语言工程中同形及兼类词语的处理策略*

俞士汶　段慧明　朱学锋
北京大学计算语言学研究所，北京 100871
Email: yusw@pku. edu. cn,　duenhm@pku.edu.cn

摘要：中文文本中的同形词（汉字相同的不同词）和兼类词（汉字相同、意义难以区分但语法功能明显不同的词）是自动分析的难点之一。《现代汉语语法信息词典》提供了区分同形词和兼类词的线索，“人民日报标注语料库”则提供了大量实例。本文解释在这两项语言工程中关于同形词和兼类词的处理策略。
关键词：自然语言处理、语言工程、同形词、兼类词、现代汉语语法信息词典、《人民日报》标注语料库。

The Processing of the homograph and the multi-category words in language Engineering

Yu Shiwen　Duan huiming　Zhu xuefeng
Institute of Computational Linguistics, Peking University Beijing　100871
Email: yusw@pku.edu.cn,　duenhm@pku.edu.cn

ABSTRACT: One of the most difficult points in automatic analysis is to identify the homographs, which are the different words but consist of same Chinese characters, and the multi-category words, which consist of same Chinese characters and are near-synonymous, but have different grammatical functions. “The Grammatical Knowledge-base of Contemporary Chinese” contains the index of the homograph and the multi-category words. “《The People’s Daily》 Tagged Corpus” also supplies a lot of the examples. This paper explains the policies of the homograph and the multi-category words processing in these two projects on language engineering.
Keyword: Natural Language Processing, Language Engineering, Homograph, Multi-Category Words, The Grammatical Knowledge-base of Contemporary Chinese, 《The People’s Daily》 Tagged Corpus

1. 引言

汉语自动分析必须克服一系列难题。中文文本中的同形词（汉字相同的不同词，如“鲜花”的“花”和“花钱”的“花”）和兼类词（汉字相同、意义难以区分但语法功能明显不同的词，如名词的“决心”和副词的“决心”）是自动分析的难点之一。从事自然语言处理研究之初，笔者已认识到，要实现高性能的自然语言处理系统，必须建设好综合型语言知识库[1]。当然，这样的知识库不可能一蹴而就。北京大学计算语言学研究所自 1986 年成立以来，一直坚持语言知识库的建设。到目前为止，已积累了一些成果[2]。这里只列举两项最重要的、也是与本文关系最密切的工作。其一，与北大中文系合作研制的《现代汉语语法信息词典》（以下有时简称《语法信息词典》）[3]。现在，这部电子词典的规模已扩充到 7.3 万词语[4]。其二，与 FUJITSU 合作的《人民日报》标注语料库[5]，现已完成 1600 万字以上语料的加工。

*得到国家自然科学基金项目“中文信息提取技术研究”（69483003 ）、973 项目（G1998030507-4）和北大 985 项目的支持。

要做这些工作，必要的基础是对现代汉语中词的辨识和词的归类有一个清晰的认识，并且要建立一套能贯彻始终的操作规则。《现代汉语语法信息词典》提供了区分同形词和兼类词的线索，“人民日报标注语料库”又提供了大量实例。本文解释《现代汉语语法信息词典》和“人民日报标注语料库”关于同形词和兼类词的处理策略。既是策略，当然主要是从语言信息处理的技术层面上考虑的，但是又不可能不涉及现代汉语中“词”及“词类”（包括兼类）的一些理论问题。笔者明白，包括词的定义在内的有关“词”及“词类”的理论问题是汉语语法本体研究的主要内容之一，并且是老大难的问题。笔者并不自信有能力驾驭这些问题。但北大计算语言学研究所毕竟完成了如此规模的语言工程，有一些心得和体会。笔者将它们整理出来，期望与学术界的同行进行深入的切磋，并得到指教。

