现代汉语动态助词“了”的自动生成研究

何晓丽，陈小荷，陈锋，钱小飞

（南京师范大学，南京 210097）

摘 要： 自然语言生成（Natural Language Generation）是自然语言处理的两大域之一，国外许多学者都在致力于 NLG 技术的研究。本文主要介绍计算机自动生成汉语中的动态助词“了”的策略。首先简单地阐述动态助词“了”的语言学研究现状；其次，具体描述动态助词“了”的生成策略；最后，谈生成的实验实现及实验结果分析。

关键词： 自然语言生成；动态助词；了

Research of the Auto-generation of Modern Chinese dynamic auxiliary “le”

He Xiaoli, Chen Xiaohe， Chen Feng， Qian Xiaofei

(Nanjing Normal University, Nanjing, 210097)

Abstract: Natural Language Generation is one of the two major fields of Natural Language Processing, Many foreign scholars are working on NLG. This article introduces the strategy of the computer automatically generated the dynamic auxiliary "le" in Chinese language. First briefly explain the status quo of linguistics research on dynamics auxiliary "le"; Secondly, describe the generation strategy specially; Finally, we discussed the realization of the experiments and analyze the experimental results.

Keywords: Natural Language Generation, dynamic auxiliary, le

1 引 言

自然语言生成是当前以计算语言学和人工智能为基础的自然语言处理领域中相当活跃的一个领域，是研究如何用计算机来生成自然语言文本的研究领域，有着极其重要的应用价值。自然语言生成的研究可以作为检验特定语言理论的一种技术手段，不断为理论语言学提供反馈，推动语言学朝纵深方向发展。

动态助词“了”一直是传统语言学界研究的热点和难点。本文主要探讨如何用自然语言生成方法自动在汉语的动词之后生成动态助词“了”。从而从计算语言学和自然语言生成角度来检验语言学界已有的对动态助词“了”的出现条件的研究成果。进一步还可以为中文信息处理服务，譬如提高机器翻译、句法结构分析的准确率等。同时在对外汉语教学中也具有重要的应用价值。

2 动态助词“了”的语言学研究

现代汉语中有两个同形的“了”字，汉语语法学界一般有动态助词“了”和语气助词“了”之分，也就是“了$_1$”和“了$_2$” 之分。所谓动态助词“了”，也就是语法学界通常所说的“了$_1$”或词尾“了”，即动词性词语或短

作者简介： 何晓丽（1981-），女，四川岳池人。南京师范大学文学院硕士研究生。E-mail: hxlsoftwind@sohu.com

语后带有的“了”，它和前面的动词性词语或短语共同构成谓词性短语。在经过分词和词性标注过的语料中，根据北京大学词性标注集的标准，一般将动态助词“了”标注为“了/u”，将语气助词“了”标注为“了/y”。（下文将动态助词“了”统一称为“了/u”，如提到语气助词“了”，则统一称为“了/y”。）

传统语言学界早就有不少学者从历时或共时方面对“了/u”做过许多开拓性的研究，并多角度地探讨了“了/u”的用法和出现的条件规律、“了/u”的隐现条件和规律，或从单个句法结构或短语结构中考察“了/u”，并发表了一些颇有分量的文章。代表人物如吕叔湘（1956；1982）、刘勋宁（1988—2002）、李兴亚（1989）、卢英顺（1991）、金立鑫（1998—2005）、任鹰（2001）、胡树鲜(2002)等。但目前对“了/u”的研究重点主要集中于单个方面的独立阐述、外部功能的探讨等方面。

“了/u”的使用情况是极其复杂的，经过多年的演变，其规则和限制更多了，而至今语言学界对“了/u”的特点和使用规则等研究仍不够齐全。总的看来，前人在探讨“了/u”的问题时，从不同角度，不同程度涉及到了“了/u”的出现条件和规律的问题，尽管已经取得不少研究成果，但大都是举例性地说明自己的观点，各自为证，缺乏对“了/u”的全面考察，因而缺乏整体的说服力，有的还需要重新考虑。

而且，前人对于“了/u”的研究，其方法多为传统的语义和语法描写法，多是定性分析，缺乏定量的描写，且仍存在不少分歧，仍有许多问题没有解决。例如，现代汉语里动词带“了/u”的实际情况如何？究竟有多少动词能带“了/u”，有无规律性？如果有规律，有哪些规律？“了/u”的出现与哪些因素相关，句子长度、能带“了/u”的动词个数、小句个数、小句位置？等等，这些问题都需要更进一步地探讨。

