gain a quantitative understanding of the linguistic aspects of the songs and how they relate to translation and cultural preservation.

The most frequently used words and expressions in Northern Shaanxi and Hua'er folk songs, in both the source language (SL) and the target language (TL), can be identified via observing word frequency in the CEPCNSHFS. AntConc's Keyword List Tool is a valuable feature for identifying phrases that are particularly common or distinctive in a specific corpus compared to a reference corpus (Anthony, 2004). The findings of the word frequency analysis, which are presented in Table 5, can be used as a foundation for deeper research on the cultural and translational elements of Northern Shaanxi and Hua'er folk songs. It provides quantitative data that can complement qualitative analyses of the songs' content and translation choices.

TABLE 5
WORD FREQUENCY LIST

Rank	Chinese		English	
	Word	Frequency	Word	Frequency
1	的	948	the	1,712
2	你	601	and	1,067
3	我	468	I	957
4	了	408	you	864
5	**呦**	294	a	829
6	那个	278	my	642
7	哥哥	272	to	601
8	**呀**	269	is	535
9	**哎**	253	in	495
10	是	228	of	467

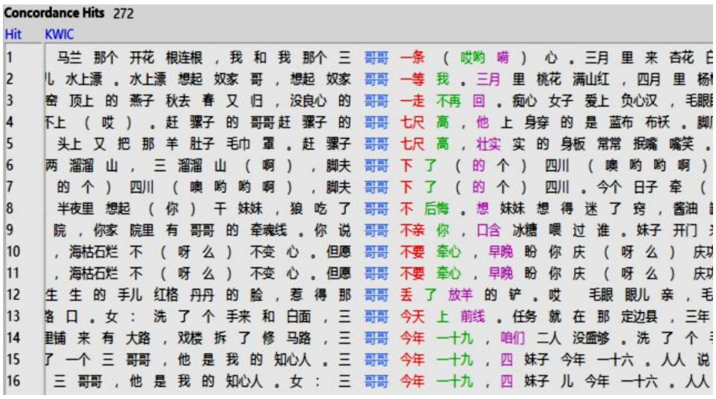
Table 5, which lists the top 10 terms from the CEPCNSHFS, highlights the frequent usage of mood auxiliary words such as "呦", "呀", and "哎" in Northern Shaanxi and Hua'er folk songs. While these words may not carry specific semantic meanings, they play a crucial role in enhancing the aesthetic appeal of folk songs. They serve to evoke emotions and establish a connection with the audience, as seen in Table 6.

TABLE 6
EXAMPLES OF MOOD AUXILIARY WORDS IN THE CEPCNSHFS

ST	TT	Corpora Source	Translation Method
（**呦**……哎咳呦……）	(**Yooooooo, Aihaiiiiiiiiii, Yoooooooooo**!)	N_cn1_3	transliteration
女：心里（**呀**）想你脸上笑， 口里（**呀**）不说谁知道。	W: I put up a smile for I think of you. I do not tell and who could know.	N_cn1_37	omission
正月里冻冰（**呦**）一春消	Ice in the first month melts in spring.	N_cn1_50	omission
一个在那山上（**呦**）一个在那沟	You're high on the hilltop and I, down in the dale.	N_cn1_51	omission
与往年不一般，不（**呀**）一般	Its old look is forever gone.	N_cn1_92	omission
九月里九重阳（**呀**）秋收忙	Autumn is a season for harvesting.	N_cn2_69	omission
谷子（**呀**）糜子（**呀**）收上场	Farmers are busy reaping and bundling.	N_cn2_69	omission
一道道的（那个）山来（**呦**）一道道水	Across mountains and rivers one after another.	N_cn2_83	omission
绿叶儿它自家展**哩**	Its green leaves will stretch out and grow.	H_cn1_22	omission
就是个铁心也软**哩**	Without doubt her cold heart will then be warmed.	H_cn1_22	omission
哎，漫一首花儿了问一句话	**Well**, let me ask by singing.	H_cn1_51	replacement
哎，花儿本是个尕俗话	**Well**, the name is a common term.	H_cn1_51	replacement

In Table 6, it becomes evident that only a minuscule fraction of terms, exemplified by "呦……哎咳呦……", were translated literally, adhering closely to their Chinese pronunciation. Occasionally, "哎" was rendered as the adverb "well", a versatile term capable of conveying various emotional nuances. However, due to the potential to disrupt the fluidity and coherence of the target text (TT) (Li, 2009), the translator chose to omit the majority of mood auxiliary words in the CEPCNSHFS in the translation process.

