北京外国语大学中国语言文学学院　编

人文丛刊

第九辑

出版社

北京外国语大学中国语言文学学院　编

人文丛刊

第九辑

學苑出版社

图书在版编目（CIP）数据

人文丛刊．第9辑 / 北京外国语大学中国语言文学学院编．— 北京 ：学苑出版社，2015.9
ISBN 978-7-5077-4864-2

Ⅰ．①人… Ⅱ．①北… Ⅲ．①人文科学－文集 Ⅳ．①C53

中国版本图书馆CIP数据核字(2015)第212738号

责任编辑：战葆红
封面设计：徐道会
出版发行：学苑出版社
社　　址：北京市丰台区南方庄2号院1号楼
邮政编码：100079
网　　址：www.book001.com
电子信箱：xueyuanpress@163.com
销售电话：010-67601101（销售部） 67603091（总编室）
经　　销：新华书店
印 刷 厂：保定市彩虹艺雅印刷有限公司
开本尺寸：889×1194　1/16
印　　张：24.5
字　　数：500千字
版　　次：2015年9月第1版
印　　次：2015年9月第1次印刷
定　　价：100.00元

编 委 会

目 录

【语言本体】

【语言教学】

【文学文化】

【博士专栏】

语言本体

中文组织名资源库的建设与应用

陈 慧

[内容摘要] 我们基于国家语言资源监测语料库，抽取中文组织名实例与上下文信息、文本外信息，建设了一个动态更新的中文组织名数据库，并应用该数据库尝试进行了一系列应用研究，分别是：面向文本分类和行业信息挖掘的组织名分布特征研究、面向机构名识别的组织名结构规则研究和字词符号使用研究，面向国家语言资源监测的组织名动态监测与榜单发布实验研究。

[关键词] 中文组织名 国家语言资源监测语料库 分布 结构规则 动态监测

一、研究背景

中文组织名是组织的专有名称。何谓“组织”？简单说来，“组织”涵盖了与个人、非正式临时群体相对应的集体。“这样一个集体应该具有如下两个特征：专有的组织目标、科层制组织系统。马克斯·韦伯认为科层制的特征是：有明确的权威等级；有一定的规章制度；成员在组织内的任务与其在组织外的生活相互分离；组织成员不具有它们所调配的物质资源。在民间团体等组织的发展越来越兴盛后，对过去的‘明确的权威性’、‘一定的规章制度’等都表现出一定的灵活性，但基本一致。[Anthony Giddens，2003]”。

据《中国语言生活状况报告》2005—2014 年历年的统计结果，这样的中文组织名在词语种数中的比例稳定在 36%左右，词语在不同年度中使用差异最大的是组织名，分别占到各年词种数的 40%—43%。了解和研究组织名对于语言信息处理、语言学、社会学、大众传媒研究等都有不可小觑的价值和意义。但是目前这些

价值和意义并未得到大家的重视。在工作中，我们深切感受到中文组织名的基础研究亟待突破。就拿中文信息处理来说，近 30 年来中文组织名识别成为了各种统计技术的“沙场”。然而从中文分词评测结果来看，中文组织名识别仍是中文分词标注工作的瓶颈。组织名识别除了要应用成熟的技术，还要应用相关的语言知识。和其他词语成分的语言研究相比，中文组织名的很多基本问题没有得到很好的研究和解决。

要做基础研究，就需要有材料。实际上我国有不少现成的中文组织名数据库，但是从长远角度考虑，我们所需要的数据库应该是能免费获取的，大规模的，来自真实语料的，有代表性的，而且如果能同时支持共时、历时的研究就更好。2001 年起我国建造了国家语言资源监测语料库，成为了我国规模最大、媒体覆盖面最广而且不断更新完善的国家语料库。国家语言资源监测语料库依据流通度对主流报纸进行抽样，采录真实、语言规范的平面媒体、网络媒体、有声媒体等媒体语料。这些媒体语料经过了文本预处理、分词标注和领域分类。该语料库经历了从 2001 年至今的持续动态更新，能实现历时稳态和实时动态研究的要求。如果我们基于这一语料库对组织名进行全方位的基础研究，也许将对上述领域起到一定的基础支撑作用。出于这样的考虑，我们基于该语料库初步建构了一个组织名资源库，在此基础上从多个维度对其进行统计、分析、考察、实验。现在我们就简要汇报我们的资源库建设情况及初步的应用成果。