本文在利用传统语言学界现有研究成果的基础上，通过学习和借鉴前人的理论、方法、经验和教训，结合大规模真实语料，利用以规则为主，统计为辅，两者结合起来的技术，从自然语言生成角度来重新考虑“了/u”的使用。因为传统语言学界关于“了/u”的使用情况的研究，与我们的生成研究在很大程度上是重合的，关键只在于一个特殊性和普遍性的问题上了。对于我们的“了/u”的生成研究，因为刘勋宁等人的研究已经开始在朝形式化方向探索了，所以他们的一些定量研究，可以稍加完善，直接为我所用。而其他人的研究，则可以引导我们通过观察语料、自省和统计等方式，进一步提取出能被程序实现的规则来。

3 动态助词“了”的语料观察与统计

语料的观察和统计是我们进行“了/u”生成研究的出发点。我们结合大规模语料库，对相当数量的动词带“了/u”的情况进行了穷尽的考察，以此来统计现代汉语里动词带“了/u”的实际情况，来归纳总结动词能带“了/u”的规律性等等。我们的规则就是从语料的统计观察和总结前人研究成果双重角度，将之形式化、优化而成的。

我们通过对经过分词和标注过词性的1200万字的1998年上半年人民日报语料的统计得出，带“了/u”例句一共有39269句。共有2738个带“了/u”动词词条，约占动词词条总数的15.68%；动词带“了/u”次数共有43557次，占“了”总出现次数的63.66%。

我们从语料中训练得到了两个总词表：带“了/u”动词总词表和未带“了/u”动词总词表。这两个词表对于我们考察动词能否带“了/u”具有重要的参照作用。词表主要由词条本身、出现次数和带“了/u”次数三项构成。其中带“了/u”频率最高的前十个动词依次为：进行、有、取得、参加、提出、提供、成、会见、形成、建立。其中，带“了/u”1000次以上的动词有3个，“进行”带“了/u”次数最多，一共有1797次，“有”和“取得”各带“了/u”1357次和1107次。而语料中从未带过“了/u”的动词按照出现频率的高低，前十个依次是：是、能、会、可以、可、达、认为、要求、就业、继续。其中“是”在语料中一共出现了63545次。

我们从98上半年人民日报语料中带“了/u”的例句中随机抽取了1963个样本。然后按照样句的长度进行排序和编号后，统计了句长（即句子长度，单位为字节）以及具有相同句长的频度。最长的句子为1270字节，最短的句长为34字节。具有不同句长的句子数共有468句，句长区间在100~300字节之间最多。

4 基于规则的计算机自动生成

4.1 主要研究内容

本文考察的“了/u”，主要限于句中的“了”，即动词词尾“了/u”。

本文考察的原则是：先排除所有的句末“了”，接着考察典型的不能带“了/u”的动词情况，然后用排除法排除这些情况，再考虑分类能带“了/u”的动词。通过考察统计，分别总结可带“了/u”、不可带“了/u”的形式规律，用以在大规模真实语料中程序实现。

我们这里的“了/u”的生成，指的是在满足带“了/u”条件的动词或动词性短语（词性标注符号为“/v”）后由机器自动加上“了/u”。哪些动词或动词性短语可以带“了/u”，将由实验中的生成规则集提供。在此基础上，根据“了/u”生成规则集进行生成测试，可以准确地统计生成效果，如准确率、召回率等。

从 “了/u”生成的总体设计的角度说，“了/u”生成包括两方面的问题：一个是根据哪些语言知识来生成，这与汉语研究有密切关系，是总结规则库的问题；另一个是怎样实现生成的过程，这是把问题形式化和设计算法的事情。

4.2 “了/u” 生成的重点和难点

(1)规则的形式化。因为我们的生成实验主要是基于规则的，所以规则库要能充分将能后带“了/u”的动词所需具备的条件形式化地描述出来，也要尽量使之容易用程序来实现。

(2) 一个句子里有多个动词的情况。例如：“他/r 站/v 起来/v 开/v 门/n 迎/v 了/u 出去/v。/w”这个句子有五个动词。那么“了/u”应该怎么加，这实际上涉及到了有的动词可加可不加“了/u”的情形。是在符合加“了/u”的动词后都加“了/u”呢？还是用其他的策略？对此，刘勋宁有“末动词说”，即“叙述紧接着发生的几个动作，为了增强节奏感，前几个动作虽已完成，但动词后可以不用‘了’”[7]。我们基本遵照“末动词说”，先使用规则来判断该在哪些动词后加“了/u”，然后检查规则的正确率，如果正确率不高或者只有50%左右，我们再考虑概率的计算。比如一个句子里平均有几个“了/u”，那么就在可加“了/u”并且在词表中概率最高的那个动词后加“了/u”。

