Modern Linguistics 现代语言学, 2024, 12(8), 489-495
Published Online August 2024 in Hans. https://www.hanspub.org/journal/ml
https://doi.org/10.12677/ml.2024.128716

Hans汉斯

基于“讲述中国故事”日语文本的专用分词词表构建研究

——以《人民网日文版》为例

谭元昌，吴　雨*

重庆交通大学外国语学院，重庆

收稿日期：2024年7月5日；录用日期：2024年8月15日；发布日期：2024年8月27日

摘　要

用日语讲述中国故事、传播中国声音是面向日本构建中国国际形象的重要一环，也是国内日语教育的主要目标之一。《人民网日文版》正是国家对日宣传的重要媒介，可用作培养讲述中国故事日语人才的教学资源。为分析该媒体的对日新闻文本中国形象宣传现状，充分挖掘其教育教学功能，收集《人民网日文本》的新闻文本并构建语料库开展量化分析，是较为有效的研究路径。准确的分词结果是量化分析日语语料的前提，但研究发现，目前的日语分词工具难以处理中国故事日语文本的精确分词，将严重影响分析结论的可靠性。因此，本研究抽取《人民网日文版》新闻文本中与中国社会、经济、文化和科技等相关的日语表述，构建适用于中国故事日语文本的专用分词词表，并评测该词表的实用效果。

关键词

讲好中国故事，日语文本，日语分词，词表构建，人民网日文版

A Study on Constructing a Custom Word Segmentation Dictionary for Japanese Texts Based on “Telling China’s Story”

—A Case Study of *People’s Daily Japanese Edition*

Yuencheong Tam, Yu Wu*

School of Foreign Languages, Chongqing Jiaotong University, Chongqing

*通讯作者。

文章引用：谭元昌，吴雨．基于“讲述中国故事”日语文本的专用分词词表构建研究[J]．现代语言学，2024，12(8): 489-495. DOI: 10.12677/ml.2024.128716

Received: Jul. 5th, 2024; accepted: Aug. 15th, 2024; published: Aug. 27th, 2024

Abstract

Telling China's story and spreading China's voice in Japanese are crucial for shaping China's international image in Japan and are also one of the primary goals of domestic Japanese language education. The People's Daily Japanese Edition is an important medium for China's stories towards Japan and can serve as a valuable resource for training Japanese language talents to tell China's story. To analyze the current state of China's image publicity in this media's news texts and to fully exploit its educational functions, collecting and constructing a corpus of these news texts for quantitative analysis is an effective research approach. Accurate word segmentation is a prerequisite for the quantitative analysis of Japanese corpora. However, research has found that current Japanese word segmentation tools struggle to precisely segment texts related to China's story, which significantly affects the reliability of the analysis results. Therefore, this study extracts Japanese expressions related to Chinese society, economy, culture, and technology from the People's Daily Japanese Edition news texts, constructs a custom segmentation word dictionary for these texts, and evaluates the accuracy and practicality of this dictionary.

Keywords

Telling Chinese Stories Well, Japanese Texts, Japanese Word Segmentation, Word Segmentation Dictionary Construction, People's Daily Online Japanese Edition

1. 引言

日本作为中国的重要邻国，是“讲述中国故事”的主要受众之一，用日语讲好中国故事是构建中国国际形象的重要环节。在新文科背景下，必须不断地丰富教学资源，在输出日语专业知识的同时，深入挖掘中国元素，增加中国文化比重，突出中国文化特色，反映当代中国的发展面貌[1]。《人民网日文版》作为中国国内主流的日语媒体，在利用新闻话语讲述中国故事、传播中国声音中扮演着关键角色。同时也是国内日语学习者构建中国故事日语文本的重要学习资源。通过收集和分析《人民网日文版》的新闻文本，对于培养日语学习者讲述中国故事的能力具有重要意义。

基于日语语料库研究方法的语料分析，若以词汇为分析单位，通常需要进行文本分词处理。然而运用现有的日语分词系统处理《人民网日文版》日语语料时，本研究发现处理结果中出现了大量单词切割和词性赋码的错误。例如，“粤港澳大湾区”一词被错误识别为“粤港”、“澳”、“大湾”、“区”四个词，并且词性信息赋码也被错误标记为“名词－固有名词－地名”，诸如此类的错误在其他相关的文本分词处理中也大量出现，无法确保后续语料分析的准确性。出现这一情况的原因在于现有的日语自动分词系统[1]均通过内置日语分词词表来实现日语文本的单词切割、词性赋码处理，然而，该类内置的日语分词词表是基于日本国内的语料资源构建和更新的，因此对国内以中国故事为主题的日语新闻文本

