# 语料库中语言知识的标记问题

李宇明

教育部 语言文字信息管理司,北京

**摘要** 语料库具有五种基本性质,知识标记是语料库的重要内容,其中语言知识标记是语料库深加工的核心内容。应广泛搜集已有的语言学文献,将其数字化为语言知识数据库,以大力支持语料库的语言知识的标记,迅速提升计算机的语言智慧。应有计划地开发基于语料库的语言研究软件,方便学者利用语料库,推进语言研究手段的现代化,促进"基于统计"的和"基于规则"的两种语言研究路向的交流与合作。

**关键词** 语料库;语言知识;标记

在信息时代快步走来之时,必须明确这样的观念:语料库是语言信息处理必不可少的基础工程,是当今语言研究和语言应用研究的重要工作平台,是语言知识及其他知识提取的十分重要的工具。

多年来,语料库问题已广受关注,不断有语料库建设的新计划出台;已有多个语料库投入使用并得到继续开发;语料库语言学已有不少研究者,并且发表了专门研究语料库的博士论文;国家语委正立项研究语料库的规范问题。尽管如此,无论是理论上还是实践上,我国的语料库建设都还处在初级阶段,许多问题都待深入研究。例如:建设哪些类型的语料库才能满足学术和应用的需求;怎样为特殊语料库(比如古代汉语语料库,少数民族语料库等)的建设创造必要的条件;对库中的语料如何深加工,语料库的建设规范和加工规范,语料库成果的最大限度的共享,语料库知识产权问题等等。

本文讨论三个问题:①语料库的性质;②语言知识标记在语料库建设中的地位;③语言知识的获取。这些讨论都与语料库的深加工有关。

## 1 语料库的性质

认识语言知识及其标记在语料库建设中的地位,需要先定义语料库。从文后的参考文献可以窥见,语料库研究已有不少成果。根据这些研究可把语料库定义为:语料库是指机器可以处理的有一定规模、结构和知识标记的自然话语材料的集合。

这一定义从五方面对语料库进行了界定或描述,这五方面其实也可以看做现代意义上的语料库的五种性质。

### 1.1 语料库可以用机器进行处理

"处理"在这里应当理解为三种含义:①利用机器自动管理语料库;②机器可以对语料进

行检索、统计等;③对语料进行自动或半自动的加工等。过去也曾有过专书引得,或是为词典编纂、语言研究等而做的语料卡片库,但它们不能用机器处理,不是现代意义上的语料库。

有学者将语料库的这一特性概括为"可以机读"。这一概括很精要,其着眼点主要在上述的含义②。"可以机读"的确是现代意义上的语料库区别过去"卡片库"等的主要特点,用机器对语料进行检索统计也的确是现代语料库的主要功能,但是,从语料库管理的宏观角度进行思考,从计算机自学习功能的开发和语料库的自动加工等当前人们努力的方向考虑,在"可以机读"的基础上提出"可以用机器进行处理"及"处理"的上述三种具体含义,应该是有价值的。

## 1.2 语料是自然话语材料

自然话语指的是处于自然状态的人类的自然语言材料,是真实存在过的。是自然话语材料,就不是人工语言(如 C++、BASIC 等的计算机高级程序设计语言),也不是经过语料收录者修改润色的语言。自然话语材料具有真实性,语料库不仅收录合乎语言规范的语料,而且也应包含语言运用中的各种失误。语料的真实性是语料库精神之所在,价值之所在。对原始语料修改润色,哪怕是对原始语料中不规范现象的修正,都会损害语料的真实性,都可能降低语料库的应用价值。例如某大型语料库,由于建库时字库技术的限制而将原始语料中的繁体字更换为简体字。就是这一技术上的语料修改,使得该语料库失去了研究 20 世纪汉字简繁演变情况的功用,且影响到字种使用情况的考察,甚至还可能影响到字频、词频的统计结果。

再如,现在可以通过网络较为方便地下载数以亿计的语料。这些语料的原始载体(首次发表时的媒体形式),有的就是电子媒体(包括网络),如网络新闻、网络小说、网络短信、BBS 话语等,但是多数是纸媒的,网络只是它的"迁移载体"。就实际状况而言,语料从原始载体向迁移载体转录的过程中,有许多都没有经过严格的校对,技术误差往往超过国家的有关标准。利用这种"失真"语料建立的语料库,做些粗放的考察还过得去,但却不很适合进行语言的精细研究。

