Modern Linguistics 现代语言学, 2022, 10(11), 2451-2456
Published Online November 2022 in Hans. http://www.hanspub.org/journal/ml
https://doi.org/10.12677/ml.2022.1011335

Hans汉斯

韩国主流媒体涉华新闻语料库的构建

韩睿鑫，刘汶轩，姜　珊，李冰仟如，杨美妍

大连外国语大学，辽宁　大连

收稿日期：2022年10月7日；录用日期：2022年10月31日；发布日期：2022年11月9日

摘　要

贯彻落实新时代精神，向世界讲好中国故事，填补中国语料库发展在韩国语语言这一方面的空白，通过构建韩国主流媒体涉华新闻语料库，着重介绍建设该语料库的方法，为分析我国在韩国主流媒体中所呈现出的国家形象，同时为语言学者提供语言学基础资源，实现语料库更深层次的建设。

关键词

语料库，涉华新闻，韩语，语料库构建

The Construction of Corpus of Korean Mainstream Media's China Related News

Ruixin Han, Wenxuan Liu, Shan Jiang, Bingqianru Li, Meiyan Yang

Dalian University of Foreign Languages, Dalian Liaoning

Received: Oct. 7th, 2022; accepted: Oct. 31st, 2022; published: Nov. 9th, 2022

Abstract

In order to implement the spirit of the new era, tell the world a good story about China, and fill the gap in the Korean language in the development of the Chinese corpus, through the construction of the Korean mainstream media news corpus about China, focus on the method of building the corpus, to analyze China's national image in the Korean mainstream media, at the same time to provide language scholars with basic linguistic resources, to achieve a deeper construction of the corpus.

Keywords

Corpus, China Related News, Korean, Corpus Construction

文章引用: 韩睿鑫, 刘汶轩, 姜珊, 李冰仟如, 杨美妍. 韩国主流媒体涉华新闻语料库的构建[J]. 现代语言学, 2022, 10(11): 2451-2456. DOI: 10.12677/ml.2022.1011335

1. 引言

语料库(corpus)，一般是指经科学取样和加工的大规模电子文本库。通过借助计算机分析工具，研究者可开展相关的语言理论及应用研究。

随着语料库语言学已成为应用语言学的常规研究方法，但使用语料库来进行焦点搜索，把不同语言的异同现象转化为直观可视的数据对比，还属于目前语言学界相对较新的研究领域，相关专业性的韩语语料库更是少之又少。

2. 语料库简介

2.1. 语料库的分类

语料库的类型很多，分类标准也有许多。有人曾经把语料库分成四种类型：1) 异质的(Heterogeneous)：没有特定原则，广泛收集并原样存储各种语料；2) 同质的(Homogeneous)：只收集同类内容的语料；3) 系统的(Systematic)：根据预先确定的原则和比例收集，使语料具有平衡性和系统性，能够代表某一范围内的语言事实；4) 专用的(Specialized)：只收集用于某一特定用途的语料。

2.2. 语料库的应用范围

随着计算机技术的迅猛发展，语料库的应用范围也日益广泛，其在建设的过程中涉及多学科，因而其应用也覆盖了多个学科和各个方面。

在语言教学方面，主要是围绕利用语料库辅助课堂教学；在建立领域辞书和词典方面，大规模语料库已成为词典编撰的前提和主要工具。不仅可以提高词典编撰的质量，节约编撰时间，还能保证所有的词义、句法信息的可靠性和准确性；在语言对比和翻译研究方面，它不仅可以为部分术语翻译提供强有力的支撑，还可以保证在翻译进入实用后的严谨性和准确性。

除此之外，还应用于一些其他领域比如语言课本的撰写和稿件的完善润色等，可见语料库经过最初的语言学领域已经可以延伸到不同领域的具体研究领域[1]。

2.3. 语料库发展现状

近年来，随着计算机技术、网络技术的发展和语料库开发应用方面的资源共享与合作，学科专业语料库的研制也呈现出快速发展的趋势。很多学科和领域都相继建设了语料库。从总体上来看，学科专业语料库的数量越来越多，规模也越来越大，但是有影响力的领域语料库还是凤毛麟角。由于各建库机构和单位之间缺乏交流与协作，语料库建设的领域相对集中，重复建设的情况很严重。此外，由于缺乏统一的规范和标准，以及知识产权方面的问题，很多专业库的利用效率非常低，也同样面临很多亟需解决的问题。

