带标注语料库中切分变异的统计分析及思考

董宇[1]，陈小荷[2]

（1. 南京师范大学文学院，南京 210097；2. 南京师范大学文学院，南京 210097）

摘 要： 本文对 1998 年 1 月《人民日报》标注语料中具有多种切分形式的词进行了统计，发现 1034 个单纯的切分变异。在此基础上从不同层面分析切分变异的出现规律，得出大量的切分变异出现在低频词、二字词中，且随着使用频率的增加，变异的从合度逐渐趋向于 1。作者从语义、语法结构和语用的角度将切分变异划分为 13 类，使得切分变异在语义、语法结构和切分形式上取得了类的一致性，不仅可以解决某个词在语料中的切分变异问题，而且可以使语料中具有相同语义和语法结构类型的切分变异获得切分形式上的统一，从而基本解决带标注语料库中切分变异的问题。

关键词： 金本位；切分变异；统计

The Statistics Analyses and Reflection on Segmentation Variation in Annotated Corpus

DongYu[1], Chen Xiaohe[2]

(1, 2.Literature College of Nanjing Normal University, Nanjing 210097)

Abstract: This passage, through the statistics of words of various segmentation forms in People Daily in January 1998, found out 1034 pure segmentation variation. On the basis of analyzing the appearing rules of segmentation variation from different levels, it was found that a large number of segmentation variations appear in words of low frequency and two-word phrases. Besides, with the increase of the using frequency, the frequency of the entire form get increasingly towards 1.The writer divided the segmentation variation into 13 categories from the angle of semantics、grammar structure and pragmatics, so that variation get a concord in category on semantics、grammar structure and segmentation form. This not only solve the problem of segmentation variation of certain word but also make the segmentation variation with the same semantics and grammar structure in corpus get unity in segmentation form. Therefore, the problem of segmentation variation is fundamentally solved.

Key Words: Gold Standard; Segmentation Variation; Statistics

1 引言

自动分词是自然语言处理的首要环节，也是一个长期困扰我们的瓶颈问题。当《信息处理用现代汉语分词规范》中“结合紧密”、“使用稳定”的定性要求不能覆盖所有的分词现象时，有学者提出了“规范+词表”的设想，使规范可以通过词表定量地加以界定。可随之而来的由于词表的静态性使得根据同一个词表会得到不同的切分形式。于是我们开始借助于“带标注语料库”，因为带标注语料库动态地反映了词语在语料中的实际情况，成了测试各种机器学习算法的训练和测试材料，被誉为“金本位”。从分词规范到“规范+词表”，再到带标注语料库，

自动分词技术处在不断地进步之中，并且取得了巨大的成绩，可是我们同时也应该看到，“金本位”的标准对带标注语料库的标注质量提出了更高的要求。

2 带标注语料库中的切分变异

分词的一致性是衡量分词语料库质量的重要标准之一。孙茂松（1999）对语料库关于分词的一致性作了如下界定：

一致性 1:在保持语义同一性的前提下,一个结构体在语料库中的分合是否始终一致(例如:“猪肉”是否始终保持一个整体,或者始终分开)；

一致性 2:与某个结构体具有相同结构类型的其他一切结构体在语料库中的分合是否与该结构体始终一致(例如:“牛肉”与“猪肉”的结构类型完全相同,“牛肉”是否跟随了“猪肉”的分合状态)。[1]

对于一致性 1，黄昌宁（2005）提出了“切分变异”这个术语：“如果一个词在一个语料库中有不止一个切分形式就叫做一个变异，它的每一个切分形式叫做一个异例（instance），每个异例由一个或多个词（token）组成。”[2]

我们按照这样的界定，选取北大经过人工校对的 1998 年 1 月的《人民日报》标注语料作为实验的样本，从中抽取具有多种切分形式的词条。一共有 45640 个词次，1780 个词条具有多种切分形式。下面是造成不同切分形式的几种类型：

