

基于语料库的中日地震新闻报道的对比研究

冷金桃

重庆交通大学外国语学院, 重庆

收稿日期: 2024年6月4日; 录用日期: 2024年7月12日; 发布日期: 2024年7月26日

摘要

本研究以中日两国权威媒体对各自国内的地震新闻报道为研究对象, 使用语料库的研究方法从形式、主题和情感三个维度分析两国媒体的报道差异。研究发现, 中国媒体在报道甘肃地震新闻时主要采用专栏等形式, 较为详尽地展示了社会各界对灾区的积极救援和支持, 情感态度积极。相比之下, 日本媒体在报道石川县地震时主要采用短讯的方式, 简要介绍地震发生的时间、地点等基本情况, 对地震灾害的描述比较客观, 且较少涉及日本社会各界对灾区的紧急救援, 情感态度较为消极。

关键词

语料库, 地震新闻, 对比研究

A Corpus-Based Comparative Study on Chinese and Japanese Earthquake Reports

Jintao Leng

School of Foreign Languages, Chongqing Jiaotong University, Chongqing

Received: Jun. 4th, 2024; accepted: Jul. 12th, 2024; published: Jul. 26th, 2024

Abstract

The study focuses on the earthquake news coverage by authoritative media in China and Japan, using corpus analysis to examine differences in form, theme, and emotion. It reveals that Chinese media mainly used columns and extensively detailed the active rescue efforts and support for the disaster area following the Gansu earthquake, with a positive emotional tone. In contrast, Japanese media provided brief and short message reports on the Ishikawa earthquake, with less emphasis on emergency response efforts and a relatively negative emotional tone.

Keywords

Corpus Analysis, Earthquake News, Comparative Study

Copyright © 2024 by author(s) and Hans Publishers Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

1. 引言

地震等灾难性事件由于其突发性、破坏力和异常性，一直备受关注，是新闻热点[1]。随着近年来灾难性事件频繁发生，如何妥善处理灾难新闻报道并引导社会舆论，已成为新闻报道领域的重要议题[2]。2023 年，中国的甘肃地震和日本的石川县地震造成了两国经济和社会建设受到严重打击。日本作为一个地震频发的国家，通过比较中日两国媒体在地震新闻报道上的差异，可以促进双方交流与借鉴，进一步提升灾难新闻报道引导社会舆论、降低社会风险、控制民众情绪以及进行心理疏导等方面的重要作用。本研究选取了 2023 年中国《人民网》和日本《共同网》对甘肃地震和石川县地震的约 50 篇新闻报道为研究对象，基于语料库的研究方法使用 Antconc 等软件从形式、主题和情感三个维度分析中日两国媒体对地震新闻报道的差异，以促进中日两国新闻媒体行业在地震新闻报道上的交流和借鉴。

2. 研究设计

2.1. 研究问题

本研究基于自建小型语料库，对比分析 2023 年《人民网》关于中国甘肃地震和《共同网》关于日本能登半岛石川县地震新闻报道的差异，并提出以下三个研究问题：1) 中日两国媒体对中国甘肃地震与日本石川县地震的新闻报道在形式维度有何差异？2) 中日两国媒体对中国甘肃地震与日本石川县地震的新闻报道在主题维度上有何差异？3) 中日两国媒体对中国甘肃地震与日本石川县地震的新闻报道在情感维度上有何差异？

2.2. 研究工具

本研究在文本处理和分析阶段使用了数款语料库软件来提高统计分析的效率和准确度。具体而言，在文本清洗与编码阶段，笔者选用了肖航研究员开发的 Corpus Word Parser 软件来进行中文文本的分词和词性编码工作[3]。Corpus Word Parser 是一款创建于 2014 年且免费使用的语料库分词和标注工具和该软件具有类似分词标注词性功能的软件“NLPPIR-ICTCLAS”系统(也可直接称为“ICTCLAS”)，都是由中国科学院计算技术研究所研制的，主要功能包括中文分词、词性标注、命名实体识别、新词识别等等[3]。

在文本分析阶段，笔者则运用了 Antconc 和“微词云”等工具来完成高频词统计、主题词聚类以及情感分析等任务。Antconc 作为语料库语言研究的重要工具近来常被用于研究大规模新闻文本的语言理论及应用研究[4]，具有较高的可信度。