二、中文组织名资源库建设

我们首先从国家语言资源监测语料库中提取得到了全部被标记的中文组织名。规模如表 1、表 2 所示：

表 1　中文组织名研究语料库中的组织名规模

	总数	种数	平均频次
词语	247,257,749	8,750,105	28.258
组织名	3,954,716	615,681	6.423
比例	1.60%	7.04%	22.73%

表 2　中文组织名研究语料库语料量统计表

年度	媒体	语料量(字节)	文本散布数	词语总数	词语种数
2002	北青报	514,664,332	341,770	83,941,419	2,263,096
	北京晚报	309,507,162	212,581	33,668,764	1,459,554

续表

年度	媒体	语料量(字节)	文本散布数	词语总数	词语种数
2006	法制日报	160,664,324	74,120	32,443,659	707,298
	环球时报	71,661,318	33,033	12,876,099	829,325
	人民日报	283,673,378	141,520	39,635,334	1,702,745
	羊城晚报	251,773,337	184,604	44,692,474	1,788,087
	总计	1,591,943,851	987,628	247,257,749	8,750,105

该语料库我们运用中科院自动化所分词标注系统进行分词标注，该系统的基本特点是：训练语料来自北京大学计算语言学研究所建设的《人民日报》6 个月的语料；系统的命名实体识别模块完全通过统计技术训练获得；整个分词系统没有词典和规则等资源的支持。该分词系统在我国 2004 年 863 分词评测中取得了优秀的成绩，且在组织名识别方面的准确率达到了国内领先水平。我们的组织名考察研究工作就建立在此分词标注系统分词的基础上是更有意义和价值的。我们的组织名资源库结构如图 1 所示：

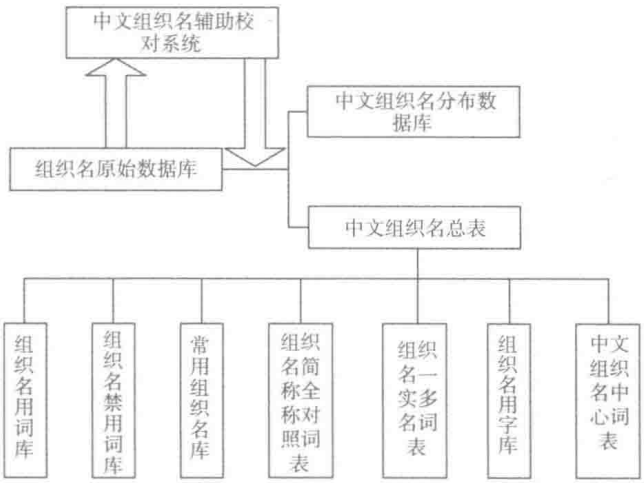

图 1　中文组织名资源库结构

由于分词系统存在一定的错误率，长远考虑，我们专门设计了一个中文组织名辅助校对系统，界面如图 2 所示。

图 2　中文组织名辅助校对系统界面（截图）

三、面向行业分析和文本分类的中文组织名宏观分布特征考察

通过组织名的分布，了解到不同媒体的关注喜好，洞察到不同行业信息，还能对文本分类有一定助益。面对多种可能的应用前景，我们做了最基础的组织名频率分布、领域分布、历时分布、报纸分布、字长分布考察，得到了以下可能有价值的分布特征：

1. 频率分布特征：平均每个中文组织名出现 6.4 次。约 63%的组织名仅出现一次；约 17%的组织名出现两次；约 4.3%的组织名频次在 10 以上。排名第一的教育科研类组织为——北大。排名第一的国际、外国体育类组织为——韩国队。

2. 领域分布特征：政治类语料出现的中文组织名最多，环境类语料的中文组织名种类最少。法制类、经济类的组织名种数在各领域类词语种数中的比例是最高的。领域独用组织名的比例较高，平均达到 56.63%的独用率。高频组织名在每个领域中都有可能出现。因此其领域特征确实不明显。

我们发现，衡量组织名是否进入通用领域，领域分布特征比频次更可靠。我们新创了“领域表征值”概念并给出操作标准，如，中国人民大学的政治领域表征值和法制领域表征值十分相近，说明中国人民大学至少在这两个领域有很强的表征能力。再如，“清华大学”和“北京大学”都在政治、教育领域特征很强，但是清华大学的政治领域特征强于教育领域特征，北京大学的教育领域特征强于政治领域特征。

