electronics

MDPI

Article

WCC-JC 2.0: A Web-Crawled and Manually Aligned Parallel Corpus for Japanese-Chinese Neural Machine Translation

Jinyi Zhang [1,*], Ye Tian [2], Jiannan Mao [3], Mei Han [4], Feng Wen [1], Cong Guo [1], Zhonghui Gao [1] and Tadahiro Matsumoto [3]

[1] School of Information Science and Engineering, Shenyang Ligong University, Shenyang 110159, China
[2] Zhuzhou CRRC Times Electric Co., Ltd., Zhuzhou 412001, China
[3] Faculty of Engineering, Gifu University, Gifu 501-1193, Japan
[4] School of Electrical and Information Engineering, Hunan University of Technology, Zhuzhou 412007, China
* Correspondence: zhangjinyi@sylu.edu.cn

Abstract: Movie and TV subtitles are frequently employed in natural language processing (NLP) applications, but there are limited Japanese-Chinese bilingual corpora accessible as a dataset to train neural machine translation (NMT) models. In our previous study, we effectively constructed a corpus of a considerable size containing bilingual text data in both Japanese and Chinese by collecting subtitle text data from websites that host movies and television series. The unsatisfactory translation performance of the initial corpus, Web-Crawled Corpus of Japanese and Chinese (WCC-JC 1.0), was predominantly caused by the limited number of sentence pairs. To address this shortcoming, we thoroughly analyzed the issues associated with the construction of WCC-JC 1.0 and constructed the WCC-JC 2.0 corpus by first collecting subtitle data from movie and TV series websites. Then, we manually aligned a large number of high-quality sentence pairs. Our efforts resulted in a new corpus that includes about 1.4 million sentence pairs, an 87% increase compared with WCC-JC 1.0. As a result, WCC-JC 2.0 is now among the largest publicly available Japanese-Chinese bilingual corpora in the world. To assess the performance of WCC-JC 2.0, we calculated the BLEU scores relative to other comparative corpora and performed manual evaluations of the translation results generated by translation models trained on WCC-JC 2.0. We provide WCC-JC 2.0 as a free download for research purposes only.

Keywords: Japanese-Chinese parallel corpus; neural machine translation; construction of the parallel corpus; manually aligned corpus

check for updates

Citation: Zhang, J.; Tian, Y.; Mao, J.; Han, M.; Wen, F.; Guo, C.; Gao, Z.; Matsumoto, T. WCC-JC 2.0: A Web-Crawled and Manually Aligned Parallel Corpus for Japanese-Chinese Neural Machine Translation. *Electronics* **2023**, *12*, 1140. https://doi.org/10.3390/electronics12051140

Academic Editor: Rui Pedro Lopes

Received: 6 February 2023
Revised: 19 February 2023
Accepted: 24 February 2023
Published: 26 February 2023

Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

The Belt and Road Initiative encompasses a wide array of countries, populations, and linguistic diversity, encompassing exchanges across multiple technical, cultural, and economic domains. As neighboring countries and key players within the Belt and Road strategy, China and Japan play a crucial role in this endeavor. In 2017, the Japan Research Center for the Belt and Road was established in Tokyo by a group of Japanese scholars. The two countries are significant trading partners; Japan is currently China's second biggest trading partner, while China is Japan's largest trading partner. As language is the communication link of cultural and industrial cooperation, the language barrier remains a challenge to enhancing communication between the two countries.

Particularly with the emergence of neural machine translation (NMT) as the leading framework for translation applications, the field of natural language processing (NLP) has undergone substantial advancements. Despite the similarities between Chinese and Japanese, the machine translation of these languages still lacks practical application, and the accuracy of the results is inadequate for practical use.

In recent times, the accelerated advancement of deep learning has been predominantly centered around technological development but with less emphasis on the underlying data

class. However, data are the cornerstone of deep learning, and without adequate data, even the most advanced technology is just castles in the air.

Linguistic researchers have recognized the significance of corpus processing methods and means. While spoken corpora are the ideal object of linguistic research, they have not yet been given sufficient attention. The current ad hoc approach to including spoken corpora in a corpus highlights the need for a dedicated Japanese-Chinese spoken language corpus for the exploration of Japanese-Chinese NMT systems in academic research.

The creation of a Japanese-Chinese spoken language corpus will generate substantial interest among researchers of Japanese-Chinese NMT, promoting advancements in the field and providing crucial information and experience for bilingual information processing and subsequent research. The objective of this research pursuit is to address the language barrier in Japanese-Chinese communication and improve mutual understanding.

In the present study, our overarching objective is to advance Japanese-Chinese NMT by constructing a Japanese-Chinese spoken language corpus. In our previous study, we collected approximately 753,000 Japanese-Chinese sentence pairs from the Web and created the Web-Crawled Corpus of Japanese and Chinese (WCC-JC 1.0) [1]. However, the results of translation tests and manual evaluations indicated that WCC-JC 1.0 was inferior in quantity and quality compared with other corpora. To address these shortcomings, we carefully analyzed the limitations of WCC-JC 1.0 and applied extensive manual alignment efforts to the collected corpus texts. As a result, WCC-JC 2.0, now the world's largest free Japanese-Chinese bilingual corpus, was created with approximately 1.4 million sentence pairs (an 87% increase in the aggregate number of sentences). The improved translation test results and manual evaluations demonstrate the effectiveness of WCC-JC 2.0 compared with its predecessor WCC-JC 1.0.

The subsequent parts of this article are logically outlined as follows. We will present related works in Section 2, describe the construction of WCC-JC 2.0 in Section 3, report the experimental framework and results in Section 4, discuss the legality of WCC-JC 2.0 in Section 5, and conclude by expounding on the scholarly contributions and charting the course for future research endeavors in Section 6.

2. Related Works

The limitations of early English word frequency dictionaries have been subject to extensive scholarly discourse in the existing body of literature. In a study by Leech et al., the authors recognized the limitations in terms of sample size and breadth of early English word frequency dictionaries [2]. To overcome these limitations, the authors investigated the written and spoken English found in the British National Corpus, culminating in the development of frequency dictionaries for the major languages spoken worldwide.

Contrary to other areas of linguistics, forensic linguistics research remains entrenched in traditional, intuition-driven, and retrospective methodologies. Ting, in his study, combined the features of parallel corpora and current forensic linguistics to examine the potential of applying parallel corpora to forensic linguistics. He explored this concept from six different perspectives [3].

Tiedemann presented a dictionary-based strategy that leveraged automatic word alignment to enhance the alignment quality compared with the synonym-based approach [4]. By adopting this approach, the quality of alignment during the construction of a parallel corpus that uses translated subtitles is significantly enhanced.

From a linguistic perspective, spoken corpora are deemed to be the primary objects of research. However, to date, such corpora remain scarce. The determination of content is also a crucial consideration when constructing a spoken corpus. Cermák successfully tackled this challenge by transforming the pivotal dissimilarities between spoken and written texts into pragmatic parameters, thereby facilitating the creation of a spoken corpus [5].

