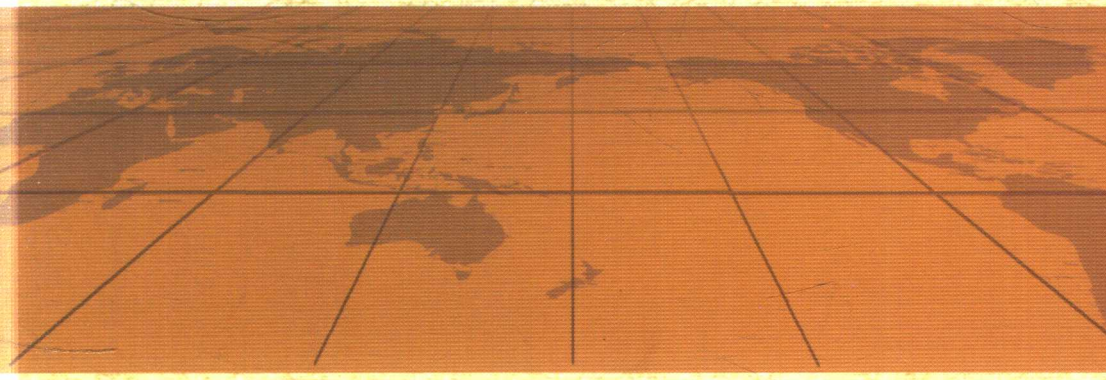
PLAUFL 解放军外国语学院
PUBLICATION FUND 出版基金

越-英-汉时事新闻框架语义研究

A Study of Vietanmese, English and Chinese Current Affair News from thePerspective of Frame Semantics

林 丽·著



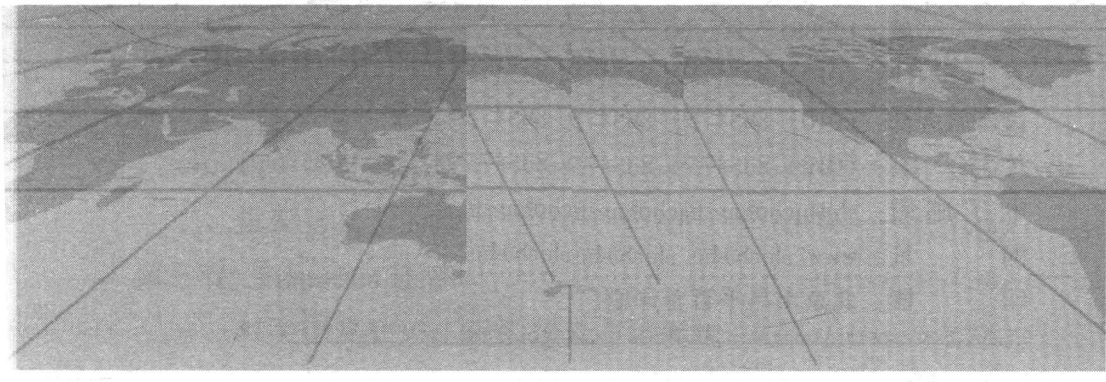
时事出版社

PLAUFL 解放军外国语学院
PUBLICATION FUND 出版基金

越-英-汉时事新闻框架语义研究

A Study of Vietanmese, English and Chinese Current Affair News from thePerspective of Frame Semantics

林　丽·著



时事出版社

图书在版编目（CIP）数据

越－英－汉时事新闻框架语义研究/林丽著. —北京：时事出版社，2017.2

ISBN 978-7-5195-0064-1

Ⅰ.①越… Ⅱ.①林… Ⅲ.①新闻—语义学—研究—越、英、汉 Ⅳ.①G210②H030

中国版本图书馆 CIP 数据核字（2017）第 009262 号

出 版 发 行：时事出版社
地　　址：北京市海淀区万寿寺甲 2 号
邮　　编：100081
发 行 热 线：（010）88547590　88547591
读者服务部：（010）88547595
传　　真：（010）88547592
电 子 邮 箱：shishichubanshe@ sina. com
网　　址：www. shishishe. com
印　　刷：北京市昌平百善印刷厂