(3) 可加可不加“了/u”的情形，这是“了/u” 生成的一个主要难点。我们的实验没有考虑这种情形，因为这种情形跟人的语言习惯等多种不定因素有关。这也是影响我们的精确率和召回率的主要因素。

4.3 生成规则描述

基于规则的生成策略是我们采用的的主要技术。所以，生成效果的关键在于规则的定义，规则定义越全面越精确，生成的效果越好。然而，对规则的把握必须以对语言现象进行充分归纳整理为前提，这是任何自然语言处理工作的根本之道。我们在总结传统语言学的已有研究成果的基础上，通过对语料库的统计和观察来增进我们的知识，完善我们的生成规则库。

我们在总结现有语言学研究成果和统计观察大规模语料库的基础上，形成了两个主要的生成规则集：“不可加‘了/u’的规则”集和“可加‘了/u’的规则”集。

4.3.1 不可加“了/u”的规则描述

“不可加‘了/u’的规则”集是主要分为以下六大类：

(1) 标点符号。

如果标注为“/v”的动词后为句末标点，如，。— … ！；？：……等，其后不加“了/u”（我们将句末的“了”均视为为“了/y”）。

(2) 副词。

如果句子中出现时间副词（如：明天、永远、随时）、否定副词（如：没有、不曾、未、并不）和疑问副词（如：能否、能不能、如何）三类副词，后面出现标注为“/v”的动词，“/v ”后便不能加“了/u”。

(3) 动词本身。

这种情况比较复杂，也比较细。主要有九种情况：否定动词（如：不、无、别、没、未、非、甭、莫）、能愿动词（如：能、会、可以、愿意、应该）、属性关系动词（如：比、当、等于、是、属于、作为、姓）、部分表主观心态和心理活动动词（如：尊敬、欣赏、彷徨、懊恼、爱惜）、例词（如：如、似、若）、动词后缀（如：~于、~化、~乎、~在）、问词（如：何谓、岂止、谁料、奈何、是否、怎样）、接谓词性宾语的动词（如：怀疑、怕、省得、允许、加以、有必要）及其他情况的动词（如：称、使、把握、凭、透）。

(4) 介词。

如果句中有“在、向、朝、往、对、为、为了、通过、据、根据”等介词，且正确标注为“/p”，则句中标

注为“/v”的动词后不加“了/u”。

（5）词语序列。

如果句子中有“是…的、只有…才…、…以前、（在）…时、（在）…之前”等词语序列，则句中标注为“/v”的动词后不加“了/u”。

（6）其它情况。

如拟声词（如：呢喃、唏嘘、嘘、呵呵、哼哼）和文言动词（如：觐、迄）等。如果句中将它们标注为“/v”，其后也不加“了/u”。

4.3.2 可加“了/u”的规则描述

“了/u”在一些情况下是必须使用的。在总结已有语言学研究成果和观察语料基础上，我们总结出了34条“可加‘了/u’的规则”，这些规则主要由词性序列构成。如果词性序列是规则中的序列，则在标注为“/v”的动词后后自动加上“了/u”。

下面举几种规则样例予以说明（未全部列出）：（注：+表示紧邻出现；（）表示此处可有可无；…表示连续的若干个词语，最少为一个词语；|表示或；#表示终止符。）

（1）/v+/m（+/q）…（+/n）…（+了/y）

语言学界大多认为汉语动词后的三种宾语（时量宾语、动量宾语和数量宾语）与词尾“了”的出现有很密切的关系，这条规则即体现了这种观点。

例如：著名/a 指挥家/n 陈/nr 佐湟/nr 、/w 陈/nr 燮阳/nr 、/w 谭/nr 利华/nr 分别/d 指挥/v 演奏/v 【了/u】 一/m 批/q 中外/j 名曲/n 。

此句有两个标注为“/v”的动词，其后又有词性序列“/m/q…/n”，所以在后一个“/v”后要加上“了/u”。这个例子跟刘勋宁的“末动词说”也是相符合的。

（2）/v+/c+/v…+/n

这条规则主要体现：当并列动词做谓语时，“了/u”一般用于第二个动词后。

例如：中国/ns 伟大/a 的/u 领导人/n 孙/nr 中山/nr 、/w 毛/nr 泽东/nr 、/w 周/nr 恩来/nr 和/c 邓/nr 小平/nr 的/u 榜样/n 极大/a 地/u 启发/v 和/c 鼓舞/v 【了/u】 非洲/ns 人民/n 。