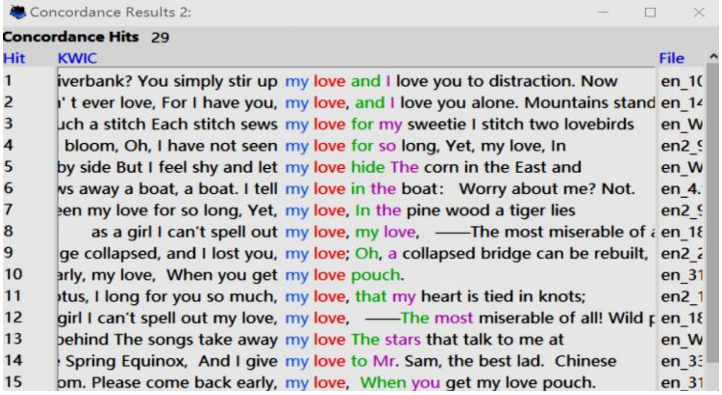
Additionally, researchers can make use of Concordance, often referred to as Key Word in Context (KWIC), a software feature that allows users to search corpora for specific words or phrases and view their contextual usage. KWIC has the advantage of generating a comprehensive list of all relevant instances along with their respective contexts. To illustrate this, let's examine the keyword "哥哥" from the ST of Northern Shaanxi and Hua'er folk songs, and the corresponding search results are provided in Figure 11.



```
Concordance Hits  272
Hit  KWIC
1     马兰 那个 开花 根连根 ，我 和 我 那个 三 哥哥 一条 （ 哎哟 嗬 ） 心 。三月 里 来 杏花
2    儿 水上漂 。水上漂 想起 奴家 哥 ，想起 奴家 哥哥 一等 我 。三月 里 桃花 满山红 ，四月 里 杨
3    窑 顶上 的 燕子 秋去 春 又 归 ，没良心 的 哥哥 一走 不再 回 。痴心 女子 爱上 负心汉 ，毛眼
4    下上 （ 哎 ）。赶 骡子 的 哥哥赶 骡子 的 哥哥 七尺 高 ，他 上 身穿 的 是 蓝布 布袄 。
5     头上 又 把 那 羊 肚子 毛巾 罩 。赶 骡子 哥哥 七尺 高 ，壮实 实 的 身板 常常 抿嘴 嘴笑 。
6    两 溜溜 山 ， 三 溜溜 山 （ 啊 ） ，脚夫 哥哥 下 了 （ 的 个 ） 四川 （ 噢 哟 哟 啊 ）
7     的 个 ） 四川 （ 噢 哟 哟 啊 ） ，脚夫 哥哥 下 了 （ 的 个 ） 四川 。今个 日子 牵 （
8     半夜里 想起 （ 你 ） 干 妹妹 ，狼 吃 了 哥哥 不 后悔 。想 妹妹 想 得 迷 了 窍 ，酱油
9     院 ，你家 院里 有 哥哥 的 牵魂线 。你 说 哥哥 不亲 你 ，口含 冰糖 喂 过 谁 。妹子 开门
10    ，海枯石烂 不 （ 呀 么 ） 不变 心 。但愿 哥哥 不要 牵心 ，早晚 盼 你 庆 （ 呀 么 ） 庆
11    ，海枯石烂 不 （ 呀 么 ） 不变 心 。但愿 哥哥 不要 牵心 ，早晚 盼 你 庆 （ 呀 么 ） 庆
12   生 生 的 手儿 红格 丹丹 的 脸 ，惹 得 那 哥哥 丢 了 放羊 的 铲 。哎  毛眼 眼儿 亲 ，毛
13   路 口 。女 ： 洗 了 个 手来 和 白面 ，三 哥哥 今天 上 前线 。任务 就 在 那 定边县 ，三年
14   里铺 来 有 大路 ，戏楼 拆 了 修 马路 ，三 哥哥 今年 一十九 ，咱们 二人 没盛够 。洗 了 个
15   了 一个 三 哥哥 ，他 是 我 的 知心人 。三 哥哥 今年 一十九 ，四 妹子 今年 一十六 。人人 说
16    三 哥哥 ，他 是 我 的 知心人 。女 ： 三 哥哥 今年 一十九 ，四 妹子 儿 今年 一十六 。人人
```