表 1 多种切分形式的类型分布

Tab.1 The pattern distribution of various segmentation form

	词次	比例(%)	词例	比例(%)
单纯切分变异	14254	31.23	1034	58.09
单纯组合型歧义	19454	42.62	422	23.71
切分变异与组合型歧义混合	10730	23.51	165	9.27
专名	1199	2.63	156	8.76
切分错误	6	0.01	3	0.17
总数	45643	100.00	1780	100.00

从词次上看，单纯的组合型歧异是造成分词多样化的最主要类别，但是从词例的数量分布上看，单纯的切分变异占比重较大。可见切分变异虽然出现的总频次不高，但分布却很分散。我们认为组合型歧异无论从语义还是结构都有显著的不同，应视为两个词，应当有两种切分形式。由于人名、地名等专名的存在也造成了一些切分形式的多样化。这些均不属于本文所讨论的切分变异的范畴。另外，切分变异语料中有极少数由于切分错误造成的切分变异，因为数量有限，我们暂时不加以讨论。除去这些，我们发现语料中有一部分组合型歧义却存在着切分变异的情况，即在组合型歧义的一个义项中出现了两种形式的切分，更有甚者，组合型歧义的两个义项都有不同的切分。在这种情况下，切分变异和组合型歧义交织在一起，使切分变异的问题又增加了难度。

现在我们将每一对组合型歧异看成是两个词，如果其中一个词出现切分变异，我们认为是一个变异，如果两个词都存在切分变异，我们认为这是两个变异。这样经过统计，我们在 1998 年 1 月的《人民日报》带标注语料库中统计出 22184 个词次的切分变异，有 1222 个变异。我们做这样的统计并不是要评判这个语料质量的高低，而是为了证明由于不同的人对词的接受度不同，因而造成切分变异现象在语料中广泛存在，值得我们重视。

3 切分变异的出现频率

由于组合型歧义中的切分变异是一种特别的情况，涉及很多由于组合型歧义误切的可能。我们现在只考虑 1034 个单纯的切分变异。经过统计，我们发现少数变异是经常出现的，频率较高，而大量的变异出现的频率比

较低，大体的情况如右图所示：

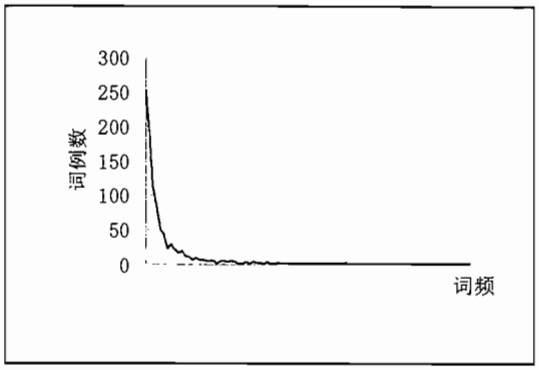

图 1 不同词频下切分变异分布

Fig.1 The distribution of segmentation variation in different word frequencies

在 98 年 1 月的语料库中，切分变异出现频次为 2 的有 253 个变异，出现频次在 10 以内的有 809 个变异，占了总变异的近 80%。当词频超过 100 之后，仅有少数切分变异，且分布非常分散。可见大量的切分变异出现在低频词中，由于使用频率较低，人们对这些词的认同度不高，因此切分时比较容易出现变异。这是切分变异主要出现在低频词的原因。当然，由于语料库的限制，很多低频词在该语料库中从未出现，或只出现了一次，没有出现切分变异的可能。可是在更大规模的语料库中，会不会发现更多不同的切分变异？我们在考察 1998 年 2 月的《人民日报》语料中的变异时，发现有大量变异是 1 月语料中所没有的，可见语料库中的切分变异是大量存在着的，远远大于我们现在所统计到的数据。

从合度是指一个切分变异从合的异例的词频与该变异所有分合异例的总词频之比。假如“绿叶”这个词在语料库中一共出现了 10 次，其中有 8 次标注成“绿叶/n”，而标注成“绿/a 叶/n”有 2 次，则“绿叶”这个变异在该语料库中的从合度是 80%。

