

汉语语法工程研究的意义、内容与方法

杨春雷

上海外国语大学英语学院, 上海 200083

yangchunlei@shisu.edu.cn

摘要

面向深层语言处理的汉语语法工程 (Grammar Engineering, GE) 不仅可弥补在自然语言处理领域亟需的系统汉语语法理论, 还可提供计算实现的平台。首先, 通过数据分析和文献梳理, 指出构建汉语GE的迫切性和重要性。然后, 从语言普遍性视角设计了面向深层语言处理的汉语GE研究内容, 尤其是重点研究汉语特殊语法现象的描写。接着, 提出跨语言学本体和计算实现两个领域的具体研究方法。最后, 介绍多个语言学层面的研究成果, 验证了本方法的可行性。

关键词: 语法工程; 深层语言处理; 中心语驱动的短语结构语法; 计算语言学

Chinese Grammar Engineering: Significance, Content and Methodology

YANG Chunlei

School of English Studies, Shanghai International Studies University,

Shanghai 200083, China

yangchunlei@shisu.edu.cn

Abstract

Chinese Grammar Engineering (GE) for deep linguistic processing will provide not only the urgently needed theoretical Chinese grammar in natural language processing, but also the platform for computational implementation. First, by data analysis and literature review, the urgency and significance of constructing the Chinese GE are pointed out. Then, the research content of Chinese GE, especially the description of Chinese special linguistic phenomena in focus, is proposed from a linguistically universal perspective and for deep linguistic processing. Next, the revised research methodology that bridges across ontological linguistics and computational implementation is put forward. Finally, the research findings on multiple levels are introduced, which proves the feasibility of the proposed methodology.

Keywords: grammar engineering (GE), deep linguistic processing, Head-driven Phrase Structure Grammar (HPSG), computational linguistics

1 引言

©2023 中国计算语言学大会

根据《Creative Commons Attribution 4.0 International License》许可出版

基金项目: 上海市浦江人才计划 (2020PJC101); 上海市I类高峰学科建设项目 (41004525/001); 中央高校基本科研业务费资助项目 (2019114030)。

对语法工程 (Grammar Engineering, 简称GE) 的理解和定义不尽相同。从偏语言本体的视角, GE作为一种基础工程, 指构建覆盖广泛的语言知识基础, 以应用于自然语言系统的研究 (Erbach and Uszkoreit, 1996)。从偏应用的视角, GE指将语言学假设编写为计算机可读取的形式, 从而可以使用计算机对假设进行验证 (Bender et al., 2011)。本研究中的GE既包含前者, 也包含后者。GE可基于不同语法理论框架。本研究以中心语驱动的短语结构语法 (Head-Driven Phrase Structure Grammar, HPSG) (Pollard and Sag, 1994; Sag et al., 2003) 为例, 介绍GE的研究意义、内容和方法。具体地说, 第2至第4节分别回答三个问题: 1) 第2节分析和讨论面向深层语言处理 (deep linguistic processing) 的汉语GE研究现状和价值, 回答“为什么 (why)”要研究的问题; 2) 第3节详细讨论汉语GE的研究内容, 回答研究“什么 (what)”的问题; 3) 第4节讨论汉语GE的完整研究流程, 回答“如何 (how)”研究的问题。然后, 第5节介绍根据本研究方法产出的汉语GE研究成果, 验证其有效性。第6节是结语。

2 为什么要开发汉语GE

2.1 基于HPSG的汉语GE研究现状明显落后

基于HPSG的汉语GE研究非常重要和紧迫。尽管多年前诸位汉语语法大家就已指出这个极具潜力的研究方向, 但至今国内的相关研究在理论、计算实现和应用等三方面都与国际水平差距明显。

首先, 理论方面, HPSG是主要生成句法理论框架之一。国际上, 无论是在语法理论还是在自然语言处理领域, HPSG研究自创立以来一直保持活力。HPSG热度不减, 而且不断发展, 有理由相信未来HPSG会获得更多关注, 更受欢迎 (郑国锋 and Whitman, 2020; Flickinger et al., 2021)。语法理论的未来应是类似HPSG这样的理论, 融合了面向表层、基于约束和模型理论的方法 (Müller, 2020)。

就汉语研究而言, 由于HPSG“不仅具有较广泛的描写语法现象的能力, 而且所做的描写也比较自然” (方立 and 吴平, 2003), 而且汉语中丰富的词汇特征在很大程度上决定了句法和语义结构, 所以HPSG的词汇主义特征特别适合汉语分析 (陆俭明, 2006a), 为语言处理带来了新曙光, 可能是一条光明大道 (陆俭明, 2006b)。今后汉语本体研究首先需加强句法格式的句法、语义的特征研究 (陆俭明, 2006b)。这正是本研究的中心任务。

然而, 虽然许多语言学大家都特别强调HPSG在汉语分析中特别重要的作用, 国内在HPSG框架内的汉语研究在内容和数量上都大幅落后于其他主要语种。国际上, 除英语外, 自上世纪80年代至今, 已陆续出版了多种语言的基于HPSG的语法理论专著, 如德语 (Nerbonne et al., 1994)、韩语 (Kim, 2004; Kim, 2016) 等。但是, 迄今还未出现一部系统的, 面向自然语言处理的汉语短语结构语法。国内计算语言学领域对于语言本体研究存在较大的忽视, 例如国务院印发的《新一代人工智能发展规划》中, 没有在最具原创性和与人工智能核心架构相关的语言及语言学研究方面给予足够的关注 (耿立波 et al., 2021)。一部完整的语法通常有上千条规则, 几万甚至十几个词条。以完全的人工模式开发这样一部语法, 大约需要几年至十几年的时间 (化柏林, 2004)。然而, 国内的大多HPSG研究还停留在理论介绍的层面, 只有少数学者使用HPSG对汉语进行过理论分析和计算实现 (高明乐, 2004; 杨春雷, 2013; 杨春雷 and Flickinger, 2014; 杨春雷, 2016; 杨春雷, 2017; Yang and van de Weijer, 2021), 从研究广度和精度上与国际先进水平尚存差距, 例如英语资源语法 (English Resource Grammar, ERG) (Copestake and Flickinger, 2000; Flickinger, 2011)。