3. 历时分布特征：每年独用的组织名种类约占当年全部组织名种数的 2/3，独用组织名总数约占全年组织名总数的 1/5。年度独用组织名一般为频次为 1—2 的组织名。如频次较高，则为当年较热门的组织名。相邻两年会重复用到的组织名只有大约 1/5。两年共用的词语一般也就是多年共用的高频词语和历时关注度较高的非高频词语。对于组织名而言，也是如此。

4. 字长分布特征：

组织名的长度很不确定，在 2—17 的范围内均有分布。

组织名字长越大，其频次越低。从种数来看，字长为 6 的组织名最丰富。从词总数来看，三字组织名总数最多。频次和字长呈正相关。

下面将对 2—4 字长的组织名的形式特征进行进一步考察分析。字长为 2 的组织名都是组织名简称。字长为 3 的组织名结构最多的是“2＋1”式，字长为 4 的组织名结构最多的是“2＋2”式，更多字长的组织名结构实际上是在此基础上发展而来的。

5. 报纸分布特征：

报纸独用的组织名约占该报纸组织名总种数的 2/3。

《北京青年报》词语总数、词语种数均为 6 份报纸之冠。组织名分布、独用组织名比例是所有报纸中最平均的；《环球时报》的规模最小，所关注的组织名范围更集中，报道范围也更集中；《人民日报》上的组织名高度集中。更关注一些重要的、官方的组织机构的社会活动；《羊城晚报》组织名独用比率大，对于其关注的组织，其报道量也并不多；《法制日报》的组织名最丰富，关注的组织范围更广泛，报道范围更广泛。其报纸上载的组织名与其他报纸的差异性很大。

四、面向分词系统的中文组织名结构规则研究

汉语是没有形态变化的语言，如果单纯用西方语言的命名实体识别方法——主要依据机器学习和统计进行组织名识别——效果并不理想。因为单纯的统计模型无法解决数据稀疏和用词随机性带来的问题。因此中文组织名的识别必须引入规则。在这方面，我们全面查阅了从《马氏文通》至今的语言学文献与中文信息处理文献，发现这方面的研究主要集中在某一类别组织名的结构规则描写（如高校名、企业名）、中心词统计分析、用词分析、组织名形式化分类。主要的问题是，规则稍嫌琐碎，难以操作。随机获取的数据不能反映真实语料中的组织名分布。我们将在前人研究基础上，基于识别结果，以提高识别精度为目标，深入剖析组织名的中心词、上下文、结构等，建立了一套组织名规则研究的初步成果。其体系详见图 3。以《中心词词表》为例，其中包括：52 个单义中心词、26 个兼类中心词、25 个简

称中心词、8 个小概率中心词、19 个非组织名中心词。我们依据中心词对全部组织名进行了形式化分类，并对每一类组织名进行了规则描述，以企业类组织名为例：

〈组织名〉::＝&&{〈地名〉}〈字号〉〈内容说明词〉〈中心词〉

〈地名〉::＝&&〈国名〉|〈名词：表地名〉|〈地名〉〈方位词〉|〈处所词〉

〈企业类中心词〉::＝&&〈中心词限定成分〉〈中心词〉

在上下文规则研究方面，我们初步选取的是“英国广播公司”、“中国证监会”和“清华大学”这三个代表性词语进行了前接续成分关系、后接续成分关系的考察。厘清了直接搭配关系之外的伪搭配和间接搭配关系。并对每一种情况进行了规则描述。

