北大语料库加工规范：切分·词性标注·注音

俞士汶 段慧明 朱学锋 孙斌 常宝宝

北京大学计算语言学研究所

(2003 年 5 月 14 日)

摘要： 北京大学计算语言学研究所为研制 2600 多万字《人民日报》基本标注语料库制订了词语切分·词性标注规范（这里简称为《规范 2001》，以《北京大学现代汉语语料库基本加工规范》为题，已经发表在《中文信息学报》2002 年第 5 期和第 6 期）。为研制 100 万字的注音语料库又制订了注音规范。这里结合《规范 2001》和注音规范，并增加若干词性标记，制订了新的《北大语料库加工规范：切分·词性标注·注音》（简称《规范 2003》），现在标记集包含 100 多个标记。遵循《规范 2003》，北京大学计算语言学研究所将继续研制新的高质量的语料库，并将加工深度逐步向前推进。

关键词：现代汉语；语料库；词语切分；词性标注；注音，规范

Specification for Corpus Processing at Peking University: Word Segmentation, POS Tagging and Phonetic Notation

YU Shiwen DUAN Huiming ZHU Xuefeng

(Institute of Computational Linguistics, Peking University, Beijing, 100871)

Abstract: The Institute of Computational Linguistics, Peking University made a specification for the word segmentation and POS tagging of its People's Daily corpus (over 26 million Chinese characters) [hereinafter: Specification 2001, which was published in the *Journal of Chinese Information Processing* (Issue 5 & Issued 6, 2002), entitled *The Basic Processing of Contemporary Chinese Corpus at Peking University – Specification*]. In addition, another specification was made for building the phonetically annotated corpus (1 million Chinese characters). Based these two specifications, we hereby present the latest *Specification for Corpus Processing at Peking University: Word Segmentation, POS Tagging and Phonetic Notation* [hereinafter: Specification 2003]. With the newly added ones, the tagset now includes more than 100 tags. Following Specification 2003, the Institute of Computational Linguistics will go on with more corpora of high quality and in-depth processing.

Keyword: Contemporary Chinese; Corpus; Word Segmentation; POS Tagging; Phonetic Notation; Specification

---

*本文有关研究得到 973 项目（G1998030504-01,G1998030507-4）、863 项目（2001AA114010,2001AA114040）和国家自然科学基金（69973005）的支持。

1. 前言

北京大学计算语言学研究所（以下简称北大计算语言所）从 1992 年起开始研究现代汉语语料库的多级加工，历时已有 10 余载，已取得的重要成果首推自 1999 年 4 月至 2002 年 4 月历时 3 年完成的 1998 年全年《人民日报》的标注语料库。该语料库包含 2600 多万汉字，对全部语料已完成词语切分和词性标注等基本加工。该项成果通过了合作单位 Fujitsu 的验收。其中 1 月份的近 200 万字的语料已在网上（www.icl.pku.edu.cn 或 icl.pku.edu.cn）公布，截至 2003 年 5 月 6 日，下载人次已达 1018。上半年的 1300 万字的语料正通过人民日报社新闻信息中心向业界转让许可使用权。此外，北大计算语言所在 2000 年完成了另外 100 万字语料的加工任务，除了词语切分和词性标注之外，还注上了汉语拼音。尽管北大计算语言所为保障加工质量倾注了智慧和诚实的劳动，却不敢妄言没有瑕疵。北大计算语言所真诚希望，用户经常反馈他们的发现、意见或疑惑。研制者一定虚心接受批评、指正，决不讳疾忌医。研制者愿意同广大用户一起努力，共同创造出有实用价值的现代汉语语料库，为语言信息处理的发展和现代汉语文化建设贡献一份力量。

语料库的加工离不开详细的、严谨的加工规范的指导。《人民日报》语料库的加工规范是 1999 年 3 月制订、2001 年 7 月修订的《现代汉语语料库加工规范——词语切分与词性标注》（以下简称为《规范 2001》）。该规范以《北京大学现代汉语语料库基本加工规范》[1]为题，已经 在《中文信息学报》2002 年第 5 期和第 6 期全文发表。在对另外选取的 100 万字语料加工的时候，又制定了注音规范。

由于业界对大规模标注语料库的需求日益旺盛，也由于国家语委制定的语言文字应用“十五”科研规划和于 2001 年秋季公布的国家 863 计划的项目指南都把语料库建设放在相当重要的地位，国家 973 计划倡议成立的“中国语言资源联盟”（ChineseLDC, Chinese Linguistic Data Consortium 之缩写）也在支持加工语料库的建设。预计今后或许会出现一个语料库开发的热潮。当然，北大计算语言所关于语料库的研究工作也在不断前进，加工规模不断扩大，加工深度越来越深，质量越来越好。随着研究工作的深入，对《规范 2001》的认识也在逐步提高。如果调整其中的一些切分、标注规则，并扩充标记集，就可以把语料加工得更精细，加工后的语料库就可以提供更丰富的语言知识，因此，对《规范 2001》又做了一次较大的修订，并与注音规范结合在一起，形成这里的《北大语料库加工规范：切分·词性标注·注音》，简称为《规范 2003》，下面行文中，有时也称“本规范”。除了注音，《规范 2003》同《规范 2001》的最大差别在于，《规范 2003》大幅度地扩充了标记集，词性标记总数由 40 个左右增加到*** 。

北大计算语言所正依据《规范 2003》对一部份语料进行试加工。主要目的在于探索基于《现代汉语语法信息词典》（简称《语法信息词典》）[2]把依据《规范 2001》加工的语料自动转换到依据《规范 2003》的格式的可能性。以后将完成按《规范 2001》加工的语料的再加工并扩充其他语料。作者在这里发表《规范 2003》，和发表《规范 2001》的初衷一样，主要是期望引起对语料库加工规范这个问题的讨论和争鸣。

《规范 2003》共分 9 章。第 1 章是前言。第 2 章借用一个例子说明语料要加工成什么样。第 3 章介绍制订《规范 2003》的基本思路。第 4 章是切分规范。第 5 章是切分和标注相结合的规范。第 6 章是词性标注规范。第 7 章是注音规范。第 8 章是加工后语料的形式化描述。第 9 章是结束语，对有关人员致以谢意。最后还有参考文献和关于标记集的附录。

2. 关于加工任务的说明

2.1 关于原始语料

1998 年全年《人民日报》语料库包含的文章都是完整的。 同《人民日报》语料库不同，服务于汉语信息处理和汉语语法研究（包括词汇、句法、语义）所选取的一些原始语料基本上以句子为单位，不刻意保证原始语料内容、知识、信息的完整性，其中一部分语料是作者或北大计算语言所自身的创作与积累。

今后，将根据研制目的、加工深度、版权等因素，综合考虑原始语料的选择。如果有条件完成对《人民日报》逐年语料的加工，其价值是不言而喻的。

2.2 关于语料的基本加工

汉语语料库的基本加工通常指词语切分与词性标注两项内容。依据《规范 2001》的 1998 年《人民日报》标注语料库的加工项目多于这两项内容，还包括专有名词（人名、地名、团体机构名称等）标注、语素子类标注、动词和形容词的特殊用法标注、短语型名称的标注等，其标记总数约 40 个。下面摘录一段原始语料如下：

咱们中国这么大的一个多民族的国家如果不团结，就不可能发展经济，人民生活水平也就不可能得到改善和提高。

加工后的语料如下所示：

咱们/r 中国/ns 这么/r 大/a 的/u 一个/m 多/a 民族/n 的/u 国家/n 如果/c 不/d 团结/a ，/w 就/d 不/d 可能/v 发展/v 经济/n ，/w 人民/n 生活/n 水平/n 也/d 就/d 不/d 可能/v 得到/v 改善/vn 和/c 提高/vn 。/w

词语之间有了空格，斜杠之后的字母是该词语的标记，其中包括词性标记（如r, n, v, a, u, m, w 等）、专有名词标记（如 ns）、动词和形容词的特殊用法标记（如 vn ）。不妨将这些标记笼统地称为“词性标记”。关于这些标记的含义请参阅本规范的第4,5,6章或附录。

2.3 关于注音

拼音标注的任务就是在这样的语料上对每一个斜杠前的词语加上汉语拼音，得到如下形式的加工结果：

咱们{zan2men5}/r 中国{zhong1guo2}/ns 这么{zhe4me5}/r 大{da4}/a 的{de5}/u 一个{yi1ge4}/m 多{duo1}/a 民族{min2zu2}/n 的 {de5}/u 国家{guo2jia1}/n 如果{ru2guo3}/c 不{bu4}/d 团结{tuan2jie2}/a ，/w 就{jiu4}/d 不{bu4}/d 可能{ke3neng2}/v 发展{fa1zhan3}/v 经济{jing1ji4}/n ，/w 人民{ren2min2}/n 生活{sheng1huo2}/n 水平{shui3ping2}/n 也{ye3}/d 就{jiu4}/d 不{bu4}/d 可能{ke3neng2}/v 得到{de2dao4}/v 改善{gan3shan4}/vn 和{he2}/c 提高{ti2gao1}/vn 。/w

如此注音，对多音词是必要的。但对单音词，则冗余量太大。只对多音词注音、对单音词不注音（另附一张词表）的标注结果如下：

咱们/r 中国/ns 这么/r 大{da4}/a 的{de5}/u 一个/m 多/a 民族/n 的{de5}/u 国家/n 如果/c 不/d 团结/a ，/w 就/d 不/d 可能/v 发展/v 经济/n ，/w 人民/n 生活/n 水平/n 也/d 就/d 不/d 可能/v 得到/v 改善/vn 和{he2}/c 提高/vn 。/w

另附加一个单音词的词表，基本格式如下：

不	d	bu4
得到	v	de2dao4
多	a	duo1
发展	v	fa1zhan3
改善	vn	gan3shan4
国家	n	guo2jia1
经济	n	jing1ji4
就	d	jiu4
可能	v	ke3neng2
民族	n	min2zu2
人民	n	ren2min2
如果	c	ru2guo3
生活	n	sheng1huo2
水平	n	shui3ping2
提高	vn	ti2gao1
团结	a	tuan2jie2
也	d	ye3
一个	m	yi1ge4
咱们	r	zan2men5
这么	r	zhe4me5
中国	ns	zhong1guo2

实际上，只要这些词语在《现代汉语语法信息词典》中，拼音信息就已经有了。若这些词语是未登录词，则要生成并校对它们的拼音。

依据《规范2003》，加工结果则如下所示（省略仅词性标记有变化的单音词词表）：

咱们/rr 中国/ns 这么/rz 大{da4}/a 的{de5}/ud 一个/mq 多/a 民族/n 的{de5}/ud 国家/n 如果/c 不/df 团结/a ，/wd 就/d 不/df 可能/vu 发展/v 经济/n ，/wd 人民/n 生活/n 水平/n 也/d 就/d 不/df 可能/vu 得到/v 改善/vn 和{he2}/c 提高/vn 。/wj

其中，rr，rz，ud，mq，df，vu，wd，wj 等都是新增加的标记。

3. 制订《规范 2003》的基本思路

《规范 2003》基本上继承《规范 2001》的理论体系和规则。根据新的认识，也有所发展。（1）词语的切分规范尽可能同中国国家标准 GB13715“信息处理用现代汉语分词规范”（以

下简称为“分词规范”）[3]保持一致。因为现在词语切分与词性标注是结合起来进行的，而且又有了一部《现代汉语语法信息词典》可以作为基本参照，所以对“分词规范”作了必要的调整和补充。