国内HPSG相关研究的数量也非常有限。据统计: 1) 在1985年1月至2022年11月间发表的共1524项HPSG相关成果中 (参见<https://hpsg.fu-berlin.de/HPSG-Bib/>), 研究汉语的成果只有24项, 仅占约1.5%。2) 在WoSCC (Web of Science Core Collection) 中所有HPSG的期刊论文中, 研究汉语的仅占1.4%。3) 中国知网 (CNKI) 中研究汉语的HPSG相关文献仅有20篇, 而韩国引文索引 (Korean Citation Index) 中研究韩语的HPSG文献为132篇。

第二, 计算实现方面, 根据欧洲专家咨询小组 (European Expert Advisory Group) 发布的报告, HPSG是计算语言学领域应用最广泛的语法理论 (Backofen et al., 1996), 至今仍是自然语言处理领域中最受欢迎的语法理论之一。例如, HPSG是计算实现语言数量最多的语法框架, 共计32种 (Müller, 2020), 其中包括8部资源型语法, 即已商用或可直接商用的语法 (Oepen et al., 2002; Bender et al., 2010; Bender and Emerson, 2021)¹。虽然已有3部中型汉语GE, 但

¹ 详见<http://github.com/delph-in/docs/wiki/GrammarCatalogue>。

在精度和规模方面还有较大改进空间。1999年，德国人工智能研究中心DFKI、斯坦福和东京的研究组织共同发起成立了基于HPSG的深层语言处理研究组织（Deep Linguistic Processing with HPSG Initiative, 简称DELPH-IN），其目标是帮助更广泛地开发语法和平台(Oepen et al., 2002)。在接下来的20年中，主要参与者来自13个国家的18家知名高校或研究机构，如剑桥大学计算机实验室、东京大学井实验室等，但还没有国内高校或机构参与。此外，国外对GE开发平台的研究，始于上世纪九十年代初，十余年前已成果显著，而国内才刚刚起步。

第三，应用方面，前面提及的资源型语法已有多种用途，既用于计算语言学研究，也用于商业软件产品；但是，由于前期理论和计算实现研究基础不足，国内相关应用还是空白。1) 在机器翻译领域，有基于ERG，应用于旅游和预约规划方面的机器会话翻译（Verbmobil）、自动客户服务应答（YY技术），以及挪威语远足宣传册的机器翻译（LOGON）。2) 在教育领域，自1993年起，由斯坦福大学语言与信息研究中心（Center for the Study of Language and Information, Standard University, CSLI）的“灵构”系统项目组开发的写作教学工具用于孟菲斯公立学校。教育应用还有英语语法检查与修改（EPGY/Redbird）、挪威语语法教学、逻辑教学等。3) 在许多其他领域也有应用，如数据库查询（惠普实验室）、选集搜索、本体获得（ontology acquisition）、视觉机器人控制、视觉问答等(Bender and Emerson, 2021)。

总之，汉语GE的发展落后已成为汉语深层自然语言处理与世界接轨的主要障碍之一。HPSG框架内，对汉语的形式化研究的主题和数量较少，视角还不够宽广，描写不够精细，尚未开发出系统全面的汉语GE，限制了汉语形式语法的深入发展和推广应用。建立面向深层语言处理的汉语GE成为一个亟待解决的问题。

2.2 汉语GE的理论和应用价值

基于HPSG的汉语GE的理论和应用价值体现在以下三方面：第一，统一使用分类特征结构（typed feature structures, TFS）进行跨层次描写和计算实现；第二，跨语言的设计考量与计算实现；第三，可作为相关研究的平台。

首先，统一使用分类特征结构作为不同语言层次的描写手段，无论是在语言描写还是在计算实现方面都有其独特的优点，包括俭省、界面和语言处理的独立性等(Sag et al., 2003)：1) 俭省：统一使用特征结构对不同层次的语符进行形式化描写，包括音系、词项、词汇规则、语法规则、普遍原则，使语法体系更精简。特定的特征只适用于特定的单位类型，可能的特征—值配对（feature-value pairing）的约束条件也只与特定的类型相关(Sag et al., 2003)。2) 界面：把音系、句法和语义信息都纳入同一性质的数据结构，即分类特征结构，从而能够使用约束条件表述不同种类信息间的关系。特征结构还会纳入语用信息，进一步扩展界面研究的范围。许多语言学家的主要研究兴趣是“把语言的各个方面都归入一种统一的理论中”(Zwicky, 1988)。在HPSG的理论框架下，建立全面系统的汉语GE，有助于探讨汉语音系、词汇、句法和语义接口的互动关系。3) 语言处理的独立性：信息描写是对不同层次上语符的客观说明。这个特点使该语法可以广泛应用于多种计算应用，如语言理解模式、生成模式等。使用特征结构的语法（包括HPSG）的优点是它们可以对语言学分析进行编码和计算处理（computationally tractable）(Partee, 1979)。