Figure 11. KWIC for “哥哥” in the CEPCNSHFS

Therefore, it has been observed that the term “哥哥 (*gege*)” appears 272 times in the ST of Northern Shaanxi and Hua’er folk songs, with its primary usage denoting a familial relationship. However, it holds a distinct meaning from the term for a biological older brother and is frequently employed by songwriters in Northern Shaanxi and Hua’er folk songs. Concordance proves to be an invaluable tool for second or foreign language learning, as it allows learners to explore vocabulary, identify common collocations, analyze syntax, and understand writing styles. In this study, it is also utilized to examine the TT of Northern Shaanxi and Hua’er folk songs, with a focus on identifying translation choices and addressing cultural nuances, as demonstrated in Figure 12:



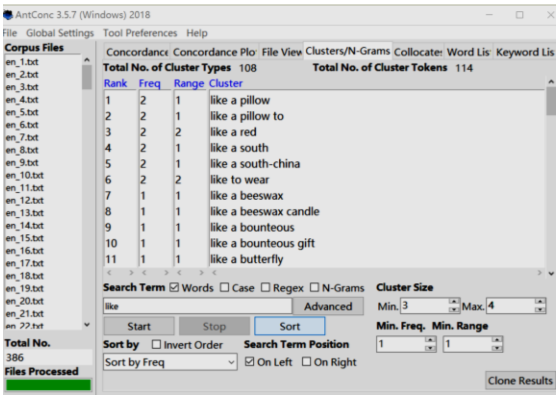
```
Concordance Results 2:
Concordance Hits  29
Hit  KWIC                                                                      File
1    iverbank? You simply stir up my love and I love you to distraction. Now   en_1(
2    ' t ever love, For I have you, my love, and I love you alone. Mountains stand en_14
3    uch a stitch Each stitch sews my love for my sweetie I stitch two lovebirds en_W
4     bloom, Oh, I have not seen my love for so long, Yet, my love, In         en2_9
5    by side But I feel shy and let my love hide The corn in the East and     en_W
6    ws away a boat, a boat. I tell my love in the boat:  Worry about me? Not. en_4.
7    een my love for so long, Yet, my love, In the pine wood a tiger lies     en2_9
8        as a girl I can’t spell out my love, my love,  ——The most miserable of a en_18
9    ge collapsed, and I lost you, my love; Oh, a collapsed bridge can be rebuilt, en2_2
10   arly, my love,  When you get my love pouch.                               en_31
11   otus, I long for you so much, my love, that my heart is tied in knots;    en2_1
12   girl I can’t spell out my love, my love,  ——The most miserable of all! Wild p en_18
13   behind The songs take away my love The stars that talk to me at          en_W
14   Spring Equinox,  And I give my love to Mr. Sam, the best lad.  Chinese   en_33
15   om. Please come back early, my love,  When you get my love pouch.         en_31
```

Figure 12. KWIC for English Translation of “哥哥” in the CEPCNSHFS

According to the search engine results, the address term “哥哥” was translated into English with the following frequencies: 5 times as “brother”, 22 times as “lover”, and 29 times as “my love”. Common expressions and address phrases are particularly prevalent in Chinese dialects, which are commonly used in the performance of Chinese folk songs (Yin, 2013). In Northern Shaanxi and Hua’er folk songs, “哥哥” is used to address the boy adored by the girl as a culturally loaded word, serving as a euphemism for a “boyfriend” (Li, 2016). Therefore, translating it as “brother” is inaccurate (Tang & Ji, 2022). Furthermore, the term “lover”, which often refers to a person with whom one has a romantic or sexual relationship but is not married in English, is not a suitable translation for “哥哥” in Northern Shaanxi and Hua’er folk songs, as per most dictionaries. Nevertheless, “my love”, “my luve”, or “my dear” may be a preferable choice for translators when rendering such address terms in Chinese folk songs, as commonly found in English poetry and folk ballads (Liu, 2013).