Aziz et al. proposed a novel approach to subtitle translation and assessed its efficacy through the implementation of two strategies for translating and compressing subtitles from English to Portuguese [6]. The results of the experiments revealed that fine-tuning

the model parameters in the translation system can lead to improvements over an unconstrained baseline. Furthermore, the final translation quality can be further enhanced by incorporating specific model components that guide the translation process. In this investigation, the translation quality was meticulously evaluated through the process of human post-editing.

Wang et al. undertook a research investigation on the refinement of word segmentation in machine translation using a corpus with word alignments that were manually annotated [7]. The corpus was annotated to align words to the smallest translation unit. The findings of the study shed light on the effectiveness of utilizing a word alignment corpus with manual annotations to improve word segmentation in machine translation.

Adolphs et al. explored the area of spoken language corpus linguistics, analyzing both monomodal and multimodal spoken language corpora [8]. They explored the obstacles faced during the design, development, and application of spoken language corpora. The authors emphasized the importance of spoken language corpora in revealing unique patterns of language use, which has significant implications for both the description of language use patterns and the field of applied linguistics.

Liu et al. proposed an unsupervised word alignment method for a manually aligned Chinese-English parallel corpus, which contains approximately 40,000 sentence pairs [9] (https://nlp.csai.tsinghua.edu.cn/~ly/systems/TsinghuaAligner/TsinghuaAligner.html (accessed on 10 November 2022)). The proposed method's effectiveness in aligning words between Chinese and English languages was demonstrated using this corpus.

Tiedemann established a vast collection of parallel corpora, OPUS [10], which encompasses over 200 languages and dialects. This corpus comprises roughly 3.2 billion sentences and sentence fragments representing over 28 billion tokens and draws data from various sources and domains. The data within OPUS can be conveniently downloaded in a uniform data format, facilitating their use in research.

The ASPEC-JC corpus was curated by means of manual translation of Japanese scientific articles into Chinese, with the articles being either under the ownership of the Japan Science and Technology Agency (JST) or preserved in Japan's largest electronic platform for academic journals (J-STAGE) [11].

The utilization of alternative translations in subtitle translation has been underexplored. To address this issue, Tiedemann proposed a methodology that leverages time-based alignment and lexical resynchronization techniques in combination with BLEU score metrics to categorize substitute translation versions into groups, employing the measures of edit distance and heuristics [12]. The implementation of this approach resulted in a substantial number of sentence-aligned translation alternatives, providing a solution to the challenges posed by spelling errors, incomplete or corrupted data files, or misaligned subtitles in the corpus of subtitles.

Wee et al. conducted research on the impact of dialogue-specific aspects, such as dialog acts, speaker, gender, and text register, and the impact of fictional dialogues on the efficacy of machine translation systems was assessed [13]. They constructed and published a corpus of multilingual movie dialogues and found that the BLEU values between categories were significantly larger than expected. As a result of this, it was hypothesized and demonstrated that adapting machine translation systems to dialogue acts and text registers could enhance their performance when translating fictional dialogue.

A significant amount of conversational data is necessary for the training and optimization of neural models used for dialogue generation, which are often obtained from chat forums or movie subtitles. However, these data can sometimes lack multiple references, making them difficult to utilize. To address this challenge, Lison et al. put forward a weighting model with the aim of enhancing the effectiveness of neural conversational models, as shown through evaluation with unsupervised metrics [14].

Wang et al. presented a new de-duplication algorithm for web pages, which utilizes TF-IDF and the distance of word vectors to enhance the performance of web page de-

duplication during the construction of a corpus of semantically annotated natural language via a cloud-based service infrastructure [15].

Levshina conducted a quantitative analysis of online film subtitles, considering them as a separate mode of communication [16]. The study contrasted movie subtitles translated into English from other languages with two prominent English language corpora representing spoken and written language—British and American English. The findings of this study, based on an analysis of n-gram frequencies, indicated that the subtitles were not significantly dissimilar from other variants of the English language, closely resembling informal spoken language in both forms of English used in the United Kingdom and the United States. Additionally, the language used in subtitles was observed to be more emotionally charged and expressive when compared with typical conversations.

Love et al. created the Spoken British National Corpus 2014, a comprehensive collection of 11.5 million words of conversational British English spoken by native speakers from various regions of the UK [17]. The paper highlights the crucial steps involved in creating the corpus, with a particular emphasis on ensuring the sensitivity of the data collection process and implementing innovative techniques.

In English for Specific Purposes (ESP) education, a corpus provides a wealth of examples for students and can assist in guiding them toward the right texts that fit their specific needs. Chen et al. investigated the challenges and solutions in creating a corpus for ESP, with the aim of transforming traditional teaching methods into a student-centered approach to learning [18].

The OpenSubtitles2018 corpus was a rich resource for multilingual parallel movie subtitle data, comprising 3.7 million subtitles across 60 different languages [19,20]. To improve the alignment quality, the authors proposed a sophisticated regression model and employed its scores to filter the parallel corpus, thus effectively eliminating low-quality alignment results. This approach improved the overall quality of the corpus, making it an even more valuable resource for cross-linguistic studies and machine translation research.

In order to meet the needs of various biomedical research, Ren et al. developed a precision medicine corpus that stores a more comprehensive collection of medical knowledge, enriching the biomedical corpus and promoting research in the field of biomedical text mining [21].

Davies presented the creation and utilization of the television shows and movies corpora, both of which are available at English-Corpora.org [22]. The television shows corpus consists of a collection of 75,000 episodes' subtitles. The subtitles cover a time span ranging from the 1950s to the 2010s and were obtained from 6 English-speaking countries, totaling 325 million words. The movies corpus contains subtitles from 25,000 motion pictures, covering 200 million words in the same 6 countries and time period. These corpora provide a unique opportunity to compare very informal language from different periods.

Doi et al. devised an innovative and extensive corpus for English-Japanese simultaneous interpretation (SI) and compared it to offline translation [23]. They analyzed the differences in terms of latency, quality, and word order. The results showed that the data from experienced interpreters were of better quality and that a larger latency negatively impacted the quality of the corpus.

Meftah et al. created the King Saud University Emotions (KSUEmotions) corpus, the inaugural Arabic emotion corpus accessible to the public, which encompasses emotional speeches from speakers originating from Syria, Saudi Arabia, and Yemen [24]. The corpus consists of emotions such as neutral, joy, sorrow, astonishment, and rage. The authors conducted a significant number of validation experiments to assess the corpus.

Xu emphasized the crucial role that big data crawl-based language-networked corpora now play in practical applications and language research [25]. The research emphasized that the utilization of linguistic databases has evolved into an essential research tool across all domains of language research, proving indispensable for natural language researchers, lexicographers, and language scientists alike.

Duszkin et al. presented the design principles of 10 parallel pair translation corpora and the parallel CLARIN-PL corpus of the Slavic and Baltic languages [26]. These corpora possess unique features, including resource selection, preprocessing, manual sentence-level segmentation, lemmatization, annotation, and metadata, which were constructed based on these design principles. The paper highlighted the importance of well-designed corpora in facilitating and advancing language research.