2. 面向语言工程的词语观

电子版《语法信息词典》由一系列关系数据库文件组成。每个文件由若干记录组成。每个记录又包含若干字段。其中，有一个命名为“词语”的字段是每个文件都必须包含的数据项，它的值就是本词典所收的对象，通常在书本式的词典中叫做“词条”。其他字段则是描述该“词语”的各种属性的数据项。既是“词典”，为何不将其所收的对象直截了当地叫做“词”或“词条”而含糊地叫做“词语”呢？这里是有所考虑的。

《语法信息词典》中“词语”这个字段的值的主体是词。研制组在词典开发过程中逐步明确了收词原则，其中义项与语法功能相结合的原则体现了本词典收词的特色[3]。“七五”期间开发的“现代汉语词语语法信息库”是《语法信息词典》的基础。当时它所收的4万词都是语言学家挑选的。当词典的规模从5万扩展到7万时，为了避免将过多的自由短语包揽进来，又补充了如下原则：若干汉字的组合是一个语法单位，并且其中有的汉字不是词，则该汉字组合可作为一个词语收进扩充词典，如：“黄雀、古堡”，其中“雀、堡”是不成词的名语素；两个以上的单纯词或语素可以组合成一个较长的词，其词性不能由组合结构推导出来，这个词也可收进扩充词典，如：“卖力”中的“卖”是动词，“力”是名词，“卖力”是“述宾结构，而“卖力”却是形容词；如果复合词的意义不是组成成分意义的简单相加，这个复合词也可收进扩充词典，如：“抠门儿”。可以说，本词典收词是严格的，对所收对象是经过仔细筛选和甄别的。

不过，《语法信息词典》所收的对象并不局限于“词”，也不应该局限于“词”。开发《语法信息词典》的目的很明确，主要是为计算机自动分析汉语的句子服务，要以真实文本为对象。而在大规模的真实文本中，一定存在其地位相当于词却不是词的语言成分。这样的语言成分有以下7类：前接成分、后接成分、语素、非语素字、成语、习用语、简称略语。

进行“词”的理论研究，当然可以将“词”模型化，完全不理会这些语言成分。但面对真实的文本，计算机自动处理程序（如机器翻译程序）却不能视而不见。为了适应语言信息处理工程实践的需要，本词典纳入了这7类语言成分。前4类是比“词”的更小的语言单位，不成词。这些成分的数量是有限的，应当尽可能都收入词典。后3类是比“词”的更大的单位，词典中只能收一部分使用频率高的。本词典还收入了标点符号。因此，从实用出发，将作为本词典的“词语”字段的值出现的各种语言成分笼统地叫做“词语”既是方便的，也是确切的。

研制者虽然从实践中“悟”出这样的处理方法是必要的，但仍难免有忐忑感。后来，参阅《现代汉语词典》的凡例，发现以推广普通话、促进汉语规范化为己任的《现代汉语词典》

[6]将所收的对象称之为“条目”。对于单字条目，实际包括单字词（如：“爱”、“葱”、“从”、“多”）、不成词的语素（如：“杜”、“民”、“遥”）以及非语素字（如：“鹌”、“鹑”）。对于多字条目，实际包括单纯词（如：“鹌鹑”、“凡士林”）、合成词（如“爱护”、“洋葱”、“从此”、“阿姨”、“桌子”）、成语（如：“哀鸿遍野”、“挂羊头卖狗肉”）、短语（如：“阿猫阿狗”、“高等教育”、“高级神经活动”、“走后门”）以及简称或缩略语（如：“人大”、 “三北”）。笔者以为《现代汉语词典》的“条目”与《语法信息词典》的“词语”在概念上是相容的。笔者还认为《语法信息词典》在选词时采取了不拘一格与宁严勿滥相并重的原则。

3.“词语”与“切分单位”的关系

上节中，曾论述了“在大规模的真实文本中，一定存在其地位相当于词却不是词的语言成分”。并将“地位相当于词”这几个字加粗，加了下划线，就是为了提示读者：“地位相当于词”的含义是什么呢？