话语都是在一定交际领域中使用的,因而有不同的使用特点。不同使用领域、不同使用特点的话语,会形成不同用途的语料库。书面语语料形成书面语料库;口语语料形成口语语料库;某一专业领域的语料形成某种专业语料库;多领域的平衡性较好的语料形成平衡语料库。当然,如果考虑到话语的时间特征,还可以区分出共时语料库和历时语料库。话语的使用领域、使用特点以及使用时间,具有为语料库分类的功能。

## 1.3 语料库应具有一定的规模

语料库的规模,指的是语料库收录语料的数量大小。语料库以统计见长,没有数量规模就不成其为语料库,就不能扬语料库之所长。当然,语料库的规模并不绝对,不同时期、不同技术条件下对语料库规模的要求并不相同。一般说来,语料库的规模受制于三个因素,或者说是如下三方面的和谐:

(1)语料获取的难度与成本。在网络资源逐渐丰富的条件下,语料搜集的难度与成本大大降低。但是如前所言,由于网络中迁移载体语料的失真,网络下载并不能代替或不能完全代替其他语料采录方式,而且关于语料库的知识产权问题还不明确,如何支付语料作者和语

料原始载体、迁移载体的费用，尚无成文规定。如果需要支付这笔费用，语料库建设的成本还是相当昂贵的。

(2)计算机的存储能力、运算速度、网络运行及其成本。信息技术的日新月异，使得计算机的存储能力、运算速度等有了长足发展，就语料库建设而言这方面已经没有大问题。语料库的价值在于使用，在因特网逐渐普及的现在和未来，利用网络进行语料库服务将成为重要的服务方式。语料库建设不能忽视网络运行的条件、成本等问题。

(3)不同类型语料库的应用需求。不同类型的语料库，规模的大小与功能的发挥(或者说与应用的需求)有一定关系。例如，用于常用字统计、常用词统计、一般语法现象研究的语料库，其规模就可以小些，而用于罕用字统计、罕用词统计、特殊语法现象研究的语料库，因数据稀疏因素的影响，其规模就需要大些。

## 1.4 语料库具有一定的结构

语料库的结构可以分为三种：

(1)存储结构。一定规模的语料，在语料库中需要有一定的存放方式，语料在语料库中的存放方式形成了语料库的存储结构。

(2)内容结构。语料与语料之间会发生一定关联，如时间关联、使用领域关联、作者关联、主题内容关联等。语料之间的关联方式形成语料库的内容结构。

(3)物理结构。广义的语料库结构，还应包括处理语料的物理系统和软件系统，这是语料库的物理结构。

语料库的存储结构和物理结构，是为内容结构服务的。不同功能的语料库，其内容结构的复杂程度有较大差别。一般来说，语料库的功能同内容结构成正比，即内容结构越复杂，语料库所发挥的功能或潜在功能就越大。

## 1.5 语料库具有知识标记

知识标记是标注在语料片段上的具有特定含义的机读符号。知识标记是供计算机阅读的内部符号，根据标记的内容可以将其分为三类：

(1)原始数据标记。对语料的知识版权信息、载体发行信息、采样方式信息等元数据(Metadata)的标记，称为原始数据标记。原始数据标记常与语料库的内容结构相关，关系到语料库的功能定位和功能发挥。有时，语料的收录者、校对者及语料的管理信息等也会记录下来，给以标记，这种标记可以归入广义的原始数据标记。

(2)语言知识标记。语料库中各种语言单位的性质的标记、语言单位间语法关系和语义关系的标记、语言片段的语用标记等等，统称为语言知识标记。语言单位的含义很广，它包括字、词、短语、句子、篇章等等，涉及字的文字学的各种属性，涉及词、短语、句子、篇章等单位的语音(或韵律)、语义、语法、语用等诸多属性，并通过这些属性反映出它们之间的语音(或韵律)、语义、语法、语用等诸方面的相互关系。

当然，语料库的不同发展阶段、语料库的不同用途、语料库建设者对语言知识的不同看法等，对语言知识进行标记的细密程度、广窄范围会不相同或很不相同。况且，为语料库标注语言知识，是需要数代人完成的非常浩大的工程，绝不可能一蹴而就。