（2）便于扩充的标记集。《规范 2001》的词性标注除了使用《语法信息词典》中的 26 个词类代码（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、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、标点符号 w）外，增加了以下 3 类标记：①专有名词的分类标记，即人名 nr，地名 ns，团体机关单位名称 nt，其他专有名词 nz，英语等其他非汉字的字符串 nx。②语素的子类标记，即名语素 Ng，动语素 Vg，形容语素 Ag，时语素 Tg，副语素 Dg 等；由于标注时只使用这些子类标记，故语素标记 g 不在标注语料库中出现。 ③动词和形容词的特殊用法标记，即名动词 vn（动词的名词用法），名形词 an（形容词的名词用法），副动词 vd（动词的副词用法），副形词 ad（形容词的副词用法）。合计约 40 个。这个标记集虽然不算大，但与《现代汉语语法信息词典》结合，它是很容易扩充的。制订《规范 2003》的实践是对这个基本思路合理性的一次验证。实际上，《规范 2003》中的标记集就是《规范 2001》的扩充。如从数词中细分出数量词，例：一个/mq；又如，将成语、习用语和简称细分为名词性的（in/ln/jn）、动词性的（iv/lv/jv）、形容词性的（ia/la/ja）、区别词性的（ib/lb/jb）、副词性的（id/ld/jd）等。还从动词中细分出联系动词 vl、形式动词 vx、助动词 vu、趋向动词 vd。子类的标记用两个以上的字母组合表示，第一个字母仍是原来的基本词类标记。标记总数达到***多个。

（3）多方面的适应性。既要适应语言信息处理与语料库语言学研究的需要，又要能为传统的语言研究提供充足的素材；既要适合计算机自动处理，又要便于人工校对。依据《规范 2003》的加工结果可以全自动地、完全准确地映射到依据《规范 2001》的加工结果（实际上就是归并子类）。在依据《规范 2001》的加工结果的基础上，借助《语法信息词典》基本上可以自动得到符合《规范 2003》的加工结果（实际上，就是细分类。要保证百分之百的正确，还有一部分工作需要人工校对）。在这个意义上，《规范 2003》和《规范 2001》是兼容的。

（4）汉语的词组本位语法体系的指导作用[4,5]。汉语的词类与句法成分之间不存在简单的一一对应关系。同一个句法成分可以由不同词性的词来充任；而具有确定词性的同一个词又可以充当不同的句法成分，词本身的形态可以没有任何变化。《现代汉语语法信息词典》是在词组本位语法体系的指导下研制的，对数以万计的词语根据其在实际语料中的语法功能优势分布决定了它们的词性（即它们所属的词类）以及各种语法属性。进行词性标注时利用了《现代汉语语法信息词典》的成果，避免了只根据词在当前句子中的句法功能就决定其词性。同时考虑到语言学界对汉语词类的划分存在不同意见，在标记集中增加了名动词 vn，名形词 an，副动词 vd，副形词 ad。增加这些标记可以为词的兼类研究提供计量根据。《现代汉语语法信息词典》对词的各种语法属性的描述为扩充标记集做好了铺垫。依据《规范 2003》加工好的语料又可以为汉语词的概率语法属性描述准备数据资源[6]。

（5）新闻语料中有大量的专有名词（人名 nr，地名 ns，团体机构名称 nt 等）。为了对专有名词的命名规律和自动识别研究提供支持，《规范 2001》已对专有名词进行了标注，并且还对由若干个词语组合而成的短语型专有名词加上方括号和类型标记（主要是 nt，nz，ns）。《规范 2003》还对汉族人名进一步区分“姓 nrf”和“名 nrg”。

（6）标注语料库同北京大学的《现代汉语语法信息词典》、《现代汉语语义词典》、《中文概念词典》、《现代汉语短语结构知识库》、《英汉对照双语语料库》等资源相结合，将形成一个综

合型的语言知识库，可以为语言信息处理研究和汉语语言学研究提供更完备的资源。

本规范分为 4 个部分：

① **切分规范**，见第 4 章。

切分规范主要规定将句子的汉字串形式切分为词语序列的原则，即规定什么样的汉字组合可以作为一个切分单位。

② **切分和标注相结合的规范**，见第 5 章。

在汉语中，像“双音节动词＋单音节名词”通常构成新的名词，对于这个新的名词，即使在词典中没有登录，也应该把它们处理为一个切分单位。因此，在本规范中，给出了一些基于词性描述的构词规则，规定了什么样的组合可以处理为一个切分单位，并给出了新组合的词的词性。

③ **标注规范**

标注规范用以确定切分单位的标记，包括一般词性标注和专有名词标注两部分。

③-1 一般词性标注，见第 6 章。

a. 《规范 2001》的标记集以 26 个词类标记为基准，名动词、副动词、名形词、副形词和专有名词的标记是在动词代码 v、形容词代码 a、名词代码 n 后增加一个小写字母，语素标记是在语素代码 g 前面增加一个大写字母。《规范 2003》对标记集作了进一步的扩充。除了将汉族人名分出“姓”和“名”外，对时间词、代词、动词、数词、量词、副词、助词、成语、习用语、简称、标点符号等基本词类还区分出一些子类（在以下行文中，如遇到不熟悉的代码，请参见附录）。

一个词若在《现代汉语语法信息词典》中已属于某一个或若干个词类，标注时不轻易增加词性。如“训练”、“强调”在《语法信息词典》中只属于动词，标注时切勿仅根据其在当前句子的功能就将它们改为名词或副词，可以标注为名动词 vn 或副动词 vd。

b. 当《语法信息词典》给某个词确定的词性确实不对或不完备时，当然也要订正或补充。例如，“行政”这个词，在《语法信息词典》中只有名词词性 n，但在《人民日报》语料中，诸如“依法行政”这样的用法并不少见，“依法行政”应该加工为“依法/d 行政/v”，那么，“行政”应当兼属动词 v。当然，《语法信息词典》并不一定要补全在语料库中出现的所有词性，也要考虑频度等其他因素。

③-2 专有名词标注，见第 4 章与第 5 章。

这里“专有名词”的含义有了拓展。短语型的地名、团体机构名称及其他专有名称在词的切分基础上用 ASCII 码的方括号括起来，并在右方括号之后标以相应的 ns，nt，nz，方括号不嵌套。

④ **注音规范** 见第 7 章。

4．切分规范

4.1 基本概念

⑴ 切分单位

“分词单位”是中国国家标准“分词规范”中的一个基本概念[3]。它是指汉语信息处理

中使用的、具有确定的语义和语法功能的基本单位。为了同“分词规范”衔接，这里仍沿用“分词单位”这个概念，不过术语改用“切分单位”，因为“分词”这个术语已在英语语法中用于表述其他概念（现在分词、过去分词等）且为大家所熟悉，而用同一个术语表达同一或邻近学科的多个概念容易引起混淆。

按照“分词规范”对“切分单位”的定义和解释，本切分规范中的“切分单位”主要是词，也包括了一部分结合紧密、使用稳定的词组。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，如在动词的离合形式

出/v　了/ul　一/m　次/q　差/Ng　。/w

中，“差/Ng”是名语素；又如在

鹧鸪/n　的/ud　鹧/x　有/v　什么/ry　意思/n　吗/y　？/w

中，“鹧/x”是非语素字。

（这两个例子中，出现了 ul，ud，ry 这些新扩充的标记，可参看附录）。

从字数考虑，对两个字的组合较宽地看作是一个切分单位，三个字的较严，四个字以上的若不是成语、习用语一般不看作是一个切分单位。

⑵ 词典词条

“词典词条”（或“词条”）指《现代汉语语法信息词典》数据库中的字段“词语”所登录的那些语言成分（包括：词、词组、语素、前接成分、后接成分、成语、习用语、简称乃至标点符号等），泛称为“词语”。这些词语都已归了类，即已经带有词性标记。

⑶ 切分单位和词条的关系

汉语中，单音节的成词语素和不成词语素、多音节的复合词和词组的边界是模糊的。本规范规定，凡收入《语法信息词典》的词条（包括：词、词组、成语、习用语、简称乃至标点符号等）一般都是切分单位。由于这些词条多达 7.3 万，对真实文本的覆盖率很高，可以保证大多数切分单位和词条是一致的，但两者之间还是有差异的。例如 5 个字以上的成语、习用语是切分单位，但未被收入《语法信息词典》。像“一百二十八”、“五分之三”、“百分之九”、“1998 年”、“10 月 30 日”这样的数词和时间词实际上是无限多的，《语法信息词典》不可能全收，只可能收少量的构成成分。反过来，像“分之”、“百分之”作为助数词收入了语法信息词典，但它们并不是切分单位。语法信息词典中包含的前接成分、后接成分、语素、非语素字都不是切分单位，尽管当它们不能与前后成分组合时也会孤立地出现在切分序列中。

当处理大规模真实文本时，不可避免地会碰到未登录词。第 5 章给出了一些合成词的构造规则。根据这些规则自动生成的或经校对者确认的切分单位，如果结合稳定，使用频度较高，以后有可能补充到《语法信息词典》中。

4.2　对《分词规范》的补充和调整

为醒目起见，以下用符号“*”标识那些《规范 2001》中补充《分词规范》的规定，用“Δ”标识《规范 2001》中那些对《分词规范》加以调整的规定。

⑴ 人名：nr

① 汉族方式的“姓”和“名”单独切分，“姓”标注为 nrf，“名”标注为 nrg 。

张/nrf 仁伟/nrg ，　欧阳/nrf 修/nrg ，　阮/nrf 志雄/nrg ，　朴/nrf 贞爱/nrg

* 汉族人除有单姓和复姓外，还有双姓，即有的女子出嫁后，在原来的“姓”前加丈夫的姓。如：陈方安生。这种情况切分、标注为：陈/nrf　方/nrf　安生/nrg ；

唐姜氏，切分、标注为：唐/nrf　姜/nrf　氏/Ng。

② 姓名后的职务、职称或称谓要分开。

江/nrf 总书记/n，　李/nrf 主席/n，　小平/nrg 同志/n，
张/nrf 教授/n，　王/nrf 部长/n，　陈/nrf 老总/n，
李/nrf 大娘/n，　刘/nrf 阿姨/n，　龙/nrf 姑姑/n

③ 对人的简称、尊称等若为两个字，则合为一个切分单位，并标以 nr。

老张/nr，　大李/nr，　小郝/nr，　郭老/nr，　陈总/nr

*④ 一些作者或艺术家的笔名或艺名，不易区分姓和名，可以作为一个切分单位。

鲁迅/nr，　茅盾/nr，　巴金/nr，　三毛/nr，　琼瑶/nr，　白桦/nr

⑤外国人或少数民族的译名（包括日本人的姓名）不予切分，标注为 nr。

克林顿/nr，　叶利钦/nr，　才旦卓玛/nr，　小林多喜二/nr，　北研二/nr，
华盛顿/nr，　爱因斯坦/nr

Δ 有些西方人的姓名中有小圆点，也不分开。 卡尔·马克思/nr

⑵ 地名：ns

安徽/ns，　深圳/ns，　杭州/ns，　拉萨/ns，　哈尔滨/ns，　呼和浩特/ns，
乌鲁木齐/ns，　长江/ns，　黄海/ns，　太平洋/ns，　泰山/ns，　华山/ns，
亚洲/ns，　海南岛/ns，　太湖/ns，　白洋淀/ns，　俄罗斯/ns，哈萨克斯坦/ns，
彼得堡/ns，　伏尔加格勒/ns