开本：787×1092　1/16　印张：20.25 字数：264 千字
2017 年 2 月第 1 版　2017 年 2 月第 1 次印刷
定价：80.00 元
（如有印装质量问题，请与本社发行部联系调换）

目 录

第一章 绪 论

第一节 研究背景

据 2014 年 3 月 12 日统计数据[①]，Internet 内容语种（Content languages for websites）世界排名前十位中包括联合国七种通用语种（英语、俄语、德语、西班牙语、法语、汉语、阿拉伯语）中除阿拉伯语外的六种。非通用语种中的日语、葡萄牙语、意大利语、波兰语也进入排名前十。这表明，一方面 Internet 内容语种呈多语化发展趋势，另一方面，除英语外，其余各通用语和非通用语之间发展差距并不显著。

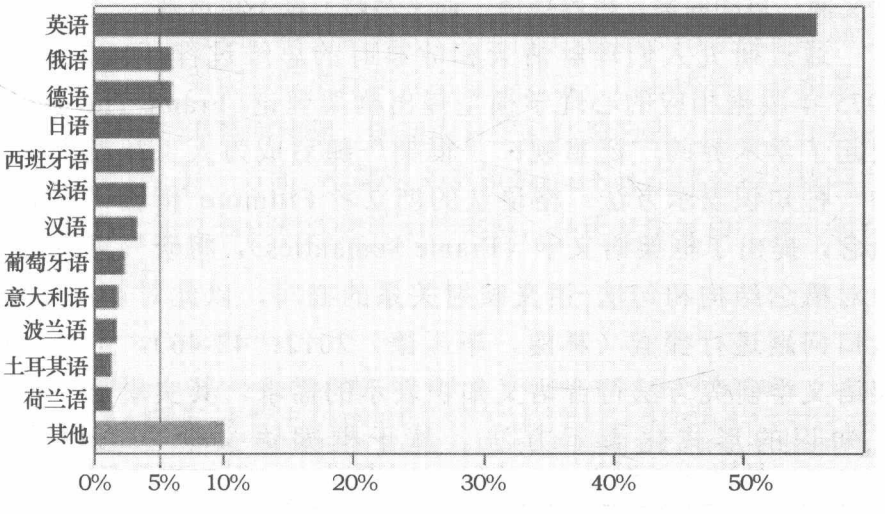


图 1—1 2014 年 3 月 12 日 Internet 内容语种统计

① http://en.wikipedia.org/wiki/Languages_used_on_the_Internet.

当前，多语种大数据信息呈现出爆炸增长态势，不论是否为通用语种，各语种网页绝对数量都相当庞大。如何利用现代语言学方法和信息处理技术对巨量的多语种新闻文本进行知识表示、知识获取，及时、准确地追踪和发现有效信息具有重要的理论价值和实践意义。

根据图灵机模型，现代计算机通常采用线性符号识别与转换对语言信息进行处理。当前脑科学研究认为，人脑处理语言信息时一般依赖存储的语义和情景知识进行并行扩散多路搜索。因此，计算机对文本的语义理解和知识表示成为目前制约语言信息处理发展的重要瓶颈之一。

对纷繁复杂的语义和情景知识进行形式化，将其表示为计算机可操作的符号，采用词汇语义知识库（Lexical Semantic Knowledge Database，LSKD）[①]的形式进行存储，是解决信息处理中语义问题的主流方法。LSKD 构建已经成为语言信息处理的核心工程，基于 LSKD 的语义分析方法对各种语言信息处理应用（信息检索、信息抽取、自动文摘、自动问答、机器翻译、词义消歧）都不可或缺。