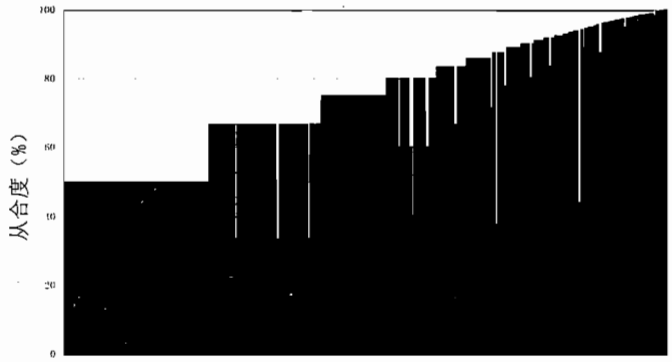

图 2 98 年 1 月切分变异的从合度

Fig.2 The frequency of the entire form of segmentation variation in January 1998

上图是 1998 年 1 月的 1034 个变异的从合度按照总词频的由低到高的走势。从上图我们可以看到，当总词频为 2 时，变异的从合度恒为 50%，之后从合度在 50%上下摆动。虽然有个别变异的从合度相对较低，但是随着词频的增加，从合度依然呈现出一个明显的趋势：从合度随着变异的使用频率的增加缓慢提高，逐渐趋向于 1。由此看来，一个词是倾向于分还是倾向于合跟它的使用频率有很大关系。此外，由上面两张图，我们可以清晰地看到少数出现频率较高的词存在切分变异的现象，这些使用相对“稳定”的词虽然表现出强烈的从合趋势，但得不到所有人的认同，出现切分变异。我们完全可以认为这些词在《人民日报》以及其他带标注语料库中都有极大存在切分变异的可能。

4 变异的分类

虽然切分变异的现象在语料中大量存在，但并不是无规律可寻，我们拟通过对这一个月的所有的变异进行分类，以期寻找到一些解决办法。分类的方法是先按照变异的语义类型加以划分，再从语法和语用的角度划分。在对变异进行语义划分时，我们参考了冯志伟教授的 ONTOL-MT-2.0 概念结构层次表，并按照需要作了一定修改。以下就是这一个月的切分变异的主要类别（C 表示从合度）：

4.1 人

4.1.1 [泛称]

a. 并列结构

【警民】(C：40.0%)　【官兵】(C：99.4%)　【男女】(C：77.8%)

b. 定中结构

【好人】(C：85.7%)　【一家人】(C：27.2%)

4.1.2 [职业]

【护法】(C：50.0%)　【助教】(C：33.3%)

4.2 **自然物**（这一类在结构上有很大的同一性，且以定中结构居多）

4.2.1 [物质]

【白霜】(C：50.0%)　【大雪】(C：96.4%)　【暖风】(C：50.0%)

4.2.2 [天地]

【大河】(C：90.0%)　【小岛】(C：37.5%)　【浅海】(C：66.7%)

4.2.3 [植物]

【大树】(C：96.0%)　【落叶】(C：80.0%)　【断木】(C：66.7%)

4.2.4 [天气]

【晴到多云】(C：88.9%)　【温暖如春】(C：20.0%)　【阴有小雨】(C：42.9%)

4.2.5 [动物]

【游鱼】(C：50.0%)　【金蛇】(C：50.0 %)　【长龙】(C：11.1%)

4.2.6 [身体]

【右腿】(C：66.7%)　【双脚】(C：25.0%)

4.3 人造物

4.3.1 [食品]

a. 并列结构

【粮油】(C：92.3%)　【烟酒】(C：28.6%)　【油盐】(C：50.0%)

b. 定中结构

【奶类】(C：44.4%)　【饮水】(C：90.9%)　【干鲜果品】(C：16.7%)

4.3.2 [用品、材料和工具]

a. 定中结构

【泰铢】(C：91.2%)　【球类】(C：75.0%)　【无绳电话机】(C：92.3%)

b. 动宾结构

【植胶】(C：80.0%)　【用料】(C：75.0%)　【用药】(C：66.7%)