① 国名不论长短，作为一个切分单位。

中国/ns，美国/ns，日本国/ns，阿富汗/ns，巴勒斯坦国/ns
中华人民共和国 ns，　美利坚合众国 ns

△ ② 地名后有“省”、“市”、“县”、“区”、“乡”、“镇”、“村”、“旗”、“州”、“都”、“府”等单字的现代行政区划名称时，不切分开，作为一个切分单位。

四川省/ns，　天津市/ns，景德镇市/ns，　沙市市/ns，　牡丹江市/ns，正定县/ns，
海淀区/ns，　通州区/ns，东升乡/ns，　双桥镇/ns　南化村/ns，华盛顿州/ns，
俄亥俄州/ns，东京都/ns，　大阪府/ns，长野县/ns，开封府/ns，平谷县/ns

△ ③ 如果地名后的行政区划有两个以上的汉字，则将地名同行政区划名称切开，不过要将地名同行政区划名称用方括号括起来，并标以 ns。

[芜湖/ns 专区/n]ns，[宣城/ns 地区/n]ns，[内蒙古/ns 自治区/n]ns，
[宁夏/ns 回族/nz 自治区/n]ns，　[深圳/ns 特区/n]ns，
[厦门/ns 经济/n 特区/n]ns，　[香港/ns 特别/a　行政区/n]ns，
[香港/ns 特区/n]ns，　[华盛顿/ns　特区/n]ns，
[广西/ns 环江/ns 毛南族/nz 自治县/n]ns，
[青海/ns 果洛/ns 藏族/nz 自治州/n]ns

④ 地名后有表示地形地貌的一个字的普通名词，如“江、河、山、洋、海、岛、峰、湖”等，不予切分。

鸭绿江/ns，　亚马逊河/ns，　喜马拉雅山/ns，　珠穆朗玛峰/ns，地中海/ns，
大西洋/ns，　洞庭湖/ns，　塞普路斯岛/ns

△⑤ 如果地名后接的表示地形地貌的普通名词有两个以上汉字，则应切开。也要将地

名同该普通名词用方括号括起来，并标以 ns。

[台湾/ns 海峡/n]ns，　[华北/ns 平原/n]ns，　[帕米尔/ns 高原/n]ns，

[南沙/ns 群岛/n]ns，　[京东/ns 大/a 峡谷/n]ns　[横断/b 山脉/n]ns

⑥ 地名后有表示自然区划的一个字的普通名词，如“ 街，路，道，巷，里，町，庄，村，弄，堡”等，不予切分。

中关村/ns，　长安街/ns，　学院路/ns，　景德镇/ns，　吴家堡/ns，

庞各庄/ns，　三元里/ns，　彼得堡/ns，　北菜市巷/ns，北海道/ns，

（注：在日本、朝鲜、韩国，“道”或许是行政区划，但中国人一般不熟悉，将它作为自然区划比较简单。若有其他类似情况，亦同样处理。）

△⑦ 如果地名后接的表示自然区划的普通名词有两个以上汉字，则应切开。也要将地名同自然区划名词用方括号括起来，并标以 ns。

[米市/ns 大街/n]ns，　[蒋家/nz 胡同/n]ns，　[陶然亭/ns 公园/n]ns

⑧ 大小地名相连时的标注方式为：

北京市/ns 海淀区/ns 海淀镇/ns [南/f 大街/n]ns [蒋家/nz 胡同/n]ns 24/m 号/q

△（3）　团体、机构、组织的专有名称：nt

① 团体、机构、组织的专有名称若作为名词登录在《语法信息词典》中，则直接标注为 nt。

联合国/nt，　中共中央/nt，　国务院/nt，　北京大学/nt

② 大多数团体、机构、组织的专有名称一般是短语型的，较长，且含有地名或人名等专名，本规范规定先切分，再组合，加方括号标注为 nt。

[中国/ns 计算机/n 学会/n]nt，　[香港/ns 钟表业/n 总会/n]nt，

[烟台/ns 大学/n]nt，　[合肥/ns 矿业/n 学院/n]nt，

[北京/ns 图书馆/n]nt，　[富士通/nz 株式会社/n]nt，

[上海/ns 手表/n 厂/n]nt，　[北京/ns 国安队/nt]nt，

北京队/nt，　雷锋班/nt

注：本规范主张像“北京队”、“雷锋班”这样的专名作为一个“切分单位”，如果处理为

[北京/ns　队/n]nt、[雷锋/nr　班/n]nt 也是可以接受的。

③ 团体、机构、组织名称的专指性是必要的，孤立的“大学、学院、图书馆”等只标为 n，不标为 nt。在一篇文章的开头，团体、机构、组织名称的专指性是明确的，后文往往使用简称。当省略了专名，只剩下普通名词时，就不再标 nt。如采访浙江省委书记的报道，记者开始一定会写明“浙江省委”，这时加工成：

[浙江/ns 省委/n]nt

后文引用省委书记的话时，尽管“省委”指的就是“浙江省委”，但只标注为：

省委/n

也就是说，本次加工只考虑局部的上下文，而不作远程相关分析。同样，“北京大学校长办公室”应加工为：

[北京大学/nt 校长/n 办公室/n]nt

若句子中只有“校长办公室”，前面没有“北京大学”，则只加工成：

校长/n 办公室/n

尽管在给定的更大的上下文环境中，该“校长办公室”是专指的，也不标为专名。

④ 尽管有③的规定，对于在国际或中国范围内的知名的唯一的团体、机构、组织

的名称即使前面没有专名，也标为nt。

联合国/nt， [世界/n 贸易/n 组织/n]nt，

国务院/nt， 外交部/nt， 财政部/nt， 教育部/nt， 国防部/nt，

[国家/n 教育/vn 委员会/n]nt， [信息/n 产业/n 部/n]nt，

[全国/n 信息/n 技术/n 标准化/vn 委员会/n]nt，

[全国/n 总/b 工会/n]nt， [全国/n 人民/n 代表/n 大会/n]nt

美国的“国务院”，其他国家的“外交部、财政部、教育部”，必须在其所属国的国名之后出现时，才联合标注为nt。

[美国/ns 国务院/n]nt， [法国/ns 外交部/n]nt， [美/j 国会/n]nt

日本有些政府机构名称很特别，无论是否出现在“日本”国名之后都标为nt。

[日本/ns 外务省/nt]nt，[日/j 经济/n 产业/n 省/n]nt，[日本国/ns 法务省/nt]nt， 通产省/nt ， 外务省/nt

⑤ 前后相连有上下隶属关系的团体机构组织名称的处理方式如下:

[联合国/nt 教科文/j 组织/n]nt

[中国/ns 农业/n 银行/n 北京/ns 分行/n]nt

[河北省/ns 正定县/ns 西平乐乡/ns 南化村/ns 党支部/n]nt

[北京大学/nt 昌平/ns 分校/n]nt

[安徽/ns 人大/j 常委会/j 办公室/n]nt

[北京大学/nt 计算/vn 语言学/n 研究所/n]nt

当下属单位名称含有专名（如“北京/ns 分行/n”、“南化村/ns 党支部/n”、“昌平/ns 分校/n”）时，也可脱离前面的上级单位名称单独标注为nt。

[中国/ns 农业/n 银行/n]nt [北京/ns 分行/n]nt

河北省/ns 正定县/ns 西平乐乡/ns [南化村/ns 党支部/n]nt

北京大学/nt [昌平/ns 分校/n]nt

如果下属单位名称不含有专名，则必须同上级单位名称捆绑在一起标注。

⑥ 团体、机构、组织名称中用圆括号加注简称时的处理方法示例。

[宝山/ns 钢铁/n （/w 宝钢/nt ）/w 总/b 公司/n]nt

[宝山/ns 钢铁/n 总/b 公司/n]nt （/w 宝钢/nt ）/w

△（4） 除人名、国名、地名、团体、机构、组织以外的其他专有名词都标以nz，具体规定如下。

① 专有名称后接单音节的语素，如表示民族的“族”、表示语言的“语”，表示文字的“文”，则不切分，标注为nz。

满族/nz， 俄罗斯族/nz， 哈萨克族/nz， 塞尔维亚族/nz， 高山族/nz，

维吾尔语/nz， 蒙古语/nz， 汉语/nz， 罗马利亚语/nz， 捷克语/nz

中文/nz， 英文/nz， 西班牙文/nz，蒙文/nz，俄文/nz

② 专有名称后接单音节的名词，如表示人种的“人”、表示奖项的“奖”，通常不切分，标以nz；也允许切分，分别标注。

满人/nz， 哈萨克人/nz， 诺贝尔奖/nz， 茅盾奖/nz，

哈萨克/nz 人/n， 高山族/nz 人/n， 安徽/ns 人/n

③ 包含专有名称（或简称）的交通线，标以nz；短语型的，使用方括号。

津浦路/nz， 石太线/nz， [京/j 九/j 铁路/n]nz，

[京/j 津/j 唐/j 高速公路/n]nz，

[北京/ns -/w 西雅图/ns 航线/n]nz

④ 历史上重要事件、运动等专有名称一般是短语型的，按短语型专有名称处理，标以 nz。

[卢沟桥/ns 事件/n]nz，[西安/ns 事变/n]nz， [五四/t 运动/n]nz

[明治/nz 维新/n]nz， [甲午/t 战争/n]/nz

⑤ 专有名称后接多音节的名词，如“语言”、“文学”、“文化”、“方式”、“精神”等， 则应切分。

欧洲/ns 语言/n， 法国/ns 文学/n， 西方/ns 文化/n，

贝多芬/nr 交响乐/n， 雷锋/nr 精神/n，

美国/ns 方式/n， 日本/ns 料理/n， 宋朝/t 古董/n

也有人认为“主义”是后接成分,且其后常接另一个后接成分“者”,因此将“主义”同其前面的专有名称合在一起作为一个切分单位（参见：5.2(2)③之 d）。

马克思主义/n, 马克思列宁主义/n, 杜鲁门主义/n,

马克思主义者/n, 列宁主义者/n, 社会主义者/n

⑥ 商标（包括专名及后接的“牌”、“型”等）是专指的，标以 nz，但其后所接的商品仍标以普通名词 n。

康师傅/nz 方便面/n, 中华牌/nz 香烟/n, 牡丹 III 型/nz 电视机/n,

联想/nz 电脑/n, 鳄鱼/nz 衬衣/n， 耐克/nz 鞋/n

⑦ 以序号命名的名称一般不认为是专有名称。

2/m 号/q 国道/n ， 十一/m 届/q 三中全会/j

如果前面有专名，合起来作为短语型专名也是可以的。

[中国/ns 101/m 国道/n]nz,[中共/j 十一/m 届/q 三中全会/j]nz

⑧ 书、报、杂志、文档、报告、协议、合同等的名称通常有书名号加以标识，不作为专有名词。由于这些名字往往较长，名字本身按常规处理。

《/wkz 宁波/ns 日报/n 》/wky ，《/wkz 鲁迅/nr 全集/n 》/wky,

杜甫/nr 诗选/n， 陆/nrf 俭明/nrg 自选集/n

《/wkz 大众/n 医学/n 》/wky，

少数收入词典的书名、报刊名等专有名称，则不切分。

红楼梦/nz， 人民日报/nz， 儒林外史/nz

⑨ 当有些专名无法分辨它们是人名还是地名或机构名时，暂标以 nz。

[巴黎/ns 贝尔希/nz 体育馆/n]ns,

其中“贝尔希”只好暂标为 nz。

⑩ 一般的命名活动常用引号表示，不看作专有名称。

迎/v 香港/ns 回归/v 京九/j 植绿护绿/l 活动/vn

第三/m 次/q 横田/ns 基地/n 噪音/n 诉讼/vn

食谱上的菜名等通常也是短语型的，若拆开了，意思差别甚远，则不切分，否则切分。即使不切分，也不看作是专有名词。

宫保肉丁/n， 木樨肉/n， 松鼠鳜鱼/n， 红烧肉/n，

鸡蛋/n 汤/n， 芝麻/n 饼/n， 鸡丝/n 面/n

△（5） 数词与数量词组

① 基数、序数、小数、分数、百分数一律不予切分，为一个切分单位，标注为 m 。

一百二十三/m， 120 万/m， 123.54/m, 五点四亿，

第一/m， 第三十五/m， 20%/m， 三分之二/m， 千分之三十/m

“几”和“零”属于基本的系数词（或位数词），因此包含“几”和“零”的基数、序数、小数、分数、百分数也不切分。

几十/m 人/n， 十几万/m 元/q， 第一百零一/m 个/q

② 约数，前加副词、形容词或后加“来、多、左右”等助数词的应予切分。

约/d 一百/m 多/m 万/m， 仅/d 一百/m 个/q， 四十/m 来/m 个/q，

二十/m 余/m 只/q， 十几/m 个/q， 三十/m 左右/m，

几十/m 人/n， 几十万/m 元/q， 近/a 20/m 年/q 来/f

两个数词相连的及“成百”、“上千”等则不予切分。

五六/m 年/q， 七八/m 天/q， 十七八/m 岁/q， 成百/m 学生/n，

上千/m 人/n， 成千上万/i 的/u 群众/n

相连的两个数字之间若插了顿号等标点符号，还是要切分。如：

五、六年——>五/m 、/w 六/m 年/q，

九、十点钟——>九/m 、/w 十点钟/t

③ 数量词组应切分为数词和量词。

三/m 个/q， 10/m 公斤/q， 一/m 盒/q 点心/n

④ 少数数量词已登录《语法信息词典》中，则不再切分，标注为 mq。

一个/mq， 一些/mq（“分词规范”中也将“一些”作为一个切分单位）

百年/mq， 半天/mq， 一会儿/mq， 一整套/mq

⑤ 表序关系的“数＋名”结构，应予切分， 如：

一/m 营/n， 二/m 连/n ， 三/m 部/n，

△（6） 时间词

① 年月日时分秒，按年、月、日、时、分、秒切分，标注为 t 。

1997 年/t 3 月/t 19 日/t， 98 年/t 10 月/t 8 日/t，

3 月/t 10 日/t 下午/t 2 时/t 18 分/t

这里应注意时间词与数量词的区分，例如：“78 年”指“1978 年”时应标注为“78 年/t”，当指数量“七十八年”时应切分标注为“78/m 年/q”。再如 两/m 个/q 月/n，三/m 天/q 时间/n 。同样，当“8 日”指一个月当中的第八天时为时间词，不予切分，标注为“8 日/t”；若表示 8 天时段，则要分开，标注为“8/m 日/q”。

若数字后无表示时间的“年、月、日、时、分、秒”等的标为数词 m。

中文/n 电脑/n 国际/n 会议/n ’/w 96/m

1998/m 中文/n 信息/n 处理/vn 国际/n 会议/n

*② “牛年、虎年”等一律不予切分，标注为：

牛年/t、 虎年/t

“甲午年、庚子、戊戌”等也不予切分，标注为：

甲午年/t， 甲午/t 战争/n， 庚子/t 赔款/n， 戊戌/t 变法/n

*③ 像“唐朝”、“宋代”等中国历史朝代的名称有专有名词的性质，标注为 tt。

如：西周/tt， 秦朝/tt， 东汉/tt， 南北朝/tt， 清代/tt

△（7） 单音节代词“本”、“每”、“各”、“诸”后接单音节名词时，和后接的单音节

名词合为代词；当后接的名词有 2 个以上音节时，应予切分。

本报/rz， 每人/rr， 本社/rz， 本校/rz，

本/rz 地区/n， 各/rz 部门/n， 贵/rz 出版社/n

(rr：人称代词， rz：指示代词)

△（8） 区别词

① 一般为切分单位，并标以词性 b。

女/b 司机/n， 金/b 手镯/n， 慢性/b 胃炎/n， 古/b 钱币/n

② 单音节区别词和单音节名词或名语素组合，作为一个切分单位，并标以名词词性 n。

雄鸡/n， 雌象/n， 女魔/n， 古币/n

*③ 也有少数单音节区别词后接双音节词常看作是一个复合词（特别是职位），则不再切分。

总书记/n， 副主任/n， 副教授/n， 总工程师/n， 副总参谋长/n

△⑼ 动词加动词或动词加形容词构成的述补结构

未收入词典的双音节述补结构，若拆开各是一个词，通常作为两个切分单位。

走/v 到/vq， 撞/v 上/vq ， 调/v 好/a， 坐/v 稳/a

若拆开了，其中至少有一个是语素，通常就不切分，作为一个切分单位。

形成/v， 鼓动/v， 说明/v， 震动/v

双音节的述补结构中间插入“得”或“不”一般应予切分，

走/v 得/u 到/vq， 走/v 不/d 到/vq， 安/v 得/u 上/vq， 安/v 不/d 上/vq

但是如果去掉“得”或“不”后，前后两个字不能组合成词的，则作为一个切分单位。

来得及/v， 来不及/v， 对得起/v， 对不起/v， 说得过去/lv， 说不过去/lv

有的去掉“得”或“不”后虽然是一个合成词，但其中至少有一个是语素，拆开了却是难以理解的，仍作为一个切分单位。

形得成/v， 形不成/v

*⑽ 四个字以上的短语，通常应切分。

总结/v 经验/n， 贯彻/v 执行/v， 调查/v 研究/v，

一/m 慢/a 二/m 看/v 三/m 通过/v

但像“生产资料/n ”、“国民经济/n”、 “生产关系/n”等若作为一个词已收入词典的就不再切分。

*⑾ 四个字的成语或习用语为一个切分单位，除标注其词类标记 i 或 l 外，还要求根据其在句子中的功能进一步标注子类。《语法信息词典》对成语（和习用语）按如下原则划分子类：

名词性	in/ln (名)
动词性	iv/lv (动)
形容词性	ia/la (形)

区别词性	ib/lb（区别）
副词性	id/ld (副)

（注：词典数据库中原来使用的符号都是大写字母，为了同语料库标注保持一致，将统一改为小写字母）。

既然《语法信息词典》已对 i 和 l 划分了子类，在大多数情况下，进一步标注子类只是复制词典中的信息而已。不过，需要说明的是，词典中子类较多，这里只区分出 5 个子类。另外，词典中静态的功能描述与实际语料中的功能认定不可能、也不应该总是一致的。例如，词典中有“修饰功能子类 im/lm”，包括副词作状语和区别词作定语的功能，而在句子中，一个成语或习用语要么作定语，要么作状语，不会同时作定语和状语，也就是说，标注时不能复制 im/lm，只能标 ib/lb 或 id/ld。这同某个词兼属区别词和副词两类，而在标注时只能标区别词或副词的道理是一样的。词典中还曾分出“补语功能”子类，指该成语或习用语作述补结构中的补语。但这样的划分，与其他子类不协调，标注时可合并到 ia/la (形)或 iv/lv (动)。即使对功能相对清晰、容易理解的 in/ln、iv/lv、ia/la 也不能认为词典中划分的子类一定同真实文本中的语法功能一致，因此最终还是要根据实际表现的语法功能决定其在当前句子中的子类。如果一时判断不了，暂时只标注 i/l 也不算错。

八拜之交/in,　胸有成竹/iv,　彬彬有礼/ia,
史无前例/ib,　平白无故/id,　绘声绘色/iv,
吃闭门羹/lv,　木头疙瘩/ln,　五大三粗/la。

请注意，上面所说的所谓“根据实际表现的语法功能决定其在当前句子中的子类”仍要遵循词组本位语法体系的约束，不能将出现在一个句子杼滨与

⑿ 超过四个字的成语或习用语，一般不予切分，暂不要求划分子类。

近水楼台先得月/i,　一年之计在于春/i,
不管三七二十一/i,　众人拾柴火焰高/i,　铁公鸡一毛不拔/l,
挂羊头卖狗肉/ i。

中间用标点符号分开的成语，则先分别标注，再用方括号括起来，标注为 i。

挂羊头，卖狗肉——>[挂羊头/i ，/wd 卖狗肉/i]i
百尺竿头，更进一步——>[百尺竿头/i ，/wd 更进一步/i]i
上不着天，下不着地——>[上不着天/i ，/wd 下不着地/i]i

#⒀ 表达一个完整概念或集合的简称或缩略语为一个切分单位，标以 j，也要求根据其在句子中的功能进一步标注子类。子类的划分原则同成语和习用语。代码有：jn（名），jv

（动），ja（形），jb（区别），jd（副）。

三好/jn，　爱委会/jn，　教科文/jb，　省直/jb，
老弱病残/jn，　农工牧副渔业/jn，　中西方/jn

*在有顿号分开的情况下，则切分:

中/jn 、/w 美/jn 、/w 日/jn,　　港/jn 、/w　澳/jn 、/w 台/jn,
港/jn 、/w 澳/jn　同胞/n，
林/jn 、/w 牧/jn 、/w 副/jn 、/w 渔/jn 等/u 副业/n

最后一个简称如果与后面一个字（语素）可合成一个词的，则不单独切分出来。

农/jn 、/w 林/jn 、/w 牧/jn 、/w 副/jn 、/w　渔业/n

国名、地名的简称并列在一起时，即使中间没有顿号也应切分开。

中/jn　美/jn　跨/v　国/n　公司/n
[京/jn　津/jn　唐/jn　地区/n]/ns
中/jn　日/jn　联合/vn　公报/n
港/jn　澳/jn　台/jn　同胞/n

用括号表示的一种特殊形式的缩略语

建（构）筑物——>建（构）筑物/jn
武术馆（校）——>武术馆（校）/jn
国（边）境——>国（边）境/jn
厅（局）长——>厅（局）长/jn

*⑭　语素和非语素字的处理

除下列特殊情况外，语素和非语素字一般不作为切分单位。

① 某些双音节离合词分开使用，其中一个是语素，可将它标注为语素。

出/v　过/uo　两/m　天/q　差/Ng,
理/v　了/ul　一/m　次/q　发/Ng,
洗/v　了/ul　一个/mg　舒舒服服/z　的/u　澡/Vg

② 单字名词或名词性语素后接单纯方位词，通常应合成为一个处所词或时间词，但为了同“分词规范”保持一致，也为了汉外机器翻译处理的方便，这里采用以下的处理方法：

a. “单字名词 + 单字方位词”的组合，切分为两个单位。

饭/n　前/f，　树/n　上/f,　包/n　里/f，　床/n　下/f

b. “名语素字+单字方位词”的结构，合为一个处所词或时间词。

桌/Ng　上/f --> 桌上/s，　午/Ng　后/f --> 午后/t，
身/Ng　上/f -->身上/s,　胸/Ng　前/f -->胸前/s

c. “省、市、县、乡、村、部、局、处、团、营、连、院、系、班”等名词后“里、上”等方位词，仍有组织、机构的意义，作为一个切分单位，标为名词。

如：　部里/n，　县里/n，　村里/n，　系里/n，　班上/n

③ 非语素字单独出现在文本中时，标注为 x。

“/wyz　鹌鹑/n　”/wyy　的/ud　“/wyz　鹌/x　”/wyy 字/n　怎么/ry 写/v　？/ww

*⑮　文本中非汉字的字符串的处理意见

① 已经约定俗成的或科学技术中已通用的符号保持原有的意义，根据其原有的意义决定相应的标记。

阿拉伯数字：121/m 号/q 房间/n

2000 年/t 8 月/t 15 日/t

单独的罗马数字：Ⅱ/m

Ⅰ X/m

X Ⅴ/m

英文字母（或字母组合）代表常用的度量单位：A 代表“安培”，

例句：

然后指针回指在 1.5A 处

正确的切分、标注为：

然后/c 指针/n 回指/v 在/p 1.5/m A/qd 处/n

又如 V 代表“伏特”；W，“瓦特”；m，米；kg，千克； 等等。

② 其他英文字母（或字母组合或语句）一律标注为 nx，如：

世界杯/n 足球赛/n A/nx 组/n 的/u 两/m 场/q 比赛/vn

（这里的 A 起代词作用）

A/nx 公司/n ，B/nx 先生/n ，X/nx 君/Ng

（这里的 A, B, X 起专有名词或代词作用）

24/m K/nx 镀金/n

（这里的 K 实际上是含纯金量的度量单位，中文用“开”，计算机将它标注为 nx，人又未校对出来，不算错，最好能保持一致。）

C/nx 是/v 光速/n

Windows98/nx

PentiumIV/nx

I/nx LOVE/nx THIS /nx GAME/nx

(这是一个英语句子，将空格分开的字符串作为一个切分单位)