通过研究人们理解情景和故事时的思维过程，Minsky 于 1975 年根据相应的心理学模型提出框架理论（Frame Theory）[②]，引起了学术界的广泛重视；“框架”随后成为人工智能界常用的一种知识表示方法；格语法的创立者 Fillmore[③]借鉴“框架”概念，提出了框架语义学（Frame Semantics），将研究重点确定为对概念结构和句法-语义映射关系的描写，以此对句法-语义接口问题进行探索（林丽，毕玉德，2012：42-46）。可见，框架语义学研究方法符合语义知识表示的需求，其实质是一种语义知识的形式化表示方法。基于框架语义学理论构建的

[①] 如无需特别说明，下文中将用 LSKD 指代“词汇语义知识库”。

[②] 参见 Minsky. M.A Framework for Representing Kownledge[C]. In Winston(ed.). *The Psychology of Computer Vision*. New York. McGraw Hill, 1975.

[③] 由于 Fillmore 的译名有菲尔墨、费尔莫、菲尔默等多个版本，本书使用英文形式以确保统一。为方便文献查阅，除译著名称外，本书中提及的其他外国学者姓名均使用相应外文形式。

FrameNet[①]以事件框架的方式确定词汇化编码的语义信息（可转化为DAML+OIL 语言）并预测这些信息如何在句法上得到投射。（俞士汶，黄居仁，2005：1-20） FrameNet 以语义框架[②]作为基本描述单元，并建立了框架与框架之间的网状层级联系，具备完善的语义表示与描述体系，是 LSKD 中设计合理、构建完备、应用广泛的典型工程。

从具体应用上看，由于框架语义学和 FrameNet 以人的认知经验作为基础，对“概念结构”，即人类关于现实世界的语义知识进行阐释和描写，因此在一定程度上具有普适性。截至目前统计，各国研究学者基于框架语义学，以 FrameNet 为蓝本构建了 19 个语种[③]的平行框架网络资源。除我国民族语言维吾尔语和藏语外，其余 17 种外语均在有具体排名的 Internet 内容语种前 36 位之列（各语种具体排名见脚注中括号内数字）。也就是说，Internet 内容语种前 36 位中，已研究构建 FrameNet 的语种比重为 47.2%。由此可见，FrameNet 的多语种扩展趋势与 Internet 内容语种的发展是一致的。可以据此推断 FrameNet 对于巨量网络信息处理是必要的。

基于以上背景，本书认为，一方面对巨量的多语种网络新闻文本知识表示、知识获取进行研究有迫切需求，而各语种，特别是非通用语在领域语料库构建、面向语言信息处理的语义研究方面尚显滞后和薄弱；另一方面，框架语义学和 FrameNet 在理论和实践两方面都可作为语义知识形式化表示的典范，在研究深度和广度上都有突出的贡献。

---

① http://framenet.iesi.berkeley.edu。国内相关研究中，FrameNet 的常见译名有“框架网络”、“框架语义网”和“框架语义知识库”。本书中以 FrameNet 特指美国加州大学伯克利分院构建的英语框架网络，其他语种构建的相应语义工程称作“**语框架网络”或其缩写(如：日语框架网络缩写为 JFN，西班牙语框架网络缩写为 SFN 等等)。

② 作为一种认知结构，语义框架具有丰富的框架元素定义，由一组语义接近的词元激活，为每一个词元提供完成框架语义标注的例句。

③ 具体为德语(3)、日语(4)、西班牙语(5)、汉语(7)、葡萄牙语(8)、意大利语(9)、波兰语(10)、波斯语(13)、捷克语(16)、瑞典语(17)、越南语(18)、希腊语(21)、丹麦语(22)、泰语(24)、保加利亚语(27)、希伯来语(29)、斯洛文尼亚语(35)、维吾尔语、藏语。数据来源：http://en.wikipedia.org/wiki/Languages_used_on_the_Internet。另外，各语种框架网络研究程度和构建规模不等。

因此，充分发挥框架语义学和 FrameNet 的已有优势，以多语种网络新闻文本知识表示、知识获取为应用导向，探索更加高效、通用的框架语义表示模式并进行框架网络构建和应用探索具有较为重要的理论和现实意义。