[1]日语自动分词系统是指可以将日文文本或句子自动切分为词单位并进行词性赋码的软件系统，MeCab、Juman 和 Chasen 等是目前常用的日语自动分词、赋码系统。

的适配性较差。现有的日语分词系统主要有 MeCab、Juman 和 Chasen 等，其中 MeCab 是现代日语分词处理效率最高的[2]。而为了满足不同研究目的的分词需求，MeCab 也为使用者提供了修改系统分词词表和添加用户个性化词表的功能[3]。

因此，在收集以中国相关的日语新闻文本并构建语料库过程中，为了提高语料处理和分析的效率和可信度，本研究将以《人民网日文版》为对象，广泛收集语料，从中抽取具有中国特色的日语表达，构建适用于“讲述中国故事”日语文本的专用日语分词词表，并对导入专用分词词表前后的分词系统进行精度评价，以分析专用分词词表的精确度和实用性。

2. 国内外研究现状

在语料库语言学研究中，需要切割语料文本，对词汇的词性、语种等信息赋码。上述处理对于词汇分析、语法研究和词性分布调查是必不可少的环节，在以往的日语语料库研究中，采用人工标记的方式对收集的文本数据进行单词切割和词性赋码，随着文本规模的增加，为处理超过百万词以上的文本数据，利用计算机和自然语言处理技术开发的日语自动分词系统开始普及。但是，现有的日语分词词表对文本环境愈发复杂的语料处理出现难以适配、分词精度下降的情况[4]。现存的主流日语分词词表 IPAdic、NAIST Japanese Dictionary 等长期未更新词表信息，后期开发的 UniDic、NEolodg 等词表虽然在固有名词的识别上准确度有所提升，但对于专门用途的语料而言，自动分词后仍需要大量纠错[5]。

在构建特定领域的语料库过程中，使用日语自动分词系统预处理语料时，可以根据自身研究目的构建专门分词词表以提高语料库的准确性和可靠性[6]。在构建专门用途的日语分词词表研究中，基于分词词表的日语自动分词系统单靠内置的系统分词词表无法满足特定领域语料的分词需求，在相关研究中，研究者通过收集近代日语的杂志、新闻等语料形成学习语料库的方式，采用人工纠错，构建适用于近代日语文本的专用日语分词词表，计算词表导入前后的准确率发现，对近代日语文本的分词精度得到了明显的改善，可以提高后续语料分析工作的效率和质量[7]-[9]。

综上所述，本研究以改善以中国故事为主题的日语文本的分词处理效果为目标，采用可操作性高的日语自动分词系统 MeCab 作为分词工具，收集《人民网日文版》的新闻语料，构建适用于中国故事日语文本量化分析的专用日语分词词表，并评估专用分词词表导入前后的分词精度。同时，通过实例考察专用日语分词词表的实用性。以期通过本研究能够为培养中国日语学生讲述中国故事的能力提供语料支持。

3. 文本挖掘和整理

为提高中国故事日语文本的分词解析精度，本研究将构建专门针对中国故事日语文本的日语分词词表，以扩大分词系统的词表规模。通过文本挖掘方法，从《人民网日文版》中挖掘语言资源。利用现有的分词系统及其内置的分词词表，排除单词切割、词性信息赋码均正确的词项，手动筛选与中国故事日语文本相关的词项。

本研究通过爬虫程序，大规模抓取《人民网日文版》经济、社会、文化和科学四个主要板块中，时间范围为 2023 年 9 月 15 日至 10 月 31 日的日语新闻报道(http://j.people.com.cn/，引用日期：2023 年 11 月 18 日)。采用每个文本独立处理再去除重复词项的方法，并通过正则表达式[2]清洗文本中包含的中英文、标点和颜文字等无效字符。最终，我们得到了形符数 247,818 词，类符数 13,417 词规模的日语语料资源。

本研究采用日语自动分词系统 MeCab 作为分词解析工具，由于 MeCab 的单词切分依赖于计算词与词之间的连接值，而连接值是由日语自动分词系统计算每个单词的“左文脈 ID (左文脉 ID)”，“右文脈 ID (右文脉 ID)”和“コスト(出现值)”所得到的，并且词项的连接值越小出现概率越高。因此，需对从