③ 其他西文（希腊文、俄文等）的处理同英文。

④ 日文假名处理同英文。日文中的汉字处理同中文，但不能保证切分的正确性。

5. 切分和标注相结合的规范

汉语中的语素是构词的基本单位。语素构成合成词的方式主要有三种：重叠、附加和复合[13]。对这些情况的切分标注作如下规定。

5.1 重叠：

汉语以重叠变化方式构词的情况，主要有 AA，AAB，ABB，AABB，A 里 AB，A 不 AB，ABAB 等形式（其中 A，B 分别代表一个汉字），若这种词形作为词条收入了《语法信息词典》，其词性是确定的。下面的讨论主要是针对词典中没有该词形的情况：

⑴ “AA”重叠形

① 单字动词重叠式 AA 作为一个切分单位，并标注为动词词性 v。

如：走走/v，听听/v

② 单字形容词重叠式 AA，有的成词，有的不成词。如后面不紧跟“的”就成词，作

为一个切分单位，通常为副词 d。

好好/d　干/v　吧/y，　　久久/d　没/df　说话/v

若后面再加“地”，不改变原有的规定，如：

轻轻/d　吊/v　起/vq　又/d　轻轻/d　地/ui　放/v　下/vq

久久/d　地/ui　没/df　说话/v

但是，如果只有紧跟着“的”或“地”才成词，则“AA 的”或“AA 地”合为一个切分单位，标注为状态词 z。

甜甜的/z　点心/n，　　削/v　得/ue　尖尖的/z，

圆圆地/z　坐/v　一/m　圈/qc

③ 单字名词重叠式 AA，为一个切分单位，并标注为名词词性 n。

人人/n，　　家家/n

④ 单字量词重叠形式 AA，为一个切分单位，并标上量词词性 q。

张张/q，　　个个/q

⑤ 单字副词重叠式 AA，为一个切分单位，并标注为副词词性 d。

常常/d，　仅仅/d

⑵ “AAB”重叠形

① VO 结构形式的双音节离合动词的“AAB”重叠形式为一个切分单位,并标为动词词性 v。

洗洗澡/v，　　挥挥手/v，　　理理发/v

② 单音节动词的重叠式 AA 加“看”合为一个切分单位，并标注为动词词性 v。

试试看/v，　　查查看/v，　　念念看/v

⑶ “ABB”重叠形

① 双音节形容词的重叠形式 ABB, 为切分单位，并标注为状态词 z。

孤单单/z，　　亮堂堂/z，　　孤零零/z

② 数量结构的“ABB”形式，不予切分，并标上数量词词性 mq 。

一个个/mq，　　一阵阵/mq，　　一团团/mq，　一辆辆/mq

⑷ “AABB”重叠形

① 二字动词的重叠形式“AABB”为一个切分单位，并标注动词 v。

比比划划/v，　　勾勾搭搭/v

② 二字形容词的重叠形式“AABB”为一个切分单位，

高高兴兴/z，　　舒舒服服/z

若后加“的”或“地”，则标注为：

高高兴兴/z　的/u，　　舒舒服服/z　地/u

③ 二字名词的重叠形式“AABB”为一个切分单位，并标注为名词 n。

山山水水/n，　　方方面面/n

④ 二字数词的重叠形式“AABB”为一个切分单位，并标注为数词 m。

许许多多/m，　　多多少少/m

⑤ 有两个意义相反的单字形容词并列而成的名词再重叠所得到的重叠形式“AABB”为一个切分单位，并标注为状态词 z 。

大大小小/z，　　高高低低/z

⑥ 凡只能处于状语位置上的重叠形式“AABB” 标注为副词 d。

原原本本/d，　　确确实实/d

⑸ “A 里 AB”和“A 不 AB”的词形

① 双音节形容词的重叠形式“A 里 AB”，为一个切分单位，并标注为状态词 z。

马里马虎/z，　　糊里糊涂/z，　　慌里慌张/z

② 用肯定加否定的形式表示疑问的动词或形容词的词组，一般切分开。

相信/v　不/d　相信/v，　　容易/a 不/d 容易/a

但是如形成“A 不 AB”的不完整形式，则不予切分，并分别标以词性 v 或 z。

相不相信/v，　　容不容易/z，　　漂不漂亮/z

⑹ “ABAB”重叠形

双音节词的重叠形式“ABAB”，都切分开，这主要包括：

① 动词的“ABAB”　　如：研究/v　研究/v，　　比划/v　比划/v

② 形容词的“ABAB”　　如：高兴/a　高兴/a，　　舒服/a　舒服/a

③ 数词的“ABAB”　　如：很多/m　很多/m，　　许多/m　许多/m

④ 状态词的“ABAB”　　如：雪白/z　雪白/z，　　碧绿/z　碧绿/z

⑤ 数量词的“ABAB”　　如：一个/mq　一个/mq

⑺ 双音节拟声词的 “ABAB”重叠形式同其他词类一样，切分开，如：

哗啦哗啦——>哗啦/o　哗啦/o

⑻ 其他形式的重叠情况

由动词形成的“V 一 V，V 了 V，V 了一 V”重叠形式，作为动词词组都切分开。

谈/v　一/m　谈/v，　　想/v　了/ul　想/v，　　读/v　了/ul　一/m　读/v

5.2　附加

⑴　前接成分＋语素或词

由“前接成分＋语素或词”构成的合成词，为一个切分单位。这又可细分为以下情况：

① “阿”＋单音节名词或名语素，组成名词，并标以 n；若该名语素是指人的专名，则标为 nr。

如：阿哥/n，　　阿华/nr

② “小”或“老”或“大”＋单音节姓氏字，组成指人专有名词，标以 nr。

如：小王/nr，　老张/nr，　大杨/nr

③ “老”或“小”＋单字基数词（二，三，……，九），组成名词并标以 n。

如：老二/n，　老六/n，　小三/n

④ 其它前接成分（“非”,“超”,“无”,“过”，……）与词构成的新的合成词，可能保持原词的词性，也可能改变词性。

如：非金属/n，　超音速/b（音速/n），　超声波/n，　无公害/v（公害/n），

无条件/d（条件/n），　过饱和/z （饱和/a）

若“非”等前接成分所管辖的范围超过一个词，则仍然切分开。

如：非/h 国家/n 工作/vn 人员/n，　　非/h 本市/rz 注册/vn 车辆/n

⑵　语素或词＋后接成分

由“语素或词＋后接成分”组成的合成词，一律为一个切分单位。 详述如下：

① #＋“儿”（#表示任意语素或词，下同）

儿化词一般为名词，如：花儿/n，画儿/n

也有例外： 一/m 捆儿/q，玩儿/v，颠儿/v，滚圆儿/z，好好儿/d， 好好儿的/z

② #＋“们”

a. 表示名词复数的“们”单独切分，并标以 k。如：

朋友/n 们/k ，孩子/n 们/k

b. 二字词中的“们”或口语中的 “们”同前面的名词的组合（可儿化）拆开了无意义，就合起来作为一个切分单位，并标以 n。如：

人们/n ， 哥儿们/n ， 爷儿们/n， 老少/n 爷们儿/n

③ 有类化作用的后接成分

a. 由后接成分“家”,“员”,“生”,“长(zhang3)”,“性”,“机”等组成的合成词，一般为名词。 如：艺术家/n ，办事员/n ，劳动者/n ，毕业生/n ，参谋长/n ，革命性/n ，磁盘机/n

b. 由后接成分“头(tou5)”,“子(zi5)”等组成的合成词，一般为名词，如：

码头/n，孩子/n， 对头/n ， 码子/n

但也有特殊情况，如： 前头/f ，后头/f

应该注意的是，具有实在意义的名词“头（tou2)”、“子（zi3)”不看作后接成分，试比较：

对头{dui4tou5}/n， 对头{ dui4tou2}/a

砖头{zhuan1tou5}/n, 砖头{zhuan1tou2}/n(义为“碎砖”),

儿子{er2zi5}/n, 继子{ji4zi3}/n(义为“过继来的儿子”)

至于“炮弹头”、“围棋子”,不论是作为一个词还是作为一个词组,其中的“头”、“子”都不是后接成分。

炮弹头{pao4dan4tou2}/n ， 围棋子{wei2qi2zi3}/n ，

炮弹{pao4dan4}/n 头{tou2}/n ， 围棋{wei2qi2}/n 子{zi3}/Ng 。

c. # ＋“化”， 一般组成动词，如：标准化/v， 多元化/v；也有例外：四化/j，理想化/a。

d. # ＋“者”,“者”前面为较短的词或短语时，它和前面的词一起合成一个切分单位，标注为 n;“者”前面为较长的短语或句子时，分开来，标注为 k 。

研究者/n， 探索者/n， 求知者/n， 屡教不改者/n

经过/p 苦苦/d 追求/v 而/c 获得/v 幸福/a 者/k

不/d 顾/v 劝告/v 而/c 执意/vd 闹事/v 者/k

④ 词加多个后接成分，仍为一个切分单位。

物理学/n， 物理学家/n， 语言学/n， 语言学界/n

⑶ 前接成分＋语素或词＋后接成分，此种形式组成的合成词，也为一个切分单位。

非党员/n， 无政府主义者/n， 超薄型/b

注意：单音节区别词与前接成分的处理方式有所不同。

5.3 复合词

“复合”方式可将两个构词成分结合成一个新词[7]。构词成分通常认为是语素。由于复

合词的构成方式和短语的构成方式是一样的，包括定中、状中、述宾、述补、主谓、联合、连动等。当语素是成词语素时，复合词与短语的界限是不清晰的。只有当构词成分中至少有一个是不成词语素时，才有把握判断新组合的结构是一个未登录词，否则存在一定的弹性。形式上，两个字的或三个字的组合可以较宽地认为是一个词。以下使用的“名”指标注为 n 的名词或标注为Ng 的名语素。“形”,“动”的含义可以类推。

由于构成方式（定中、状中、述宾、述补等等）对一个结构是否应当看作是词有重要影响，同样的“名＋名”可能是定中结构，也可能是述宾结构，又由于目前尚不能依靠自动分析，因此这里需要较多的人工介入。