3. 韩语语料库相关研究现状

3.1. 韩国语料库发展的回顾与发展趋势

在语料库语言学不断兴起的时候，韩国也在 20 世纪 80 年代开始了语料库建设的脚步。80 年代后期，

韩国社会学界对电脑语料库建设的热衷开始空前高涨。当时以延世大学、韩国科学技术院、高丽大学等学术机构为中心的团体开始试图构建语料库体系，这样的努力一直在韩国延续，直到1998年韩国政府启动了名为“21世纪世宗计划”的活动，韩国语料库的建设工程才真正迎来了一个重大的转折点。所谓“21世纪世宗计划”是当时韩国文化观光部支持下的一项公共事业，它旨在构建韩国自己的语料库和电子词典体系，以此作为在学术和产业当中可以普遍使用的信息资源，然后再将其进行普及，并运用于各项基础研究事业当中[2]。如今，韩国的世宗计划已经整体完成，这使韩国拥有了一个世界级的大型语料库，其中包括的词量达到了3亿词次。

当21世纪世宗计划打造语言资料库的同时，为使语言资源能够在教育、研究、产业等方面广泛利用韩国把目光的焦点同样对准了语言工程软硬件构造上面，并把这一系列的动作命名为“国语信息化”运动，把这些行为推到了一个新的学术名称上来。每年夏天，韩国都会举行所谓国语信息化研讨会这样一种短期教育。而“国语信息学”在韩国就是指以韩国语为基础建立的语料库语言学。

如今韩国也同世界大多数国家一样处在信息化社会当中，知识和信息的处理变得尤为重要，在此其中，语言信息的处理也渐渐被人所关注，所以，作为语言资源重要基础的语料库的重要性也大大提高了。现在，不仅研究语言的学者们积极参与建设，就连电算学者们也在不断倾注热情在这一方面投入自己的努力。

3.2. 韩国语料库发展的现状

21世纪世宗计划从1998年开始，历时10年最终打造完成。这一工程成果丰硕，内有2亿词次的语料库相比英国91~94年建成的英国国家语料库1亿词次数量还多，可以说达到了世界级的水准[3]。这样大规模语料库能够建成也科学地证明了韩国语言是一种优秀的语言。该计划为了构建基础的语言资料数据库，专门把报纸、杂志、小说等各种资料进行了计算机处理，实现了利用计算机即时翻译、纠正语法和文章要素错误等机能。借助世宗计划韩国还开发出了包含60万词汇量的电子词典，这不但让信息检索、文本分析与制作、自动翻译等成为可能，多国语言电子词典的出现也给外国人学习韩国语带来了很大的福音。

4. 韩国主流媒体涉华新闻语料库构建方法

韩国主流媒体涉华新闻语料库的构建主要分为两步，首先是语料的获取，其次是语料的加工，即对语料进行预处理。

4.1. 韩语新闻语料获取

韩语新闻语料的获取主要包含三个部分，分别是语料的搜集、筛选和整理。

4.1.1. 语料搜集

为保证语料的真实性和准确性，我们通过韩国新闻网站 bigkinds 获取韩国新闻。韩国新闻网站bigkinds中收录了韩国几乎所有新闻社的新闻报道，并且可以通过bigkinds中的检索功能进行涉华新闻的收集。我们选取韩国四家较权威的新闻社，分别是中央日报、朝鲜日报、东亚日报、韩民族日报。时间上是大致以2018到2021年的新闻为主。

4.1.2. 语料筛选

在语料的筛选过程中，为便于快速筛选出涉华新闻，我们在新闻检索时输入关键词“중국(中国)”，那么凡是含有关键词“중국(中国)”的新闻都会出现。在此基础上，为确保语料的准确性，减少误差，将第一遍筛选出来的新闻再进行人工筛选，即把每篇新闻大体读一遍，了解文章内容后，筛选出与中国相

关的新闻。除此自外，我们对新闻的类型及主题不做限制，新闻包含政治、经济、文化、科技等多个领域，以便于获得更加完整、覆盖面广的涉华新闻语料库。以 2018~2020 年为例，含有关键词的新闻共有 36,521 篇，而经过人工进一步筛选之后获得 15,325 篇新闻。