第二，汉语GE从跨语言视角研究汉语，通过主要比较英汉，兼顾日、德、韩等语言的句法语义特征的异同，探索更深层次的，具有普遍意义的词汇分类体系和短语及语法规则。理想的自然语言处理（natural language processing, 简称NLP）程序不仅能处理语言的内部结构，也能处理不同语言类型的自然语料，但如何描写语言差异对NLP领域的学者来说是个巨大挑战。因此，语言学家需要从NLP的角度构建普遍适用的语言学知识体系。如今的NLP系统大多只适用于有限的几种语言，因此跨语言的语言学知识对于NLP的重要性日益凸显。

跨语言视野下的语法研究可避免局限在单一语言中“闭门造车”，导致“南辕北辙”，或者“各照隅隙”，导致“鲜观衢路”。凭借标准化的分类特征结构等描写手段、语法规则体系、开发平台、工具和研究方法（见第4节），特定语言GE的研究者即使“闭门造车”也可“出门合辙”，或者即使“各照隅隙”，仍可“兼观衢路”，以提供具有更广泛借鉴意义的理论和应用成果，并为跨语言的计算实现与应用（如机器翻译等）奠定兼容性好且高效的研究基础。

第三，汉语GE可作为信息科学相关研究的平台。汉语GE将提供编制汉语可计算语法的基础框架，从而使利用可计算语法检验形式化语法分析的效力成为可能。如果把语言系统比作一座建筑，本体专题研究像是建筑的组成部分，如穹顶、门、窗等。这些部分是否合格取决于它们安装在建筑上是否合适。一个看上去华丽且规模宏大的穹顶可能让整座建筑不协调，甚至有

垮塌的风险。事实上，在GE的过程中，我们常发现，把某些语言学形式化分析编译后，虽然可成功自动剖析特定现象的语料，但同时会导致无法自动剖析更多本已成功剖析的语料，或产生大量冗余计算等后果，导致整体自动剖析效率反而明显下降。GE过程中牵一发而动全身的情况随着语法规模的不断扩大会越来越明显。在一个完整的语法系统中考虑某个语法现象的最佳解决方案虽然难度更大，但更有价值。当前亟待解决的问题是，提供这样一个完整的系统作为检验的平台，而该系统需要一个精确的语法体系作支撑。汉语GE及其计算实现将从根本上满足这两种需要。

下面，将以HPSG为理论基础和描写手段，具体讨论构建面向深层语言处理的汉语GE的内容与方法。

3 汉语GE研究什么

把一部“手写 (hand-written)”的，即HPSG理论语法转换成计算语法需要复杂的考量(Melnik, 2007)。汉语GE的研究内容包括本体语言学和计算实现两部分。本体研究方面的任务是建立基于HPSG的汉语短语结构语法 (Chinese Phrase Structure Grammar, 简称CPSG)。前文讲到，GE必须有系统理论语法的支撑，否则GE就成了无米之炊。虽然自然语言处理是多边缘的交叉学科，但应以语言学为主(冯志伟, 2005)。中文信息处理的现状和任务是，“语言研究已成为信息工程科学发展的瓶颈。就中文信息说，眼下特别要加强词汇句法语义研究，集中精力解决好‘句处理’问题。在中文信息处理方面，目前大多偏重于工程研究，理论研究不多”(陆俭明, 2000)。

针对以上问题和发展方向，汉语GE具有以下三个特点：1) 立足于挖掘和梳理汉语语言事实，建立覆盖广泛的形式语法体系，立足于基础“理论研究”；2) 与HPSG同属“高度词汇化”，“句法语义兼重”的语法体系(Kim, 2000)；3) 以小句为研究的基本单位，属于“句处理”层面的研究，并在描写机制和计算实现方面具有向语篇扩展的潜力。

构建汉语GE应从语言普遍性 (linguistically universal) 视角进行理论分析和计算实现。所谓“普遍”有两层意思：1) 遵循普遍语法的根本观点，并从普遍关注的语言现象入手。梳理语言研究中普遍关注的语法现象及其在汉语中的相应表现。2) 采用普遍使用的形式化描写和计算实现的手段。这样做可以增强语言研究的全局观，避免陷入单一语言所谓“特殊性”的窠臼，增强研究成果在不同语言之间相互借鉴的可能性，从而方便后期跨语言的相关应用研究，如机器翻译等。

汉语GE包括音系、词库、句法和语义描写及其计算实现。仍以把语言系统比作建筑为例，音系仿佛是这座建筑的部分外貌，词库是地基，句法是骨架，语义是功能。汉语GE具体包括以下研究内容：

第一，汉语GE的音系研究内容包括但不限于：借鉴标X-杆理论分析汉语音节结构的方法(Van de Weijer, 2012; Van de Weijer and Zhang, 2008)；使用修正的分类特征结构描写汉语普通话的音系；把音调特征纳入HPSG系统；综合使用分类特征结构和约束条件描写变调等汉语音系现象(Yang and van de Weijer, 2021) (见图3)。

第二，汉语GE的词库建设：在权威词典基础上，建立更细致深入的汉语词类的分类特征结构系统。如前文所言，HPSG是高度词汇化的语法，词汇跨类 (lexical cross-classification) 信息非常丰富。这些词汇信息不是一一列举出来，而是通过多层继承层阶 (multiple inheritance hierarchies) 和词汇规则组织起来，从而能够从词库底层开始，一层层累积，产生复杂的词汇特征。通过扩展多层级的词库，删减词汇和句法规则，并按照一般方式从语义特征中产生连接模式 (linking pattern)。