(3)其他知识标记(简称 c 类标记)。语言与知识的关系是辩证的，一方面，语言文字是

人类知识的最为重要的载体，另一方面，语言的运用也离不开各种知识，包括各种常识和与具体交际相关的专业知识。因此，计算机能够处理语言，只有语言知识是不够的，还必须有相关的常识和专业知识。刘开瑛(2000)指出："在信息领域中 80％以上的信息是以语言文字为载体的。"这就是说，语料库其实也是人类的知识库，是知识发掘的重要操作对象。正因如此，语料库也需要对语料标注 c 类标记。当然，利用语料库进行知识挖掘的学术研究刚刚开始，怎样在语料上标注 c 类标记，还没有相对成熟的经验。

## 2 语言知识标记在语料库建设中的地位

知识标记是语料库的五种性质之一，但却是语料库开发深度及其应用价值的重要标志，语料库的深度开发，主要体现在为语料库标注各种知识标记。不同类型的知识标记，意味着语料库建设的不同水平或不同用途。例如，原始数据标记是语料的外部标记，只具有原始数据的语料库是"生语料库"，生语料库可以进行语言单位的简单查询与统计。生语料库查询与统计功能的大小，取决于原始数据标记的类型的多少。与原始数据标记不同，语言知识标记是语料的内部标记，代表着计算机对语言的理解水平。具有语言知识标记的语料库是"熟语料库"，熟语料库可以进行语言的多领域的复杂查询与统计，而且用这种语料库武装的计算机，会拥有一定的语言智能。c 类标记也是语料的内部标记，这种知识标记与语言知识标记相互结合，将大大提高语料库的使用价值，使计算机拥有更多的语言智能和其他智能。可以这样说，语料库所拥有知识标记类型的多少，对知识标注的精度和深度，决定了语料库的功能大小和价值大小。因此，知识标记历来都是语料库建设的核心问题。

尽管各种知识标记对于语料库都有重要意义，但是就当前语料库建设的实际和学术的发展状况来看，语言知识的标记是需要着重解决的问题。这是因为：

(1)原始数据标记体现着语料库建设的基本规范。当前语料库建设普遍存在忽视原始数据标记以及原始标记不规范等问题，带来了语料库功能不能充分发挥、不能很好共享使用等多种弊端。但是，对语料进行原始数据标记，并没有科学难度，只是对原始数据标记是否重视、数据类型的设计是否周全、语料选取及工程实施能否很好实现设计方案、选用什么样的符号进行标记、标记方式能否具有最大的共享性等问题。但是，标记语言知识的难度却超乎寻常，这种难度主要还不是技术和成本之类的问题，而是科学难题。

标记语言知识的科学难题主要在两方面：①当前的语言研究，特别是语义研究和语用研究，还无法满足机器的语言理解需要。②对语言知识如何标记，还在探索之中。例如，英语已经用树库的方式标记了一些语料，如 Penn 树库等。汉语也进行了一些树库尝试，如清华大学的汉语测试树库、美国宾州大学的 UPenn 树库、台湾中研院的树库项目等。但是，树库对汉语的效用是否像对英语的效用那样好，是值得划一个问号的。再如，语义特征和语义关系异常复杂，而且不同的语言学流派或语言学视野(如传统词汇学、配价语法、格语法、语义特征分析、语义指向分析等)又有各自的语义学框架，如何整合各家学说科学有效地标记语义？当然，人们对标记语用知识的探索更少，怎样在语料库中标记语用知识，所知也更少。

(2)语料库也是一种数据库，但不是一般的数据库。只具有原始数据标记和 c 类标记的语料库，不能算是真正的语料库。语料库具备了语言知识标记，方能从中获取关于语言的各种信息，方能使计算机理解语言。语言知识标记在语料库中占有中心地位。

(3)当前我国的语料库建设,正处在给语料标注各种语言知识的关键时期。词语切分有了较大进展,并逐步取得共识;编制了一些有实用价值的电子词典;按照一定的语法体系为切分单位标注了词性;探讨了消除歧义等语义理解的若干问题;提出了汉语语块的标注体系等。因此,就当前语料库建设的实际和学术的发展状况来说,语言知识标记问题是大家共同关心且需要急切解决的问题。