Cluster is another useful tool in AntConc that can be employed for research on Northern Shaanxi and Hua’er folk songs. A word cluster is defined as a set of words connected forward and backward in the text (Scott, 2008). In other words, it refers to instances where two or more word forms are repeated in the text with an intentional relationship of combination or placement. Therefore, when using AntConc’s search item-Cluster to conduct a context-sensitive search, a word or word forms can serve as the central focus.

For example, metaphors and similes are frequently employed in Northern Shaanxi and Hua’er folk songs as rhetorical devices to convey people’s feelings and emotions. Consequently, the word “like” can be used as a trigger word to search for its Cluster Types in the CEPCNSHFS. The minimum frequency is set to 1 time, and the word cluster size is set to 3 to 4 words. Figure 13 presents the cluster window for “like”:


AntConc 3.5.7 (Windows) 2018
File Global Settings Tool Preferences Help
Corpus Files
Concordance Concordance Plot File View Clusters/N-Grams Collocates Word List Keyword List
Start
Stop
Sort
Advanced
Clone Results

en_1.txt
en_2.txt
en_3.txt
en_4.txt
en_5.txt
en_6.txt
en_7.txt
en_8.txt
en_9.txt
en_10.txt
en_11.txt
en_12.txt
en_13.txt
en_14.txt
en_15.txt
en_16.txt
en_17.txt
en_18.txt
en_19.txt
en_20.txt
en_21.txt
en_22.txt

Total No. 386
Files Processed

Total No. of Cluster Types 108 Total No. of Cluster Tokens 114

Rank	Freq	Range	Cluster
1	2	1	like a pillow
2	2	1	like a pillow to
3	2	2	like a red
4	2	1	like a south
5	2	1	like a south-china
6	2	2	like to wear
7	1	1	like a beeswax
8	1	1	like a beeswax candle
9	1	1	like a bounteous
10	1	1	like a bounteous gift
11	1	1	like a butterfly

Search Term ☑ Words □ Case □ Regex □ N-Grams
like
Cluster Size Min. 3 Max. 4
Min. Freq. 1 Min. Range 1
Sort by □ Invert Order
Sort by Freq
Search Term Position ☑ On Left □ On Right

Figure 13. Cluster Window for “Like” in the CEPCNSHFS

The data reveals that certain phrases, such as “like a pillow”, “like a pillow to”, “like a red”, “like a south”, “like a south-China”, and “like to wear”, are frequent occurrences within the cluster. When comparing the source text (ST) and target text (TT), we observed that these clusters, especially “like a pillow” and “like a pillow to”, were utilized in the translation of this Hua’er folk song, as illustrated in the following example.

Example 1:

ST:

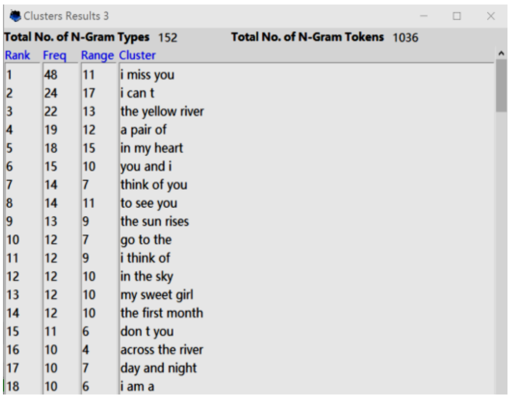
瞌睡遇了枕头了，
好连好是遇下的；
花儿搁成媳妇了，
这是老天爷施下的。(H_cn2_41)

TT:

Like a pillow to rest one’s weary head,
Blessed with many a fortune;
My sweetheart has now become my wife,
Like a bounteous gift from heaven.