⑴ 二字名词

① “名＋名”的定中结构， 一般为一个切分单位。

牛肉/n， 铝锅/n， 敌营/n

② “动＋名”如果是定中结构，一般为一个切分单位。

炒菜/n， 烤肉/n， 绑腿/n， 来函/n， 恋人/n

③ “动＋名” 如果是述宾结构，则是短语，应切分开。

我/r 喜欢/v 吃/v 烤肉/n 。/w （这里“烤肉”是定中结构）

我/r 来/v 烤/v 肉/n 吃/v 。/w （这里“烤肉”是述宾结构）

但有些使用频度稳定的述宾结构习惯上已经被看成词，则处理成一个切分单位（离合词），标注为动词 v，如：吃饭/v， 洗澡/v， 讲话/v。

④ “形＋名”的定中结构，若中间不能插“的”或插“的”后意义改变，则作为一个切分单位；否则，应予切分。

红茶/n ， 苦瓜/n ， 红花/n（一种药材）

小/a 床/n ， 白/a 花/n ， 红/a 花/n

⑵ 三字名词

① “动（双音）＋名（单音）”的定中结构，一般为一个切分单位。

消耗品/n, 证明信/n， 救济粮/n， 控制阀/n

如果“动（双音）＋名（单音）”是述宾结构，则是短语，应切分开，例如，“他负责控制水、电、气”中的“控制水”是述宾结构，只能切分为“控制/v 水/n”。

② “名（双音）＋名（单音）”结构，通常为一个切分单位，但弹性较大，若前面的双音节名词与后面的单音节名词组合后意义不变，也可以分开。

牛仔服/n， 电流表/n， 热带鱼/n， 河北/ns 人/n, 手表/n 厂/n

③ “名（单音）＋名（双音）”结构，通常为一个切分单位，但弹性较大，若前面的单音节名词与后面的双音节名词组合后意义不变，也可以分开。

手指甲/n， 马尾巴/n, 电/n 暖壶/n

④ “形（单音）＋名（双音）”的定中结构，处理原则同二个字的“形＋名”组合。

小媳妇/n， 老姑娘/n

白/a 砂糖/n， 香/a 橡皮/n， 甜/a 点心/n

⑤ “形（双音）+名（单）”的定中结构，处理原则同④。

美丽岛/n， 贫困/a 县/n， 富裕/a 村/n

⑶ 其他

① 单纯方位词＋名（单音）的定中结构，为一个切分单位。所组成的合成词一般是

处所词，但在某些特殊情况下可能是名词或时间词。

前院/s，　　里屋/s，　　后街/s

左肩/n，　　旁权/n，　　前天/t，　　后天/t

② 明显带排行的亲属称谓要切分开，分不清楚的则不切开。

三/m 哥/n，　　大婶/n，　　大/a 女儿/n，　　大哥/n，　小弟/n，　　老爸/n

6 标注规范

6.1　词性标注与《语法信息词典》的关系

根据《语法信息词典》，对于那些不兼类的词，在切分的同时就可以确定其词性。标注规范重点描述在特定的上下文环境下如何唯一确定兼类词的正确的词性。

⑴ 尽管自动标注的依据是《语法信息词典》，但由于还需要“多选一”和确定“未登录词”的词性，因此自动标注的正确性最终还要依靠人工鉴别。

⑵ 由于上下文的信息充分，文本中的词性标注相对于词的归类要容易些，但在北大的语法体系内应坚持词类的多功能性,主要防止的倾向是仅仅根据一个词在当前句子中所实现的功能来确定其词性。如果将主宾语位置上的词一律定为名词，那是不恰当的。

⑶ 由于词典的空间限制，不仅存在未登录词问题，已登录的词也存在兼类不完备的问题。如有些名词可兼量词（“一/m 船/q 水/n”的“船”就是量词），词典中可能只描述它可以临时作量词，而未明确规定它兼属量词类，这时仍应以文本中的实际功能决定其词性。又如“新”，词典中只确定它是形容词，也有人认为“新同学”中的“新”是区别词，标注成“新/b 同学/n”也是可以的。这样将充分发掘每个词形可能兼有的词性。至于以后是否把新增加的词性收入词典则还要考虑其他因素。

6.2 常见兼类词的词性选择

由于文本数据的特点，机器无法区分同形异音词与同形同音异类词，可以笼统地把汉字相同兼属多个类的词称为兼类词。下面说明兼类词的一些标注原则。

⑴　n-q 兼类情况。

汉语中的一些名词（主要是单音节名词）可以兼作量词，对于这些词，依据上下文来确定句子中的词的词性。

① 数词＋ n-q ＋ n，取 q。

一/m　车/qr　煤/n，　　三/m　桶/qr　水/n

另外，汉语中有一部分名词临时作量词且只能前接数词“一”，对于这种情况，也是应该把它标为量词 q。

做/v　了/u　一/m　桌子/qr　菜/n,

生/v　了/u　一/m　肚子/qr　气/n

② “这”,“那”,“每”等指示代词＋ n-q＋n，取 q。

这/rz　床/qe　被子/n,　　这/r　门/qz　功课/n

③ 其它情况，一般取 n。

上/v　车/n，　进/vq　门/n，　　买/v　车/n，　　送/v　桶/n　去/vq　工地/s

⑵　a-v 兼类情况

①　若该词在句子中带了真宾语，则标为 v。

他/rr　跟/p　她/rr　没/df　红/v　过/uo　脸/n,

繁荣/v　市场/n,　　端正/v　态度/n

② 若该词受“很”一类程度副词修饰，则标为 a。

这/r 花/n 很/dc 红/a, 市场/n 很/dc 繁荣/a

③ 若该词修饰名词作定语，则一般应标为 a。

繁荣/a 的/ud 景象/n, 红/a 颜料/n, 巩固/a 的/ud 国防/n

④ 若该词作动词的补语，则应标为 a。

放/v 明白/a 一些/mq

涨/v 红/a 了/ul 脸/n

(3) v-n 兼类情况

实际上指的是广义兼类现象[2]。当该词表示一种动作时，后面带真宾语，则是 v；当它指称人或物时，则是 n。

编辑/v 科技/n 文献/n

她/rr 是/v 责任/n 编辑/n

要/v 锁/v 上/vq 门/n

忘/v 了/ul 买/v 一/m 把/qe 锁/n

及时/ad 报告/v 首长/n

一/m 份/qe 重要/a 报告/n

(4) p-v 兼类情况

这类词主要有“在”,“到”,“比”,“朝”,“跟”,“给”等，它们的区分主要依据以下方法：

① 从词的语法功能与分布考虑，若该词（包括带“着、了、过”的情况）单说或单独做谓语，则为动词。

“你/rr 爸爸/n 在/v 不/df 在/v ？/ww ” “在/v 。/wj”

北京/ns 到/vq 了/y ， 新加坡/ns 我/rr 到/vq 过/uo

别/df 老/d 跟/v 着/uz, 咱们/rr 比/v 一/m 比/v

② 对“p-v＋其他成分” 的结构，若单说或单独作谓语，则其中的 p-v 为动词；若不是单说也不是单独作谓语，而是作状语或补语，则其中的 p-v 为介词。试比较：

动 词	介 词
他/rr 不/df 在/v 教室/n	他/rr 在/p 教室/n 自习/v
他/rr 在/v 不/df 在/v 家/n ？/ww	我们/rr 走/v 在/p 校园/n 的/u 小路/n 上/f
在/v	
列车/n 已/d 到/vq 了/ul 北京/ns	老王/nr 到/p 北京/ns 出差/v 去/vq 了/ul
到/vq 没/df 到/vq 站/n ？/ww	从/p 东/f 到/p 西/f 共/d 长/a 30/m 米/qd
到/vq 了/ul	
狗/n 总/d 跟/v 着/uz 主人/n	我/rr 常/d 跟/p 他/rr 学/v 日语/n
葵花/n 向/v 太阳/n	运动员/n 正/d 跑/v 向/p 终点/n

(5) p-c 兼类情况

常见的词有“和”、“跟”、“同”、“与”，这些词的词类排歧主要依据下列原则：在句子中，如果这些词的前后成分不能互换位置或者在这些词的前面可以加修饰成分，则这些词为介词；如果这些词的前后成分可以互换位置即互换位置后句子的意思基本不变并且在这些词的前面不能有修饰成分，则这些词为连词。

我/rr 跟/c 他/rr 都/d 是/v 大学生/n
你/rr 别/df 跟/p 他/rr 跑/v
我/rr 跟/p 他/rr 请教/v 问题/n

注意：下面的句子是有歧义的（括号内信息为判定标准）。

我/rr （已经/d） 和/p 他/rr 见面/v 了/y 。/wj
我/rr 和/c 他/rr （已经/d） 见面/v 了/y 。/wj

因此，需结合前后上下文信息，才能确定正确的词类标记。

⑹ b-d 兼类情况

① 若此词作状语，则为副词。

我们/rr 会/v 共同/d 进步/v
自动/d 取消/v 订单/n

② 若此词作定语，或与“的”组成“的”字结构，则为区别词。

共同/b 目标/n 是/v 完成/v 这/rz 项/qe 任务/n
这/rz 个/qe 玩具/n 是/v 自动/b 的/ud

⑺ c-d 兼类情况

这类兼类词，主要有“不过”、“尽管”、“但”、“可”等。一般来说，若该词在句子中修饰谓语（形容词、动词）则为副词。若此词主要连接句子和子句，表示子句之间转折、让步等语义组合关系，则为连词。试比较：

① 不过

我/rr 受/v 了/ul 点/qb 伤/Ng ，/wd 不过/c 不/df 要紧/a
他/rr 不过/d 随便/ad 谈谈/v

② 但

雨/n 停/v 了/y ，/wd 但/c 地上/s 还/d 很/dc 湿/a 。/wj
但/d 见/v 门上/s 贴/v 着/uz 一/m 副/qj 对联/n 。/wj

③ 可

大家/rr 虽然/c 累/a ，/w 可/c 都/d 很/dc 愉快/a 。/wj
她/rr 待/v 我/rr 可/d 好/a 了/y 。/wj

④ 尽管

尽管/c 天/n 下/v 着/uz 雨/n ，/wd 他/rr 还是/d 出发/v 了/y 。/wj
你/rr 尽管/d 说/v ，/wd 别/df 怕/v 。/wj

6．3 关于标记 vn，vd，an，ad

这 4 个标记分别是动词 v 和形容词 a 的特殊用法标记。当将文本中的一个词标为 vn、vd 或 an、 ad 时，首先认为它们是动词或形容词，只不过它们在语句中表现了特殊的语法功能。有时也赋予这些标记以中文名称，如名动词等，只是为了方便。

⑴ 有一部分双音节动词，当它在句法结构中具有以下 4 种语法功能之一时，标为 vn：

a. 作形式动词“有”的宾语。
b. 充当了形式动词或其他准谓宾动词的准谓词性宾语。
c. 直接充当体词性短语的中心语。
d. 不加助词“的”，直接充当体词性短语的修饰语。

领导/n　对/p　这/rz　件/qe　事/n　有/vx　考虑/vn
进行/vx　一/m　次/qv　深入/a　的/ud　考察/vn
予以/vx　严肃/a　处理/vn
加以/vx　整理/vn
语法/n　研究/vn　很/dc　重要/a
必须/d　改进/v　训练/vn　方法/n
这个/rr　研究/vn　思路/n　很/dc　新颖/a

需要注意，动词直接作主语或谓宾动词的宾语，仍标为 v，不标作 vn。

考察/v　是/v　必要/a　的/ud
我们/n　来/v　的/ud　目的/n　就是/v　考察/v　考察/v
需要/v　考察/v
需要/v　考察/v　实际/a　情况/n