汉语GE可借鉴如《现代汉语词典》(中国社会科学院语言研究所词典编辑室, 2016)、《现代汉语语法信息词典》(俞士汶 et al., 1996) 等知识工程类的词典、动词配价信息方面的成果，如袁毓林(1998; 2010)。但是，在此基础上，还要在以下两方面做大量细致深入的工作：1) 根据句法和语义特征丰富词类：与主要语言GE中包括的几千种词类相比，现有的汉语词类系统明显不足以描写复杂的汉语自然语料。大量句法信息实际是储存在多层的词汇-句法界面的构式中，而不仅在底层词库 (见第5节第二点关于数词词库的介绍)。2) 构建词汇跨类系统：描写汉语中普遍存在的兼类词 (词汇跨类)，有助于应对语法语法功能标记不丰富的特点对自然语言处理的挑战。例如，从论元结构看，“推荐”这一词形分属于一元、二元和三元动词 (如“我推荐他参赛”) 三个词汇小类；从控制关系看，其三元动词细类属于宾控动词 (即传统的兼语动词)；

此外，该小类还能构成主语共享类型的动结式（如“我推荐成功了”）。

词汇层面GE的研究内容包括但不限于：（相对）穷尽性描写具有鲜明句法语义特征的词汇细类，如按照可数性、生命性等特征区分的名词类别；按照配价区分的动词类别，以及离合词、控制动词、提升动词、中动词、状态变化动词等动词细类；形容词中的修饰性与谓词性形容词等细类。此外，应描写汉语的形态标记，以及派生与屈折等变化的词汇规则。

第三，汉语GE的句法描写：句法是附加在语句上的约束条件，决定了语句的语法正确性和信息组合成语义结构的方式。句法描写的任务是形式化描写基本句法原则和规则，如配价原则、中心语特征原则、中心语—修饰语规则、中心语—补语规则、中心语—指定语规则、并列和一致规则等。重点研究形式化描写汉语的特殊语法现象，如汉语句子的左缘结构（包括话题、焦点和主语等）、时体貌、复谓结构（包括兼语式、连动式、趋向动词结构等）、“把”字句、“被”字句、名物化、无主句、限定成分序列、“VP+不/没+VP”结构、双宾结构、定语——表语形容词转化、无连接词并列小句、复杂数词短语（见图4的计算实现结果）等。

第四，汉语GE的语义描写：1) 在词汇语义层面，应深入考察汉语词汇语义和句法结构间的关系，建立具有普遍语言学意义的语义特征。当代句法理论好像都认同一点，即根据词语的意义通常可以预测句子结构(Wasow, 1985)。例如，许多近义词体现出非常迥异的句法特征。如likely后带不定式补语（如Pat is likely to win），但probably却不能（如*Pat is probably to win）。2) 在句子语义方面，根据“弗雷格原则”，句法很大程度上决定了信息组合成语义结构的方式，为语义“搭建脚手架（scaffold）”(Bender, 2013)。在基于HPSG的GE中，可用最小递归语义（Minimal Recursion Semantics, 简称MRS）(Copestake et al., 2005)为句法结构匹配语义描写。语义层面的研究内容包括但不限于：语义原则（语义承袭和语义组合原则等）、匹配论元结构的语义角色、长距离约束中的语义、特定句法结构（如控制、提升、动结式等）中的语义角色、数词短语语义关系的描写（见图5的计算实现结果）等。

第五，计算实现：使用类别描写语言（Type Description Language, TDL）(Krieger and Schäfer, 1994; Copestake, 2002)，对前文介绍的各语言层次的本体描写和假设进行编码，并在语法平台（如Answer Constraint Engine, ACE）上计算实现。根据计算实现的反馈和排错，修改完善语言学描写和假设。

下节将关注GE的具体研究方法。

4 如何开发汉语GE

本节将简要介绍GE的研究方法，即如何把语言学假设进行编码，成为计算机可以读取的形式，并借助计算机对语言学假设进行验证和完善。汉语GE跨领域、跨学科、系统性强且涵盖内容广泛，涉及语言学的本体研究和计算语言学，既包括运用语言学知识和研究方法分析语法现象，建立语法系统，又包括运用相关软件和常用开发工具，对语法系统进行计算实现。如何把汉语语言学的本体研究和计算科学结合起来，是近20年来语言学界和计算科学界共同关注的热点和难点。研究者要具备系统的语言学和相关计算机科学的知识与能力，同时要能把两者融会贯通。

无论是理论语言学还是计算语言学，两个领域受各自研究条件的限制，都很难独立高效开展GE研究。一方面，传统的语法研究，受人脑计算能力的限制，无法处理大规模语料，从而无法及时获得语言学假设应用于分析自然语料时的效率，作为判断假设合理性的重要根据。此外，如第2.2节第三点所言，GE是系统工程，牵一发而动全身。语言学家很难顾及到针对某种语言现象的假设是否与整个语法体系兼容。这些问题可通过使用计算机解决。因此，GE中，语言学家需要熟悉计算机可识别的编码和相关软件。另一方面，有计算机辅助的GE同样离不开语言学家的专家知识。例如，基于自然语言的语言学假设（即形式化的语法描写）决定了计算语言编码的结果。后者很大程度上是对语言学假设的转写，虽然转写需要一定的方法和技巧，但不涉及面对语言现象的规律总结，因此并没有语言学意义上的创新。此外，平台和工具操作的对象是依据语言学假设转写的代码，一般不会提供原创性的语言学问题解决方案。

因此，汉语GE必须综合使用传统的依靠纸笔的方法和有计算机辅助的方法，并在语言学本体的形式化描写和计算实现的编码之间建立直接联系，构成在语言学本体研究和计算实现两个领域间相互关联的循环系统。Bender等(2011)提出了有计算机辅助的GE的具体研究环节，如图1所示（作者译）。其中长方形环节由语言学家完成；椭圆形环节由人机交互完成；钻石形环