## 3 语言知识的获取

### 3.1 语言知识获取的两种途径

标记语言知识,本质上依赖标注者所具有的语言知识,因此,语言知识的获取是本源上的问题。语料库出现之后,语言知识的获取有两条基本途径:基于语料库的(A 途径)和 B 非基于语料库的(B 途径)。把是否利用语料库作为语言知识获取途径的划分标准,是因为通过语料库获取语言知识有重要的特点和广阔的发展前景:

其一,A 途径的基本研究方法是统计分析,有人称之为"基于统计"的。其实严格地说,统计并不是完全自足的,统计的设计需要已有的语言学知识支撑,统计的结果需要进行语言学分析,而且 B 途径也早已运用统计分析方法。可以肯定地说,统计并不是 A 途径的专利。但是毋庸赘言,A 途径的确将统计的作用发挥到了极为充分的地步,它所发现的语言学规律,基本上是由统计数据支持的。

其二,A 途径所处理的语言材料,不仅规模巨大,而且由于语料库是真实文本的集合,语料中包含有现有规范之外的语言现象。这种从海量真实文本中通过统计获得的数据,必然会有一些不同于 B 途径所获得的语言知识,因为 B 途径处理的语言材料不仅在规模上无法同语料库相比,而且它所处理的语料也往往是经过选择的,较为规范的。

其三,A 途径对语言知识的表述方式,便于机器阅读。

B 途径通过仔细观察语言现象来获取语言规律,定性研究多于定量研究,与 A 途径比较多地利用统计分析相比,有人将其称为"基于规则"的或"基于内省"的。B 途径这种非基于语料库的语言研究,已经有几千年的传统,取得了令许多人文科学艳羡的丰硕成果,正是这些研究,奠定了当今语言信息处理的基础。然而非常遗憾的是,这些语言学知识,并没有在我国的语料库建设中发挥应有的作用,其原因大致有以下几点:

(1)B 途径所获取的语言学成果比较分散,缺乏必要的积聚整合。例如现代汉语语法,其成果主要储存在语言学论文中,至今没有反映整个现代汉语语法研究成果的"语法长编"。现代汉语的语法论著和教科书,主要提供的是现代汉语的语法体系或基本架构,并不能起到"语法长编"的作用。因此,任何人都难以在短时间内囊括这些成果并将其用于语料标注。

(2)以往语言学成果的表述方式,适合"人读"而不适合"机读"。将适合人读的语言学知识转换为适合机读的,需要具有语言学和计算机科学双重素养者的创造性的工作。当前,具有这种双重素养的学者还不多,愿意进行这种转换工作的人也不多。

(3)有些语料库的建设者,对 B 途径所获取的语言知识并不重视,以为"基于统计"的研究可以获得足够的语言知识,而"基于规则"的研究并不能对机器的语言处理提供太多的帮助,起码是在当前不能提供太多的帮助。这种认识也许可成一说,但很可能是偏颇的。A、B

两种获取语言知识的途径，以及这两种途径所得到的语言知识，应该是相辅相成的，相得益彰的。

## 3.2 已有语言学知识的“计算机化”

标记语言知识，就本源上说，依赖标注者的语言知识，但是，也受制于语言知识所能“计算机化”的程度。当前，语料库深加工的最重要的方略，亦即标记语言知识最重要的方略，笔者以为是整理B途径所获取的汗牛充栋的成果，使之尽快“计算机化”。这项工作的关键是建立语言知识数据库。语言知识数据库的建设及其开发利用，大约需要做如下一些工作：

(1)广泛搜罗相关的语言学文献，分学科汇总编目，录入计算机，形成语言学文献目录库。

(2)依照国家法律或国际通例，妥善处理即将入库文献的知识产权问题。

(3)将搜罗到的语言学文献数字化，由纸媒载体转化为可以进行数据加工的数字载体。在文献的载体转化过程中，既要充分发挥计算机的功能，又要严格校对，确保质量。

(4)将数字化的语言学文献入库，配之以合适的管理软件，形成语言学文献数据库。

(5)设计合适的语言学框架，为梳理语言知识作准备。这种语言学框架的主要特点在于巨大的包容性，保障文献中的所有语言学知识都能找到一个存放位置。

(6)依据设计的语言学框架，编制知识点目录及其相应的标记符号清单。

(7)运用已订的标记符号将语言学文献数据库中的全部知识点进行标注，分类入库。标注入库要尽量利用计算机，最好是研发专用的人机互助软件，或以人为主机器辅助，或以机器为主人来辅助。