Liu et al. constructed a comprehensive spoken English corpus, the Spoken BNC2014, which contains 11.5 million words of conversational data collected from native British English speakers across the UK [27]. As one of the UK's largest spoken English corpora, this corpus is a valuable resource for linguistic work and natural language processing.

Using Mark Davies' mega-corpora, Ha's study examined the lexical profile of informal spoken English [28]. The findings indicated that vocabulary knowledge at the 3000 and 5000 word frequency levels was required to comprehend 95% and 98% of the words in general scripted dialogues, respectively. The research revealed that soap operas demand less vocabulary than TV shows and movies.

Additionally, previous research has extensively cited a multitude of relevant studies [1].

For the purposes of organization and clarity, these studies have been sorted chronologically. A summarized table (Table 1) is also provided to categorize the related works based on three aspects: (1) spoken corpus; (2) corpus construction; and (3) corpus applications.

Table 1. Summary of related previous studies.

Classification	Related Previous Studies
Spoken corpus	[2,4–6,8,12–14,27,28]
Corpus construction	[3,10,11,16–24,26]
Corpus applications	[7,9,15,25]

The prior research has highlighted the crucial significance of spoken corpora, corpus construction, and their applications in driving forward the field of NLP. These insights motivated us to create a Japanese-Chinese spoken language corpus for NMT, which can have significant practical implications for overcoming the challenge of limited corpus resources.

3. Building of the Japanese-Chinese Spoken Language Corpus

The corpus we aim to build, WCC-JC 2.0, comprises dual-language subtitles in Japanese and Chinese, much like its predecessor, WCC-JC 1.0. In this section, we delve into the rationale behind our decision to utilize manual alignment in the construction of the corpus as well as provide a comprehensive overview of the steps involved in the corpus construction process. Furthermore, we outline the major challenges encountered during the alignment phase.

3.1. Why Manual Alignment

There are six primary reasons that we chose manual alignment for our corpus construction:

1. File format problems: Some of the subtitle files are not in ASS format but rather in SRT format, which lacks certain information such as keywords. As a result, these files could not be processed by the code program and required manual alignment.
2. Special effects lyrics problems: Due to the dynamic effect, lyric subtitles are marked character by character, and there may not be a one-to-one correspondence between the number of words in the Chinese and Japanese sentences. Deleting all sentences within two characters could impact the quality of the spoken language data. As shown in Table 2, the subtitles give each character in this lyric a special effect, including position and color. It is difficult to extract the content of a lyric because it is too complex in

the subtitle files. Due to these factors, the song lyrics required manual extraction and alignment.

3. Mistranslation problems: As indicated in Table 3, it is evident that there are mistranslations present in the subtitle text, as seen in the incorrect positioning of the upper and lower sentences. This issue is not unique to this work and is present in all similar subtitle texts. Therefore, manual alignment of the subtitle content is necessary for accurate error correction.
4. Keyword problems when extracting: The wrong positioning of keywords in the subtitle file presents a challenge in proper language recognition. As a result, manual extraction of subtitle content is required to ensure accurate translation. The keyword "Default" serves as a prime example of this issue, as it is unable to be properly distinguished as either Japanese or Chinese by the program. To correct these errors, manual alignment of the subtitle content is necessary.
5. One-to-many and many-to-one problems: The absence of punctuation marks in subtitle files presents a challenge for sentence alignment techniques. As evidenced by Table 4, where the Chinese subtitle text consists of three sentences while the corresponding Japanese subtitle text consists of four sentences, manual alignment is necessary. The aforementioned sentence underscores the insufficiency of solely depending on previous sentence alignment research outcomes for subtitle files, thereby necessitating a manual approach to ensure precise alignment of the subtitle content.
6. Timeline problems: The subtitle files may also contain timeline errors, where the data are flawed and cannot be aligned correctly by the code program. As demonstrated in Table 5, the Chinese and Japanese timelines did not match. Manual alignment is deemed necessary in such instances to rectify any errors. Timeline errors pose a significant challenge to the automatic alignment of subtitle content and call for manual intervention to ensure accuracy.

Table 2. Examples of special effects lyric subtitles.

Chinese Lyric Subtitles with Special Effects. **English Translation: This is the time for change.**
Dialogue: 0,0:01:40.76,0:01:42.66,OPCN,,0,0,0,fx,\an5\pos(758,1041)\fscx100\fscy100\bord3\blur0 \1vc(H152BC9,H49BFFC,H152BC9,H49BFFC)\3c&HFFFFFF&此
Dialogue: 0,0:01:40.76,0:01:42.66,OPCN,,0,0,0,fx,\an5\pos(812,1041)\fscx100\fscy100\bord3\blur0 \1vc(H152BC9,H49BFFC,H152BC9,H49BFFC)\3c&HFFFFFF&刻
Dialogue: 0,0:01:40.76,0:01:42.66,OPCN,,0,0,0,fx,\an5\pos(865,1041)\fscx100\fscy100\bord3\blur0 \1vc(H152BC9,H49BFFC,H152BC9,H49BFFC)\3c&HFFFFFF&正
Dialogue: 0,0:01:40.76,0:01:42.66,OPCN,,0,0,0,fx,\an5\pos(919,1041)\fscx100\fscy100\bord3\blur0 \1vc(H152BC9,H49BFFC,H152BC9,H49BFFC)\3c&HFFFFFF&是
Dialogue: 0,0:01:40.76,0:01:42.66,OPCN,,0,0,0,fx,\an5\pos(960,1041)\fscx100\fscy100\bord3\blur0 \1vc(H152BC9,H49BFFC,H152BC9,H49BFFC)\3c&HFFFFFF&
Dialogue: 0,0:01:40.76,0:01:42.66,OPCN,,0,0,0,fx,\an5\pos(1001,1041)\fscx100\fscy100\bord3\blur0 \1vc(H152BC9,H49BFFC,H152BC9,H49BFFC)\3c&HFFFFFF&改
Dialogue: 0,0:01:40.76,0:01:42.66,OPCN,,0,0,0,fx,\an5\pos(1055,1041)\fscx100\fscy100\bord3\blur0 \1vc(H152BC9,H49BFFC,H152BC9,H49BFFC)\3c&HFFFFFF&变
Dialogue: 0,0:01:40.76,0:01:42.66,OPCN,,0,0,0,fx,\an5\pos(1108,1041)\fscx100\fscy100\bord3\blur0 \1vc(H152BC9,H49BFFC,H152BC9,H49BFFC)\3c&HFFFFFF&之
Dialogue: 0,0:01:40.76,0:01:42.66,OPCN,,0,0,0,fx,\an5\pos(1162,1041)\fscx100\fscy100\bord3\blur0 \1vc(H152BC9,H49BFFC,H152BC9,H49BFFC)\3c&HFFFFFF&际