2.3. 数据收集与编码

本研究从互联网上分别选取了 2023 年末，中国《人民网》和日本《共同网》关于中国甘肃和日本能登半岛地震的新闻报道。《人民网》和《共同网》都是中日两国的权威新闻媒体，在各自国内都有庞大

的读者群体,以两大媒体的新闻报道为研究对象,具有较高的可靠性和代表性,能反映中日出两国在报道本国地震新闻时的差异。首先,笔者在《人民网》和《共同网》新闻客户端搜索框中以“甘肃地震”和“日本地震”为关键词,形成“共同网-日本地震”和“人民网-甘肃地震”两个语料库。经初步计算,“共同网-日本地震”语料库共 11,957 个字符数(不计空格),“人民网-甘肃地震”语料库共计 26,449 字符数(不计空格)。

完成对文本数据的初步收集和整理后,语料中仍然纯在大量的非文本元素、非法字符以及重复内容等无关信息,如果直接用于语料分析,会严重影响研究的准确性和严谨性[5],因此笔者对收集来的两份文档进行了文本数据的预处理工作,流程如下:

1) 数据清洗

- a) 去除非法格式标签:使用正则表达式等方法移除文本中的非法标签,以便提取纯文本内容。
- b) 非法字符处理:替换或删除文本中的非法字符,比如特殊符号、非汉字数字字符等。
- c) 去除重复、无关内容:删除重复的文本片段以及记者姓名、时间等无关信息,减少分析时的噪音。

2) 分词

分词是指将文本拆分成词语或句子的过程。本研究使用教育部语言文字应用研究所肖航研究员开发的中文语料库分词和标注程序 *Corpus Word Parser* 来完成语料的分词和词序标注工作。在实验性操作过程中,笔者发现该程序的分词功能对地名、人名和日期等专有名词的识别效果不甚理想,便基于实验性操作结果手动创建并加载用户自定义词表,提升文本分词的准确性。

3) 词性标注

词性标注是为了识别文本中的词语的词性,如名词、动词、形容词、介词、地名、能源动词、代词、叹词、标点符号、地名、数词、机构名、联系动词、时间名词、方位名词等约 40 余种标注类型。肖航(2014)开发的中文语料库分词和标注程序 *Corpus Word Parser* 中自带词性赋码功能和软件使用的词性赋码集按照程序所预设的词性赋码集,使用自动赋码功能对分词后的文本进行实验性自动赋码。根据实验性自动赋码结果,笔者手工校对,修改明显的错误,对于有争议的结果,笔者与其他研究者充分讨论研究,最终达成一致意见,最终完成了语料的词性赋码工作。

在上述文本清洗和编码的基础上,剔除个别重复和无法访问的报道,最后筛选出共计 49 (中国甘肃地震新闻 21 篇,日本地震新闻 28 篇)篇新闻报道作为研究对象。最终形成了《人民网》和《共同网》关于 2023 年中日两国地震新闻报道的小型自建语料库。

3. 研究结果与讨论

3.1. 中日地震新闻报道的形式维度分析

通过将《人民网》关于中国甘肃地震新闻报道的分词后语料文件导入 *Antconc3.6* 版本,选择“词表”标签选项卡,我们得知该文件共有 23,305 个形符,共计 2958 个类符。该文件包含了 21 篇新闻报道,因此可以计算出《人民网》关于中国甘肃地震的新闻报道平均每篇约有 1109 字。通过类似方法,我们再次用 *Antconc3.6* 分析了《共同网》关于日本地震新闻报道的分词后语料文件,结果显示“共同网-日本地震分词”文件包含 9704 个形符,共计 1139 个类符。这一文件中包含了 28 篇新闻报道,因此《共同网》关于日本地震的新闻报道平均每篇约有 364 字。通过上述对中日地震新闻语料的形符和类符分析,可以明显看出语料中关于中国甘肃的地震新闻报道不仅在规模上是日本地震新闻报道的约 2 倍,而且就单篇新闻的字符数来说,关于中国甘肃地震的新闻报道也远大于对日本地震新闻的报道,约为日本地震新闻的 3 倍。