节完全由计算机自主完成。

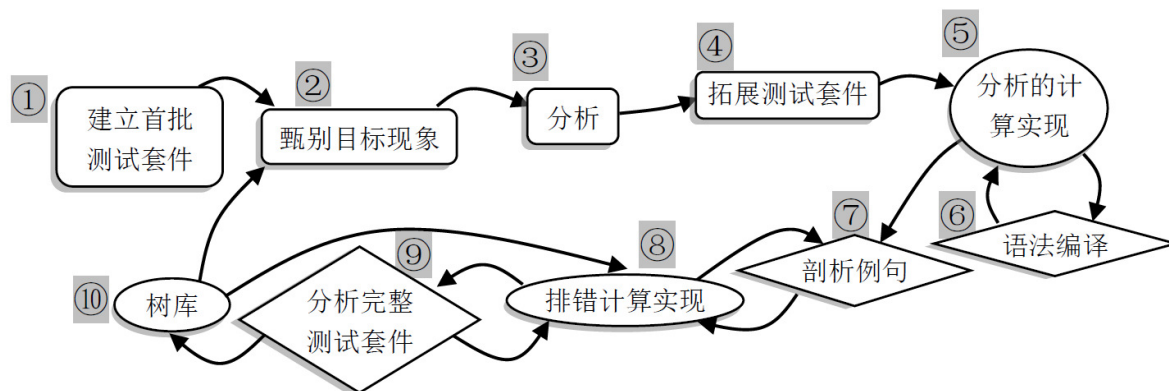


图 1: 有计算机辅助的语法工程工作流程(Bender et al., 2011)

其中，环节①、④和⑨中的“测试套件”是一组自然语句，它们或由语言学家人工选取并覆盖较广泛的语言现象，或摘自较可靠的语料来源（如经典文本，教材等）。此外，环节⑩“树库”指的是利用语言学家的专业知识，对有歧义的句法自动剖析结果进行更精细的排歧标注，按照歧义的程度进行排序。GE可按照这个排序，在剖析、生成和翻译时给出优先结果，提升准确率。

不过，根据我们开发和完善汉语GE的经验，发现该流程在以下六个方面需要改进：

1) 环节①“建立首批测试套件”和环节②“甄别目标现象”应交换位置。图1显示，要先建立并处理首批测试套件，然后从处理结果中发现问题，但尤其是在完善特定语言现象时，语言学家可基于专家知识，或者仅基于一个自动剖析的不当结果发现问题。2) 环节④“拓展测试套件”是可选的，可并入环节③“分析”，使整个流程更简明。由于在分析语料的过程中，很可能会不断补充新语料至测试套件，因此可将④视为一个可持续循环的过程（即图2中的环节3“分析语料”+“扩充语料”）。而且，语言学家不必等到把扩充后的完整测试套件分析完再开始计算实现，而是可借助计算实现平台和工具，先做一些计算实现的尝试，初步检验假设是否可行。3) 环节⑤“分析的计算实现”、⑥“语法编译”和⑦“剖析例句”应依次进行而不能从⑤跳至⑦。因为任何修改都要先确认与整个语法系统能够兼容，即修改后的语法能成功编译，修改才会生效。4) 环节⑦“剖析例句”和⑨“分析完整测试套件”是并行的，而不是依次进行。既要通过单句分析验证和完善语法假设，又要通过批处理及时检验和评估该假设对整个语法系统的影响。这两个环节应不断交叉重复进行。基于批处理结果的回归测试非常重要。如果发现某个假设显著降低了整部语法的效力，开发者应谨慎考虑其正确性。5) 尽管树库对最终提高语法的精确度非常重要，但在GE的初始阶段，或在针对特定语言现象的GE中，如果目标现象（如数词短语结构）的剖析结果不涉及歧义，则不必建设树库，因此该环节是可选环节（见图2中用虚线表示的环节11“树库”）。如果语法整体上还不够完善，会产生许多不合理或冗余的结果，此时建设树库会不可避免地造成人工和计算能力的浪费。6) 在图2中增加了三处带箭头的虚线，标识可选操作步骤。除了前文第2点谈及的环节3“分析语料”的“扩充语料”外，开发者为了排错，需要在环节5“将假设编码”，甚至是环节9“批量剖析”回到环节3“分析语料”，重新分析语料，检视之前的假设。

基于以上讨论，我们提出GE的混合模式工作流程，如图2所示。

图2显示，语言学专业知识贯穿整个流程；同时，计算实现的专业知识也至关重要。一方面，从“界定语言现象”到“提出假设”的诸环节的工作由语言学家完成，语言学知识的重要性显而易见。此外，在依据“单句剖析”和“批处理”结果进行“排错”的环节，语言学专业知识对于提出解决方案也起关键作用。另一方面，如果不具备计算实现的知识，提出语言学假设时，则无法预想计算实现过程中可能出现的问题或困难。这常会导致计算实现效果不佳，甚至最终无法计算实现。GE开发者应考虑至少以下三方面的内容：1) 是否有可匹配语言学假设的特征进行计算实现的编码。添加新特征的解决方案有较高的准入门槛。只有当有充分的理论依据和语料数据支撑证明，新特征对语言描写有普遍意义，而不是为应付特定语言所谓的“特异性”而采取