中文文本基本上是“按句连写”的，汉字一个接一个地排列，词与词之间没有间隔标记。因此，计算机对中文作深层自动分析的前提是将句子切分成较小的单位，也就是要把“字”的天然序列转换成带有主观因素的“词语”的序列。即使某些系统声称其分析过程并没有独立的切分环节，但无论如何总是要从句子中辨识出“词语”的。对“词语自动切分”的困难，从事中文信息处理的学者早就有充分的认识。因此，“七五”期间就有一批学者制订了国家标准 GB13715“信息处理用现代汉语分词规范” （以下简称为“分词规范”）[7]。概览现有的汉语信息处理系统，基本上都是在遵循这个规范的同时，又根据实际的需要，对这个规范做了或多或少的调整或补充。

北大计算语言所正在实施的一项大规模语言工程就是对总量达 2600 多万字的《人民日报》语料进行加工，第 1 期计划的加工项目包括词语切分、词性标注，并标出专有名词（包括短语型专有名称）。为保证这项工程的顺利实施，制订了《现代汉语语料库加工——词语切分与词性标注规范与手册》（以下简称《规范与手册》）[8][9]。“七五”期间，制订“分词规范”时，尚无一部可供参照的电子词典，现在，《语法信息词典》可作为基本参照。现在的加工是将词语切分与词性标注结合起来进行，因此《规范与手册》同“分词规范”比较，是有发展的，更便于操作。不过，《规范与手册》仍沿用了“分词规范” 中的一个基本概念“分词单位”。“分词规范”将“分词单位”定义为“信息处理中使用的、具有确定的语义和语法功能的基本单位”。 显然，“分词规范”同样不把“分词单位”局限于“词”。 为了不同英语语法中的“分词”混淆，《规范与手册》只是改用“切分单位”以替代“分词单位”。

本文不介绍《规范与手册》对“切分单位”的详细规定。无论是“分词规范”，还是《规范与手册》都认为“切分单位”主要是词，也可以是比词大的单位，如成语、习用语和简称略语。在某些特殊情况下孤立的语素或非语素字也可能出现在切分序列中，如动词的离合形式：

洗/v 了/u 一/m 个/q 澡/Ng 。/w

这里将“澡/Ng”划为名语素；又如：

蟋蟀/n 的/u 蟋/x 字/n 和/c 蟀/x 字/n 单独/d 有/v 意思/n 吗/y ？/w

“蟋/x”和“蟀/x”是非语素字。

这样看，“语料库加工”工程中的“切分单位”同《语法信息词典》中的“词语”、《现代汉语词典》中的“条目”在概念上都是相容的。既然成语、习用语、简称略语、语素、非语

素字等语言单位像词一样都可以出现在切分序列中，其地位也就相当于“词”。由工具软件进行自动切分通常要有一部“机器词典”。由于《语法信息词典》收录的“词语”已超过 7.3 万，对中文真实文本有很高的覆盖面，而且《语法信息词典》又包含了每个“词语”的词性等丰富的语法属性，因此《规范与手册》规定《语法信息词典》中的“词语”一般都是切分单位，这不仅使得自动切分易于实现，而且，当人对自动切分的结果进行校对时，也有了基本的参照，尽可能避免“见仁见智”主观因素的干扰。

《规范与手册》定义的“切分单位”同《语法信息词典》中的“词语”之间还是有差异的。

(1) 5 个字以上的成语、习用语、简称、地名或外族人名是切分单位，但未被收入现在的《语法信息词典》。不过，这个区别不是本质的。将 5 个字以上的切分单位作为“词语”单独建一部同样内容的词典，或者加长《语法信息词典》的“词语”字段，对开发和使用都没有困难。

(2) 尽管现在对存储空间的限制已大大缓解，也并非凡是在文本中出现过的切分单位都要收入词典。假如词典收了“条几”（一种家具的名称），就会对语句“一条几斤重的鱼”、“ 一条几米长的绳子”的自动切分造成干扰。

