基于语料统计的以“不”开头双字分词不一致研究

程月　季娜　洪鹿平

（南京师范大学文学院，南京　210046）

摘　要： 大规模语料库中分词不一致现象普遍存在，并影响语料库的建设质量。在对熟语料进行分析统计的基础上，着重研究以“不”开头的双字结构，深入分析该结构分词不一致的产生原因。从全新的角度以集合的概念进行详细分类，并得出造成组合型歧义和分词变异的一系列原因。

关键词： 分词不一致；“不”开头的双字；组合型歧义；分词变异

Corpus Based Study on Segment Inconsistency of Two-character Chinese Words Starting with"不"

Cheng Yue, Ji Na, Hong Luping

(School of Chinese Language and Literature, Nanjing Normal University, Nanjing 210046)

Abstract: The phenomenon of segment inconsistency is universal in large-scale corpus, and affects the quality of corpus establishment. we lied our emphases on the structure of two-character Chinese words starting with “不” after calculating and analyzing the processed corpus statistic. Then we analyzed the reasons that led to the segment inconsistency .The structure was classified in detail using the set theory, and we acquired a series of reasons that producing Combinatorial ambiguity and segment variation.

Keywords: segment inconsistency; two-character Chinese words starting with “不”; Combinatorial ambiguity; segment variation

引言

分词是汉语自动分析中必不可少的第一道工序，分词不一致问题是自动分词中面临的一大难题，直接关系到语料库的建设。1988 年国家审定颁布的《信息处理用现代汉语分词规范（国家标准）》（以下简称《规范》）从信息处理的实际要求出发，根据现代汉语的特点和规律，确定了一系列具体的分词规则，使得分词工作有标准可依，对自动分词和人工校对都起到了积极的作用。但是《规范》中对有些语言现象界定比较模糊，校对者对于“分词单位”的语感不稳定，时有疏忽错漏存在，因而目前语料中分词不一致的现象依旧普遍存在。

迄今为止，人们关心的是正确切分算法的研究，对分词结果不一致现象研究的报道不多。文献[2]列举了导致分词语料库出现不一致的主要类型结构。文献[3]描述了从熟语料中自动获取文本切分知识。采用基于机器学习的方法，提高语料库分词的加工质量。文献[4]针对真实语料中分词结果前后不一致的现象，提出了基于规则库的校对方法与策略，设计了人机交互的一致性校对系统。这些文献对于分词不一致现象以及产生原因缺乏系统

作者简介：程月（1980—），女，江苏连云港，助教（南京师范大学中北学院），硕士在读，chengyue@njnu.edu.cn

性的探究。

本文针对 1998 年上半年的人民日报语料做了详细统计，发现以“不”字开头的双字分词不一致字串占所有分词不一致字串总数的 3.17%，例句的比例更是达到 7.56%之高。因此，本文着重研究以“不”开头双字分词不一致情况，并对这一结果进行了细致的分类，根据这些分类总结出了一系列导致分词不一致的原因。本文的初衷：一来由于以“不”开头双字所占比例相对较大，具有一定的代表性，合理分析该种结构对于其它结构的分析和认识有很大帮助；二来打算以此为切入点尝试寻找合理的分析方法来研究分词不一致现象，将来扩展到更加复杂的情况。帮助总结规律让计算机基于统计和规则自动识别分词不一致的各类情况并且能做出相应的处理。

1 前期数据统计及考察工作

语料库分词的一致性是指在相同的语境下对同一字符串切分结果是相同的，我们在不考虑一些专有人名、地名的情况下，观察语料，造成分词不一致的主要有两种情况。我们把由多义型歧义字段的切分结果导致的分词不一致叫做组合型歧义，其余那种按照构词法或信息处理用分词加工规范等标准，可以统一该字串切分形式的叫做分词变异。

为了研究分词不一致现象，我们针对 1998 年 1—6 月的人民日报语料做了详细统计，分词不一致字串共 3124 个，例句共 311229 个。观察语料，发现以“不”字开头的双字不一致字串有 99 个，占 3.17%，例句有 23525 个，占 7.56%，其中分词变异和组合型歧义交错复杂，并且高频的分词不一致字串居多，比如“不断”例句 3106 个，“不同”例句 2390 个等等。