4.1.3. 语料整理

语料的统一整理方式是保留新闻的新闻社、标题、正文内容，每篇新闻放在一个文档中，格式为.txt，并且按照新闻发表时间的年月日进行分类整理。这也便于我们在进行语料库的检索时能够快速找到某一日期的韩语新闻，并根据不同的新闻社、时间而进行语料的分析。例如在语料库检索过程中，输入某一关键词，即可获得所有所含关键词的新闻中的语句，同时会显示该语句所在的新闻标题、新闻社、时间等信息。

4.2. 韩语新闻语料预处理加工

语料，顾名思义就是我们平时所说的文本，带有文字描述性的文本都可以归类于语料。但这种原始文本无法直接用来训练模型，需要进行前期预处理。语料预处理方法主要包括数据清洗、分词、词性标注、去停用词等，总共分为五步。

4.2.1. 基本预处理

对语料进行加工的第一步就是对语料进行基本预处理，即去除数字、标点、符号、标签以及英文字母切换小写等。例如“△기존 협정문에 포함된 투자자 - 국가분쟁해결(ISDS)에서의 투자자 보호규범 개선 등을 제기할 예정이다”这句话经过第一步基本预处理之后变为“기존 협정문에 포함된 투자자 - 국가분쟁해결 isds 에서의 투자자 보호규범 개선 등을 제기할 예정이다”。

4.2.2. 拼写检查

韩语语料预处理的拼写检查主要分为两个部分，分别是以词典为基准纠正错别字和恢复缩略语原型。例如：I’m not happy -> I am not happy

拼写检查中使用한국어 맞춤법 검사 라이브러리(韩国语拼写法检查库)和论文中使用过的外来语词典。

4.2.3. 形态素分析、词性标注

形态素分析指对分析每个单词的词性，并在单词后标注出该单词的词性。我们使用的是 kakao 的 Khaii，这是 Python 为基础的形态素分析仪中性能最好的一种。

例如：한겨레/NNG 한중/NNG 서비스/NNG 투자/NNG 후속/NNG 협상/NNG 개시/NNG 사드보/NNP 복/NNG 실질/NNG 중단/NNG 재차/MAG 요구/NNG

4.2.4. 动词原形恢复

动词原形恢复主要参照以下原则：

1) NNG|NNP|NNB + XSV|XSA --> NNG|NNP|NNB + XSV|XSA + 다

2) NNG|NNP|NNB + XSA + VX --> NNG|NNP + XSA + 다

3) VV --> VV + 다

4) VX --> VX + 다

例如，“요구했다”将变为“요구하다”，即将动词过去式变为动词原型。

4.2.5. 去停用词

停用词是指在一个句子中没有具体的含义，只是起到衔接句子以及增强语气的作用的词。这些词对

文本分析也没有任何帮助，因此我们需要对分词后的数据做停用词处理。我们使用的韩国语停用词词典为：https://www.ranks.nl/stopwords/korean。

4.3. 语料库高频词统计

语料库构建好之后，可以对新闻语料中每个单词出现的频率进行统计。我们使用

‘# count frequency

count_frequency()’

来生成语料库的高频词。

以2018年为例，“중국(中国)”一次出现次数为37,970次，“미국(美国)”出现次数为24,355次，“북한(朝鲜)”出现次数为11,197次。

从高频词内容来看，像“중국(中国)”“미국(美国)”“조선(朝鲜)”“한국(韩国)”等国家以及“대통령(总统)”“회담(会谈)”“정상(元首)”“트럼프(特朗普)”等政治类词语出现次数最多，可见韩国主流媒体涉华新闻中政治类新闻偏多，韩国主流媒体对政治的关注度较高。其次“관세(关税)”“무역(贸易)”“전쟁(战争)”“기업(企业)”等经贸类词汇出现较多，涉华新闻中经济贸易相关的文章也较多，经济贸易是韩国主流媒体的第二大关注点。