从新闻报道的形式维度来看,《人民网》在报道国内甘肃灾难新闻时,通常采用少量报道、篇幅较长、内容丰富的通讯、专栏等新闻形式来呈现国内灾难新闻;而《共同网》在报道日本国内石川县地震新闻时,则倾向于采用大量报道、篇幅较短的快讯、短讯等内容形式进行呈现。

在文本清洗过程中,笔者使用中文语料库分词和标注程序制作了中日地震新闻语料的词性标注文件。将中国甘肃和日本地震新闻报道的词性赋码文件分别导入到 Antconc3.6 版本中,按照中文语料库分词和标注程序所提供的词性赋码集,分别查找计算中国甘肃地震和日本地震新闻语料中的词性占比,并绘制了中日地震新闻语料的词性饼状图,见图 1 和图 2。

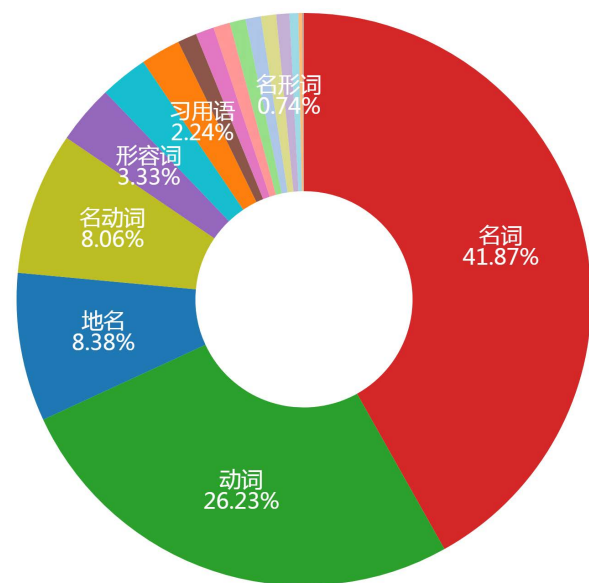


Figure 1. Part-of-speech diagram of Gansu earthquake news
图 1. 甘肃地震新闻的词性饼状图

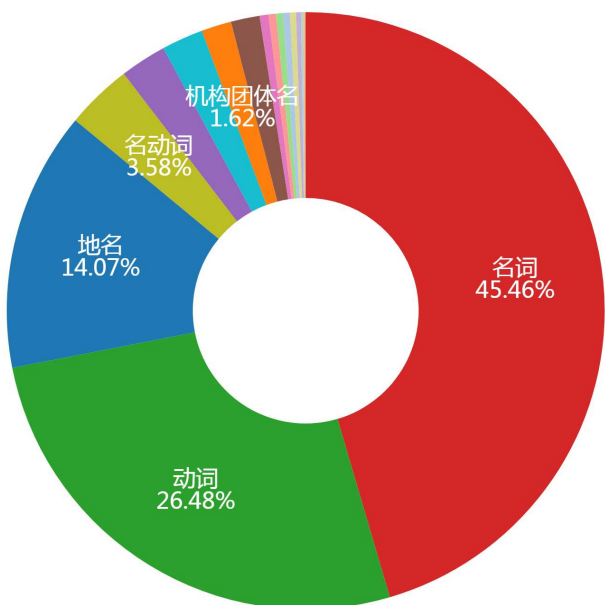


Figure 2. Part-of-speech diagram of Japanese earthquake news
图 2. 日本地震新闻的词性饼状图

通过对中日地震新闻语料的词性分析发现, 名词和动词是最主要的词类, 分别超过了 40%和 20%。此外, 甘肃地震新闻语料和日本地震新闻语料中名词和动词的比例差异不大。排名第三和第四的是地名和名动词。在甘肃地震新闻语料中, 地名占比约为 8%, 而在日本地震新闻语料中, 地名的比例超过了 14%, 这一差异显著。至于名动词, 甘肃地震新闻语料中占比为 8%, 而日本地震新闻语料中占比仅为 3%。