The opening line of this song, “瞌睡遇了枕头了”, which signifies that individuals achieve what they sought or experience unexpected good fortune, actually portrays the hero’s joy in realizing his dream. The translator enhanced the song’s meaning by incorporating the preposition “like” in the first and fourth lines to establish a connection between these four lines. Whether it’s “Like a pillow to rest one’s weary head” or “My sweetheart has now become my wife”, both convey the idea of being “Blessed with many a fortune” or simply “Like a bounteous gift from heaven”. It can be concluded that the translation of this Hua’er folk song not only faithfully captures the essence of the original text but also effectively interacts with readers in the target language.

AntConc software categorizes word cluster tables based on the full text or corpus as N-grams, which can be employed to identify and explore clusters throughout the entire CEPCNSHFS. Figure 14 presents the N-gram interface.



Total No. of N-Gram Types 152 Total No. of N-Gram Tokens 1036

Rank	Freq	Range	Cluster
1	48	11	i miss you
2	24	17	i can t
3	22	13	the yellow river
4	19	12	a pair of
5	18	15	in my heart
6	15	10	you and i
7	14	7	think of you
8	14	11	to see you
9	13	9	the sun rises
10	12	7	go to the
11	12	9	i think of
12	12	10	in the sky
13	12	10	my sweet girl
14	12	10	the first month
15	11	6	don t you
16	10	4	across the river
17	10	7	day and night
18	10	6	i am a

Figure 14. The N-Gram Interface of the TT in the CEPCNSHFS

The N-grams results, which analyze the frequency of word sequences, offer the researcher an opportunity to examine the word structure and its translation across the entire corpus. These results reveal that the cluster “I miss you” appears most frequently, occurring 48 times in the target text of the CEPCNSHFS. In fact, “missing someone” is a prevalent theme in such love songs in Northern Shaanxi and Hua’er folk songs, and people frequently express their deep affection by repeatedly uttering “I miss you”. For example,

Example 2:

ST:
男：前半夜**我想你**吹不熄个灯，
后半夜**想你**就翻不了个身。
女：三天（呀）不见哥哥的面，
口含上砂糖也不甜。 (N_cn1_47)
TT:
M: I couldn’t put out light before sleep when **I miss you**,
I couldn’t turn over in bed in dream when **I miss you**.
W: I couldn’t taste the sweetness of sugar,
During the three days I didn’t see you.

This song is a superb example of a Northern Shaanxi folk song that perfectly conveys romantic feelings. In Chinese language, the night is often divided into two parts-“the earlier part of the night (前半夜)” and “the latter hours of the night (后半夜)”, signifying the passage of time. The first two lines are translated as “I couldn’t put out the light before sleep when I miss you,/I couldn’t turn over in bed in a dream when I miss you”. This conveys that the man is unable to sleep through the entire night due to the intensity of his longing for his beloved girl in the song.

In response to the boy, in the last two lines, the girl expresses that if she were separated from her beloved for even three days, she wouldn’t be able to enjoy the sweetness of sugar. This usage is akin to the expression “Not seeing each other for a day feels like being separated for three autumns (一日不见如隔三秋)”, frequently used in Chinese literary works to denote the deep longing and missing someone even after a short separation between protagonists. The translator rearranged the third and fourth lines of the song to ensure they were more suitable and comprehensible to the TT reader.

To convey this sense of longing, Hua’er folk songs employ various rhetorical devices, as exemplified in examples 3 and 4.

Example 3:
ST:
我问花儿啥痛呢？
啥也不痛想人呢！
想得深嘛想得浅？
肠子想成一根线。
心肠肝花早想烂，
肋巴想成了蜜蜂箭。
想得我浑身啪啦啦颤，
三天没吃一叶儿面。(H_cn2_84)
TT:
I asked my sweet girl, “What’s wrong with you?”
“My heart aches, for I miss you so much,” she said.
I asked her, “How badly do you miss me?”
“I miss you to the very core, and my gut has turned into a string.
I miss you so much that I ache from heart to liver,
Even my ribs have turned thin like honeybee stingers;
And I miss you so much that I cannot help trembling,
And I have not eaten for three long days.”