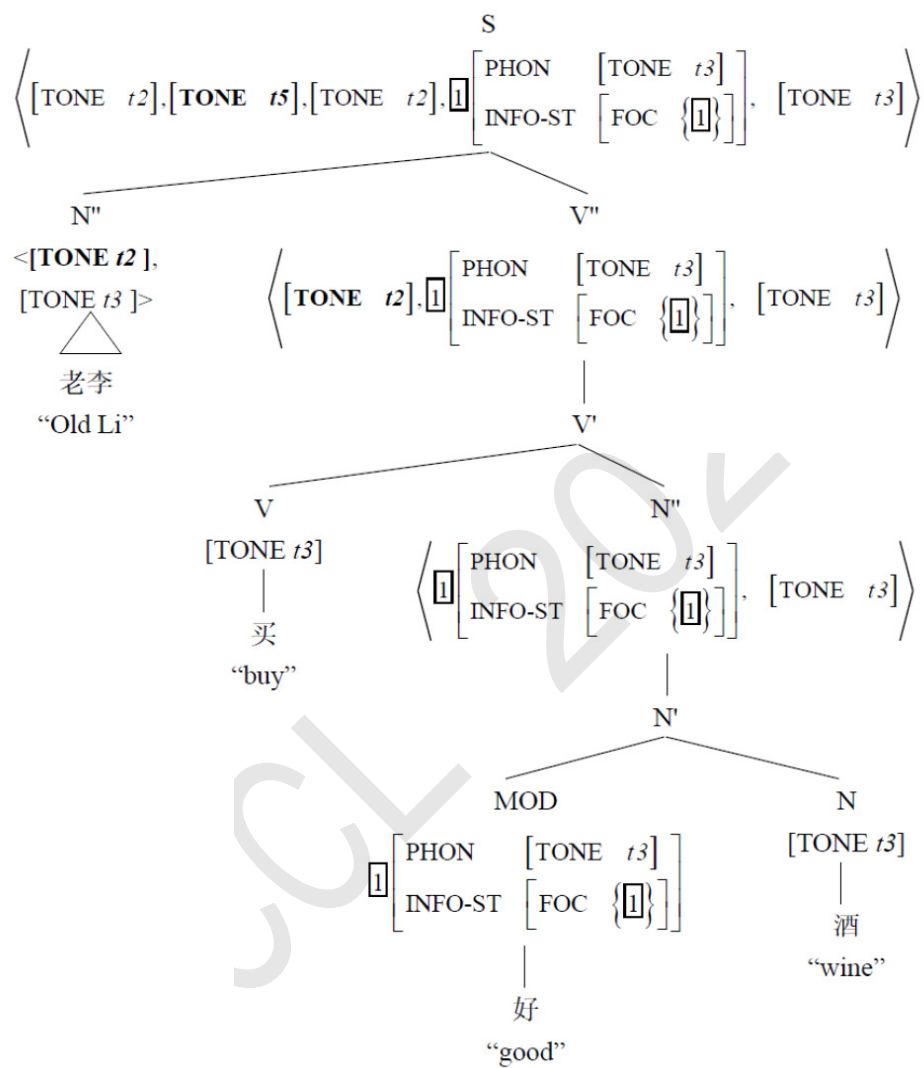


图 3: 汉语普通话三声变调的HPSG分析和描写(Yang and van de Weijer, 2021)

类，如个位数之前的“零”（即零_{j_1}）与“百万”之前的“零”（即零_{j_5}）（见图4中自动剖析的结果）。2）普遍句法规则：中心语——补语（head-comp）和中心语——指定语（head-spec-noncl）规则的交替使用，体现了X-标杆理论的合理性和语言的递归本质。3）约束条件：特殊位数词中心语（如“万”）的指定语（specifier）的饱和特征（SATURATED）限定了其必须满足该条件才能继续组合，避免生成过度概括的数词短语。

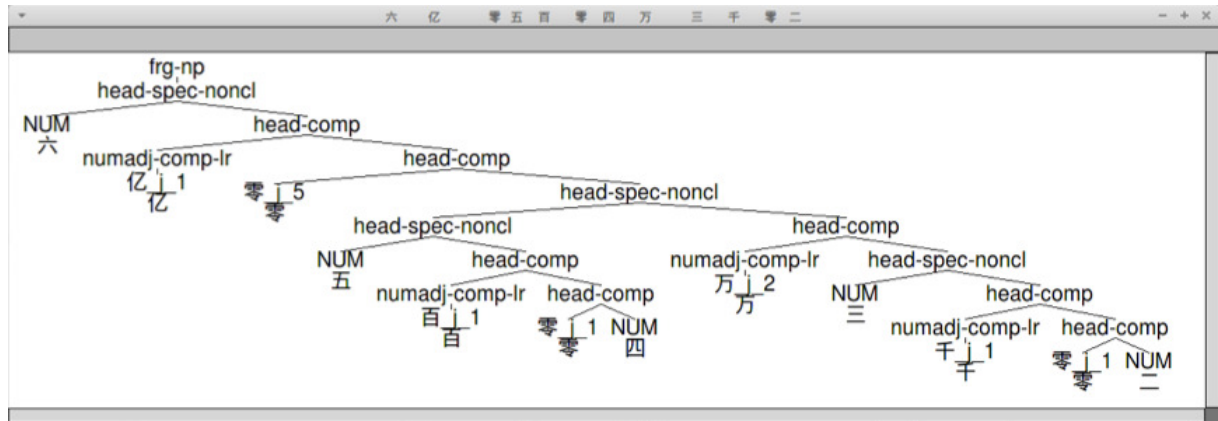


图 4：“六亿零五百零四万三千零二”的句法自动剖析结果

第三，在语义层面，GE还能匹配句法结构的语义信息。受空间所限，无法完整且清晰地展示图4中的复杂短语的语义自动剖析结果，因此以较简短的“五十万四千”为例，展示其语义自动剖析结果，如图5所示。整个数词短语的语义通过乘法（即times_rel）和加法（即plus_rel）两种语义关系把各构成成分的语义信息（即CARG的值）组合得出。可以看出，基于HPSG的汉语GE体现了该理论框架“句法语义兼重”的特点。