4.3.3 [土地、道路和设施]

a. 动宾结构

【用地】(C：57.1%)　【用房】(C：66.7%)

b. 定中结构

【大棚】(C：96.2%)　【热土】(C：75.0%)　【高速公路】(C：92.7%)

4.3.4 [精神活动产物]

【陈言】(C：80.0%) 【大字】(C：85.7%) 【书系】(C：66.7%)

4.4 抽象物

4.4.1 [情况]

a. 并列结构

【祸福】(C：50.0%) 【苦乐】(C：20.0%) 【急难】(C：66.7%)

b. 定中结构

【省情】(C：50.0%) 【重压】(C：97.5%) 【毒情】(C：16.7%)

4.4.2 [学问]

【文化学】(C：50.0%) 【银行法】(C：66.7%)

4.5 事情

4.5.1 [事情和事态]

【快事】(C：50.0%) 【好事】(C：96.4%) 【盗窃案】(C：80.0%)

4.5.2 [集会与比赛]

【本赛】(C：33.3%) 【三中全会】(C：96.3%)

4.6 时间

4.6.1 [时点]

【秒钟】(C：25.0%) 【腊月廿九】(C：50.0%)

4.6.2 [时段]

【百年】(C：98.6%) 【本周末】(C：50.0%) 【世纪末】(C：33.3%)

4.6.3 [时间属性]

【古今】(C：16.7%) 【夏商周】(C：25.0%)

4.6.4 [顺序]

【前后】(C：97.0%) 【前列】(C：88.9%)

4.6.5 [时间先后]

【产后】(C：75.0%) 【产前】(C：50.0%) 【编后】(C：66.7%)

4.7 空间

4.7.1 [场所与区域]

a. 并列结构

【镇区】(C：95.2%) 【厂矿】(C：90.0%) 【村镇】(C：92.3%)

b. 定中结构

【旧城】(C：40.0%) 【小街】(C：50.0%) 【行政区域】(C：83.3%)

4.7.2 [左右、前后、内外、方向]

a. 并列结构

【南北】(C：92.9%) 【中外】(C：97.4%) 【前后】(C：97.0%)

b. 定中结构

【外市】(C：50.0%) 【南端】(C：75.0%) 【笔下】(C：91.67%)

4.7.3 [企业、机构与组织]

【京剧院团】(C：66.7%) 【教科文部】(C：66.7%) 【农村司】(C：50.0%)

4.7.4 [地名 机构名]

a. 并列结构

【黄淮】(C：92.9%) 【亚欧】(C：20.0%) 【港澳】(C：45.0%)

b. 定中结构

【黎明俱乐部】(C：66.7%) 【济南站】(C：50.0%) 【乌鲁木齐市】(C：75.0%)

4.8 数量

4.8.1 [数词+量词]

【一道】(C：94.4%)　【两节】(C：83.3%)　【万顷】(C：50.0%)

4.8.2 [计量单位]

【千米】(C：60.0%)

4.8.3 [整体 部分]

【全矿】(C：66.7%)　【一体】(C：93.8%)　【余下】(C：75.0 %)

4.8.4 [大小 多少]

a. 并列结构

【点滴】(C：66.7%)　【大中小】(C：50.0%)

b. 定中结构

【小额】(C：33.3%)　【全额】(C：60.0%)　【大幅度】(C：74.1%)

4.9 行为动作

行为动作是一个比较复杂的类别，我们发现在行为动作具体的小类中，语法结构没有相似性，因此这一类基本上从语法结构的角度来划分（下同）。

a. 并列结构

【拨打】(C：80.0%)　【争抢】(C：50.0%)

b. 状中结构

【长啸】(C：66.7%)　【深知】(C：80.0%)　【倍感】(C：66.7%)

c. 述补结构

【听到】(C：96.1%)　【走过】(C：95.5%)　【乐得】(C：75.0%)

d. 述宾结构

【解愁】(C：80.0%)　【张口】(C：50.0%)　【拉车】(C：33.3%)

e. 主谓结构

【心动】(C：50.0%)　【谁知】(C：80.0%)　【志在】(C：34.3%)