(8)对入库的各知识点进行科学梳理，形成语言学知识库。

(9)对语言学知识库中各知识点的内容“计算机化”，形成便于机读的语言学知识，并将计算机化了的语言学知识加载进语料库，形成语料库的带有标记的语言知识，从而生成计算机的语言智慧。

相信，包括语言学文献目录库、语言学文献数据库和语言学知识库的语言学知识数据库的建立，及其数据的计算机化，一定会大幅度提高语料库的语言知识标记水平，促使计算机语言处理上升一个新阶段。

当然，语言学知识数据库不只是对语料库深加工具有重大作用，而且也会促成语言学的大跨越，因为这一数据库的建立，将分散的语言学知识进行了积聚整合，对今后的研究和对语言学研究者的培养，提供了全面的文献支持和坚实的知识平台，并可以利用计算机方便地获得这一支持和利用这一平台。

## 3.3 提倡利用语料库的语言学研究

A、B两种语言学知识的获取途径，在一段时期内都会有人采用。但是，随着语料库的逐渐完善和基于语料库的语言研究软件的不断开发，越来越多的人会加入到利用语料库进行语言研究的行列。因为：

(1)语料库为观察和分析语料提供了方便。语料库所提供的海量真实语料，比人工查阅语料丰富快捷，比内省的语料全面具体，而且可以利用软件较快地进行统计分析。就语料的搜集、观察而言，语料库具有极大的优越性。

(2)语言学不仅需要定性地描写规律,而且也需要对各种语言单位的数量、分布、组合等获取统计数据,通过定量研究来认识语言。对语言进行定量的统计研究,获取语言的各种统计信息,语料库肯定是较为理想的工具。

(3)语言学研究的目的是多方面的,其中的一个重要方面,就是促进语言的信息化。促进语言的信息化需要利用语料库,需要对语料库进行语言学的深加工。

(4)利用语料库进行语言研究不仅有如上便利,而且一定会发现语言的新特点、新规律,可能产生语言学的新方法、新流派。

这些学术方便和学术魅力,一定会吸引更多的语言研究者利用语料库。为方便利用语料库的语言学研究,应当有计划地开发机助的语言研究软件。这些软件起码包括语料检索软件、语料自动下载归档软件、统计分析软件、结果检验软件等。语言学知识数据库、标记了诸多语言知识的语料库和系统的语言研究软件等"三大法宝",足以使语言研究手段现代化。用这些现代化手段装备语言学家,一定会使语言研究获得长足进展。利用语料库(和其他数据库)从事语言研究,将逐渐成为语言知识获取的主要途径,其成果也较易计算机化。所谓"基于统计"的和"基于规则"的两类研究,也肯定将逐渐接近,相互借鉴,相互促进,最终合而为一。