1998 年 1—6 月的人民日报语料做如下统计，“不”开头双字不一致字串简写为“不 X”：

	语料库	“不 X”	“不 X”所占百分比
例句数	311229	23525	7.56%
分词不一致字串数	3124	99	3.17%

其中高频“不 X”前十三个分别是（出现例句多于 500）：

分词不一致字串	例句数	分词标记	占“不 X”类百分比
不断	3106	/d /d/v	13.20%
不同	2390	/a /an /d/v /d/p	10.16%
不能	2277	/v /d/v	9.68%
不是	1780	/v /c /d/v	7.57%
不仅	1736	/c /d/d	7.38%
不少	1715	/m /d/a /d/v	7.29%
不要	794	/d /d/v	3.38%
不得	668	/v /d/v	2.84%
不足	645	/a /an /v /vn /d/a	2.74%
不可	609	/v /d/v	2.59%
不会	583	/v /d/v	2.48%
不够	548	/a /d/v	2.33%
不过	534	/d /c /d/v	2.27%
合计	17385		73.90%

由表中可以看出“不 X”高频出现的词语相对集中在了几个词上（前十三个合计占到了 73.90%），它们的分词标记比较繁多复杂，其中大部分分词不一致是由组合型歧义和分词变异共同造成的。特别是“不断”、“不同”、“不能”在我们语料中出现的频率相当高，我们有必要把这些分词不一致的情形分析清楚，以利于自动文本处理下一步工作。

2“不X”结构分类

为了将“不X”结构的分词不一致情况研究透彻，我们仔细观察了98年1月的语料，逐句逐例分析，力图寻找更新更细的方法来分析和处理此类问题。此语料里共有54个以“不”字开头的双字不一致字串，接下来我们对观察到的54个“不X”结构做具体分类。

我们将“不X”结构切分从分的时候“X”的词性标记集合记做A，例如：“不同”切分从分情况下“同”被标记成/v或/p，此时我们得到集合A={/v，/p}。将切分从合的时候“不X”的整体词性标记集合记做B，例如：“不同”切分从合情况下“不同”被标记成/a或/an，此时我们得到集合B={/a，/an}。比较A与B，可以将“不X”结构做如下分类：

1、A=B

代表性的不一致字串有：

1-1 不得、不等、不服、不顾、不合、不及、不见、不解、不怕、不忍、不容、不胜、不无、不休、不依、不语、不知、不止（A=B={/v}）

1-2 不好、不快、不灵、不平（A=B={/a}）

2、A∩[1]B=Φ[2]且B={/d}

代表性的不一致字串有：

2-1 不定、不断、不尽、不禁、不愧、不料、不许、不要、不用、不住（A={/v} B={/d}）

2-2 不大、不难(A= {/a} B={/d})

3、A∩B=Φ且/a∈B

代表性的不一致字串有：

3-1 不当、不够、不满、不配、不通、不准、不便、不明{（A={/v}，B∩{/a，/an,/ad}≠Φ）

4、A∩{/p}≠Φ

代表性的不一致字串有：

4-1 不对、不和 (A={/p}，B={/a})

4-2 不同 (A={/v，/p }，B={/a，/an })

5、B∩{/c,/y,/m}≠Φ

代表性的不一致字串有：

5-1 不管、不如、不说（A={/v}，B={/c}）

不独（A={/d}，B={/c}）

不成（A={/v}，B={/y}）

5-2 不是（A={/v}，B={/v，/c}）

5-3 不只、不过（A={/v}，B={/d ，/c }）

5-4 不少（A={/v，/a}，B={/m}）

采用这样的分类方式可以将“不X”切分的不同情况进行细致分析，因考虑到切分从分的时候“不”字永远是被标注为“/d”，所以我们将不考虑切分从分时的“不”字，而着重分析比较“X”与“不X”的词性标注的差别。以上所举的代表性的不一致字串已经将1月语料中观察到的54个“不X”进行了归类，这样的归类有助于下面对分词不一致产生原因的详细研究。