Table 2. *Cont.*

Japanese Lyrics Subtitles with Special Effects.
Dialogue: 0,0:01:40.76,0:01:42.66,OPJP,,0,0,0,fx,\an5\pos(746,981)\fscx100\fscy100\bord3\blur0 \1vc(H873235,HCB8D58,H873235,HCB8D58)\3c&HFFFFFF&今
Dialogue: 0,0:01:40.76,0:01:42.66,OPJP,,0,0,0,fx,\an5\pos(800,981)\fscx100\fscy100\bord3\blur0 \1vc(H873235,HCB8D58,H873235,HCB8D58)\3c&HFFFFFF&こ
Dialogue: 0,0:01:40.76,0:01:42.66,OPJP,,0,0,0,fx,\an5\pos(853,981)\fscx100\fscy100\bord3\blur0 \1vc(H873235,HCB8D58,H873235,HCB8D58)\3c&HFFFFFF&そ
Dialogue: 0,0:01:40.76,0:01:42.66,OPJP,,0,0,0,fx,\an5\pos(907,981)\fscx100\fscy100\bord3\blur0 \1vc(H873235,HCB8D58,H873235,HCB8D58)\3c&HFFFFFF&
Dialogue: 0,0:01:40.76,0:01:42.66,OPJP,,0,0,0,fx,\an5\pos(960,981)\fscx100\fscy100\bord3\blur0 \1vc(H873235,HCB8D58,H873235,HCB8D58)\3c&HFFFFFF&変
Dialogue: 0,0:01:40.76,0:01:42.66,OPJP,,0,0,0,fx,\an5\pos(1014,981)\fscx100\fscy100\bord3\blur0 \1vc(H873235,HCB8D58,H873235,HCB8D58)\3c&HFFFFFF&わ
Dialogue: 0,0:01:40.76,0:01:42.66,OPJP,,0,0,0,fx,\an5\pos(1067,981)\fscx100\fscy100\bord3\blur0 \1vc(H873235,HCB8D58,H873235,HCB8D58)\3c&HFFFFFF&る
Dialogue: 0,0:01:40.76,0:01:42.66,OPJP,,0,0,0,fx,\an5\pos(1121,981)\fscx100\fscy100\bord3\blur0 \1vc(H873235,HCB8D58,H873235,HCB8D58)\3c&HFFFFFF&と
Dialogue: 0,0:01:40.76,0:01:42.66,OPJP,,0,0,0,fx,\an5\pos(1174,981)\fscx100\fscy100\bord3\blur0 \1vc(H873235,HCB8D58,H873235,HCB8D58)\3c&HFFFFFF&き

Table 3. Examples of mistranslations.

Chinese Subtitle Text	**Japanese Subtitle Text**
银色魔术师 Sliver Heart English Translation: Silver Magician Sliver Heart	今宵も 華麗に見参
今晚也华丽登场了 English Translation: Gorgeous show tonight too	銀の魔術師 シルバーハートじゃ

Table 4. Examples of one-to-many and many-to-one errors.

Chinese Subtitle Text	**Japanese Subtitle Text**
原来 你们触犯了禁忌吧 English translation: So, you have broken the taboo, right?	そうか お前ら
你们触犯了人体炼成 这一炼金术最大的禁忌 English Translation: You have violated the greatest taboo of alchemy: human alchemy	禁忌を犯したな
阿尔 阿尔冯斯 English Translation: Al Alphonse	錬金術師最大の禁忌 人体錬成をやったんだな
	アル アルフォンス

Table 5. Examples of timeline errors. The timeline has been bolded.

Chinese Subtitles	**Japanese Subtitles**
Dialogue: 0,**0:03:47.81,0:03:50.03**,*Default, NTP,0000,0000,0000,,索敌系统捕捉到敌人 English Translation: Solicitor system captures the enemy	Dialogue: 0,**0:03:47.74,0:03:49.69**,*Default, NTP,0000,0000,0000,,索敵シ ステムが敵を捕捉
Dialogue: 0,**0:03:53.55,0:03:55.47**,*Default, NTP,0000,0000,0000,,后方有 大型舰船 English Translation: Large ships in the rear	Dialogue: 0,**0:03:53.57,0:03:55.17**,*Default, NTP,0000,0000,0000,,後方に 大型艦艇

Despite the challenges posed by the inaccuracies and complexities of subtitle files, manual alignment was deemed the most suitable approach for building a high-quality Japanese-Chinese parallel corpus. This decision was based on our previous experience and the recognition of the limitations of existing methods in effectively addressing the various issues present in the subtitle data. The manual alignment process ensures the attainment of a high-quality parallel corpus, as demonstrated in our results.

3.2. Web Crawling

During the construction period, in line with the practice of WCC-JC 1.0, we employed Scrapy (https://scrapy.org/ (accessed on 15 October 2022)) to acquire subtitle files from websites (http://assrt.net/ and https://bbs.acgrip.com/(both accessed on 10 June 2022)). These websites contain a large number of bilingual subtitle files of TV series, animation, movies, and other media types. In addition, since some subtitle files were not directly available for download and had to be extracted from MKV video files, we downloaded a large number of MKV video files from one website (https://subs.kamigami.org/(accessed on 10 June 2022)) and extracted the subtitle Advanced SubStation Alpha (ASS) files from the MKV video files with SubtitleEdit software (https://github.com/SubtitleEdit (accessed on 10 June 2022)). This approach allowed us to gather a substantial amount of data for the WCC-JC 2.0 construction project.

3.3. Extraction of Bilingual Sentences

During the construction of WCC-JC 2.0, the majority of subtitle files acquired were in ASS format. As with the subsection of the same name in our previous paper [1], the extraction of subtitle content was performed. It will not be repeated here, and it should be noted that we added the extraction of the song OP and ED this time, a part that was not carried out in the previous study.

The raw parallel corpus was then preprocessed by converting traditional Chinese symbols to simplified Chinese symbols and standardizing Japanese katakana characters to full-width format using the zenhan library (https://pypi.org/project/zenhan/ (accessed on 10 October 2022)).

3.4. Manual Alignment

After preprocessing the text data, we began the manual alignment process in August 2020 and continued it until November 2022. We used Excel to make the alignment work easier and more efficient. The manual alignment process was a time-consuming and demanding task, but we persisted and were patient, resulting in the processing of more than 1.8 million sentence pairs, as shown in Figure 1. The texts were then sorted, and duplicate sentence pairs were removed to obtain a filtered parallel corpus, which consisted of approximately 1.4 million sentence pairs.