（3）/v+上来|下来|进来|出来|回来|过来|开来|起来|上去|下去|进去|出去|过去|回去|开去|起去|到去|到来+#

这条规则主要体现了动趋结构跟“了/u”结合的紧密性。

例如：不一会儿/l ，/w 就/d 有/v 数十/m 个/q 身/Ng 背/v 行囊/n 、/w 行色匆匆/i 的/u 旅客/n 围/v 了/u 上来/v 。/w

（4）~走|跑|动|倒|翻|病|疯|死|见|懂|完|通|穿|透|落|掉…（+/n）

这条规则主要体现了动结式动词跟“了/u”结合的紧密性。即如果标注为“/v”的动词本身的末字是“走|跑|动|倒|翻|病|疯|死|见|懂|完|通|穿|透”等表示结果意义的字，要在这个动词后加“了/u”。

例如：这/r 钱/n ，/w 这/r 衣/Ng ，/w 这/r 鞋/n ，/w 这/r 书/n ，/w 浸透/v 【了/u】 天津/ns 人民/n 对/p 彝族/nz 同胞/n 的/u 深情/n ，/w

4.3.3 规则的组织

很容易发现，一个句子有时可能同时适用于多条规则。在基于规则的生成系统中，如何解决规则间的冲突是一个非常重要的问题。我们通过对规则库中的规则进行有序组织来解决这一问题。首先，规则在规则库中按照不同的层次存放，同一层次的规则排列在一起，这样便于日后规则库的优化和维护。其次，不同层次的规则对生成产生作用的重要性不同，我们将最重要的层次放在规则库的最前面，而将次要的层次放在后面，由于系统运行是按照由前向后的顺序使用规则，这便保证了最重要的规则发挥最大的作用，同时也解决了不同层次的规则间的冲突问题。再者，为了减少同一层次的规则间的冲突，我们对每一类型尽可能细分，定义尽可能多的规则，这样每一条规则的覆盖面较小，彼此之间的冲突也就减少了。如果某些句子在某些规则作用下产生了异常效果，则必须缩小该规则的适用范围使之不再对这些句子有效。[1]

4.4 生成算法及生成结构图

[1]李锦乾，张冬荣，姚天方：自然语言生成中的句子结构优化处理，计算机应用研究，1998年第1期，第55页。

主要操作步骤基本如下：

第一步，打开语料。

第二步，生成两个词表，词表主要由词本身、出现次数和带“了/u”次数三项构成。（1）带“了/u”动词词表。（2）未带“了/u”动词词表。

第三步，扫描并去掉语料中的所有词性标记为“/u”的“了”字。

第四步，加载从大规模训练语料中训练出来的“未带‘了/u’的动词总词表”。

第五步，加载规则，先加载“一定要加‘了/u’的规则”，再加载“不可加‘了/u’规则集”。

第六步，生成最后文件生成结果。

第七步，比较第一次原标记和生成结果标记。将生成与原文不一致的句子重新生成一个错误文本，并标注出错误类型，以便进一步分析。

第八步，对其他一些特殊结构进行处理。例如等等。

第九步，进行封闭和开放测试，首先要计算测试语料词性标注的正确率，再计算出“了/u”生成结果的正确率和召回率。

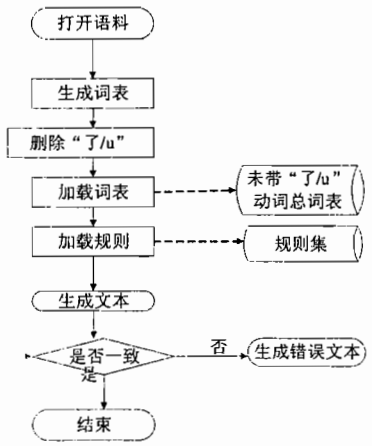

图 1　汉语“了/u”生成结构图

Fig.1　the generation structure of　"le"

5　实验结果分析及改进方向

我们从经过分词和词性标注过的《人民日报》1998 年上半年语料中随机抽取了约 1524 个词条的语料作为封闭测试语料；将词法分析器自动标注的约 2357 个词条的新闻语料，作为开放测试语料。测试结果如表 1：

表 1 动态助词“了”生成测试结果

Tab.1 The test results of the generation of dynamics auxiliary "le"

测试项目 / 测试类型		应标数（条）	标过数（条）	标对数（条）	标错数（条）	精确率（%）	召回率（%）
封闭测试	baseline	57	33	29	24	87.88%	50.88%
	测试结果	57	49	42	22	85.71%	73.68%
开放测试	baseline	59	31	24	34	77.42%	40.68%
	测试结果	59	58	41	30	70.69%	69.49%