Example 4:
ST:
白日里想你心痛烂，
夜夜晚夕者梦见。(H_cn1_9)
TT:
I miss you to death in the daytime,
And dream of you night after night.

In these two Hua’er folk songs, the emotions of “I miss you” are conveyed through the use of exaggerated expressive techniques. In example 3, the boy asks her beloved girl, “How badly do you miss me?”, and she responds, “I miss you to the very core, and my gut has turned into a string./I miss you so much that I ache from heart to liver,/Even my ribs have turned thin like honeybee stingers;/And I miss you so much that I cannot help trembling,/And I have not eaten for three long days”. The translator faithfully reproduced the intense affection of the ST for the TT reader.

Similarly, in Example 4, the feeling of missing someone is vividly described as “白日里想你心痛烂” (My heart aches when I miss you during the day) and “夜夜晚夕者梦见” (I dream of you night after night). As a result, these lines are translated liberally as “I miss you to death in the daytime,/And dream of you night after night”, with “心痛烂”

being transformed into “I miss you to death”. This transformation maintains the exaggeration while preventing any potential misunderstanding by the TT reader.

In summary, this section demonstrates the application of AntConc for conducting various research investigations on the translation of Northern Shaanxi and Hua’er folk songs through a corpus-based approach. These research techniques include the utilization of tools such as the Keyword List Tool, Concordance, Cluster, and N-grams. Additionally, researchers have the capability to explore language characteristics like type-token ratio (TTR), lexical density, high-frequency phrases, and average sentence length by employing tools like ParaConc, AntConc, WordSmith, or Python. These analytical approaches offer valuable insights into the translation of these folk songs and provide a deeper understanding of the linguistic and stylistic choices made by translators.

V. Conclusion

The corpus methodology holds significant promise as it prompts researchers to pose novel questions with both theoretical and practical implications. It facilitates the identification of contextual significance for specific words, phrases, and texts while capturing numerous instances of language use. Indeed, the corpus-based approach has gained popularity in the study of translating Chinese folk songs, particularly Northern Shaanxi and Hua’er folk songs, which are known for their strong national identity. The research presented here was motivated by the absence of an existing parallel corpus for Northern Shaanxi and Hua’er folk songs translation studies.

This article provided an introduction to the corpus methodology in translation studies, with a specific emphasis on its application in researching the translation of Chinese folk songs. We developed the CEPCNSHFS, a valuable resource for translators, researchers, and students. Folk song translators can use the CEPCNSHFS to obtain precise English translations of specific words, phrases, or collocations, improving their own translations. Researchers can employ the corpus to investigate various aspects of folk songs translation, from vocabulary and grammar to discourse features and translation strategies. Students can use the CEPCNSHFS to practice translation and learn more about Northern Shaanxi and Hua’er folk songs. The CEPCNSHFS also plays a vital role in preserving and transmitting Northern Shaanxi and Hua’er folk songs and promoting traditional Chinese culture and folk music globally.

In this context, the Chinese-English Parallel Corpus of Northern Shaanxi and Hua’er Folk Songs (CEPCNSHFS), introduced in this article, is poised to play a significant role in the field of Northern Shaanxi and Hua’er folk song translation research by offering substantial empirical evidence. We hope that this exploration will help rectify existing imbalances and address limitations within the current landscape of Chinese folk songs translation research.

References

[1] Anthony, L. (2004). AntConc: A learner and classroom friendly, multi-platform corpus analysis toolkit. *Proceedings of IWLeL*, 7-13.

[2] Atkins, S., Clear, J., & Ostler, N. (1992). Corpus design criteria. *Literary and Linguistic Computing*, *7*(1), 1-16. https://doi.org/10.1093/llc/7.1.1

[3] Baker, M. (1993). Corpus linguistics and translation studies implications and applications. *Text and Technology: In Honour of John Sinclair* (pp. 233). John Benjamins Publishing Company. https://doi.org/10.1075/z.64.15bak

[4] Baker, M. (1995). Corpora in translation studies: An overview and some suggestions for future research. Target. *International Journal of Translation Studies, 7*(2), 223-243. https://doi.org/10.1075/target.7.2.03bak

[5] Baker, P. (2010). Corpus methods in linguistics. In *Research Methods in Linguistics* (pp. 93-113). Bloomsbury Publishing.