第二节　研究内容

本书的研究目标是将框架语义学理论应用到非通用语（特别是缺乏严格意义上形态变化的孤立语）语义分析中，参照 FrameNet 工程构建面向领域的多语框架网络并基于此进行事件抽取应用探索，研究内容主要分为理论研究、工程实践和应用探索三个方面。

理论研究方面：通过系统研究框架语义学作为语义分析理论模型的的特点和优势，深入探究 FrameNet 作为语义知识工程所具有的多语种可扩展性和领域延伸性，分析其针对自然语言处理（NLP）[①]方面的不足之处，提出面向领域的多语框架语义表示（Domain-oriented Multilingual Frame Semantic Representation，DOMLFSR）[②]模式，确定该模式的整体架构及核心内容。

基于框架语义学的越南语词汇语义研究也是本书的研究内容之一。由于越南语和汉语同为孤立语的典型代表，本书也将从语言类型特点角度着重分析其在框架语义构造式系统方面的共性。

工程实践方面：将“越南语-英语-汉语”作为“非通用语-中介语-通用语”多语模式的一个研究实例，以时事新闻语料作为特定领域开展研究。设计并开发多语种 Web 新闻语料抓取软件，构建了越-英-汉语领域语料库；对各语种语料进行预处理并统计出高频动词词元；采集领域词元并进行语义分类，根据《同义词词林》（扩展版）为每一词元进行语义分类赋码，由此建成三语领域高频动词词元库；

① 如无需特别说明，下文中将用 NLP 指代“自然语言处理”。

② 如无需特别说明，下文中将用 DOMLFSR 指代“面向领域的多语框架语义表示”。

提出基于 FrameNet 1.5 数据的半自动框架库映射方法，通过三语领域高频动词词元库中的英语动词词元激活 FrameNet 相应框架进行复用和整合；设计并实现辅助建库、标注工具，构建领域越-英-汉框架网络（Domain-oriented Vietnamese-English-Chinese FrameNet，DOV-E-CFN）[①]，其主要工作包括搭建一定规模的时事新闻领域框架体系，确立框架关系，构建标注例句库和词汇库，统计词元配价模式等。

另一方面，基于 FrameNet 的越南语句法-语义基础资源构建也是工程实践的重要目标之一。原因在于越南语在 Internet 内容语种世界排名中列第 19 位[②]，目前使用人数超过 9000 万[③]，其重要性在亚洲，特别是东南亚地区更为明显。越南语信息处理在 2000 年后才正式开始起步，面向 NLP 的基础资源的建设还较为滞后。

应用探索方面：基于越-英-汉时事新闻框架网络（DOV-E-CFN），将框架语义分析方法与传统事件抽取方法进行对比，论证了框架语义标注方法在事件抽取中应用的可行性及优势，并分别基于核心依存图（kernel dependency graph，KDG）[④]和框架语义格标（frame element kasus，FK）探索其在新闻文本事件信息抽取中的应用。

本书的具体章节安排如下：

第一章为绪论，主要对本书的研究背景、研究内容及方法、学术创新、研究意义、结构和使用资源进行介绍；

第二章为框架语义研究综述；

第三章为面向领域的多语框架语义表示（DOMLFSR）模式；

第四章为越-英-汉时事新闻框架网络的体系构建；

---

① 如无需特别说明，下文中将用 DOV-E-CFN 指代领域越-英-汉框架网络的构建实例“越-英-汉时事新闻框架网络”(Vietnamese-English-Chinese FrameNet on the Current Affair News)。

② http://en.wikipedia.org/wiki/Languages_used_on_the_Internet

③ 越南学者胡秀宝（Hồ Tú Bảo）和梁芝梅（Lương Chi Mai）的报告 *Về xử lý tiếng Việt trong công nghệ thông tin*（信息技术中的越南语处理）中专门谈及越南语语言信息处理的现状和问题，其中一个重要问题是缺乏必要的基础资源的建设。参见 http://list25.com/the-25-most-influential-languages-in-the-world/