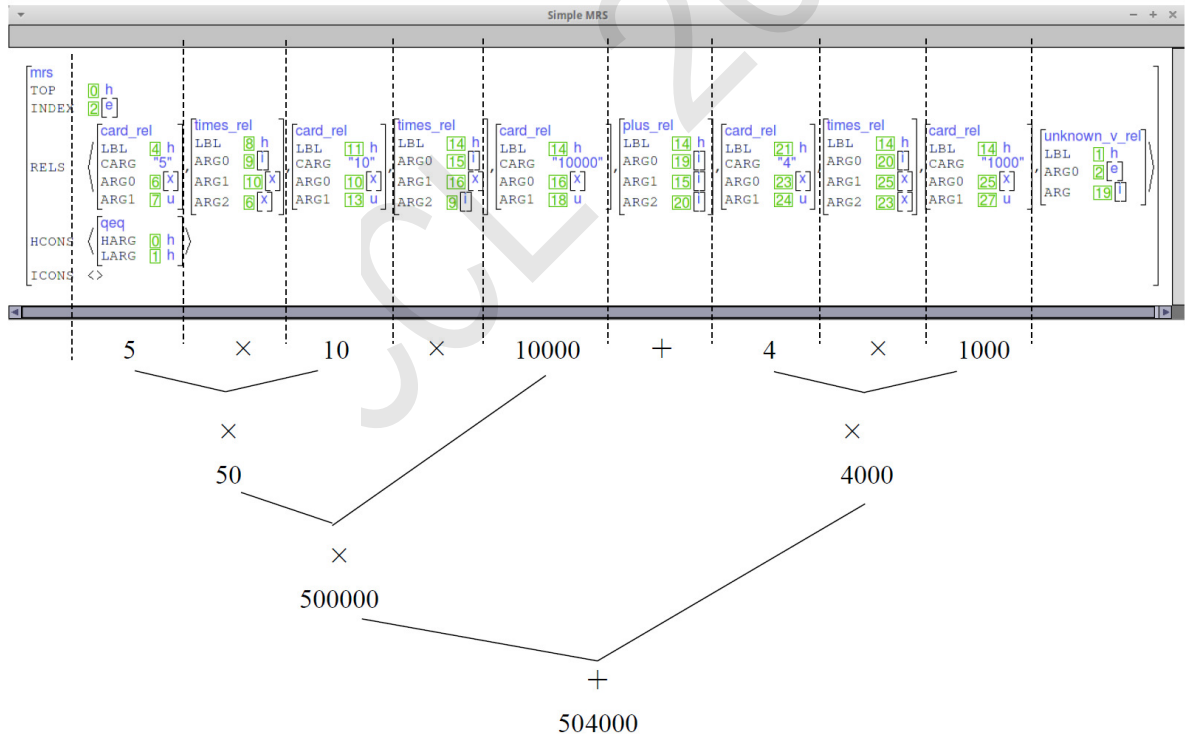


图 5：“五十万四千”的语义自动剖析结果

6 结语

汉语GE的研究内容和方法均已比较成熟，在理论和应用上都是可行的。现阶段的汉

语GE成果验证了陆俭明、方立等语言学家提出的HPSG特别适合汉语分析的观点。其理论模式和验证方法既忠实于语法现象的精确描写，又兼顾解释生成，并着眼于成果应用，具有良好的发展前景。

基于本文讨论的研究内容与方法的汉语GE，可弥补面向NLP的汉语理论和计算语法的空缺，促进汉语理论语言学和深层语言处理的发展，以及国内外相关学者的对话，并为汉语自然语言处理提供平台，为GE和相关学科深入研究奠定基础，包括语言生成、语言教育(Da Costa, 2021)、机器翻译等。无论是深入研究，还是商业应用，其成效都取决于精细的语法基础，即GE本身的质量。GE质量高，后续工作则水到渠成，反之则无从谈起。