(3) 像“一百零八”、“五分之三”、“第二”等在语法上被认为是一个数词，像“一九九八年”、“八月”等被认为是一个时间词[10]。但这样的词无限多，任何一部词典不可能全收。《语法信息词典》只收了它们的构成成分，如：“一”、“二”、“三”、“十”、“百”、“亿”、“第”、“半”、“多”、“来”、“分之”、“年”、“月”等。这些构成成分有的本身也可以是切分单位，如：“一”、“二”、“三”；但有的却不可能单独成为切分单位，如：“分之”、“第”。

(4) 在《语法信息词典》的“词语”字段中出现的前接成分、后接成分、语素、非语素字也不是切分单位。《规范与手册》规定它们应与左邻右舍组合成切分单位，只有当它们不能与前后成分组合时，才会孤立地出现在切分序列中。

4. 同形词语的处理策略

自动分析程序要区分文本中的同形词(即汉字相同的词)，可以从词典中获取相关的知识。《现代汉语语法信息词典》对真实文本中复杂的同形词的情况（见图 1）采取了相应的的区分办法。

在《语法信息词典》中，“词语”、“拼音”、“词类”这几个字段是一定要有的。将这三个字段连在一起作为数据库的主关键项（Primary key），除了“同字同音同类”的情况外，图 1 中同形词的其他情况都是可以区分的，均作为不同记录收入词典。为了进一步区分同字同音同类的情况，另设一个字段：“同形”。对于同字、同音、同类但是应算不同“词语”或“条目”的情况，在“同形”字段中填上字母 A，B，C 等。对于同字、同音、同类、同一个“词语”或“条目”的不同义项的情况，在“同形”字段中填上数字 1，2，3 等。

由于拼音字段较长，为了提高同形词的处理效率，在“同形”字段中也用 A，B，C 等区分同字同类不同音的情况。不同音，自然是不同的词。总之，“同形”中的 A，B，C 等表示不同的词语，数字 1，2，3 等表示同一个词语的不同义项。当需要字母与数字并存时，则将字母置于数字之前。从《语法信息词典》中选择一部分同形词，将有关字段示例如下（表 1），其中“备注”字段用于存储词语的释义或用例。备注栏中的“～”代表同一行中的该词

语。这样，利用语法词典中的词语、词类(仅 2 个字节)、同形(仅 2 个字节)这 3 个字段，就能区分同形不同类的词语，同形同类不同音的词语，同形同类同音的不同词语，同形同类同音的同一个词语的不同义项。

同字
- 同音
  - 同类：不同词：抄写的“抄”，抄近道的“抄”，都属于动词。
  - 同类：同一词的不同义项：“去北京”的“去”，“去果皮”的“去”。
  - 不同类：属连词的“和”与属介词的“和”是不同的词，动词的“锁”与名词的“锁”也是两个意义有联系的不同的词；属广义兼类现象。“同等”兼属区别词与副词，仅因为语法功能不同，两者的意义并无区别，这是狭义兼类现象。
- 不同音
  - 同类：和稀泥的“和”(huo4)，和一盘棋的“和”(he2)。表示加在一起的“合计”(he2ji4)，有磋商、盘算意思的“合计”(he2ji5)。
  - 不同类：属连词的“和”(he2)，与属动词的和稀泥的“和”(huo4)。