④ 如无需特别说明，下文中将用 KDG 指代“核心依存图”。

第五章为越-英-汉时事新闻框架网络例句标注与词元库构建；

第六章为越-英-汉时事新闻框架网络应用探索；

第七章为结语，总结全书研究取得的主要成果和不足，并对下一步研究进行展望。

第三节　研究思路

本书总的研究思路是通过文献调研和对比分析方法，对框架语义学理论模型和 FrameNet 语义工程进行述评，总结其特点与优势，同时指出其针对 NLP 任务方面的不足，提出 DOMLFSR 模式，构建框架元素层级系统和框架语义构造式系统作为其核心成分等。在此理论模型基础上，参照 FrameNet 构建原则和路线，结合基于语料库的经验归纳方法对文本进行统计分析，研究构建 DOV-E-CFN（越-英-汉时事新闻框架网络）并进行应用探索。

DOV-E-CFN 的整体研究思路如图 1-2 所示。

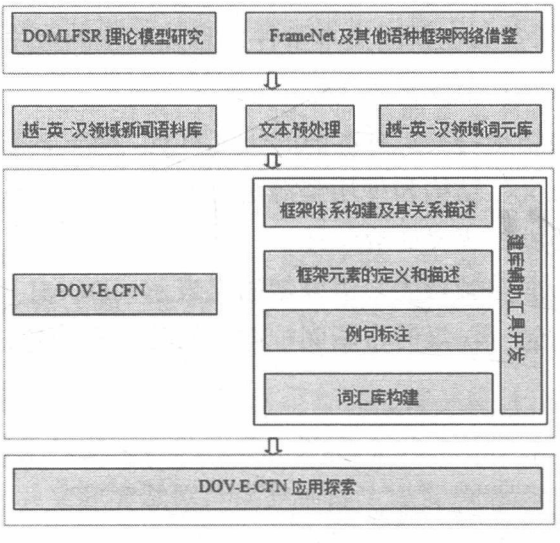


图 1-2　越-英-汉时事新闻框架网络（DOV-E-CFN）整体研究思路

第四节 学术创新

1．本书从知识表示层面剖析了框架语义学原理，在此基础上系统地对比了 FrameNet 相较于同类语义工程突出的多语种可扩展性与领域延伸性，并据此提出 DOMLFSR 模式，旨在为不同语种构建领域框架网络提供一定程度上的理论依据。

2．较为全面地对面向语言信息处理的越南语语义研究进行回顾和梳理，探索了越南语词元和例句的框架语义分析和语义结构知识表示方法，针对越南语特点构建了核心依存图分析模型。

3．提出多语种领域语料库和领域词元库构建方法和半自动框架库映射方法，设计并实现辅助建库、标注工具，为面向领域的多语框架网络构建提供技术参考。

4．在语料库资源建设方面，设计并开发多语种 Web 新闻语料抓取软件，并在英语、汉语、俄语、朝鲜语、马来西亚语等多个语种上进行了应用，取得了成功。目前已建成 2000 万词级的“越南语通用语料库”[①]。

5．将语义知识库构建与基于真实语料的例句标注相结合，构建的越-英-汉时事新闻框架网络既能丰富和发展框架语义理论，也注重构建面向 NLP 的语言资源，并以时事新闻文本语句为例，运用框架语义标注方法抽取特定事件信息，为初步服务于应用实践奠定基础。