**参考文献**


1 曹右琦.辉煌二十年——中国中文信息学会二十周年学术会议.北京:清华大学出版社,2001

2 陈群秀.一个在线义类词库:词网 WordNet.1998 语言文字应用,1998(2):93~99

3 陈小荷.现代汉语自动分析——Visual C++实现.北京:北京语言文化大学出版社,2000

4 陈玉泉,陈宣,陆汝占.内涵时态逻辑的语义解释系统.见:自然语言理解与机器翻译.北京:清华大学出版社,2001.38~44

5 段慧明,松井久仁於,徐国伟等.大规模汉语标注语料库的制作与使用.语言文字应用,2000(2):72~77

6 董振东.汉语分词研究漫谈.语言文字应用,1997(1):107~112

7 董振东.语义关系的表达和知识系统的建造.语言文字应用,1998(3):76~82

8 冯志伟.自然语言处理中的歧义消解分析.语言文字应用,1996(1):55~60

9 桂诗春,宁春岩.语言学方法论.北京:外语教学与研究出版社,1997

10 国家技术监督局.中华人民共和国国家标准 GB/T 13715-92《信息处理用现代汉语分词规范》.北京:中国标准出版社,1993

11 何婷婷.语料库研究:[博士学位论文].武汉:华中师范大学,2003

12 侯敏.计算语言学与汉语自动分析.北京:北京广播学院出版社,1999

13 胡明扬.现代汉语通用语料库的建库原则和设想.语言文字应用,1992(3):49~56

14 黄昌宁.关于处理大规模真实文本的谈话.语言文字应用,1993(2):1~10

15 黄昌宁.统计语言模型能做什么.语言文字应用,2002(1):77~84

16 黄昌宁,童翔.汉语真实文本的语义自动标注.语言文字应用,1993(4):18~25

17 黄昌宁,张普主编.自然语言理解与机器翻译.北京:清华大学出版社,2001

18 黄昌宁,李涓子.《语料库语言学》,北京:商务印书馆,2002

19 黄曾阳.HNC(概念层次网络)理论——计算机理解语言研究的新思路.北京:清华大学出版社,1998

20 靳光瑾.计算机理解汉语需要语法理论支持.语言文字应用,1998(2):82~87

21 亢世勇,常宝宝,刘海涓等.全信息标注语料库的开发与应用.见:辉煌二十年——中国中文信息学会

二十周年学术会议文集.北京:清华大学出版社,2001.125～130
22 亢世勇,刘海涧.基于数据库的现代汉语词类优势语法功能统计研究.见:辉煌二十年——中国中文信息学会二十周年学术会议文集.北京:清华大学出版社,2001.166～170
23 刘开瑛.汉语自动分词评测技术研究.语言文字应用,1997(1):101～106
24 刘开瑛.中文文本自动分词和标注.北京:商务印书馆,2000
25 刘连元.现代汉语语料库研制.语言文字应用,1996(3):2～8
26 鲁川.汉语语法的意合网络.北京:商务印书馆,2001
27 马大猷.语言信息和语言通信.北京:知识出版社,1987
28 盛玉麒.信息网络时代中日韩语文现代化国际学术研讨会论文集.香港:香港文化教育出版社,2000
29 孙宏林.信息处理用汉语语义词典的描述方法.语言文字应用,1993(1):56～61
30 孙茂松.汉语自动分词研究的若干最新进展——清华大学相关工作简介.见:辉煌二十年——中国中文信息学会二十周年学术会议文集.北京:清华大学出版社,2001.20～41
31 孙茂松,王洪君,李行健等.信息处理用现代汉语分词词表.语言文字应用,2001(4):84～89
32 王惠,詹卫东,俞士汶.现代汉语语义词典规格说明书.汉语语言与计算学报,2003(6):159～176
33 熊澄宇.信息社会 4.0——中国社会建构新对策.长沙:湖南人民出版社,2002
34 许嘉璐.现状与设想——试论中文信息处理与现代汉语研究.中国语文,2000(6)
35 尹斌庸,方世增.词频统计的新概念和新方法.语言文字应用,1994(2):69～75
36 于江生,刘扬,俞士汶.中文概念词典规格说明书.汉语语言与计算学报,2003(6):177～194
37 俞士汶.语法知识在语言信息处理研究中的作用.语言文字应用,1997(4):81～87
38 俞士汶,朱学锋.关于汉语信息处理的认识及其研究方略.语言文字应用,2002(2):51～58
39 俞士汶.北京大学语言知识库概况.汉语语言与计算学报,2003(6):119～120
40 俞士汶,段慧明,朱学锋等.北大语料库加工规范:切分·词性标注.汉语语言与计算学报,2003(6):121～158
41 语用所计算语言学室.信息处理用现代汉语词类标记集规范.语言文字应用,2001(3):16～20
42 约翰·辛克莱.关于语料库的建立.王建华译.语言文字应用,2000(2):63～71
43 袁毓林.语言的认知研究和计算分析.语言文字应用,1996(1):61～67
44 斋藤秀纪.日语语料库的构建.见:信息网络时代中日韩语文现代化国际学术研讨会论文集.香港:香港文化教育出版社,2000.256～261
45 张普.关于大规模真实文本语料库的几点理论思考.语言文字应用,1999(1):34～43
46 张普.信息处理用动态语言知识更新的总体思考.语言文字应用,2000(2):42～49
47 张普.关于汉语语料库的建设与发展问题的思考.电子文本.2003
48 周强,詹卫东,任海波.构建大规模的汉语语块库.见:自然语言理解与机器翻译.北京:清华大学出版社,2001.102～107
49 周强.汉语句法知识的自动获取研究.见:曹右琦.2001.156～165

# 语料库中语言知识的标记问题

作者：李宇明
作者单位：教育部语言文字信息管理司,北京