4.10 社会活动

a. 并列结构

【编印】(C：66.7%)　【搭建】(C：66.7%)　【审验】(C：75.0%)

b. 状中结构

【滥杀】(C：66.7%)　【电慰】(C：75.0%)　【环游】(C：50.0%)

c. 述补结构

【盗走】(C：75.0%)　【驶来】(C：33.3%)　【修好】(C：60.0%)

d. 述（介）宾结构

【办事】(C：98.6%)　【吃饭】(C：96.4%)　【传道】(C：66.7%)

4.11 属性

a. 主谓结构

【价廉】(C：33.3%)

b. 状中结构

【常见】(C：91.7%)　【很多】(C：93.0%)　【极富】(C：16.7%)

c. 述（介）宾结构

【当红】(C：50.0%)　【有害】(C：90.0%)

d. 并列结构

【重大】(C：99.7%)　【丰稔】(C：50.0%)　【升平】(C：50.0%)

4.12 连接成分

【而是】(C：99.4%)　【为了】(C：99.8%)　【就是说】(C：75.0%)

4.13 特殊用法

4.13.1 [词缀＋词根]

【多式】(C：50.0%) 【非公有制】(C：90.9%) 【反作用】(C：50.0%)

4.13.2 [动词＋着]

【有着】(C：87.4%) 【跟着】(C：89.5%) 【意味着】(C：98.0%)

4.13.3 [动词（形容词）重叠式]

【蹦蹦跳跳】(C：50.0%) 【点点头】(C：66.7%) 【小小的】(C：94.4%)

4.13.4 [离合式结构]

【提个醒】(C：50.0%)

4.13.5 [可能补语]

【算不得】(C：50.0%) 【摸得着】(C：50.0%) 【过得硬】(C：83.3%)

4.13.6 [动词＋于]

【有感于】(C：50.0%) 【植根于】(C：50.0%) 【取决于】(C：96.9%)

4.13.7 [成语、习用语]（由于人们对成语、习语的认同度是不一样的，因而造成变异）

【翻两番】(C：33.3%) 【何乐而不为】(C：50.0%) 【东方不亮西方亮】(C：66.7%)

4.13.8 [术语简称]

【双争】(C：50.0%) 【改扩建】(C：50.0%) 【输变电】(C：75.0%)

单纯依靠词频和语法结构来决定某个词是分还是合是不科学的。经过这样的分类，我们可以在语义和语法结构相同的类别里找到相同的切分形式。我们在处理标注语料的切分变异时就可以按照它们所属的类别加以统一。这样就解决了孙茂松老师所提出的“猪肉”、“牛肉”即“一致性 2”的问题。

5 结语

在语料中找到每一个切分变异，并统一它们的切分不是难事。关键是我们如何统一具有相同结构类型的切分变异。我们通过对 1998 年《人民日报》语料中切分变异的统计得出如下结论：

1．切分变异主要集中在低频词，分布分散，情况复杂。切分变异从理论上应该考虑从合的形式，因为随着使用的增加，从合度逐渐趋于 1，但是应该考虑特殊情况，特殊要求。

2．将切分变异按照语义和语法结构加以分类，相同结构类型的词我们倾向于处理成相同的切分形式。这样，不仅可以很好地解决现有的切分变异问题，统一语料的切分单位，对于未知的切分变异也可以按类处理限的。同时这样处理的好处是我们可以暂时把“语法词”和“心理词”的划分搁置在一边，并且可以按照不同的需求按类调整分词颗粒度的大小。

参考文献：

[1] 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.

[2] 黄昌宁，林娟，孙承杰.何谓金本位[A].孙茂松，陈群秀.自然语言理解与大规模内容计算[C].北京：清华大学出版社，2005.11～19.