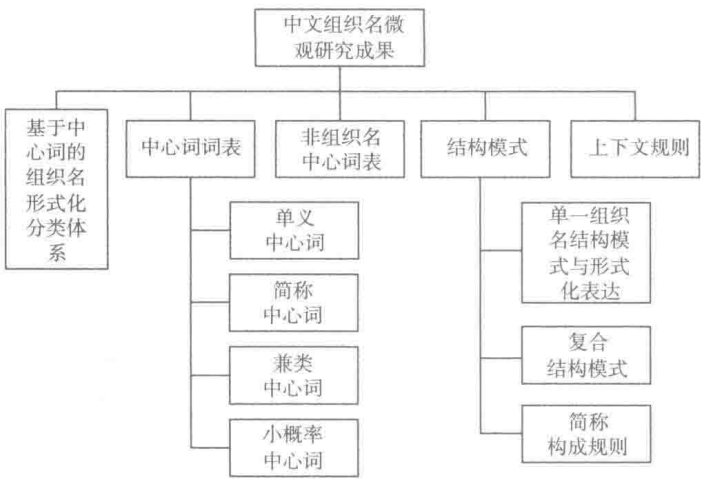

图 3　中文组织名结构规则研究成果

五、中文组织名识别结果字词符号考察

我们基于识别结果，将组织名包含的通用汉字、其他字符、词性分布、词语使用、地名、字号、内容说明成分等行了初步的全面考察，考察内容如图 4 所示。

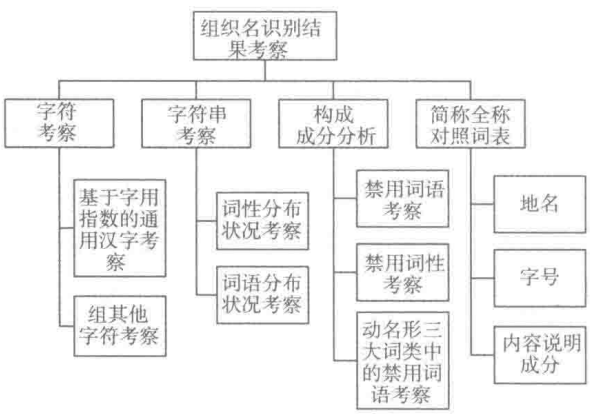

图 4　组织名识别结果考察内容

数据库中共出现了 4883 个通用汉字，其中构成组织名简称最多的 10 个字为：

中、大、航、军、北、共、部、行、铁、盟。我们对组织名表中前 60 万种[①]组织名识别结果，进行了二次分词，并根据分词结果的性质不断过滤错误的识别结果。根据字词符号考察结果操作我们的过滤程序发现，如果在组织名识别结果中引入禁用词性这一资源，能自动过滤 85475 种中文组织名。引入禁用字符串这一资源，能继续自动过滤 43930 种中文组织名。随后我们又对三大实词中的禁用词语进行了一一排查。

在以上结构规则研究和字词符号研究之后，我们提出了一个组织名识别流程，以说明中文组织名规则知识在识别中的具体应用。

六、中文组织名动态监测

我们研制了一套年度中文组织名获取流程，如图 5 所示：

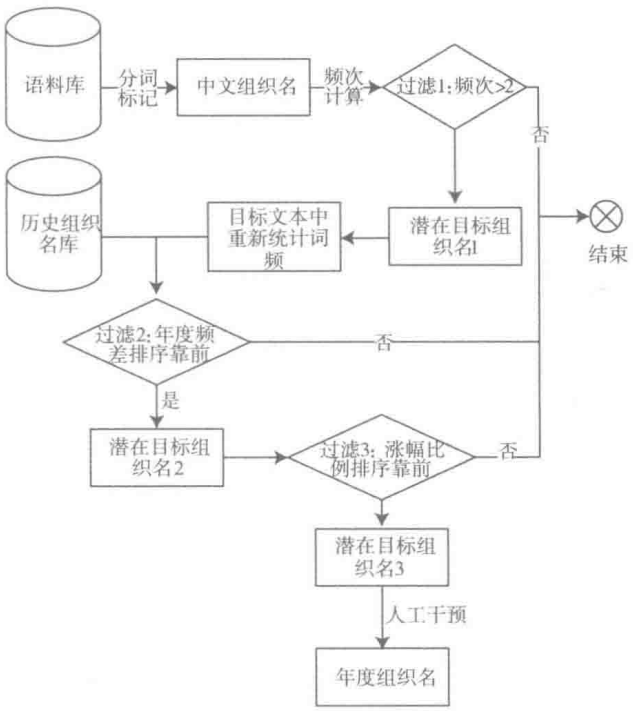

图 5　年度中文组织名获取流程

然后我们选取 2005、2006 年六份报纸语料为语料进行实验，获得年度前十五位的政府组织名如下：民进党、国土资源部、铁道部、交通部、朝阳法院、中央综治委、红四方面军、国家药监局、丰台法院、国家安监总局、药监局、全总、红一方面军、市安监局、深圳市公安局。