图 1　先按音分开的同形词的不同情况

针对表 1，从计算机自动处理的角度，将同形词先按词类区分，得到图 2。

同字
- 同类 v
  - 同音
    - 不同词：“抄写”的“抄”，“同形”为“A”；“抄近道”的“抄”，“同形”为“B”。“见面”的“会”，“同形”为“A”；“付账”的“会”，“同形”为“C”；“理解”或“可能”的“会”，“同形”为“B”。
    - 同一词的不同义项：“去北京”的“去”，“同形”为“1”；“去果皮”的“去”，“同形”为“2”。“理解”的“会”，“同形”为“B1”；“可能”的“会”，“同形”为“B2”。
  - 不同音：“和棋”的“和”(he2)，“同形”为“A”；“和稀泥”的“和”(huo4)，“同形”为“B”。
- 不同类
  - 同音：属连词 c 的“和”同属介词 p 的“和”，拼音皆为“he2”。属名词 n 的“锁”同属动词 v 的“锁”，拼音皆为“suo3”。
  - 不同音：属连词 c 的“和”，全拼音为“he2”；属动词 v 的“和”（“和稀泥”），全拼音为“huo4”。

图 2　先按词类分开的同形词的不同情况

表 1　部分同形词语示例

词 语	全拼音	词 类	同 形	兼 类	备 注
保管	bao3guan3	v	1		保存
保管	bao3guan3	v	2		担保
抄	chao1	v	A		～写
抄	chao1	v	B		～近道
和	he2	c		pv	他～她是夫妻
和	he2	p		cv	他不～她讨论
和	he2	v	A	cp	～棋
和	huo4	v	B		～稀泥
合计	he2ji4	v	A		共计
合计	he2ji5	v	B		磋商、盘算
会	hui4	n		v	会议
会	hui4	v	A	n	见面
会	hui4	v	B1	n	理解

词 语	全拼音	词 类	同 形	兼 类	备 注
会	hui4	v	B2	n	可能
会	hui4	v	C	n	付帐
锁	suo3	n		v	一把～
锁	suo3	v		n	～门

5. 面向语言工程的词类观

任何一门科学都要对研究对象进行分类。现代汉语语法的研究对象包括现代汉语词汇，词语分类自然是汉语语法的研究内容。一般说来，语言信息的自动处理也离不开语法意义上的词类。为了建立一个科学的实用的词类体系，众多的语言学家付出了才智与辛劳，已经取得了成绩[11]，当然新的意见还会不断出现[12]。

指导《语法信息词典》研制和大规模标注语料库开发的汉语词类的观点总结如下。

(1) 汉语的词是需要分类的，也是可以分类的。事实上现在已有多种词类体系存在。由于词类与人对语法的理解与解说有着密切的关系，用于句法自动分析的词类体系与人用来解释语言现象的词类体系也应该是衔接的，因此，对现有的体系不宜完全推倒，另起炉灶。应该继承已有的成果。根据语法功能分布的原则，建立了一个面向语言信息处理的现代汉语词语分类体系。其中基本词类有 18 类：名词 n，时间词 t，处所词 s，方位词 f，数词 m，量词 q，区别词 b，代词 r，动词 v，形容词 a，状态词 z，副词 d，介词 p，连词 c，助词 u，语气词 y，叹词 e，拟声词 o，附加类别有 7 类：成语 i，习用语 l，简称 j，前接成分 h，后接成分 k，语素 g，非语素字 x。

(2) 将词语的语法功能分布，特别是语法功能的优势分布作为词语分类的依据，无论对句法自动分析还是对人研究语法都是有价值的。属于同一类的词语有相当多的共同的语法功能。朱德熙先生认为，“名词在一定条件下可以做谓语”，但“体词的主要功能是作主语、宾语，一般不作谓语”；虽然“动词和形容词可以做主宾语”，但“谓词的主要功能是作谓语”[10]。朱先生一方面揭示了汉语词类的多功能现象，另一方面也指出了多功能并非均匀分布。在没有条件对大规模语料进行统计的年代，这种对词类的语法功能优势分布的把握是独具匠心的。

(3) 由于这个词类体系主要是供计算机自动分析和自动生成汉语句子的，分得较细，但考虑到机器与人的衔接，基本词类又不能太多。现在的基本词类有 18 类，其中大多数词类已在此前的语法著作中见到过。新增加的“区别词”和“状态词”是“词组本位”语法体系在 80 年代后期和 90 年代初期提出的。为了适应机器自动处理的需要，又将“时间词”、“处所词”和“方位词”从名词中分化出来（与名词不同，这 3 类词的共同特点之一是可以直接作状语）。还从中文文本的实际出发，增加了 7 个附加类别。