通常只在该动词所在的短语结构的层次内决定将它标注为 v 还是 vn。例如，

① 我们/rr 调查/v 目的/n　是/v　了解/v　实际/a　情况/n　。/wj

② 大规模/d 调查/v 语言/n 的/u 实际/a 使用/vn 情况/n 是/v 一/m 项/qe 重要/a 的/ud 基础/n 工作/vn 。/wj

③ 通过/p 调查/v

④ 通过/p 调查/v 语言/n 的/u 实际/a 使用/vn 情况/n

⑤ 进行/vx 调查/vn

⑥ 进行/vx 大规模/b 调查/vn

⑦ 通过/p 语言/n 实际/a 使用/vn 情况/n 的/ud 大规模/b 调查/vn

⑧ 通过/p 语言/n 实际/a 使用/vn 情况/n 的/ud 大规模/d 调查/v

以上 8 个例子中对“调查”的标注都是正确的。

在①中，“我们”和“调查”首先结合成主谓结构，然后再修饰“目的”。如果在“目的”之前加一个“的”，结构更清晰，读起来更流畅。不过在书面语中，这个“的”常被省掉。如果认为“的”加在“我们”和“调查”之间，“调查”和“目的”先构成定中结构，则“调查”应标注为 vn。这里有歧解。

在②中，“调查”或者先同“大规模”构成状中结构，或者先同“语言的实际使用情况”构成述宾结构，都要标成 v。

在③中，“调查”本身作介词“通过”的宾语。在语法体系内，介词可以带谓词性宾语。“调查”是动词的理由可在④中找到。

在④中，“调查”先同“语言的实际使用情况”构成谓词性的述宾结构，再作介词“通过”的宾语。

在⑤中，“调查”作形式动词“进行”的准谓词性宾语，当然标成 vn。

在⑥中，“进行”的准谓词性宾语“调查”可以带定语，“大规模”应该标成区别词。

在⑦和⑧中，对“大规模调查”的标注是不一样的。为什么前面说它们都对呢？首先，⑦是对的。因为从整体上看，“语言实际使用情况的大规模调查”是体词性的，将其中心语“大规模调查”也标成体词性的定中结构，不会引起争议。而在⑧中，“大规模调查”却被标成了谓词性的状中结构。这样标算不算错？理论上有没有困难？前面所说的“介词可以带谓词性宾语”是“词组本位”语法体系的一个重要论点，而这里认为“体词性短语的中心成分可以是谓词性成分”则是“词组本位”语法体系坚持的另一个更重要的、更显示其理论特色的论点。对

此，朱德熙先生早有阐述[8]。坚持这个论点，可以比较方便地分析下面的句法结构。

需要/v　支持/v

需要/v　支持/v　有/v　创造性/n　的/ud　探索/vn

需要/v　群众/n　的/ud　支持/vn

需要/v　群众/n　的/ud　大力/d　支持/v

“群众的支持”是体词性短语，其中心语“支持”标成了 vn。对 vn 的完整理解应当是：“支持”首先是动词，但在这个具体的句法位置上起名词的作用。“群众的大力支持”也是体词性短语，其中心语“大力支持”是谓词性的，是状中结构。在这个结构层次中“支持”是动词 v，“大力”是副词 d。由于“大力”只有一个副词词性，这样分析就不会有困难。如果主张“体词性短语的中心成分只能是体词性成分”，“支持”固然可标注为 vn，但“大力”作为副词是不能修饰体词性成分的。类似的，还有：

钢/n　产量/n　的/ud　逐步/d　增加/v

这里的“逐步”也只有一个副词词性。

当上下文信息不充分时，标注可能出现歧解。如上面①中的“调查”标为 v 或 vn 都不能算错。在⑦和⑧中，“大规模调查”也有两种都可以接受的标注结果。

在“现场考察是重要的”中的“现场考察”是有歧义的。有两种标法。

现场/s　考察/v　是/v　重要/a　的/ud

（去/vq　现场/s　考察/v　工艺/n　流程/n　是/v　重要/a　的/ud）

现场/s　考察/vn　是/v　重要/a　的/ud

（进行/v　一/m　次/qv　现场/s　考察/vn　是/v　重要/a　的/ud）

如果缺少更多的上下文，只对“现场考察是重要的”进行标注，则认为这两种标注都是正确的。

注：“现场”的词性是处所词 s，处所词可以作状语修饰动词，也可以作定语修饰名词。

⑵　当动词直接作状语时，标注为 vd 。

他/r　讽刺/vd　说/v

主任/n　强调/vd　指出/v

若动词后加“地”作状语，仍标为 v。

他/r　讽刺/v　地/u　说/v

主任/n　强调/v　地/u　指出/v

⑶　部分形容词在语料中具有以下 3 种语法功能之一时，标注为名形词 an。

a. 作了“有”的宾语，

b. 充当了准谓宾动词的准谓词性宾语，

c. 直接充当体词性短语的中心语。

他/r　有/v　很多/m　苦恼/an

这里/s　有/v　奥妙/an

维护/v　环境/n　的/u　整洁/an

交通/n　安全/an　是/v　第一/m　要/vu　注意/v　的/ud

需要注意，形容词直接作主语或谓宾动词的宾语，仍标为 a，不标作 an 。

需要/v　努力/a

需要/v　进一步/d　努力/a

⑷　形容词直接作状语时，标注为 ad 。

认真/ad　学习/v　邓小平理论/n

深入/ad　研究/v　语法/n　有利/a　于/p　自然/a　语言/n　处理/vn　技术/n　的/ud　进步/vn

形容词后接“地”作状语时，那形容词仍标注为 a 。

我们/r　应当/v　深入/a　地/u　研究/v　语法/n

6．4 关于若干词类新增加的子类标记

《规范 2003》还要求对某些词类，在语料库中进一步标注出它们的子类。对有些词，如果暂时不能准确地标注出子类，允许只标注到基本词类。

名词：　姓 nrf，名 nrg，

例：曹/nrf　操/nrg，　诸葛/nrf　亮/ nrg，　张/nrf　李/nrf　秀兰/ nrg

时间词：　时间专名 tt，例：秦朝/tt，　元代/tt，　清朝/tt

数词：数量词 mq，例：一个/mq　苹果/n，　一些/mq　葡萄/n，　俩/mq　馒头/n

一个个/mq，　一阵阵/mq，　一团团/mq，　一辆辆/mq

量词：个体量词 qe，集体量词 qj，度量词 qd，容器量词 qr，种类量词 qz，成形量词 qc，不定量词 qb，倍率量词 ql，　时量词 qt，动量词 qv

代词：人称代词 rr，指示代词 rz，疑问代词 ry

谓词性指示代词 rzw，例：就/d　这么/rzw　吧/y　。/wj

谓词性疑问代词 ryw，例：怎么样/ryw　？/ww

动词：不及物动词 vi，联系动词 vl，趋向动词 vq，形式动词 vx，助动词 vu

副词：程度副词 dc，否定副词 df

助词：助词“的”ud，　助词“地”ui，　助词“得”ue，　助词“着”uz，

助词“了”ul，　助词“过”uo，　助词“所”us

标点符号：逗号 wd，　句号 wj，　问号 ww，　叹号 wt，　分号 wf，　顿号 wu，　冒号 wm，引号 wy，左引号 wya，右引号 wyy，括号 wk，左括号 wkz，右括号 wky

7．关于注音的说明

7.1 注音范围

只对文本中的汉字串注音，不考虑语料中出现的非汉字（包括阿拉伯数字、标点符号、英语字母、数学符号、日语假名及其他字符等等）。当数字“1999”作为一个切分单位在语料中出现时，不注音；当“1999 年” 作为一个切分单位在语料中出现时，只对汉字“年”注音。同样，对“5 万”只注“万”的音。对“卡拉 OK”只注“卡拉”的音。在本规范中，只考虑国标 GB2312 中的汉字，不考虑其他字符集中的汉字。

7.2 汉语音节

汉语音节见《现代汉语词典》中的音节表。用附在音节拼音后的数字 1,2,3,4,5 分别代表普通话的四声和轻声。虽然是对文本注音，但却由《语法信息词典》中的全拼音字段提供拼音信息。当一个词（单音节或多音节）在词典中若有多个读音时，则要根据上下文从中选择一个正确的读音。

7.3 变调与变音

单音节词“一”、“七”、“八”、“不”都有变调。“不”的读音为“bu4”，但在实际口语中的发音还有“bu2”(例如，“不是”，有时读“bu2shi4”)。本次加工对不同位置上出现的“不”一律只标“bu4”。 口语中,还有更复杂的变调或变音现象（参看《现代汉语词典》的“凡例”），

本规范不考虑实际语音流中的变调或变音。

7.4 儿化音

如“花儿”的注音是“huar1”,“一点儿” 的注音是“yi1dianr3” 。尽管儿化的“一点儿” 的实际发音是“yi1diar3”。

8. 加工后语料的形式表达

同时为自动加工软件和质量检查软件考虑，可以将经过加工（切分、词性标注、注音）的语料中一个基本单元（两个基本单元之间用空格字符隔开）作如下的形式定义：

<基本单元>≡<WORD>{<PINYIN>}/<POS>｜<WORD>/<POS>

<WORD>≡<汉语词语>

<汉语词语>≡<词典的词>｜<未定义词>

<词典的词>≡《现代汉语语法信息词典》收录的词语

<未定义词>≡《现代汉语语法信息词典》未收录的词语

<PINYIN>≡<汉语音节>

<汉语音节>≡ba1｜ba2｜ba3｜ba4｜ba5｜…

<POS>≡<汉语词性标记>

<汉语词性标记>≡Ag｜a｜ad｜an｜Bg｜b｜c｜Dg｜d｜dc｜df｜e｜f｜h｜i｜ia｜ib｜id｜in｜iv｜j｜ja｜jb｜jd｜jn｜jv｜k｜l｜la｜lb｜ld｜ln｜lv｜m｜mq｜n｜nr｜nrf｜nrg｜ns｜nt｜nx｜nz｜o｜p｜Qg｜q｜qb｜qc｜qd｜qe｜qj｜ql｜qr｜qt｜qv｜qz｜Rg｜r｜rr｜ry｜ryw｜s｜Tg｜t｜tt｜u｜ud｜ue｜ui｜ul｜uo｜us｜uz｜Vg｜v｜vd｜vi｜vl｜vn｜vq｜vu｜vx｜w｜wd｜wf｜wj｜wk｜wky｜wkz｜wm｜ww｜wp｜ws｜wt｜wu｜wy｜wyy｜wyz｜x｜y｜z

这样的语料也可以用 XML 语言表述。

（1） 标记说明

- 标签说明

根据《规范 2003》，引入三个 XML 标签：text、 w、 cu，这三个标签的含义分别如下：

text 标签： 其内容是一个切分标注语料文件，<text>标记语料的开始，</text>标记语料的结束。

w 标签： 其内容是一个词，<w>标记词的开始，</w>标记词的结尾。

cu 标签：其内容是一个短语型专有名称或者中间有标点符号的成语，<cu>表示 g 该类内容的开始，</cu>表示该类内容的结束。

这三个标签之间的包含关系为：text 标签包含 w 标签和 cu 标签，cu 标签包含 w 标签。

- 属性说明

词的词性、拼音信息以及复合单位（短语型专有名称或者中间有标点符号的成语）的功能类别通过为标签设置相应属性的方式来表示。属性设置情况如下：

1) 为 w 标签设置两个属性：

pos 属性，属性值是词的词性，是一个必有属性，即每个词必须有这个属性。

pinyin 属性，属性值是词的拼音串，是一个可选属性，只有多音词才需要这个属性。例如：多音词“的”可表示为 <w pos=”u” pinyin=”de5”>的</w>。

2) 为 cu 标签设置一个属性：

cat 属性，属性值是该复合单位的类别：

例如：“巴黎贝尔希体育馆”可表示为<cu cat=″ns″><w pos=″ns″>巴黎</w><w pos=″nz″>贝尔希</w><w pos=″n″>体育馆</w></cu>。