参考文献

- Rolf Backofen, Tilman Becker, Jo Calder, Joanne Capstick, Luca Dini, Jochen Dorre, Gregor Erbach, Dominique Estival, Suresh Manandhar, Anne-marie Mineur, Gertjan van Noord, Stephan Oepen, and Hans Uszkoreit. 1996. The eagles formalisms working group-final report. Report, German Research Center for Artificial Intelligence (DFKI).
- Emily M Bender and Guy Emerson, 2021. *Computational linguistics and grammar engineering*, pages 1105–1153. Language Science Press, Berlin.
- Emily M. Bender, Scott Drellishak, Antske Fokkens, Laurie Poulson, and Safiyah Saleem. 2010. Grammar customization. *Research on Language and Computation*, 8(1):23–72.
- Emily M. Bender, Dan Flickinger, and Stephan Oepen. 2011. *Grammar engineering and linguistic hypothesis testing: Computational support for complexity in syntactic analysis*, pages 5–29. CSLI Publications, Stanford.
- Emily M. Bender. 2013. *Linguistic Fundamentals for Natural Language Processing: 100 Essentials from Morphology and Syntax*. Morgan-Claypool, San Rafael, CA.
- Ann A Copestake and Dan Flickinger. 2000. An open source grammar development environment and broad-coverage english grammar using hpsg. In *The Second Conference on Language Resources and Evaluation (LREC-2000)*, pages 591–600.
- Ann Copestake, Dan Flickinger, Carl J. Pollard, and Ivan A. Sag. 2005. Minimal recursion semantics: An introduction. *Research on Language and Computation*, 3(4):281–332.
- Ann Copestake. 2002. *Implementing Typed Feature Structure Grammars*. CSLI Publications, Stanford.
- Luis Morgado Da Costa. 2021. *Using Rich Models of Language in Grammatical Error Detection*. Thesis.
- Gregor Erbach and Hans Uszkoreit. 1996. Grammar engineering: Problems and prospects. In *Saarbrücken Grammar Engineering Workshop*.
- Dan Flickinger, Carl Pollard, and Thomas Wasow, 2021. *The evolution of HPSG*, pages 48–87. Language Science Press, Berlin.
- Dan Flickinger, 2011. *Accuracy vs. robustness in grammar engineering*, volume 201, pages 31–50. CSLI Publications, Stanford.
- Jong-Bok Kim. 2000. *The Grammar of Negation: A Constraint-based Approach*. CSLI Publications, Stanford.
- Jong-Bok Kim. 2004. *Korean Phrase Structure Grammar*. Hankook Publishing, Seoul.
- Jong-Bok Kim. 2016. *The Syntactic Structures of Korean: A Construction Grammar Perspective*, volume 1. Cambridge University Press, Cambridge.
- Hans-Ulrich Krieger and Ulrich Schäfer. 1994. Tdl – a type description language for constraint-based grammars. In *Proceedings of the 15th international conference on computational linguistics (COLING-94)*, page 893–899.
- Nurit Melnik. 2007. From “hand-written” to computationally implemented hpsg theories. *Research on Language and Computation*, 5(2):199–236.

- Stefan Müller. 2020. *Grammatical Theory: From Transformational Grammar to Constraint-based Approaches*. Language Science Press, Berlin.
- John A Nerbonne, Klaus Netter, and Carl Jesse Pollard. 1994. *German in Head-driven Phrase Structure Grammar*, volume 46. Center for the Study of Language and Information, Stanford.
- Stephan Oepen, Dan Flickinger, Jun-ichi Tsujii, and Hans Uszkoreit. 2002. *Collaborative Language Engineering: A Case Study in Efficient Grammar-based Processing*. CSLI Publications, Stanford.
- Barbara Hall Partee, 1979. *Montague grammar and the well-formedness constraint*, pages 275–313. Academic Press, London.
- Carl Pollard and Ivan A. Sag. 1994. *Head-Driven Phrase Structure Grammar*. University of Chicago Press, Chicago.
- Ivan A. Sag, Thomas Wasow, and Emily M. Bender. 2003. *Syntactic Theory: A Formal Introduction*. CSLI Publications, Stanford.
- Jeroen Van de Weijer and Jisheng Zhang. 2008. An x-bar approach to the syllable structure of mandarin. *Lingua*, 118:1416–1428.
- Jeroen Van de Weijer, 2012. *Using local constraint conjunction to discover constraints: the case of Mandarin Chinese*, pages 255–264. De Gruyter, Berlin/Boston.
- Thomas Wasow, 1985. *Postscript*, pages 193–205. Center for the Study of Language and Information, Stanford.
- Chunlei Yang and Jeroen van de Weijer. 2021. An hpsg approach to chinese syllable structure and tone sandhi. *Lingua*, 258:103048.
- Arnold M. Zwicky. 1988. Zwicky谈有关接口语法等问题. 国外语言学, 9(4):145–151.
- 方立, 吴平. 2003. 中心语驱动短语结构语法评介. 语言教学与研究, (5):31–43.
- 中国社会科学院语言研究所词典编辑室. 2016. 现代汉语词典 (第7版). 商务印书馆, 北京.
- 俞士汶, 朱学锋, 王惠, 张芸芸. 1996. 现代汉语语法信息词典规格说明书. 中文信息学报, 10(2):1–22.
- 冯志伟. 2005. 自然语言处理的学科定位. 解放军外国语学院学报, 28(3):1–8.
- 化柏林. 2004. 语法开发平台, 我们落后了. 中国计算机用户, (33):51–52.
- 杨春雷, Dan Flickinger. 2014. 汉构: 面向深层语言处理的语法工程. 现代图书情报技术, (3):57–64.
- 杨春雷. 2013. 兼语式的深层语言处理: 从语言学设计到计算实现. 外国语, (3):50–59.
- 杨春雷. 2016. 基于hpsg的汉语词库和语法规则系统构建. 现代图书情报技术, (7):129–136.
- 杨春雷. 2017. 面向语用消歧的量化约束条件系统: 从语言学设计到计算实现. 数据分析与知识发现, 1(11):1–11.
- 耿立波, 酆格斐, 詹卫东, 杨亦鸣. 2021. 中国计算语言学研究现状与展望. 语言科学, (5):491–499.
- 袁毓林. 1998. 汉语动词的配价研究 (*Hanyu dongci de peijia yanjiu*). 江西教育出版社, 南昌.
- 袁毓林. 2010. 现代汉语配价语法研究. 商务印书馆, 北京.
- 郑国锋, John Whitman. 2020. 语言学, 语言学流派, 语言学教育: 康奈尔大学语言学系主任约翰·惠特曼(John Whitman) 教授访谈录. 外国语, (5):121–125.
- 陆俭明. 2000. 汉语言文字应用面面观. 语言文字应用, (2):4–8.
- 陆俭明. 2006a. 句法语义接口问题. 外国语, (3):30–35.
- 陆俭明. 2006b. 要重视特征的研究与描写. 长江学术, (1):80–86.
- 高明乐. 2004. 题元角色的句法实现. 北京大学博士论文.