注：①精确率=标对个数/标过个数;召回率=标对个数/应标个数。

②“了/u”标注的精确率和召回率的 baseline 测试方法：从未带过“了/u”的动词一概不予考虑。设平均每个句子能有 1 个“了/u”，将这个“了/u”分派给带“了/u”概率最高的动词。

另外，还有几个需要说明的问题。

(1)就精确率而言，封闭和开放测试都取得了较好的效果，说明算法的性能较稳定。

就封闭测试而言，产生错误的原因主要有三：一是我们没有考虑可加可不加“了/u”的情况，所以对“了/u”的自由隐现情况难以处理；二是对于语料中“了”字被误标为 y 的情况，尚未做进一步处理；三是由于语料自身的一些分词标注错误，导致我们算法中所要利用的一些词例、词性知识，无法正确获取，因而出错。例如，“十分”应标注为副词 d，语料中却错误标注为/m：今年/t　的/u　各项/r　任务/n　具有/v　十分/m　重要/a　的/u　意义/n　。/w

(2) 规则的覆盖率问题。

我们对开放测试中每条规则的贡献大小作了统计，包括精确率和覆盖面。覆盖率的计算公式为：某规则覆盖率=符合规则并可加“了/u”条数/语料中符合规则词性序列的总条数*100%。如果规则的覆盖率越大，则表明这条规则可以更好地限制。

我们举几个在语言学上比较一致认同的规则，对其在生成“了/u”时的贡献作了统计。如表 2 所示：

表 2　部分规则对生成“了/u”的贡献统计表

Tab.2 Statistics of some rules's contribution on the generation of dynamics auxiliary "le"

规 则 描 述	语料中符合规则词性序列的总条数	符合规则并可加“了/u”条数	覆盖率	测试精确率
/v+/m（+/q）…（+/n）…（+了/y）	29	19	65.52%	81.82%
动趋结构动词+…	24	17	70.83%	64.71%
动补结构动词+…	19	10	52.63%	70.00%

(3) “未带‘了/u’的动词总词表”共有 13636 个，有 11382 个动词与“不能加‘了/u’规则集”中的动词不重合。不重合的动词中主要有六种情形：①是“不能加‘了/u’规则集”中的动词的补充，例如“评议”、“服从”等就属于表主观心态的动词。②词性标注错误。首先是语料中把该标为“了/u”处标注为“了/y”（如“增多/v 了/y　数量/n”中的“了”应该标注为“了/u”）之外；其次，把不是动词的词（如“他们、或、队、那么、你、女孩子”等）标注为了动词“v”；③可构成离合词的动词。如“理”、“守”、“赴”等。④语料中该动词之后跟的是语气助词。⑤动词本身是动补结构。如“阖上”、“汇出”等。⑥同义复词。如“嘱咐”、“编纂”等。

根据我们人工考察发现，大部分不重合的动词是不能带“了/u”的。但也有部分动词在在某些情形下可以带“了/u”。如离合词“赴/v 了/u　宴/n”、“守/v 了/u　寡/ng”、“理/v 了/u　发/n”等；动补结构动词带宾语时，如“阖上/v　了/u　眼睛/n”、“汇出/v 了/u　钱/n”等；部分同义复词如“编纂/v 了/u　一/m 本/q 字典/n”等。

我们进一步的实验，需要充分考虑好可加可不加“了/u”的情形，并充分利用统计知识，进一步完善规则，增加规则的覆盖面，这是我们需要努力改进的主要方向。

参考文献：

[1]顾阳，《生成语法及词库中动词的一些特性》[J]，国外语言学，1996（3）

[2]黄友能，《可移植的自然语言生成系统中知识库的设计》[J]，北京交通大学学报，2004（5）

[3]贾佩山，《自然语言生成技术及其应用》[J]，中文信息学报，1997（1）

[4]金立鑫，《词尾‘了’的时体意义及其句法条件》[J]，世界汉语教学，2002（6）

[5]刘勋宁，《现代汉语词尾“了”的语法意义》[J]，中国语文，1988（5）

[6]刘勋宁，《现代汉语句尾“了”的语法意义及其与词尾“了”的联系》[J]，世界汉语教学，1990（2）

[7]刘勋宁，《现代汉语的句子构造与词尾“了”的语法位置》[J]，语言教学与研究，1999（3）

[8]王晓娜，《从外国学生的语病看运用动态助词“了”的制约因素》[J]，外语与外语教学，1995（6）

[9]杨国文，《从计算机生成汉语的角度看汉语语法研究》[J]，中国语文，1992（2）

[10]杨国文，《自然语言生成研究的动态与方向》[J]，当代语言学，1998（2）

[11]杨惠芬，《动态助词“了”的用法》[J]，语言教学与研究，1984（1）