3分词不一致产生原因的具体分析

3.1 分类1-1、1-2的情况分析

此类“不X”的不同切分相对应语言单位的语法功能相同，产生切分不一致的原因主要是《规范》里面对于

[1] 该符号表示集合相交

[2] 该符号表示空集

从分还是从合界限模糊，“结合紧密，使用稳定”这个尺度很难把握。这类词要表达的核心意思还是对于“X”的否定，从分还是从合，X 的意义都没有发生改变，因此他们的不一致只是单纯的分词变异，每种切分“不 X”都是表达同一种意义，语法功能也相同。分类 1-1 举例：

(1) 得/u 多/a ！/w ”/w 船舶业/n 不/d 等/v 不/d 靠/v ，/w 不/d 喊/v

(2) 起/v 的/u 作用/n 。/w 他们/r 不等/v 不/d 靠/v ，/w 或/c 自动/d

表达的意思都是对于“等待”这个动词意义的否定。分类 1-2 举例：

(1) 对不起/v ，/w 我/r 中文/nz 很/d 不好/a ，/w 不/d 可以/v 说/v 很多/m

(2) ：/w 我/r 的/u 产品/n 再/d 不/d 好/a ，/w 还是/c 有人/r 买/v ，/w

表达的意思都是对于“好”这个形容词意义的否定。

细细分析也有稍微复杂一些的情况，“不 X”中“X”在不同词例中表现的词性一样，但是词义会有差别，几种词义都表现为相同的词性，在同一种词义下从分还是从合对 X 词义没有影响。比如“不快”有时“快”是“速度快”的含义，有时是“愉快、快乐”的含义，但是这两种的含义都是形容词词性的，“不 X”到底是“/a”还是“/d/a”照例是属于分词变异的问题。举例：

（1）写/v 不可/v ，/w 不/d 吐/v 不快/a ，/w 嬉笑怒骂/l ，/w 皆/d 是/v

（2）下岗/v 当然/d 是/v 令/v 人/n 不快/a 的/u 事/n 。/w 所以/c ，/w

（3）平畴沃野/i ，/w 但/c 发展/v 并/d 不/d 快/a。/w 淮河/ns 隔阻/v 了/u 与/c

（4）条件/n 的/u 约束/vn ，/w 进展/v 不快/a 。/w 现在/t ，/w 农业/n 银行/n

其中（1）（2）“愉快、快乐”；（3）（4）“速度快”，有时候从分有时候从合。

3.2 分类 2-1、2-2 的情况分析

“不 X”切分在从分、从合两种情况下的词性和词义差别比较大，属于组合型歧义。分类 2-1 举例：

（1）尽管/c 地震/v 过后/t 余震/n 不/d 断/v ，/w 但/c 邮路/n 却/d 始终/d

（2）的/u 表演/vn 深深/d 打动/v ，/w 不断/d 报/v 以/p 经久不息/i 的/u 热烈/a

这部分“不 X”当充当句子主要谓词成分的时候切分从分，当用来修饰动词或动词短语的时候切分从合。仅拿“不断”举例，一共出现词例 447 例，其中从合/d 有 418 例，均是充当副词修饰动词成分或把字结构，从分/d/v 有 29 例，均是以“断”来作为小句中的主要谓词成分。 分类 2-2 举例：

（1）操持/v 。/w 蓉蓉/nr 还是/v 长/v 不/d 大/a ，/w 并且/c 下意识/n 地/u 承认/v

（2）说/v ，/w 我/r 刚/d 来/v 不大/d 懂得/v 咱/r 这儿/r 的/u 规矩/n

这部分“不 X”当出现在句末标点或句末语气词之前时候“X”均表达的是形容词含义，“不”是对形容词“X”的否定。当出现在动词之前时候，“不 X”有整体作为副词的功能，修饰其后的动词。仅拿“不大”举例，一共出现词例 44 例，其中从合/d 有 8 例，均是出现在动词之前，从分/d/a 有 36 例子，均是出现在句末标点或句末语气词之前。