[6] Chiche, A., & Yitagesu, B. (2022). Part of speech tagging: A systematic review of deep learning and machine learning approaches. *Journal of Big Data*, 9(1), 1-25. https://doi.org/10.1186/s40537-022-00561-y

[7] Guo, H. J., & Zhou, Q. Q. (2019). Ji yu ying-han ke pu ping xing yu liao ku de fan yi han yu “bei” zi ju yu yi yun te zheng yan jiu [A study of the semantic and rhyming features of translated Chinese “Bei” sentences based on an English-Chinese popular science parallel corpus]. *Foreign Language Teaching Theory and Practice*, *(02),* 83-90. https://doi:CNKI:SUN:GWJX.0.2019-02-012.

[8] Guo, Q. (2020). Shan nan min ge yin yue xian zhuang yu yu liao ku jian she [Current status of folk song music in Southern Shaanxi and corpus construction]. *Schools of Arts (01),* 130-134+205. https://doi:CNKI:SUN:YSBJ.0.2020-01-022

[9] Hu, K. B., & Tao, Q. (2009). Explicitation in the Chinese-English conference interpreting and its motivation-A study based on parallel corpus. *Journal of PLA University of Foreign Languages, (04),* 67-73. https://doi:CNKI:SUN:JFJW.0.2009-04-015

[10] Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. In *Proceedings of Machine Translation Summit X: papers* (pp. 79-86). Phuket, Thailand.

[11] Lapshinova-Koltunski, E., Hardmeier, C., & Krielke, P. (2018, May). ParCorFull: A parallel corpus annotated with full coreference. In *11th Edition of the Language Resources and Evaluation Conference* (pp. 423-428). European Language Resources Association (ELRA).

[12] Li, L. B. (2009). Zai shi zhong ling ting ge de hui yin-Ping “Xi bei hui xiang” jian lun Shan bei min ge de fan yi [Listening to the echo of songs in poetry: A review of “Voice from the Northwest-Folk songs of Northern Shaanxi” and a discussion on the translation of Northern Shaanxi folk songs]. *Jiaoxiang-Journal of Xi’an Conservatory of Music, (03),* 78-82. https://doi:CNKI:SUN:JXXA.0.2009-03-021

[13] Li, S. S. (2016). Cong chan shi xue kan Shan bei min ge wen hua fu zai ci de ying yi guo cheng-Yi Wang Hongyin “Xi bei hui xiang” wei li [Exploring the translation process of cultural-laden terms in Northern Shaanxi folk songs from the perspective of

hermeneutics: A case study of Wang Hongyin’s “Voice from the Northwest”]. *Intelligence, (13),* 205-206. https://doi:CNKI:SUN:CAIZ.0.2016-13-178

[14] Liu, Y. L. (2013). Wen hua chuan zhen zai shan bei min ge fan yi zhong de yun yong fen xi [Analysis of the application of cultural realism in the translation of Northern Shaanxi folk songs]. *Youth Literator, (30),* 119. https://doi:CNKI:SUN:QNWJ.0.2013-30-096

[15] Sinclair, J. (1999). *Corpus, concordance, collocation*. Oxford University Press.

[16] Scott, M. (2008). *Oxford WordSmith tools 5.0 manual*. Oxford University Press.

[17] Tang, Z. J., & Ji, X. M. (2022). San mei lun zai Xintianyou ying yi shi jian zhong de ti xian [The Manifestation of the “Three beauties” theory in the English translation practice of Xintianyou]. *Journal of Yulin University, (03),* 26-30. https://doi:10.16752/j.cnki.jylu.2022.03.005

[18] Wang, H. Y. (2009). *Voice from the northwest-Folk songs of Northern Shaanxi*. Culture and Art Publishing House.