为了解释这种差异的原因,笔者仔细研读了甘肃地震新闻语料和日本地震新闻语料,发现甘肃地震报道多以专题、专栏的形式出现,而日本地震报道更多采用快讯、简讯的形式。日本地震新闻以简洁明了的方式传递信息给读者,因此地名的出现频率稍高,而快讯和简讯还强调客观叙述,因此动名词使用较少。综上所述,中日两国媒体对甘肃地震和石川县地震采取的不同报道形式直接影响了两者所使用的词汇类型。

3.2. 中日地震新闻报道的主题维度分析

为了比较中日两国媒体关于中国甘肃地震和日本地震的主题内容维度的差异,笔者采用了 Antconc3.56 软件,对出现频率前 50 位词语进行统计,并剔除无意义虚词。基于词语的出现频率和共现概率,利用在线文本分析平台“微词云”,计算单个词语的 TF-IDF 值。TF-IDF 全称为“词频-逆文档频率”,是一种常用的文本分析和挖掘加权技术。TF-IDF 被用于评估一个词语对于语料库中文件集或其中一份文件的重要程度,以衡量词语在一个文件集或查询库中的其中一份文件中的重要性。其基本思想是:若某词在一篇文章中频繁出现而在其他文章中罕见,认为该词具有很好的区分和代表性,适合用于生成主题词云图。最终制作的主题词云图见图 3 和图 4。



Figure 3. Theme word cloud diagram of Gansu earthquake news
图 3. 甘肃地震新闻的主题词云图



Figure 4. Theme word cloud diagram of Japanese earthquake news
图 4. 日本地震新闻的主题词云图

通过图3和图4的对比可以观察到,《人民网》对于甘肃地震的新闻报道主要关注于地震的救援和应急工作,而《共同网》对于日本地震的报道则更加注重地震所引发的次生灾害,如海啸、余震以及房屋倒塌等灾难的报道。可以看出,《人民网》在国内地震报道中,更侧重于激励民众的情绪,强调社会和政府采取的救灾行动,突显出“一方有难,八方支援”的人道主义精神。而在日本地震报道中,《共同网》则以客观视角为主,主要对地震带来的客观灾害进行简要描述,缺乏对当地救灾措施的详细描述,人道主义精神的描绘并不突出,与国内地震报道相比略显不足。

为了进一步揭示中日地震新闻报道主题维度的微观差异,笔者将“甘肃地震-分词”和“日本地震-分词”语料库文件分别导入到文本在线分析平台“微词云”中,平台使用相关算法并基于词汇的 TF-IDF 值,制作完成了图 3 甘肃地震新闻词汇搭配强度和图 4 日本地震新闻词汇搭配强度。

通过对图 5 和图 6 的分析,我们可以进一步验证对通过主题词云图所做的解释。在图 5 中,甘肃地震新闻中的词汇强度搭配显示出搭配强度排名前三的词汇关系分别是“地震-灾区”、“物资-灾区”以及“基金会-公益”。而在图 6 中,日本地震新闻的词汇搭配关系中,排名前三的词语搭配分别是“半岛-地震”、“气象厅-地震”和“海啸-地震”。由此可见,中日两国地震新闻的高强度搭配词是存在较大差异的。在甘肃地震新闻报道中,排名靠前的搭配关系显示出报道侧重于展示抗震救灾工作以及社会各界对灾区的援助,而高强度的日本地震新闻词汇搭配则主要集中在地震本身及其引发的次生灾害。从图 5 中甘肃地震新闻语义网络图中可以看到,“群众-物资-灾区-地震”这四个词在语义网络中处于突出的中心位置,其余节点与这四个词直接或者间接相关。而图 6 中的日本地震语义网络图则显示出“气象台-地区-地震-海啸-时间-半岛”这六个词在语义网络中处于显著位置。