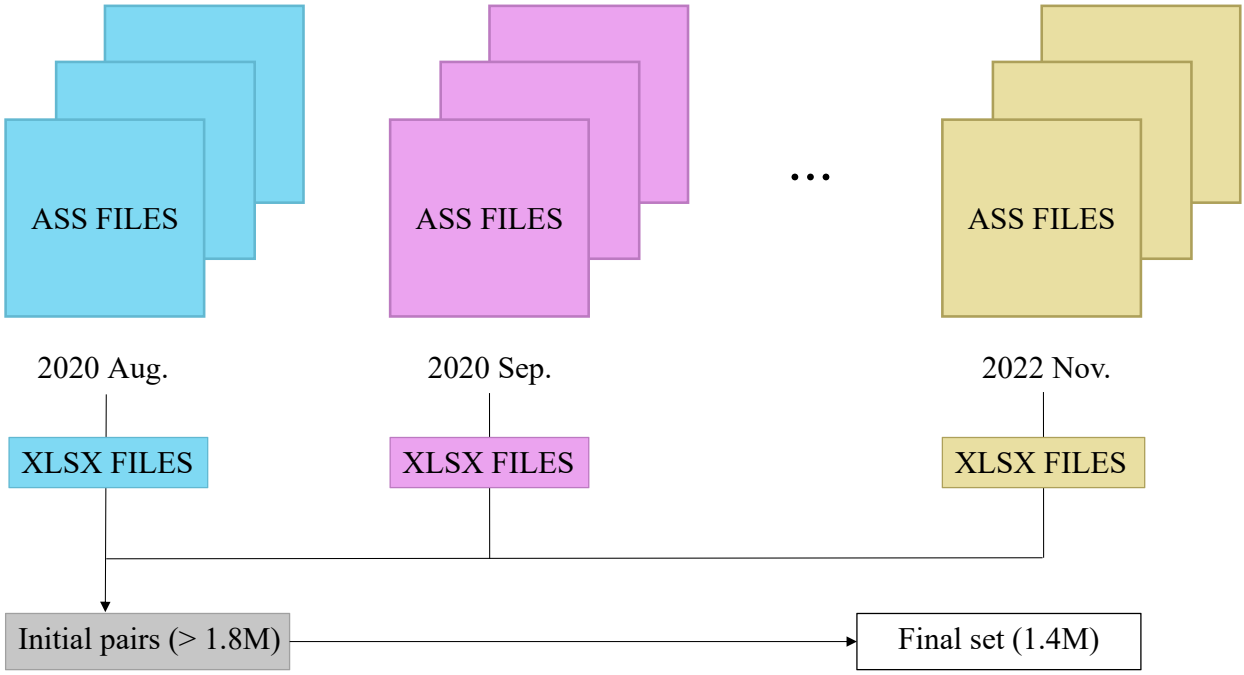

Figure 1. The timeline of manually aligning WCC-JC 2.0.

3.5. Corpus Segmentation

In conclusion, the final step of our corpus construction process involved randomly selecting a portion of the filtered parallel corpus for use as development and test sets. In accordance with established practices for constructing NMT parallel corpora, we chose to randomly extract 2000 sentence pairs of at least 10 characters each from the filtered corpus to serve as the development and test sets, with the remainder serving as the training set.

Figure 2 shows all of the process of building the WCC-JC 2.0 corpus, and it is separated into four main steps: (1) web crawling; (2) extraction; (3) manual alignment; and (4) corpus segmentation. The procedures described in Sections 3.2–3.5 encompass the aforementioned steps.

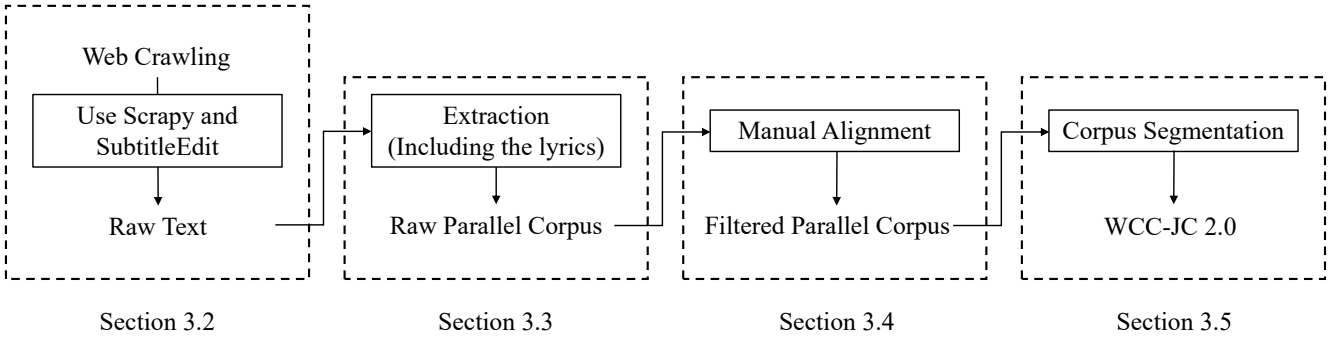

Figure 2. The main steps to build the corpus from the raw subtitle files.

The tabular data in Table 6 display the quantity of sentences contained in three distinct corpora: ASPEC-JC [11], OpenSubtitles [19], and the newly constructed corpus WCC-JC 2.0.

For the data with storage capacity in Table 6, ASPEC-JC >WCC-JC 2.0 > OpenSubtitles. The underlying cause for this discrepancy can be attributed to the varying sentence lengths in the different corpora. As shown in Figure 3, the distribution of sentence length in the WCC-JC 2.0 corpus reveals that, on average, Chinese and Japanese sentences in WCC-JC

2.0 have lengths of 11.69 and 15.02 characters, respectively. This can be attributed to the fact that WCC-JC 2.0 consists of spoken subtitles, which are generally shorter in nature.

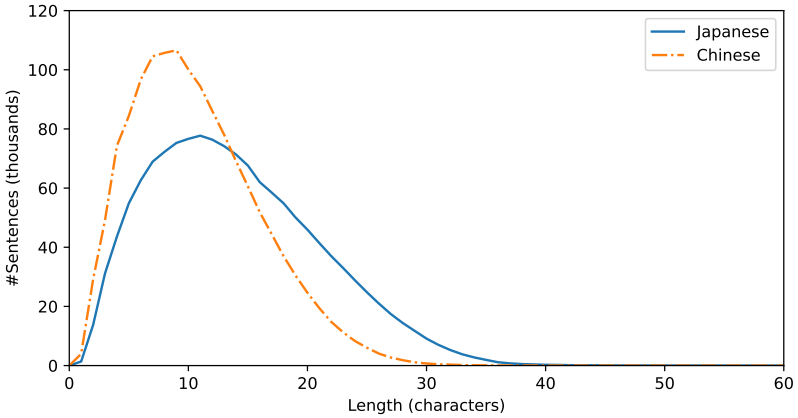

Figure 3. Sentence length distribution for WCC-JC 2.0. The length of more than 60 characters is exceeded by 5 Chinese sentences and 73 Japanese sentences.

Table 6. An elaborate account of the bilingual corpora in Japanese and Chinese.

Contents	Number of Chinese and Japanese Bilingual Sentences		
	ASPEC-JC (184.8 MB)	OpenSubtitles (72.4 MB)	WCC-JC 2.0 (101.1 MB)
Training set	672,315	1,087,295	1,398,441
Development set	2090	2000	2000
Test set	2107	2000	2000

The WCC-JC 1.0 and WCC-JC 2.0 corpora were created using different methods, with WCC-JC 1.0 being generated through automatic crawling and WCC-JC 2.0 created through manual alignment. In order to achieve optimal translation results, the training data from WCC-JC 1.0 and WCC-JC 2.0 were merged, and duplicates were removed, as described in Section 4.2.2.

4. Experiment and Evaluation

To verify the effectiveness of the corpus, a series of comparative experiments was carried out. The NMT system used in the experiments was configured as described in Section 4.1. In Section 4.2, the BLEU scores were calculated for ASPEC-JC, OpenSubtitles, and the newly constructed corpus using their respective NMT models. Furthermore, we manually evaluated the translation results of the test data W using JPO scores.