① 之所以只取前 60 万，主要是因为后面的一万多条词语都是错误的识别结果。

除了整体监测，在这些数据的基础上我们也能很方便地实现对特定组织名的动态监测。如，可以通过年度频次等统计数据绘制其历时走势图，以了解某一组织名的历时分布状况。根据不同组织名的历时走势图，我们可以得到“持续高度关注型”中文组织名和“年度高度关注型”组织名。前者如“教育部”，后者如“中国女足”。动态监测的目标一般重点在“年度高度关注型”中文组织名上，但“持续高度关注型”组织名则反映了媒体历时稳定高度的关注情况。

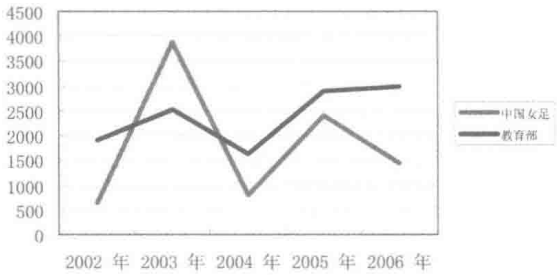

图 6 “中国女足”、“教育部”历时走势图

七、结语

中文组织名研究是一个值得深入和拓展的课题，而本文的研究只是一些初步的工作。下一步，我们将进一步研究名词、动词、形容词中的“禁用词”，结合组织名结构和语义词典，研究名词、动词、形容词在中文组织名结构中的条件限制和搭配规则，完善禁用词表，对组织名结构进行再分类，并将规则形式化，供中文信息处理使用。同时我们还将面向社会应用对组织名的动态监测和榜单发布进一步深入研究下去。

参考文献

[1]Anthony Giddens. 社会学(第四版). 北京大学出版社，2003

[2]中国语言生活状况报告：2006. 商务印书馆，2007

[3]黄昌宁，赵海. 汉语分词：十年回顾. 中文信息学报，2007(3)

[4]俞理明. 汉语缩略研究. 四川大学博士论文，2002

[5]詹卫东. 面向中文信息处理的现代汉语短语结构规则研究，清华大学出版社，1999

[6]张普. 关于动态语言知识更新与流通度研究. 语言文字应用，2001(4)

[7]张小衡，王玲玲. 中文机构名识别与分析. 中文信息学报，1997(4)

副词“直”的语义研究

池 宇

[内容摘要] 本文针对《现代汉语词典》中对副词“直”的义项，对“直”的用法进行考察，得出副词“直”除具有“一直”“径直”这一义项外，还可以表示程度、结果、反复等语义。并对各个语义的语法结构、语义限制和语用习惯进行了分析。

[关键词] 副词 直 语义

“直”的副词用法在汉语中使用比较普遍，特别是在口语中，但除了《现代汉语词典》外，《现代汉语八百词》和《实用汉语语法》都没有对“直”的副词用法进行说明。而相关的研究文章也比较少见。