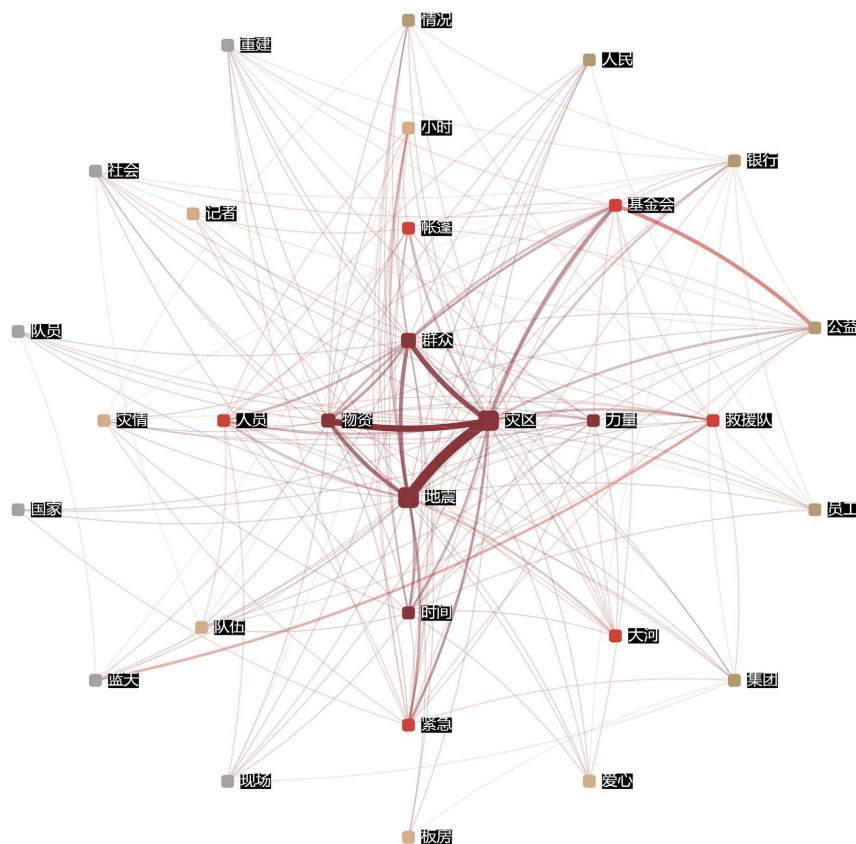


Figure 5. Semantic network diagram of Gansu earthquake news
图 5. 甘肃地震新闻的语义网络图

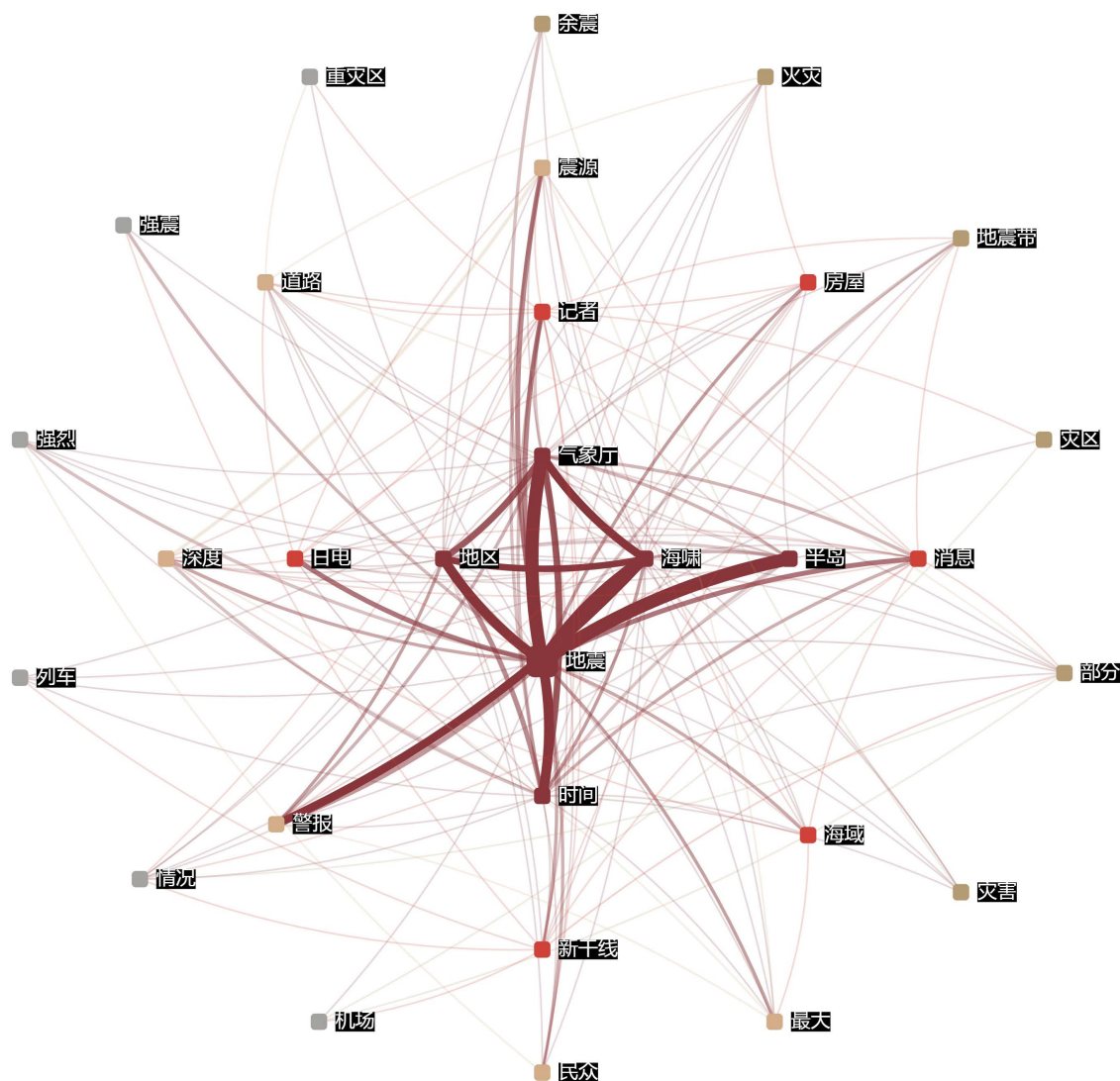


Figure 6. Semantic network diagram of Japanese earthquake news
图 6. 日本地震新闻的语义网络图

中日地震新闻报道语料在语义网络中地差异与词云图和词汇搭配强度图所显示地差异都指向《人民网》在报道国内甘肃地震新闻时侧重展示救援和社会援助等相关救灾主题而《共同网》在报道日本地震新闻时则侧重关注地震所造成地损失和引起潜在安全风险，有意无意地忽略了对人道主义救援方面地关注。从内容主题维度来看，《人民网》在报道中国国内地震灾难新闻时，倾向于描绘出积极乐观的图景，更多地报道救灾和积极应对的措施。而在《共同网》报道日本国内的地震灾难新闻时，往往构建起多灾多难和灾害严重图景，相对较少报道救援相关信息。

3.3. 中日地震新闻报道的情感维度分析

笔者在对中日地震新闻语料的主题语义分析过程中,发现《共同网》对日本的地震新闻报道中有意无意的忽略了人道主义精神的关注。笔者认为,词语使用的差异同样也会影响作者所传达出的情感态度。为了分析语料中的情感态度,笔者借助在线文本分析平台“微词云”来实现。首先,“微词云”计算出所有词汇的词性、出现次数、TF-IDF 值,见图 7。

单词	词性	次数	条数	词频	TF-IDF
信息中心	自定义词	1	1	0.000369	0.000811
地震	名词	203	121	0.074963	0.030881
日本	地名	149	108	0.055022	0.025359
发生	动词	100	82	0.036928	0.02139
石川县	地名	83	70	0.03065	0.019832
能登	动词	66	53	0.024372	0.018667
海啸	名词	46	27	0.016987	0.017856
气象厅	名词	45	34	0.016617	0.015857
地区	名词	41	29	0.01514	0.015461
时间	名词	39	27	0.014402	0.015138

Figure 7. Part of the vocabulary of emotional TF-IDF values
图 7. 部分词汇的情感 TF-IDF 值