[2] 正则表达式是指一种用于字符串操作的逻辑公式，由特定字符和组合构成，用于表达对字符串的过滤逻辑。

文本挖掘所获词项进行“左文脉 ID”、“右文脉 ID”和“出现值”的赋码。

4. 专用分词词表构建与扩充

构建专用的日语分词词表需要收集准确使用该词汇的文本，形成文本数据集，并从实际使用的文本数据集中抽取对应词汇构建词表[10]。基于该方法，本研究抓取《人民网日文版》2023 年 9 月 15 日至 10 月 31 日的新闻文本作为文本数据集，并导入 MeCab 日语自动分词系统中，每个文本独立地进行分词和赋码处理。观察分词结果发现，出现单词切分错误的词项通常与“讲述中国故事”主题相关，且由于单词切分不准确，导致词性赋码均出现错误，主要集中在名词大类。

分词结果中出现错误的词项是由于在系统分词词表 IPAdic 中并没有记录相应的词项，使得在单词识别过程中无法准确被标记为一个完整词单位，同时其词性标记结果等信息也未能完整地作为分词结果一同被输出。针对此问题本研究的解决思路是将每个文本独立分词的结果去除单词切分且词性赋码正确的词项，对单词切分和词性赋码错误的词项逐一纠错，整理成专用分词词表。表 1 即为该处理过程中，MeCab 日语自动分词系统中分词词表的添加格式(以“重慶市”为例)。

Table 1. Format of word segmentation dictionary in MeCab
表 1. MeCab 中分词词表格式

表層形	左文脈 ID	右文脈 ID	コスト	品詞	品詞細分類 1	品詞細分類 2	品詞細分類 3	活用型	活用形	原形	読み	発音
重慶市	1293	1293	0	名詞	固有名詞	地域	一般	*	*	重慶市	ジュウケイシ	ジュウケイシ

由于中国故事日语文本中出现单词切分、词性赋码的词性均属于名词大类，同时考虑定量分析中常见的词性分布分析的需求，本研究对错误词项的纠正集中在“表層形”，“品詞”，“品詞細分類 1”，“品詞細分類 2”，“品詞細分類 3”，“原形”，“読み”，“発音”，以及用作识别词与词之间界限的“左文脉 ID”，“右文脉 ID”和“出现值”，其中“左文脉 ID”、“右文脉 ID”和“出现值”的具体赋值在 MeCab 系统文件中根据词项的具体词性信息有具体的赋值参考[3]，以便 MeCab 能够准确识别新增词汇，从而初步构建成适用于“讲述中国故事”日语文本的专用分词词表，图 1 为部分分词词表。

表層形	左文脈ID	右文脈ID	コスト	品詞	品詞細分類1	品詞細分類2	品詞細分類3	活用型	活用形	原形	読み	発音
重慶市	1293	1293	0	名詞	固有名詞	地域	一般	*	*	重慶市	ジュウケイシ	ジュウケイシ
渝中区	1293	1293	0	名詞	固有名詞	地域	一般	*	*	渝中区	ユチュウク	ユチュウク
淘宝	1288	1288	0	名詞	固有名詞	一般	*	*	*	淘宝	タオバオ	タオバオ
湖北省	1293	1293	0	名詞	固有名詞	地域	一般	*	*	湖北省	コホクショウ	コホクショウ
武漢	1293	1293	0	名詞	固有名詞	地域	一般	*	*	武漢	ブカン	ブカン
殷商王朝	1288	1288	0	名詞	固有名詞	一般	*	*	*	殷商王朝	インショウオウチョウ	インショウオウチョウ
商代	1288	1288	0	名詞	固有名詞	一般	*	*	*	商代	ショウダイ	ショウダイ
衛国	1288	1288	0	名詞	固有名詞	一般	*	*	*	衛国	エイコク	エイコク
河南省	1293	1293	0	名詞	固有名詞	地域	一般	*	*	河南省	カナンショウ	カナンショウ

Figure 1. The custom word segmentation dictionary (partial presentation)
图 1. 专用分词词表(部分呈现)

上述研究过程中初步构建的专用分词词表共 443 词，通过观察词表中词项可以发现，基于原内置系统分词词表 IPAdic 致使 MeCab 未能准确识别的单词可以归类为中国的地名、行政机构、企业、传统文化和特定的科技术语等相关日语表达。为此，现有分词词表的这一特性，进一步引入该类别的其他日语

