

ICS 03.080.99;35.240.30
CCS A 10



**NATIONAL STANDARD OF THE
PEOPLE'S REPUBLIC OF CHINA**

GB/T 40035-2021

**Basic requirements for bilingual parallel corpus processing
services**

双语平行语料加工服务基本要求

Issued on: April 30, 2021

Implemented on: November 1, 2021

**Issued by: State Administration for Market Regulation;
Standardization Administration of the PRC**

Table of Contents

1 Scope	1
2 Normative references	1
3 Terms and definitions	1
4 General	2
5 Basic requirements	3
5.1 Service provider	3
5.2 Corpus Processing Staff	3
5.3 Service environment	3
5.4 Processing content	3
5.5 Processing results	3
5.5.1 Completeness	3
5.5.2 Accuracy	3
5.5.3 Availability	4
5.5.4 Normative	4
5.6 Corpus Processing Tools	4
5.6.1 Reliability	4
5.6.2 Ease of use	4
5.6.2.1 Localization interface	4
5.6.2.2 Operation function	4
5.6.2.3 Help system	5
5.6.2.4 Efficiency	5
5.6.3 Compatibility	5
6 Processing flow	5
6.1 Pretreatment	5
6.1.1 Corpus preparation	5
6.1.2 Cleaning	5
6.1.3 Deduplication	5
6.1.4 Desensitization	5
6.2 Corpus Alignment	6
6.3 Corpus review	6
7 Service content	6
7.1 Demand communication	6

7.2 Customer Agreement	6
7.3 Project Management	6
7.4 Processing link	6
7.5 Delivery content	7
7.6 Warranty period	7
7.7 Service evaluation and improvement	7
8 Data Security	7
8.1 Data backup	7
8.2 Document Management and Log	7
14 Reference	15

Foreword

This document is in accordance with the provisions of GB/T 1.1-2020 "Guidelines for Standardization Work Part

1. Structure and Drafting Rules of Standardization Documents" Drafting. Please note that some of the contents of this document may involve patents. The issuing agency of this document is not responsible for identifying patents. This document was proposed and managed by the National Technical Committee for Language and Terminology Standardization (SAC/TC62). Drafting organizations of this document. China National Institute of Standardization, China Translation Association, Shanghai Yizhe Information Technology Co., Ltd., Shanghai Youyi Information Division Technology Co., Ltd., GTCOM Technology Co., Ltd., Beijing Yueer Information Technology Co., Ltd., Suzhou Lianyue Technology Co., Ltd., Sichuan Yanqiao Information Technology Co., Ltd., Beijing Baidu Netcom Technology Co., Ltd., Shenyang Yayi Network Technology Co., Ltd., Shanghai Zhishanhe Network Division Technology Co., Ltd., Beijing Language and Culture University, Beijing University of Posts and Telecommunications. The main drafters of this document. Liu Zhiyang, Zhang Jing, Ye Jian, Chai Ying, Huang Baorong, Luo Huifang, Meng Yongye, Zhu Li, Zhang Xuetao, Wang Haitao, Zhu Xianchao, Han Lintao, Zheng Chunping, He Zhongjun, Yu Limei, Zhang Chunliang, Gan Keqin, Zhang Baolin. Basic requirements for bilingual parallel corpus processing services

1 Scope

China's national standard for the basic requirements of bilingual parallel corpus processing services. It specifies the terms and definitions, the general provisions and the basic requirements. A parallel corpus is a body of text in two languages with the segments aligned to each other, and it is the raw material of machine translation and of professional translation memory. Building one is a service industry in its own right: text is collected,

cleaned, segmented into sentences, aligned across the languages, checked, and annotated. The quality of what results determines the quality of every model trained on it, and the failure modes are specific - misaligned segments teach a model to translate one sentence as another, and a corpus scraped without attention to its provenance carries the machine-translated text of earlier systems, so a model trained on it learns their errors. Standardising the process is how a buyer of such a service can specify what they are getting.

This document specifies the basic requirements, processing procedures, service content and data security of bilingual parallel corpus processing services. This document is applicable to the digital bilingual corpus processing service that takes the original text and the translation as the object and the text as the expression form, and other digital The corpus processing of the text can also be used as a reference, and it is also suitable for the evaluation of corpus alignment tools.

2 Normative references

There are no normative references in this document.

3 Terms and definitions

The following terms and definitions apply to this document.

3.1 Text Data formed by the arrangement of characters, symbols, words, phrases, paragraphs, sentences, tables or other characters, used to express meaning, and its interpretation is basically It depends on the reader's knowledge of a certain natural language or artificial language. [Source: GB/T 4894-2009, 4.1.1.2.4]

3.2 Corpus Language materials or information.

3.3 Bilingual parallel corpus bilingualparallelcorpus A corpus composed of two languages and aligned in parallel in chapters, paragraphs, sentences or other levels (3.2).

3.4 Original sourcelanguagetext Source language text (3.1). [Source: GB/T 19363.1-2008, 3.4, with modification]

3.5 Translation targetlanguagetext Target language text (3.1). [Source: GB/T 19363.1-2008, 3.5, with modification]

3.6 Client Accept individuals or organizations that provide products or services according to their requirements. [Source: GB/T 19000-2016, 3.2.4, with modification]

3.7 Metadata Descriptive data about the content, quality, condition, and other characteristics of the data.