结合图 7 中展现的词汇的出现频率、搭配强度和 TF-IDF 值, 笔者使用微词云工具对中日地震新闻报道的词语进行了文本语料的情感分析, 并基于软件算法自动分析了词汇所在句段的情感态度, 得到了文本的情感态度饼状分布图, 见图 8 和图 9, 从而以句段为基本单位掌握了文本整体的情感态度特征。通过图 7 和图 8 的比较可以发现, 《人民网》关于甘肃地震的新闻整体上采用了正面态度进行报道, 而《共同网》对日本地震的新闻则主要呈现绝对负面态度。这一结果验证了笔者在形式维度分析时所提出的猜想: 即中日地震新闻报道中词汇搭配的差异会影响文本的情感倾向。通过从形式维度、主题维度和情感维度三方面对中日地震新闻语料的分析和探讨, 报道的最表层形式将首先决定新闻报道所采用的词汇分布和词性, 不同的词性比例搭配进一步影响报道呈现出的整体情感态度。

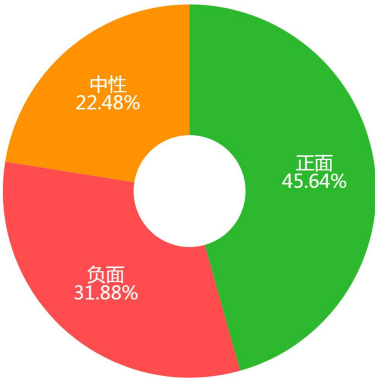


Figure 8. Diagram of the overall sentiment tendency of Gansu earthquake news
图 8. 甘肃地震新闻整体情感倾向

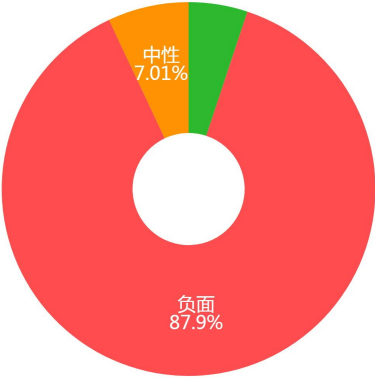


Figure 9. Diagram of the overall sentiment tendency of Japanese earthquake news
图 9. 日本地震新闻整体情感倾向

从情感维度来看,《人民网》在报道国内灾难新闻时,通常采用淡化灾难的处理手法。虽然灾情信息有时不可避免,但不过分渲染,同时强调防灾、救灾和减灾的信息,并确保对国内灾情的报道保持平衡。相比之下,《共同网》在报道日本国内地震新闻时更倾向于采用强化灾难的处理方式,以凸显灾情的严重性。在修辞和风格上,这种报道手法往往带有更强烈的情感色彩,而对于救援行动的信息则简单带过,呈现出一种淡化的趋势。

4. 总结

本研究利用自建中日地震新闻小型语料库,采用 Corpus Word Parser、Antconc、ROSTCM6、微词云等语料库和文本分析工具,从形式、主题内容和情感三个维度来探讨中日两国媒体在报道各自国内地震新闻时的文本差异。研究结果显示,中国媒体在报道甘肃地震新闻时主要采用专栏等形式,较为详尽地展示了社会各界对灾区的积极救援和支持,情感态度积极;而日本媒体在报道石川县地震时主要采用短讯的方式,简要介绍地震发生的时间、地点等基本情况,对地震灾害的描述比较客观,且较少涉及日本社会各界对灾区的紧急救援,报道范围仅限于地震本身,情感倾向较为消极。中日两国关于地震报道的差异不仅在于文本形式和内容主题,也反映出了中日两国在地理环境和地震认知上的显著差异。

参考文献

- [1] 易保树. 基于语料库的英语新闻意识形态意义分析[J]. 安徽农业大学学报(社会科学版), 2011, 20(1): 123-127.
- [2] 熊文新. 新闻报道主观性的语言学透视——一种结合语料库驱动和批评话语分析的方法[J]. 现代传播(中国传媒大学学报), 2022, 44(5): 22-32.
- [3] Shen, Y. (2008). ROST Content Mining System: Software for Content Mining and Analysis. *International Applied Linguistics*, 20, 16-22.
- [4] Froehlich, H. (2015). Corpus Analysis with Antconc. *The Programming Historian*, 15, 55-63.
<https://doi.org/10.46430/phen0043>
- [5] 孙媛. 报刊英语语料库研究现状及应用分析[J]. 语文学刊(外语教育与教学), 2011(7): 15-16+31.