Previous studies on Japanese-Chinese translation have indicated that the character level and transformer models are the most effective models [1]. As a result, we opted to employ the character level and transformer models in our subsequent experiments.

4.1. Configuration of the NMT System

In our subsequent experiments, we employed Fairseq [29] to train our model. We utilized the predefined transformer-iwslt-de-en [30] architecture provided by Fairseq. This architecture comprises six encoder layers and six decoder layers, each equipped with eight encoder attention heads. The NMT model employed in this study was based on the transformer architecture, with an embedding dimensionality of 512. In order to ensure comparability, we maintained the other hyperparameters to be consistent with prior studies, including a dropout probability of 0.1, the Adam optimizer with beta coefficients set to 0.9 and 0.98, a learning rate of 1×10^{-7}, a maximum sequence length of 4096 tokens, a maximum number of 150,000 updates, and a batch size of 128. During inference, a beam search with a width of five was employed to generate translation outputs. Japanese and Chinese sentences were tokenized using MeCab (http://taku910.github.io/mecab

(accessed on 10 October 2022)) for Japanese and Jieba (http://github.com/fxsjy/jieba (accessed on 10 October 2022)) for Chinese, respectively, as they are written without spaces.

One frequently used metric for evaluating the quality of machine-translated text is the Bilingual Evaluation Understudy (BLEU) score [31]. To compute the BLEU scores for our experiments, we utilized the "fairseq-score" command, which conducted word-level evaluation after word segmentation.

4.2. Evaluation

For the manual evaluation stage, we enlisted the expertise of native Chinese and Japanese speakers. The Chinese evaluators were individuals who held a master's degree or higher and had studied in Japan, in addition to passing the Japanese Language Proficiency Test (JLPT) at the N2 level or higher.

4.2.1. Evaluation of Translation Results

In addition to the alignment evaluation, we further assessed the quality of the translations produced by the WCC-JC 2.0 corpus. Our assessors employed the adequacy criterion of the Japanese Patent Office (JPO) to appraise the extent of content transference in the translations. The JPO criteria provides a five-point scale, where five represents the best score and one represents the worst score (https://www.jpo.go.jp/system/laws/sesaku/kikaihonyaku/tokkyohonyaku_hyouka.html (accessed on 10 October 2022)). The JPO adequacy criterion is outlined in Table 7 and is also demonstrated in Section 4.2.2.

Table 7. The specificities of the JPO adequacy criterion.

Scores	Grading Criteria
5	All of the crucial information has been correctly transmitted. (100%)
4	Nearly all of the important information has been correctly transmitted. (80∼100%)
3	Over half of the crucial information has been correctly transmitted. (50∼80%)
2	A portion of the important information has been correctly transmitted. (20∼50%)
1	Almost none of the important information has been correctly transmitted. (∼20%)

4.2.2. Machine Translation Performance Evaluation, Manual Evaluation Results, and Analysis

In Tables 8 and 9, the test data from ASPEC-JC are designated as "A", the test data from OpenSubtitles are "O", and the test data of the previous version of WCC-JC, WCC-JC 1.0, are referred to as "W 1.0". To evaluate the generalization capability of the translation models, we also included the test data from the updated WCC-JC 2.0, which are referred to as "W 2.0". Finally, "WCC-JC 1.0 + 2.0" represents the merged and de-duplicated training data from both WCC-JC 1.0 and WCC-JC 2.0. The transformer model, which is a preconfigured architecture provided by Fairseq, specifically the transformer-iwslt-de-en variant, shall henceforth be referred to as "Transformer" for the sake of brevity and convenience in our exposition.

Based on the results presented in Tables 8 and 9, it can be concluded that the training data of WCC-JC 1.0 + 2.0 were found to produce the best results for all test data (with the exception of the ASPEC-JC test data A), followed by WCC-JC 2.0. This superiority is attributed to the notable increase in the number of sentence pairs in WCC-JC 2.0, as well as the mismatch between the content of scientific papers in ASPEC-JC and the content of spoken language texts.

Table 8. Results of Japanese→Chinese translation experiments at character level (BLEU scores).

Data\Method	**Character Level Transformer**			
Test data→	A	O	W 1.0	W 2.0
ASPEC-JC	**34.5**	1.2	3.2	3.5
OpenSubtitles	0	2.1	0	0.6
WCC-JC 1.0	3.2	3.9	15.9	13.4
WCC-JC 2.0	6.0	4.3	18.3	18.1
WCC-JC 1.0+2.0	6.5	**4.6**	**20.2**	**22.8**

Table 9. Results of Chinese→Japanese translation experiments at character level (BLEU scores).

Data\Method	**Character Level Transformer**			
Test data→	A	O	W 1.0	W 2.0
ASPEC-JC	**44.8**	1.7	4.2	3.7
OpenSubtitles	0.1	4.0	2.3	2.0
WCC-JC 1.0	3.5	3.9	17.1	14.0
WCC-JC 2.0	5.0	**4.3**	20.0	20.0
WCC-JC 1.0+2.0	5.6	**4.3**	**21.1**	**22.9**

The results indicate that WCC-JC 2.0 and WCC-JC 1.0 + 2.0 achieved BLEU scores above 20 on the W 1.0 and W 2.0 test data, demonstrating the quality of WCC-JC 2.0 as a corpus of spoken language while taking into account that spoken language sentences inherently have multiple interpretations (with multiple translations). These results also highlight the generalizability of WCC-JC 2.0 to test data from other corpora.

It is worth noting that the best BLEU values obtained by WCC-JC 2.0 were significantly higher compared with those of WCC-JC 1.0. This suggests that WCC-JC 2.0 is a more advanced corpus than WCC-JC 1.0 and is thus ready for practical applications.

For the WCC-JC 1.0 + 2.0 test data, W 2.0, the highest BLEU scores of 22.8 and 22.9 were achieved in the Japanese-to-Chinese (J→C) and Chinese-to-Japanese (C→J) language pairs, respectively. To further evaluate the translation results of the W 2.0 test data, a manual evaluation was performed using the JPO scores. Figures 4 and 5 present the results of the manual evaluation for the J→C and C→J language pairs, respectively.

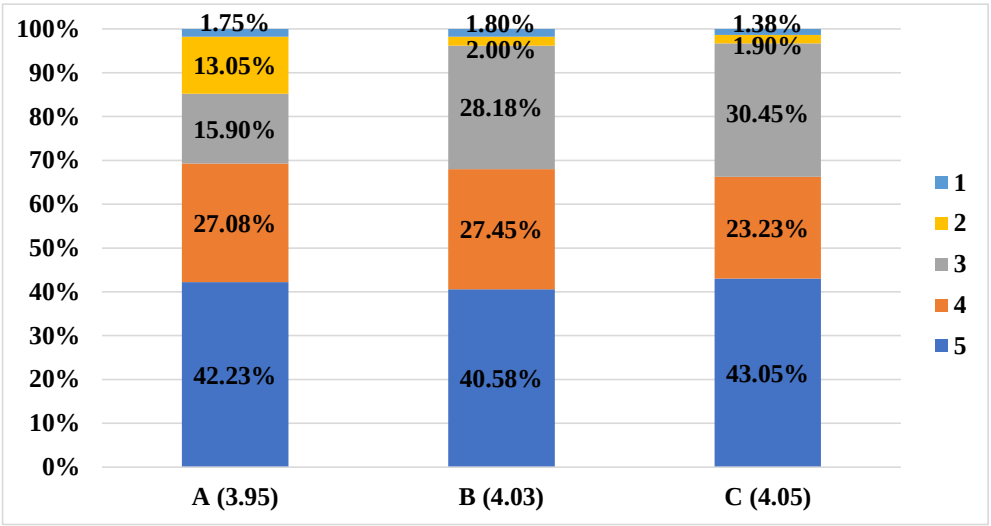

Figure 4. Manual evaluation outcomes for Japanese→Chinese translations.