3.3 分类 3-1 的情况分析

这一类的“不 X”中的“X”都是既有动词用法又有形容词用法（或形容词的名词用法、形容词的副词用法），当“X”在句中做动词用的时候切分从分“/d/v”，是根据《规范》中所述“分开不违背原义”的原则。当“X”在句中做形容词（或形活名、形活副）用的时候切分从合，是根据《规范》中所述“结合紧密，使用稳定”的原则。既然二者切分从分还是从合具有不同的句法功能或语法功能，我们把这样的情况归为组合型歧义。最具有代表性的是“不当”，举例如下：

（1）准/a 了/y ，/w 如果/c 宣传/v 不当/a ，/w 也/d 不/d 会/v 产生/v

（2）高/a 层次/n 上/f 实干/v ，/w 不/d 当/v ‘/w 口号/n 干部/n ’/w 、/w

“不当”切分从合还是从分主要看它在句中是起到形容词的语法功能还是动词的语法功能，之所以说“不当”最具有代表性，是因为“当”在两种情况下的读音也不一样，形容词时候读作“dàng”，动词时候读作“dāng”，我们明显感觉到是组合型歧义。同样，这一类中的其它词例也是如此，再看如下例子：

（1）国家/n 资金/n 投入/vn 强度/n 不够/a ；/w

（2）管理/v 不力/a ，/w 开发/v 深度/n 不够/a 。/w

在以上的例句中，“不够”在句子中是形容词充当主谓短语中的谓语，所以从合，再例如

（3）地区/n 对/p 标语/n 的/u 管理/vn 不 /d 够/v 规范/a ，/w 标语/n 中/f 常/d
（4）农业/n 和/c 农村/n 经济/n 结合/v 不 /d 够/v 紧密/a ，/w 科技/n 投入/vn 偏/v
在（3）、（4）两句中，“不够”中的“够”是动词充当状语成分，所以从分。仅拿“不够”举例，一共出现词例71例，其中从合/a有42例，均是形容词充当谓语情况。

3.4 分类4-1、4-2的情况分析

这类“不X”切分从分的时候“X”比较特殊，有些独特的用法，可以做介词使用，显然做介词使用的时候是组合型歧义的表现，例如：
（1）当代人/n 的/u 需要/n 时/Ng ，/w 不/d 对/p 后人/n 的/u 满足/v 需要/n 能力/n
（2）可是/c 不对/a 了/y 。/w 年前/t ，/w 瓜田/n
这类“不X”当其后出现代词或名词表示介词意义的时候“X”被标注为/p，（1）就是这样的例子，属于组合型歧义。4-1中出现的词例“不对”、“不和”的情况比较简单（A={/p}，B={/a}），除了“X”做介词用法的时候切分从分“/d/p”以外，其余都是按照《规范》“结合紧密，使用稳定”的原则切分从合。4-2的情况稍有变化，如果将这类“X”做介词的组合型歧义提取出来，剩下的即为A={/v}，B={/a，/an}的情况，可以参看分类3-1的分析。所以这类词在分析的时候/d/p情况不是难点，它们主要是复杂在其他的切分情况。拿“不同”举例，一共出现词例331例，其中/d/p仅有1例，/d/v有4例，/a有331例，/an有3例，所以重点还是去研究分类3。