5. 韩国主流媒体涉华新闻语料库构建意义

5.1. 促进当代语言学研究

5.1.1. 基于语料库的构建

语料库语言学即运用计算机的强大功能对各种语料进行检索，采用实证研究的方法对语言材料做语音、词汇等多方面的分析，并从语境和功能的角度构建语言意义的理论框架，更加真实、全面、准确地描述语言[4]。其兴起于20世纪80年代，并持续发展，如今已经成为语言学的重要部分。语料库的建立和发展为语言研究提供了更多的真实自然语料，为人类认识语言的本质开辟了新道路。

而所构建的韩国主流媒体涉华新闻语料库是一个从无到有的历史性进展，从韩国涉华新闻的提取，代码的编写，程序运行外部环境的建立，都是技术性的突破。其为韩语研究者提供了新的调研渠道，并且有助于词典、翻译、语言教学、文体、语言对比和计算语言学等研究领域。

5.1.2. 基于韩国主流媒体涉华新闻的语料来源

由于国际新闻的生产与传播受到世界政治、经济、文化以及意识形态等多方面因素的影响和制约，因此它与一般国内新闻的生产和传播具有不同特点，更加地复杂和多样化，因此，有关它的研究具有较为重大的理论和现实意义[5]。

通过收集조선일보(朝鲜日报)，중앙일보(中央日报)，동아일보(东亚日报)，한겨레(韩民族)四个韩国主流媒体中涉及中国的新闻，有助于研究在韩国意识形态下，其认知语言、文化负载语言等的语言特殊性。同时对中韩、韩中翻译研究具有重要意义。在收集了韩语新闻的原文文本并同时收集从中文翻译成韩文的文本之后，通过对两种文本的对比，可以分析研究特定历史、文化和社会环境中的翻译规范，探索韩语和汉语的翻译规律[6]。

5.2. 助力向世界讲述中国故事

5.2.1. 创新意义

韩国主流媒体涉华新闻语料库的构建是具有创新意义的。研究韩国主流媒体眼中的中国形象究竟如何，前人对于媒体眼中的中国形象常常局限于个别文章或某一种言论，难以进行科学的考察。构建韩国

主流媒体涉华新闻语料库，讲韩国主流媒体作为主要信息来源，能够覆盖到尽可能广的在韩新闻读者，新闻内容更具客观性，可信度更有保障，从而保证以最客观的角度分析韩国媒体对中国形象的认知。

从以往的经验来看，以数据统计、传播学角度归纳韩国媒体眼中的中国形象的方法较多，韩国媒体构建中国形象用的规律性话语无从体现，也无法将结果中的表达直观的具体到某一篇言论。语料库的构建可以让某一规律性话语或结论在细节处有更多体现。构建韩国主流媒体涉华新闻语料库以语言学为基础，综合运用编程手段整合新闻资源，可以多角度剖析韩国媒体树立中国形象的方法与手段，是一种创新的资源和语言的整合方式。

5.2.2. 如何讲好中国故事

随着中国在国际舞台中话语权逐渐增强，与周边国家缺乏深入沟通、对自身国家形象缺乏准确把握、选择错误的形象传播途径等都可能成为造成国与国之间冲突矛盾的原因。通过语料库的构建可以大致把握韩国主流媒体中国的形象，可有效修正国际社会对中国形象的误解、塑造追求和平以及负责任大国形象。韩国与中国隔海相望，了解韩国主流媒体对中国形象的定义，有助于我们更好地了解自身定位并积极调整。

致　谢

感谢指导本次项目的朱伟老师和郑家琦研究生学长。

基金项目

本项目由大连外国语大学学生创新创业训练项目资助。

参考文献

[1] 黄水清, 王东波. 国内语料库研究综述[J]. 信息资源管理学报, 2021, 11(3): 4-17+87.

[2] 이승재[韩]. 21 세기 세종계획의 개요와 금후의 활용 방안[J]. 일어일문학연구, 2005.

[3] 국립국어원[韩]. 21 세기 세종계획백서[J]. 국립국어원, 2007.

[4] Tognini Bonelli, E. (2001) Corpus Linguistics at Work. John Benjamins Publishing Company, Amsterdam/Philadephia. https://doi.org/10.1075/scl.6

[5] 廖七一. 语料库与翻译研究[J]. 外语教学与研究: 外国语文双月刊, 2000(5): 380-384.

[6] 毕玉德, 赵岩. 基于新闻语料库的朝韩词汇对比研究[J]. 东北亚外语研究, 2016(3): 35-41.