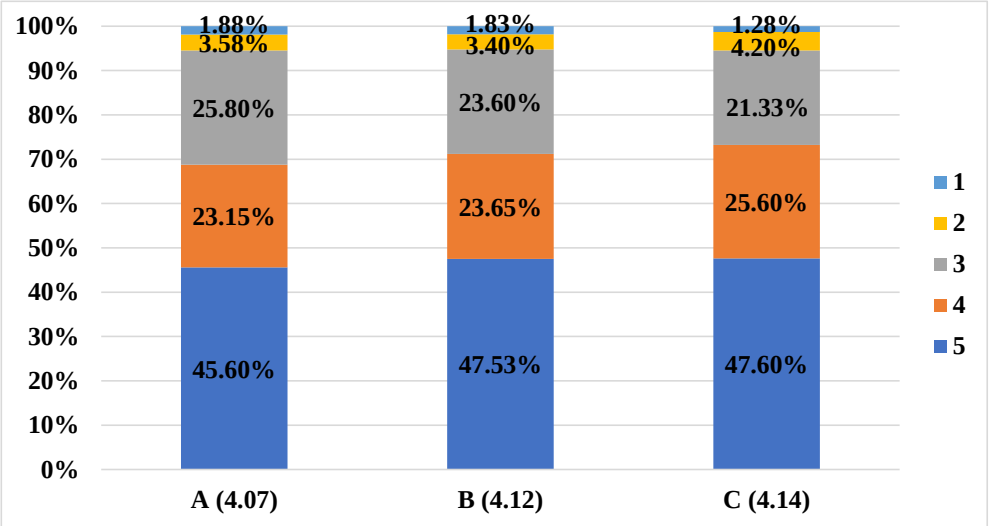

Figure 5. Manual evaluation outcomes for Chinese→Japanese translations.

We carried out the manual evaluations of the translation outputs for both the J→C and C→J directions using the W 2.0 test set of WCC-JC 1.0 + 2.0. Three groups, labeled as A, B, and C, were established, with each group consisting of two evaluators. The criteria for evaluation were detailed in Section 4.2.1. Table 10 presents the individual evaluations of each evaluator.

Table 10. Details of the human evaluators involved in the study.

	Group A		**Group B**		**Group C**	
Age	33	26	34	30	29	34
Gender	Male	Male	Female	Male	Female	Male
Profession	Laboratory Researcher	Doctoral Student	Assistant Professor	Doctoral Student	Publisher's Personnel	Assistant Professor
Proficiency Levels in Chinese and Japanese	Outstanding	Experienced	Outstanding	Experienced	Experienced	Outstanding

The average values were calculated and denoted as numerical values in parentheses following the group names. The JPO scores averaged across the J→C and C→J experiments were 3.95, 4.03, and 4.05 as well as 4.07, 4.12, and 4.14, respectively. The bar chart indicates that group C achieved slightly higher scores compared with the other groups, as reflected by the percentage of evaluation values for each group. The results of the manual evaluation indicated relatively good translation performance.

5. Publication of Datasets and Copyright Regulations

The acquisition and dissemination of data extracted from the internet presents the issue of potential copyright infringement, as such data may contain copyrighted works belonging to others. Our previous study addressed related copyright concerns.

We have sought advice from professional legal experts and have determined that WCC-JC 2.0 is compliant with the copyright regulations of both China and Japan.

6. Conclusions

The WCC-JC 2.0 corpus presented in this study is a significant contribution to the field of Japanese-Chinese spoken language corpora. The corpus, which comprises approximately 1.4 million sentence pairs of Japanese-Chinese bilingual data, was constructed through a large-scale collection of Japanese-Chinese bilingual sentences from subtitles and subsequent manual alignment. The WCC-JC 2.0 corpus is distinguished by its large size and public

accessibility, making it one of the most extensive Japanese-Chinese bilingual corpora available. Additionally, its uniqueness stems from its focus on spoken language data, which are primarily derived from subtitle files.

The effectiveness of the WCC-JC 2.0 corpus was evaluated through Japanese-Chinese translation experiments and manual evaluations. Future work will focus on unifying the WCC-JC series of corpora and exploring data augmentation techniques, which could further enhance the quality of the corpus and provide more opportunities for research in the field. The WCC-JC 2.0 corpus and the advancements made in this study have the potential to significantly reduce the knowledge gap in Japanese-Chinese NMT and further drive the development of this field.

Author Contributions: Conceptualization, J.Z.; methodology, J.Z., Y.T. and T.M.; software, J.M. and T.M.; validation, J.Z., Y.T., J.M., M.H., C.G. and Z.G.; formal analysis, J.Z. and Y.T.; investigation, J.Z., Y.T., J.M. and T.M.; resources, J.Z., Y.T., M.H., C.G., Z.G. and T.M.; data curation, J.Z., Y.T., J.M., M.H., C.G., Z.G. and T.M.; writing—original draft preparation, J.Z.; writing—review and editing, J.Z., Y.T., M.H. and T.M.; visualization, J.Z., J.M. and T.M.; supervision, J.Z., F.W. and T.M.; project administration, J.Z., F.W. and T.M.; funding acquisition, J.Z. and F.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the General Young Talents Project for Scientific Research grant of the Educational Department of Liaoning Province (Grant No. LJKZ0267), the Research Support Program for Inviting High-Level Talents grant of Shenyang Ligong University (Grant No. 1010147001004), the 2021 Shenyang Ligong University Research and Innovation Team Development Program Support Project (Grant No. SYLUTD202105), and the 2020 Program for Liaoning Excellent Talents (LNET) in University.

Data Availability Statement: A demo (200,000) of the WCC-JC 2.0 presented in this research is openly available on Github (https://github.com/zhang-jinyi/Web-Crawled-Corpus-for-Japanese-Chinese-NMT (accessed on 6 November 2022)). If you wish to obtain the entire dataset, kindly contact the email address specified on the Github page. However, please be aware that the data are exclusively for research purposes.