[3]MeCab 的系统文件 left-id.def，right-id.def 有每个单词词性的左文脉 ID、右文脉 ID 的具体赋值参考，出现值统一赋值为 0。

表达，对初步构建的专用分词词表进行词表规模扩充，扩大专用分词词表的覆盖范围，以实现提高中国故事日语文本的单词切分精度和词性赋码准确度的研究目标。经过词表规模扩充后单词数达到 969 词，后将专用分词词表以 MeCab 用户词表的形式导入分词系统。后续分词处理前可以同时选择系统分词词表和用户词表的方式解析文本数据，将提高该类文本的分词精度。

5. 词表的实用性考察

构建该专用分词词表的目的在于提高目标文本的分词精度，以提高语料的实用性。因此需要考察专用分词词表的实际使用效果，实用性考察主要分两部分，一是专用分词词表的精度评价，收集并构建与专用分词词表具有同类语言特征的语料形成评价语料库，比较导入专用分词词表前后的分词准确率。二是对实际新闻文本分词处理并分析专用分词词表的实用效果。

本研究专用分词词表精度评价运用评价语料库方法，收集同一新闻媒体《人民网日文版》2023 年 8 月的日语新闻文本作为对照语料，构建语料库规模为 25,858 词的评价语料库。由于专用分词词表的核心要素集中在单词切分和词性信息赋码，因此，主要将词境界、词性分类、发音形作为精度评价解析维度。使用准确率计算公式(公式如下)计算并比较专用分词词表导入前后的准确率，并观察专用分词词表在导入后的解析精度变化，导入专用分词词表前的准确率计算结果见表 2 和表 3。表 2 数据显示，每解析 10,000 词出现单词切分错误数达到 359 词，词性分类和发音形赋码的错误数分别达到了 418 词和 235 词。

$$\text{准确率} = \frac{\text{解析总数}-\text{解析错误数}}{\text{解析总数}} \times 100\%$$

Table 2. Parsing accuracy before importing the custom word segmentation dictionary
表 2. 专用分词词表导入前的解析准确率

	词境界	词性分类	发音形
准确率	96.41%	95.82%	97.65%

而专用分词词表导入 MeCab 后，表 3 数据表明每解析 10,000 词出现单词切分的错误词数减少至 110 词，词性分类和发音形赋码的错误词数分别减少至 92 词和 54 词。三个维度的准确率分别从 96.41%上升至 98.90%，95.82%上升至 99.08%，97.65%上升至 99.46%。分析准确率的数值变化可以发现，专用分词词表导入后，MeCab 的解析精度提升效果明显，如表 3 括号中的数值所示。

Table 3. Parsing accuracy after importing the custom word segmentation dictionary
表 3. 专用分词词表导入后的解析准确率

	词境界	词性分类	发音形
准确率	98.90% (+2.49)	99.08% (+3.26)	99.46% (+1.81)

本研究在分析专用分词表使用效果时，采取评测实际新闻文本分词效果的方式，对比得出专用分词词表在导入后有效提升了对“讲述中国故事”日语文本单词切分的准确性。收集《人民网日文版》的日语新闻文本，计算专用分词词表导入前后 MeCab 的解析数据，并统计在不同语料文本规模情况下专用分词词表导入前后的正确解析词数及其变化。其词数变化如图 2 所示，随着语料库类符数的增加，专用分词词表导入前后的正确解析词数差距愈发明显。

为分析专用分词表使用效果，本研究还通过例句解析的方式，抽取《人民网日文版》新闻文本例句，对比专用分词词表导入前后的解析情况，解析结果如表 4，可以发现专用分词词表导入后单词切分准确度比导入前更高。因此，适用于“讲述中国故事”日语文本的专用分词词表在实际的日语文本定量分析

中具有实用性和可靠性。

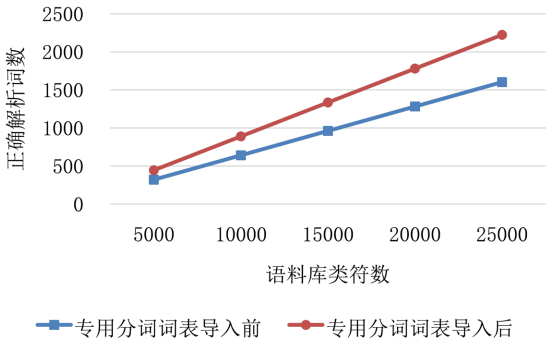

Figure 2. Changes in the number of correctly parsed words before and after importing the custom word segmentation dictionary with different corpus sizes
图 2. 不同语料规模情况下专用分词词表导入前后的正确解析词数变化