第五节 研究意义

本书具有理论研究和应用实践两方面的意义：

1．理论意义：丰富和深化了框架语义理论，在通用层面分析其作为知识表示方式的优势及针对 NLP 任务的不足，提出 DOMLFSR

[①] 基金项目：中国-东盟研究中心(广西科学实验中心)课题“越南语通用语料库建设”(20120146)。

模式。结合越南语词汇语义研究，较为系统地将 Fillmore 的框架语义学思想运用于越南语词汇语义分析，使越南语语言本体研究在语义分析和描写方面更加深入和全面。

2．应用实践意义：在深入挖掘 FrameNet 多语种可扩展性和领域延伸性基础上，探讨了具有一定普适性的“面向领域的多语框架网络构建方法”，同时构建了较为丰富的越-英-汉框架语义资源，如时事新闻词元库和例句标注库。“标注例句库”提供了高精度的搭配和句法-语义联系方面的基础数据，“词元库”能够提供词语在单个义项上的搭配信息，揭示共有框架下不同语种在语义和句法表现上的异同。据此可以构建信息抽取模板，标注数据达到一定量级后可作为 NLP 统计模型的训练数据服务于信息抽取、问答系统、词义消歧、机器翻译及其他应用。

对越南语语言教学和语言习得来说，DOV-E-CFN 揭示了三种语言领域语料库例句的句法语义性质，且检索方便，是教材编写和越汉双语、越英汉三语词典编纂的理想资源，有利于学生观察语言共性与差异，提高学习和研究效率。

第六节　使用资源

1．FrameNet 1.5 数据[①]

该资源包括 1019 个框架，11829 个词元，定义了 8252 种框架元素和 1507 种框架关系，对其中 65.2%的词元进行了例句标注。

2．越南语通用语料库

语料来源主要是越南国防网[②]、越南人民军队网[③]、越地报[④]、BBC 越南新闻[⑤]等网站新闻报道，目前规模约为 13500 个文本文件（99M）。

① https://framenet.icsi.berkeley.edu/fndrupal/framenet_data
② http://quocphong.vn/
③ http://www.qdnd.vn/qdndsite/vi-VN/43/Default.aspx
④ http://www.baodatviet.vn/
⑤ http://www.bbc.co.uk/vietnamese/

3．文本预处理工具

英语：英语语料的预处理使用 CLAWS POS tagger[①]，选择的词性标注集为 UCREL CLAWS5 Tagset。

汉语：汉语语料的分词和词性标注使用汉语词法分析系统 ICTCLAS[②]。

越南语：越南语语料预处理使用越南语文本处理工具包 vnToolkit 3.0[③]，包括分词软件 vnTokenizer、词性标注软件 vnTagger。

4．哈工大信息检索研究室《同义词词林》（扩展版）[④]

① http://ucrel.lancs.ac.uk/claws5tags.html
② http://ictclas.nlpir.org/
③ http ://www.loria.fr/ lehong/tools/
④ http://ir.hit.edu.cn/

第二章　框架语义研究综述

知识表示研究作为人工智能的核心课题与语义分析密切相关。为解决信息处理中的语义问题，语言学界和计算机学界都做了许多努力。框架语义研究是其中将语言学研究与知识工程实践有机结合的典范。因此对框架语义研究进行述评是本书研究的理论基础，其特点和优势是本书重要的选题依据，其不足和发展趋势是本书努力的方向和突破口。本章首先对框架语义学和 FrameNet 进行简介，对 Fillmore 关于框架语义学的重要文献进行综述，并从语言学研究和知识工程实践两个视角分析其相关研究背景以及各自特点和优势。

第一节　框架语义研究概览

框架语义学和 FrameNet 是理论模型和工程实践的关系。作为语义理论模型，能支撑大型工程实践并保持长久生命力的并不多见，框架语义学因此被视为典范。

框架语义学于 1970 年代开始形成，它提供了一种描写词项（lexical item）意义及语法结构（grammatical construction）意义的方法，被视作经验主义语义学之一。Fillmore 早期将其称作“理解语义学”（The Semantics of Understanding），原因在于框架语义体现了概念及意义对人类经验的高度依赖，在语言实际中可以理解为描述词项的意义应当借助于语义框架[①]（semantic frame）（Fillmore，1985）。

[①] 语义框架，简称“框架”。