(4) 建立一个科学的分类体系固然不容易，但更艰巨的更细致的任务是要将数以万计的词语归类。《语法信息词典》完成了 7.3 万多词语的归类。

(5) 归类的实践证明，汉语中确实有一部分词处于两个不同类词的交集中。对于这种“兼类现象”，存在不同的处理策略。至于具体采用哪一种处理策略，既要顾及理论的完善，又要考虑尽可能减少词典信息的冗余量。

(6) 对“词类”不宜寄托过多的期望。由于归入同一类的词语仍有许多互相区别的语法特点，对属于同一类的词语进一步采用属性描述是必要的，也是容易操作的。这正是《语法信息词典》的重心所在。

(7) 不能因为同类词语没有某个层次上的一致的语法功能，就否认现在的分类体系的价值。仅从压缩词典的信息冗余量考虑，词类的作用也是很大的。例如，在名词库中就不必考虑“能否作补语”、“能否受程度副词修饰”等属性。

6. 关于“兼类”的基本处理策略

汉语的词有两种“兼类”现象：广义兼类和狭义兼类。

指物件的名词“锁”与指动作的动词“锁”，指职务的名词“编辑”与指行为的动词“编辑”，说它们各是两个不同的词，或者适应计算机处理的需要说它们在广义上兼属名词与动词，都没有争议，类似的，指药品的名词“毒”与指性质的形容词“毒”，指物件的名词“机械”与指性质的形容词“机械”也是广义兼类现象。本文要讨论的不是广义兼类问题。

“热爱集体”中的“集体”与“集体参加”中的“集体”，“突出重点”中的“重点”与“重点讨论”中的“重点”，意义没有什么差别或者说一般人很难把它们的差别表述清楚，但语法功能显然不同，《语法信息词典》将它们处理成兼属名词和副词，相对于广义兼类，这种词义没有明显区别而语法功能却显著不同的兼类现象叫做狭义兼类。本文主要讨论这种兼类现象及其处理策略。

首先需要明确“兼类”的含义。语言学家在建立其词类体系时，主观上都期望划归同一类的词的语法功能要有足够多的共同点，同时不同的词类又要有足够多的不同点。然而当建立了 A 类词与 B 类词的鉴别准则之后，便从词的全集中划分出集合 A 和集合 B，很可能有或多或少的一部分词既符合 A 类词的鉴别准则，又符合 B 类词的鉴别准则，这部分词构成集合 A 与集合 B 的交集 C。集合 A-C，集合 B-C 仍然各是 A 类词，B 类词。在现有的若干种汉语词类体系中都承认这种现象的存在，但如何处理 C 就有不同的见解。

当发现兼类现象并将兼类现象抽象为集合 A，集合 B 及其交集 C 之后，为了确定交集 C 的归属，就需要重新调整词类体系。设新的词类集合用小写字母 a, b, c 表示。处理交集 C 的策略只有如下 3 种。

(i) a=A-C, b=B-C, c=C;

(ii) a=A, b=B;

(iii) a=A, b=B-C; 或 a=A-C, b=B;