[19] Wang, H. Y. (2014). *Chinese folk songs and their English translation*. The Commercial Press.

[20] Wang, K. F. (2004). *Shuang yu dui ying yu liao ku: yan zhi yu ying yong* [Bilingual parallel corpus: Development and application]. Foreign Language Teaching and Research Press.

[21] Wang, K. F., Qin, H. W., Xiao, Z. H., & Hu, K. B. (2021). *Zhong guo ying han ping xing yu liao ku yan jiu* [Studies based on a super-sized English-Chinese parallel corpus]. Foreign Language Teaching and Research Press.

[22] Wang, X. L. (2021). Building a parallel corpus for English translation teaching based on computer-aided translation software. *Comput Aided Des Appl*, *18*(S4), 175-185. https://doi.org/10.14733/cadaps.2021.S4.175-185

[23] Wu, Y. L. (2008). *Zhong guo Hua’er tong lun* [Comprehensive discussion on Chinese Hua’er songs]. Ningxia People’s Publishing House.

[24] Wang, Z. B. (2021). *Lyrics of Northern Shaanxi folk songs: Translation and exegesis*. Nankai University Press.

[25] Yin, L. P. (2013). Cong gong neng dui deng li lun kan ming ge de fan yi [Examining folk song translation from the perspective of functional equivalence theory]. *Journal of Yan’an Vocational and Technical College, (05),* 55-56. https://doi:CNKI:SUN:YAXB.0.2013-05-025

[26] Yuan, W. (2020). Ji yu e-han zheng zhi wai jiao ping xing yu liao ku de duan yu dui ying dan wei chou qu yan jiu [Study on the extraction of phrase correspondence units based on the Russian-Chinese political and diplomatic parallel corpus]. *Journal of PLA Foreign Languages Institute, (05)*, 38-45. https://doi:CNKI:SUN:JFJW.0.2020-05-006

[27] Yang, X. L., & Caroline, E. K. (2016). *Hua’er-Folk songs from the silk road*. The Commercial Press.

[28] Yang, X. L., Ding Y., Wang J., & Caroline, E. K. (2022). *Voice from the silk road-Hua’er of northwest China*. The Commercial Press.

[29] Zhang, T. X. (2010). Ji yu Web jian suo de shanbei min ge yu liao ku she ji [Design of a Northern Shaanxi folk song corpus based on Web retrieval]. *Modern Electronic Technology, (22),* 38-39+41. https://doi:10.16652/j.issn.1004-373x.2010.22.040

[30] Zhang, W. H. (2020). “Zhuangzi” han ying ping xing yu liao ku de chuang jian: tu jing yu yi yi [Creation of a parallel Chinese-English corpus for “Zhuangzi”: Approach and significance]. *Foreign Languages and Cultures, (04),* 125-132. https://doi:10.19967/j.cnki.flc.2020.04.012

[31] Zhou, X. M. (2017). Zhuang zu dian ji duo yu ping xing yu liao ku jian she yu ying yong yan jiu [Construction and application research of multilingual parallel corpus of Zhuang ethnic classics]. *Overview of Social Sciences, (10),* 137-139. https://doi:10.16745/j.cnki.cn62-1110/c.2017.10.03

Yan Lin is a lecturer at the Department of Foreign Languages, Ningxia Normal University of China. She is currently pursuing her Ph.D. in Translation and Interpretation at the Department of Foreign Languages, Faculty of Modern Languages and Communication, Universiti Putra Malaysia. Her primary areas of interest in research are cross-cultural studies, corpus translation studies and the translation of Chinese literature.

Hazlina Abdul Halim is an associate professor at the department of Foreign Languages, Faculty of Modern Languages and Communication, Universiti Putra Malaysia. Her main research interests are in the areas of French applied linguistics and translation studies. Orcid.org/0000-0003-3599-9195

Farhana Muslim Mohd Jalis is a senior lecturer at the Department of Foreign Languages, Faculty of Modern Languages and Communication, Universiti Putra Malaysia. Her main research interests are in the areas of German as foreign language, comparative linguistics, and cultural studies.