Acknowledgments: The authors of this research extend their gratitude to the anonymous reviewers and editors for their valuable feedback and assistance. Furthermore, they would like to acknowledge the contributions of the numerous individuals who provided support in the development and evaluation of WCC-JC 2.0.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. Zhang, J.; Tian, Y.; Mao, J.; Han, M.; Matsumoto, T. WCC-JC: A Web-Crawled Corpus for Japanese-Chinese Neural Machine Translation. *Appl. Sci.* **2022**, *12*, 6002. [CrossRef]
2. Leech, G.; Rayson, P.; Wilson, A. *Word Frequencies in Written and Spoken English: Based on the British National Corpus*; Taylor & Francis: Abingdon, UK, 2001. [CrossRef]
3. Ting, J. On Construction of Forensic Parallel Corpus. *J. Chongqing Univ. Social Sci. Ed.* **2005**, *4*, 94–97.
4. Tiedemann, J. Synchronizing Translated Movie Subtitles. In Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08), Marrakech, Morocco, 26 May–1 June 2008.
5. Čermák, F. Spoken Corpora Design: Their Constitutive Parameters. *Int. J. Corpus Linguist.* **2009**, *14*, 113–123. [CrossRef]
6. Aziz, W.; de Sousa, S.C.M.; Specia, L. Cross-lingual Sentence Compression for Subtitles. In Proceedings of the 16th Annual conference of the European Association for Machine Translation, Trento, Italy, 28–30 May 2012; pp. 103–110.
7. Wang, X.; Utiyama, M.; Finch, A.; Sumita, E. Refining Word Segmentation Using a Manually Aligned Corpus for Statistical Machine Translation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 25–29 October 2014; pp. 1654–1664. [CrossRef]
8. Svenja, A.; Ronald, C. *Spoken Corpus Linguistics: From Monomodal to Multimodal*; Routledge: London, UK, 2015. [CrossRef]
9. Liu, Y.; Sun, M. Contrastive Unsupervised Word Alignment with Non-Local Features. In Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, Austin TX, USA, 25–30 January 2015; pp. 2295–2301.
10. Tiedemann, J. OPUS—parallel corpora for everyone. In Proceedings of the 19th Annual Conference of the European Association for Machine Translation: Projects/Products, Riga, Latvia, 30 May–1 June 2016.

11. Nakazawa, T.; Yaguchi, M.; Uchimoto, K.; Utiyama, M.; Sumita, E.; Kurohashi, S.; Isahara, H. ASPEC: Asian Scientific Paper Excerpt Corpus. In Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2016), Portorož, Slovenia, 23–28 May 2016; pp. 2204–2208.
12. Tiedemann, J. Finding Alternative Translations in a Large Corpus of Movie Subtitle. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), Portorož, Slovenia, 19 May 2016; pp. 3518–3522.
13. van der Wees, M.; Bisazza, A.; Monz, C. Measuring the Effect of Conversational Aspects on Machine Translation Quality. In Proceedings of the COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, Osaka, Japan, 11–16 December 2016; pp. 2571–2581.
14. Lison, P.; Bibauw, S. Not All Dialogues are Created Equal: Instance Weighting for Neural Conversational Models. In Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue, Saarbrücken, Germany, 15–17 August 2017; pp. 384–394. [CrossRef]
15. Wang, S.; Zhang, Q.C.; Zhang, L. Natural language semantic corpus construction based on cloud service platform. *Int. Conf. Mach. Learn. Cybern.* **2017**, *2*, 670–674.
16. Levshina, N. Online film subtitles as a corpus: An n-gram approach. *Corpora* **2017**, *12*, 311–338. [CrossRef]
17. Love, R.; Dembry, C.; Hardie, A.; Brezina, V.; McEnery, T. The spoken BNC2014: Designing and building a spoken Corpus of everyday conversations. *Int. J. Corpus Linguist.* **2017**, *22*, 319–344. [CrossRef]
18. Chen, Z.; Huang, M. Corpus Construction for ESP. *Int. J. Corpus Linguist.* **2017**, *1*, 35–38.
19. Lison, P.; Tiedemann, J.; Kouylekov, M. OpenSubtitles2018: Statistical Rescoring of Sentence Alignments in Large, Noisy Parallel Corpora. In Proceedings of the the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan, 7–12 May 2018.
20. Lison, P.; Tiedemann, J. OpenSubtitles2016: Extracting Large Parallel Corpora from Movie and TV Subtitles. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), Portorož, Slovenia, 23–28 May 2016; pp. 923–929.
21. Ren, X.; An, X.; Fan, S. Corpus Construction of Precision Medicine. In Proceedings of the 2020 10th International Conference on Bioscience, Biochemistry and Bioinformatics, New York, NY, USA, 19–22 January 2020; pp. 74–77. [CrossRef]
22. Davies, M. The TV and Movies corpora: Design, construction, and use. *Int. J. Corpus Linguist.* **2020**, *26*, 10–37. [CrossRef]
23. Doi, K.; Sudoh, K.; Nakamura, S. Large-Scale English-Japanese Simultaneous Interpretation Corpus: Construction and Analyses with Sentence-Aligned Data. In Proceedings of the 18th International Conference on Spoken Language Translation (IWSLT 2021), Bangkok, Thailand, 5–6 August 2021; pp. 226–235. [CrossRef]
24. Meftah, A.H.; Qamhan, M.A.; Seddiq, Y.M.; Alotaibi, Y.A.; Selouani, S.A. King Saud University Emotions Corpus: Construction, Analysis, Evaluation, and Comparison. *IEEE Access* **2021**, *9*, 54201–54219. [CrossRef]
25. Xu, J. Application Research of Cognitive Linguistics Based on Big Data Internet Corpus Construction. *J. Phys. Conf. Ser.* **2021**, *1861*, 012028. [CrossRef]
26. Duszkin, M.; Roszko, D.; Roszko, R. New Parallel Corpora of Baltic and Slavic Languages—Assumptions of Corpus Construction. In *Proceedings of the Text, Speech, and Dialogue*; Ekštein, K., Pártl, F., Konopík, M., Eds.; Springer International Publishing: Cham, Switzerland, 2021; pp. 172–183.
27. Liu, S. *Overcoming Challenges in Corpus Construction: The Spoken British National Corpus 2014, by Robbie Love*; Routledge: New York, NY, USA, 2020; 202p. ISBN 978-1-138-36737-1. [CrossRef]
28. Ha, H. Vocabulary Demands of Informal Spoken English Revisited: What Does It Take to Understand Movies, TV Programs, and Soap Operas? *Front. Psychol.* **2022**, *13*, 1–7. [CrossRef] [PubMed]
29. Ott, M.; Edunov, S.; Baevski, A.; Fan, A.; Gross, S.; Ng, N.; Grangier, D.; Auli, M. fairseq: A Fast, Extensible Toolkit for Sequence Modeling. In Proceedings of the the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations), Minneapolis, MI, USA, 2–7 June 2019; pp. 48–53. [CrossRef]
30. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is All You Need. In Proceedings of the the 31st International Conference on Neural Information Processing Systems (NIPS'17), Red Hook, NY, USA, 2017; pp. 6000–6010.
31. Papineni, K.; Roukos, S.; Ward, T.; Zhu, W.J. Bleu: a Method for Automatic Evaluation of Machine Translation. In Proceedings of the the 40th Annual Meeting of the Association for Computational Linguistics, Philadelphia, PA, USA, 7–12 July 2002; pp. 311–318. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.