在《现代汉语词典》[①]中，“直”作为副词中的义项有三个，分别是：

1. 一直；径直，直接

2. 一个劲儿；不断地

3. 简直

作为副词的第一个义项，意义比较容易确定，本文着重对第二、第三两个义项的划分进行讨论和分析。

本文将分析副词“直”的结构搭配与语义的关系，确定不同结构所对应的语义关系，并分析其语用的规则和特点。

本文采用例句除注明外，均选自北大现代汉语语料库和北语现代汉语语料库。

一、“直”的句法结构

本文对几千条副词“直”的例句进行筛选分析，总结出“直”在语法搭配上主要有以下几种形式。

① 现代汉语词典(第五版)，商务印书馆，2005

1.1VP/AP＋得＋(NP)＋直＋程度

(1)初冬的北京已经非常寒冷，穿着薄薄戏服的我们冻得直打哆嗦，导演不喊过，我们就得一遍又一遍地重来。

(2)袁绍听到曹操救了白马，气得直跳脚。

(3) 从未参加过国际大赛的白旭红开始有点紧张，在检录处急得直哭，朝着教练程国立使劲招手。

(4) 代明到处“打工”，割两斤草得一分钱，还到建筑工地推水泥灰，累得直想哭。

(5) 赛后，刘翔在赛场上激动得直翻跟斗！自那以后，刘翔真正爱上了这项运动，且更加努力了。

(6) 后来听说他是合资企业的总经理，吃惊得直伸舌头。

例(1)，作为动词“冻”的补语，直打哆嗦是作为“冻”的程度出现的。例(2)中，袁绍由于听到曹操解了白马之围，非常生日，“跳脚”是其生气的程度。

例(3)(4)(5)(6)都是形容词结构，例(4)中，“想哭”是用来突出累的程度。而“翻跟头”和“伸舌头”也分别是“激动”和“吃惊”的程度。

1.2vp＋直＋vp

(7)听了浑身直起鸡皮疙瘩。

(8)那位女孩正帮刘斌做饭，边做边和刘斌挤眉弄眼，让我直以为是刘斌的女朋友。

1.3 拟声词＋直

(9)此时正是酷夏，小孩的胳膊上散发出阵阵恶臭，招来了一群群嗡嗡直叫的苍蝇。

(10)我满脸通红，心扑扑直跳，不敢看他们的脸，转身就走。

(11)桌上的玻璃器皿被子弹打得啪啪直响，一个女佣惊叫着，埃利奥特和罗斯伏在大理石地板上，子弹在他们周围嗖嗖飞过，把他们头顶上方的石灰打得直掉，雨点般地洒在他们身上。

(12)那人挣扎而起，疼得哇哇直叫。

(13)那饺子，一咬滋滋直冒油，真香啊！

以上四类结构，直的语义均是表示由一个刺激而导致产生程度很高的反应。

无论这个反应是单次动作还是重复性动作。这些“直”都不重读，重读的是最后的反应性动词。

1.4 直＋ 状语＋VP＋得

(14)此仇不报，非丈夫也，但须谋定而动，于是寻了个隐僻所在，花了好几个月

功夫，将一路“碧针清掌”直练得出神入化，无懈可击。

(15)因此，他爱上了唱歌，不仅自己唱，还带动了身边的队友，直闹腾得大伙儿竞相添置卡拉 OK 设备，纷纷加入歌唱大军。

(16)在气候恶劣的青藏高原，大风暴的确太可怕了！狂风肆虐，石走沙飞，直刮得遮天蔽日，天昏地暗。

(17) 就因为一次骑车路过南喜元家门前时按了几下车铃，就被南喜元抽了几个耳光，以后又三天两头找茬殴打南进喜，直打得他远逃外县亲戚家，隐姓埋名好几年。

(18)晚上睡觉时，像躺在冰窖里。我冷得实在受不了，便爬起来跳跃，直跳得身上冒汗，再爬上床去睡，就这样，一夜要反复折腾好几次。

“直＋VP＋得”结构在口语中几乎不使用，仅仅出现于某些曲艺形式的台词中，带有古典色彩。

在书面语中，例句相对于“直”的其他用法，也相对较少，而且出现在武侠或古代白话文中。

例(14)中，某人练功，练了好几个月，最终达到了出神入化的境界。“直”后面的成分是说明练功的结果。

例(15)中“他”鼓动队友唱歌，最终大家都跟他一起爱上了唱歌，并购买设备。“直”后面的内容是“他”不断带动鼓励的结果。

例(16)(17)(18)中，“直”后面的成分都是表示由于前面的事件或行为导致的结果。

1.5 直＋VP

副词“直”

(19)“你如果敢把我能‘传音入秘’的事告诉众奴，我非把你剁成血浆不可！”“是，是，是……”飞奴被吓得直说“是”。

(20)她强抑住那份恐惧，拉着方婷的手直道歉。“对不起，对不起，我好晕，人很不舒服，所以——方婷，对不起，我真的不是故意的。”

(21)我还听到有一个女孩才不过第一次见到学长，整天就直夸他有多好、多高大、多帅气呢！

(22)蓝马婆听虎又怒啸，越发心寒，不住口直劝小孩快去。

(23)当他把这个好消息告诉我时，我直埋怨他瞎闹。

1.6 直＋像/似/作

《现代汉语词典》在副词“直”作“简直”讲时使用的例句是：

(24)痛得直像针扎一样难受。

在这一句中，针扎一样难受，是痛的程度，但是像“针扎一样难受”只是一个比