3.5 分类5-1、5-2、5-3、5-4的情况分析

这类“不X”从合有一些独特的用法，可以做连词或语气词或数词使用。分类5-1相对简单，举例：
（1）院/Ng 值勤/v 首席/v 医生/n 。/w 不管/c 白天/t 晚上/t ，/w 晴天/n 雨天/n
（2）你/r 不/d 管/v ，/w 他/r 不/d 管/v ，/w 明天/t 这/r 事/n 也许/d
除去类似于（1）做连词的情况，剩下的就只有一种切分方式，所以只单纯存在组合型歧义。
分类5-2、5-3、5-4相对复杂，均是分词变异与组合型歧义共存的情况。将5-2中从合时候的组合型歧义/c提出后，剩下的即为A={/v}，B={/v}，和分类1-1一致，也就是分词变异的问题。将5-3中从合时候的组合型歧义/c提出后，剩下的即为A={/v}B={/d}，和分类2-1一致。至于分类5-4更特殊一些，我们发现“不少”被切分为“/m”时候不像5-1、5-2、5-3中的“/c”“/y”那样单纯是组合型歧义，而也有分词变异的情况夹杂在里面。请看如下例子：
（1）和/c 愚昧/a ，/w 残疾/n 孩子/n 不/d 少/a 。/w 但/c 在/p 这些/r 地区/n
（2），/w 但/c 局部/n 亮/a 点/n 不少/m ，/w 既/c 为/v 新/a 一/m
（3）一草一木/i ，/w 上任/v 干部/n 确保/v 不/d 少/v 公家/n 一/m 砖/n 一/m 瓦/q
(1)和（2）中的“不少”的语法功能和意义都相同，但是词性标注的时候却相差很大，按照我们的定义，这里属于分词变异。而（1）（2）与（3）之间明显是组合型歧义，所以此种情况也是组合型歧义和分词变异均有。

4 结语

以上是用一种全新的研究视角来分析分词不一致的产生原因，依据此种方法，我们观察了更大规模的语料，基本能将观察到的“不X”结构做相应归类，比如，高频表里面出现的“不能”、“不可”、“不会”可以顺利归入A=B={/v}的分类1-1，“不仅”可以顺利归入A={/d}，B={/c}的分类5-1等等。由此可见，该种分类分析研究方式也适合处理更大规模的语料。

基于以上的分析，我们可以借助计算机自动识别分词变异与组合型歧义，甚至可以基于规则和统计让计算机将分词变异处理成一致，比如对于/v∈A这类情况的“不X”有这样的规则：“并”或者“也”等副词＋“不X”，切分从分，例如：
（1），/w 在/p 此时/t ，/w 并/d 不/d 知/v 工程/n 的/u 质量/n
（2）公开/ad 招标/v 投标/v ，/w 并/d 不/d 准/v 转包/v 。
（3）发展/vn 能力/n ，/w 再/d 也/d 不/d 用/v 政府/n 为/p 俺/r 的/u 吃/v
另外，我们在观察语料的时候还发现有些分词变异的情况可以予以保留，比如一些文言的对仗格式，例如：
（1）党纪/n 不/d 许/v，/w 国法/n 不/d 容/v 。/w

（2）无/v 石/Ng 不/d 奇/Ag 、/w 无/v 石/Ng 不/d 语/v

由于时间仓促，我们仅把重点放在了“不X”这一大类上展开研究。尽管如此，我们的研究工作是十分细致深入的，通过分析我们发现用这样的视角来划分并进行归类研究，具有一定的可操作性。为分词不一致的研究工作提供了一种思路。这只是我们工作的开始，在后续的工作中我们将按照本文的方法和原则，尝试用聚类的方式，使得更多不一致现象得到概括和总结。总之，人工分析的目的是为了总结规律让计算机做自动处理，为了实现这个目的，我们还要继续探索之路。

参考文献：

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

[2] 孙茂松.谈谈汉语分词语料库的一致性问题[J].语言文字应用，1999，(2)：88-91.

[3] 钱揖丽，郑家恒.文本切分知识获取及应用[J].计算机工程与应用，2003，1：63-64.

[4] 杜永萍，郑家恒.分词及词性标注一致性校对系统的设计与实现[J].电脑开发与应用，2001.10：16-18.

[5] 刘开瑛.中文文本自动分词和标注[M].北京:商务印书馆，2000.

[6] 苗玺，郑家恒.中文语料库分词不一致的分类处理研究[J] 山西大学学报(自然科学版)，2001，(10).

[7] 刘江，郑家恒，张虎.中文文本语料库分词一致性检验技术的初探[J] 计算机应用研究，2005，(09)