（2）XML 描述的 DTD 定义

一般而言，每种类型的 XML 文档都有相应的文档类型定义(DTD)，DTD 严格规定了文档中可以出现那些标签、属性以及标签之间、标签属性间的关系，以及标签和属性的取值约束。针对上述标签和属性设置，每个语料文件可以包含一个如下的文档类型定义(每行后面位于//之后的内容是对该行含义的一个文字说明，不是 DTD 的组成部分)：

```
<! DOCTYPE text [              // DTD 定义开始，文档类型名是 text
    <! ELEMENT text (w | cu)*>      //定义标签 text，内容可以是多个 w 标签或 cu 标签
    <! ELEMENT cu w+ >          // 定义标签 cu，内容可以是多个 w 标签
    <! ELEMENT w (#PCDATA) > // 定义标签 w，内容是一个字符串
    <! ATTLIST cu cat (ns|nt|nz|i) #REQUIRED> // 为标签 cu 定义属性 cat，cat 属性是必
                                              // 有属性，值可以是 ns、nt、nz 或 i
    <! ENTITY %tagset      // 定义参数实体 tagset，规定词性属性的取值范围
        “(Ag｜a｜ad｜an｜Bg｜b｜c｜Dg｜d｜dc｜df｜e｜f｜h｜i｜ia｜ib｜id｜in
        ｜iv｜j｜ja｜jb｜jd｜jn｜jv｜k｜l｜la｜lb｜ld｜ln｜lv｜m｜mq｜n｜nr
        ｜nrf｜nrg｜ns｜nt｜nx｜nz｜o｜p｜Qg｜q｜qb｜qc｜qd｜qe｜qj｜ql｜qr
        ｜qt｜qv｜qz｜Rg｜r｜rr｜ry｜ryw｜s｜Tg｜t｜tt｜u｜ud｜ue｜ui｜ul｜
        uo｜us｜uz｜Vg｜v｜vd｜vi｜vl｜vn｜vq｜vu｜vx｜w｜wd｜wf｜wj｜wk｜
        wky｜wkz｜wm｜ww｜wp｜ws｜wt｜wu｜wy｜wyy｜wyz｜x｜y｜z)”>
    <! ATTLIST w pos %tagset; #REQUIRED //为标签 w 定义属性 pos，pos 属性是必
                                        // 有属性，值可以 tagset 中的任何一个值
    pinyin NMTOKEN #IMPLIED> //为标签 w 定义属性 pinyin，pin 属性是可选
                             // 属性，值可以是字母数字串
]> // DTD 定义结束
```

（3）语料 XML 描述例子

通常，一个 XML 文档第一行是版本声明、编码声明，然后是文档类型定义声明，如果文档类型定义是外部 DTD，可以在此处指出，如果 DTD 内容简单，可以采用内部 DTD 声明，语料的 DTD 声明比较简单，这里采用内部 DTD 声明：

```
<? xml version=”1.0” encoding="gb2312" ?>
<! DOCTYPE text [
    <! ELEMENT text (w | cu)*>
    <! ELEMENT cu w+ >
    <! ELEMENT w (#PCDATA) >
    <! ATTLIST cu cat (ns|nt|nz|i) #REQUIRED>
    <! ENTITY %tagset “(Ag｜a｜ad｜an｜Bg｜b｜c｜Dg｜d｜dc｜df｜e｜f｜h｜i
```

```
        |ia|ib|id|in|iv|j|ja|jb|jd|jn|jv|k|l|la|lb|ld|ln|
        lv|m|mq|n|nr|nrf|nrg|ns|nt|nx|nz|o|p|Qg|q|qb|qc|qd
        |qe|qj|ql|qr|qt|qv|qz|Rg|r|rr|ry|ryw|s|Tg|t|tt|u|
        ud|ue|ui|ul|uo|us|uz|Vg|v|vd|vi|vl|vn|vq|vu|vx|w|wd
        |wf|wj|wk|wky|wkz|wm|ww|wp|ws|wt|wu|wy|wyy|wyz|x|y|
        z)">
        <! ATTLIST w pos %tagset; #REQUIRED
                        pinyin NMTOKEN #IMPLIED>
]>
<text>
    <w pos="rr">咱们</w>  <w pos="ns">中国</w>  <w pos="rz">这么</w>
    <w pos="a" pinyin="da4">大</w>  <w pos="ud" pinyin="de5">的</w>
    <w pos="mq">一个</w>  <w pos="a">多</w>  <w pos="n">民族</w>
    <w pos="ud" pinyin="de5">的</w>  <w pos="n">国家</w>
    <w pos="c">如果</w>  <w pos="df">不</w>  <w pos="a">团结</w>
    <w pos="w">，</w>  <w pos="d">就</w>  <w pos="df">不</w>
    <w pos="vu">可能</w>  <w pos="v">发展</w>  <w pos="n">经济</w>
    <w pos="w">，</w>  <w pos="n">人民</w>  <w pos="n">生活</w>
    <w pos="n">水平</w>  <w pos="d">也</w>  <w pos="d">就</w>
    <w pos="df">不</w>  <w pos="vu">可能</w>  <w pos="v">得到</w>
    <w pos="vn">改善</w>  <w pos="c" pinyin="he2">和</w>
    <w pos="vn">提高</w>  <w pos="w">。</w>
</text>
```

9. 结语

在《中文信息学报》发表“北京大学现代汉语语料库基本加工规范”时，已经对那些为制订《规范 2001》作过贡献的专家学者致以诚挚的谢意，这里不再赘述。

在制订《规范 2003》的过程中，刘群、王惠、吴云芳、张化瑞等同仁提出过一些好的意见和建议，在此表示感谢。

衷心欢迎专家、学者和用户对本规范以及北京大学计算语言学研究所开发的标注语料库的缺点与错误继续提出批评和指正。

参考文献

[1] 俞士汶、段慧明、朱学锋、孙斌，北京大学现代汉语语料库基本加工规范，《中文信息学报》，2002 年 第 5 期，49-64，第 6 期 58-65

[2]俞士汶等，《现代汉语语法信息词典详解（第二版）》，北京：清华大学出版社， 2002 年 12 月

[3]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，1994 年第 1 版

[4]朱德熙，语法讲义，北京：商务印书馆，1982 年
[5]朱德熙，语法答问，北京：商务印书馆，1985 年
[6]俞士汶、段慧明、朱学锋，汉语词的概率语法属性描述，《语言文字应用》，2001 年，第 3 期，21-26
[7]陆志韦等，《汉语的构词法》，科学出版社，1964 年
[8]朱德熙，《现代汉语语法研究》，北京：商务印书馆，1980

附录 按代码的字母顺序排列的标记集

码，右空 2 格后的 74 个代码是扩充的代码。标记集中共有 106 个代码）
（注：在以下“代码”列中，右边的代码是 26 个基本词类代码，右空 2 格后的 74 个代码是扩充的代码。标记集中共有 106 个代码）

代码	名称	帮助记忆的诠释
Ag	形语素	形容词性语素。形容词代码为 a，语素代码 g 前面置以 A。
a	形容词	取英语形容词 adjective 的第 1 个字母。
ad	副形词	直接作状语的形容词。形容词代码 a 和副词代码 d 并在一起。
an	名形词	具有名词功能的形容词。形容词代码 a 和名词代码 n 并在一起。
Bg	区别语素	
b	区别词	取汉字“别”的声母。
c	连词	取英语连词 conjunction 的第 1 个字母。
Dg	副语素	副词性语素。副词代码为 d，语素代码 g 前面置以 D。
d	副词	取 adverb 的第 2 个字母，因其第 1 个字母已用于形容词。
dc	程度副词	
df	否定副词	
e	叹词	取英语叹词 exclamation 的第 1 个字母。
f	方位词	取汉字“方” 的声母。
g	语素	绝大多数语素都能作为合成词的“词根”，取汉字“根”的声母。由于实际标注时，一定标注其子类，所以从来没有用到过 g。
h	前接成分	取英语 head 的第 1 个字母。
i	成语	取英语成语 idiom 的第 1 个字母。
ia		形容词功能成语
ib		区别词功能成语
id		副词功能成语
in		名词功能成语
iv		动词功能成语
j	简称略语	取汉字“简”的声母。
ja		形容词功能简称
jb		区别词功能简称
jd		副词功能简称
jn		名词功能简称
jv		动词功能简称
k	后接成分	
l	习用语	习用语尚未成为成语，有点“临时性”，取“临”的声母。

	la		形容词功能习用语
	lb		区别词功能成语
	ld		副词功能习用语
	ln		名词功能习用语
	lv		动词功能习用语
m		数词	取英语 numeral 的第 3 个字母，n，u 已有他用。
	mq	数量词	在语法信息词典中归入数词库的数量短语。
	Ng	名语素	名词性语素。名词代码为 n，语素代码 g 前面置以 N。
n		名词	取英语名词 noun 的第 1 个字母。
	nr	人名	名词代码 n 和“人(ren)”的声母并在一起。
	nrf	姓	
	nrg	名	
	ns	地名	名词代码 n 和处所词代码 s 并在一起。
	nt	机构团体	“团”的声母为 t，名词代码 n 和 t 并在一起。
	nx	非汉字串	
	nz	其他专名	“专”的声母的第 1 个字母为 z，名词代码 n 和 z 并在一起。
o		拟声词	取英语拟声词 onomatopoeia 的第 1 个字母。
p		介词	取英语介词 prepositional 的第 1 个字母。
	Qg	量语素	
q		量词	取英语 quantity 的第 1 个字母。
	qb	不定量词	
	qc	成形量词	
	qd	度量词	
	qe	个体量词	
	qj	集体量词	
	ql	倍率量词	
	qr	容器量词	
	qt	时量词	
	qv	动量词	
	qz	种类量词	
	Rg	代语素	
r		代词	取英语代词 pronoun 的第 2 个字母,因 p 已用于介词。
	rr	人称代词	
	ry	疑问代词	
	ryw	谓词性疑问代词	
	rz	指示代词	
	rzw	谓词性指示代词	
s		处所词	取英语 space 的第 1 个字母。
	Tg	时语素	时间词性语素。时间词代码为 t,在语素的代码 g 前面置以 T。
t		时间词	取英语 time 的第 1 个字母。
	tt	专名时间词	用于标注中国历史朝代的时间词。
u		助词	取英语助词 auxiliary 的第 2 个字母,因 a 已用于形容词。
	ud	助词“的”	
	ue	助词“得””	
	ui	助词“地	

	ul	助词“了”	
	uo	助词“过”	
	us	助词“所”	
	uz	助词“着”	
	Vg	动语素	动词性语素。动词代码为 v。在语素的代码 g 前面置以 V。
v		动词	取英语动词 verb 的第一个字母。
	vd	副动词	直接作状语的动词。动词和副词的代码并在一起。
	vi	不及物动词	
	vl	联系动词	
	vn	名动词	指具有名词功能的动词。动词和名词的代码并在一起。
	vq	趋向动词	
	vu	助动词	
	vx	形式动词	
w		标点符号	
	wd	逗号	
	wf	分号	
	wj	句号	
	wk	括号	
	wky	右括号	
	wkz	左括号	
	wm	冒号	
	wp	破折号	
	ws	省略号	
	wt	叹号	
	wu	顿号	
	ww	问号	
	wy	引号	
	wyy	右引号	
	wyz	左引号	
x		非语素字	非语素字只是一个符号，字母 x 通常用于代表未知数、符号。
y		语气词	取汉字“语”的声母。
z		状态词	取汉字“状”的声母的前一个字母。