Table 4. Comparative analysis of example sentences before and after importing the custom word segmentation dictionary
表 4. 专用分词词表导入前后的例句对比分析

	例句 1	例句 2
文本	港珠澳大橋は粤港澳大湾区「相互接続」の成果	三星堆博物館を見学して古蜀文明に触れ
专用分词词表导入前	港|珠|澳大|橋|は|粤港|澳|大湾|区|「|相互|接続|」|の|成果	三|星|堆|博物館|を|見学|し|て|古|蜀|文明|に|触れ
专用分词词表导入后	港珠澳大橋|は|粤港澳大湾区|「|相互|接続|」|の|成果	三星堆|博物館|を|見学|し|て|古蜀|文明|に|触れ

日语新闻文本是培养讲述中国故事日语人才的教学资源。通过定量分析词汇、句型和语篇，可以高效提炼出准确的日语表述，构建“讲述中国故事”日语文本的专用分词词表。通过提高日语新闻文本的分词精度，能够确保语料分析的准确性及作为学习资源的可信度。

6. 结语

用日语讲好中国故事是对外传播中国形象的重要环节，也是当前日语教育研究的主要目标之一，量化分析“讲述中国故事”日语文本则是重要的研究手段，中国故事日语文本的量化分析依靠日语自动分词系统完成语料分词和词性赋码工作，为后续的语料分析提供保障。本研究针对目前日语自动分词系统难以适用于该类日语文本的困境，利用文本挖掘方法，以《人民网日文版》为对象，重点抓取其中具有中国故事特点的日语表达，构建适用于“讲述中国故事”日语文本的专用分词词表，以提高自动分词精度。

经数据对比、例句分析等评测发现，本研究构建的专用分词词表有效提升了分词精度，并在语料库规模扩大、确保更高的量化分析精确性方面呈现积极影响。今后本研究将持续收集相关语料，进一步扩展专用分词词表的覆盖范围，以期为培养日语学习者用日语讲述中国故事提供更多具有现实意义的资源支撑。

基金项目

重庆交通大学外国语学院研究生科研创新项目“讲好中国故事”背景下日语学习语料库建构与应用研究资助(编号：WYS23221)。

参考文献

[1] 尤芳舟. 新文科背景下日语课程思政建设的思考[J]. 外语学刊, 2021(6): 78-82.

[2] 毛文伟. 日语自动词性赋码器的信度研究[J]. 外语电化教学, 2012(3): 10-14.

[3] 工藤拓. 形態素解析の理論と実装[M]. 京都: 近代科学社, 2018.

[4] 伝康晴, 小木曽智信, 小椋秀樹, 山田篤, 峯松信明, 内元清貴, 小磯花絵. コーパス日本語学のための言語資源－形態素解析用電子化辞書の開発とその応用－[J]. 日本語科学, 2007(22): 101-123.

[5] 坂本美保, 川原典子, 久本空海, 高岡一馬, 内田佳孝. 形態素解析器『Sudachi』のための大規模辞書開発[C]//言語資源活用ワークショップ発表論文集. 東京: 国立国語研究所, 2018: 118-129.

[6] 伝康晴.多様な目的に適した形態素解析システム用電子化辞書[J].人工知能学会誌.2009(24):640-646.

[7] 小木曽智信, 小椋秀樹, 近藤明日子. 近代文語文を対象とした形態素解析辞書の開発[C]//言語処理学会第 14 回年次大会発表論文集. 東京: 言語処理学会, 2008: 225-228.

[8] 小木曽智信, 伝康晴, 渡部涼子, 近藤明日子. 現代語コーパスの利用による近代語形態素解析の精度向上[C]//言語処理学会第 15 回年次大会発表論文集. 神戸: 言語処理学会, 2009: 801-804.

[9] 小木曽智信, 小町守, 松本裕治. 歴史的日本語資料を対象とした形態素解析[J]. 自然言語処理, 2013(20): 727-748.

[10] 小木曽智信, 小椋秀樹, 田中牧郎, 近藤明日子, 伝康晴. 中古和文を対象とした形態素解析辞書の開発[J]. 情報処理学会, 2010(4): 1-5.