当集合 A，集合 B 影射到具体的词类后，只能选择其中的一种策略，至于到底选择哪一种策略，一般说来，这只是怎样处理更恰当的问题，不存在正确或错误的区分。

采用第(i)种策略，就是要将交集 C 另立一类，也要另取一个名字，当一个词类体系相对稳定之后，通常不愿这么做，《语法信息词典》也没有这么处理。

采用第(ii)种策略，就是承认交集 C 中的词既是 A 类词，又是 B 类词，即交集 C 中的词都是兼类词。上述处理名词与副词的交集的办法就是基于这种策略，《语法信息词典》对区别词与副词的交集、形容词与动词的交集也采用了第(ii)种策略。之所以对这些情况采用第(ii)种策略，是因为在句法结构中区分这些兼类词的根据是明显的，相对容易。以形容词与动词的交集为例。“繁荣市场”中的“繁荣”是动词，因为它带了宾语，且不能受“很”修饰；“市场（很）繁荣”中的“繁荣”是形容词，因为它不能带宾语，且可以受“很”修饰。一个词可能兼属两类词，即有两种不同的词性，这是指机器词典中的静态情况。但在文本中，对其进行标注时，就需要根据上下文环境，确定它的唯一的词性。

采用第(iii)种策略，就是只把交集中的词划归两类中的某一类，而从另一类中剔除。《语

法信息词典》处理名词与处所词的交集中的“中国”、“教育部”等词就是采用了这种策略，即只将它们划归名词。由于人们常把处所词看作名词的一个子集，对处所词的关注不够多，这种处理没有引起争议。由于“中国”、“教育部”等词又可以当处所词用，为了不丢失这个信息，在名词库中设立了“处所”字段。

不同的意见来自对动词与名词的交集、动词与副词的交集、形容词与名词的交集、形容词与副词的交集的处理。在北大的两项大规模语言工程中，都采用了第(iii)种策略。笔者拟另行撰文阐述采用第(iii)种策略的原由。

7. 结语

自动分析自然语言所碰到的困难主要在于不确定性。同形词带来了不确定性。一词多类或同一词类具有多功能同样增加了不确定性。仅在词类划分这个层次上处理“兼类”，无论采取何种策略，都没能减少不确定性。减少不确定性的办法只有向计算机灌输更多的知识。《现代汉语语法信息词典》和大规模语料库的标注正是在这个方向上所做的努力。当然，仅有语法知识还是不够的，还要增加语义知识和语用知识。

参考文献

[1] 朱学锋、俞士汶：“自然语言处理与语言知识库”，见罗振声、袁毓林主编，《计算机时代的汉语汉字研究》，清华大学出版社，1996 年。

[2] 俞士汶、段慧明、朱学锋：“综合型汉语知识库及其在汉语教学中的应用”，第四届全球华人教育资讯科技大会主题报告，2000 年 5 月，新加坡，GCCCE2000 Proceedings。

[3] 俞士汶、朱学锋、王惠、张芸芸“《现代汉语语法信息词典详解》”，北京：清华大学出版社，1998 年。

[4] 俞士汶、朱学锋、王惠：“《现代汉语语法信息词典》的新进展”，《中文信息学报》，2001 年第 1 期。

[5] 段慧明、松井久仁於、徐国伟、胡国昕、俞士汶：“大规模汉语标注语料库的制作与使用”，《语言文字应用》，2000 年 2 期。

[6] 中国社会科学院语言研究所词典编辑室：《现代汉语词典》（修订本），北京，商务印书馆，1997。

[7] 中国国家标准 GB13715《信息处理用现代汉语分词规范》，见刘源等著《信息处理用现代汉语分词规范及自动分词方法》，北京：清华大学出版社，第 1 版，1994 年。

[8] 俞士汶主编：《现代汉语语料库加工——词语切分与词性标注规范与手册》，北京大学计算语言学研究所，1999 年 4 月。

[9] 俞士汶、朱学锋、段慧明：“大规模现代汉语标注语料库的加工规范”，《中文信息学报》，2000 年，第 6 期。

[10] 朱德熙：“《语法讲义》”，北京，商务印书馆，1983。

[11] 胡明扬：“现代汉语词类问题考察”见：胡明扬主编《词类问题考察》，北京，北京语言学院出版社，1996 年。

[12] 陈小荷：“从自动句法分析角度看汉语词类问题”，北京语言文化大学（感谢陈博士在此稿尚未发表之前提供给了我们参考），1998。