几种常用汉语词典收词的统计分析[①]

张永奎 齐新战

(山西大学计算机科学系,太原 030006)

摘要:本文对目前常用的三部汉语词典即《现代汉语词典》,《现代汉语通用字典》,《同义词词林》的收词进行了初步分析.根据三部词典词汇的同现情况将词划分为不同的词集,并对这些词集进行抽样(或逐词)分析和分类讨论.另外,对三部词典所收词的同现比值进行了定量计算.统计分析的结果可为机用词典的编纂与修订提供参考依据.

The Statistic And Analysis of Words Included in Several chinese Dictionaries

Zhang Yongkui and Qi Xinzhan

(Dept. of Computer Science, Shanxi University, Taiyuan 030006)

Abstract: In this paper, we give a preliminary analysis of the words which were included in three Chinese dictionaries. According to the co-occurrence of these words, we classified them into different sets, and analyzed some words (or every word) in these sets.Futhermore,we calculated the co-occurrence ratio of the words in these dictionaries. The results of analysis can provide a reference for the editing and recising of machine tractable dictionaries.

1 引 言

现代汉语机用语义词典是现代汉语信息处理的重要工具之一.机用词典的建造又是自然语言处理的一个瓶颈问题,尤以语义词典的建造最为突出.如何获取足够的知识,以建立处理大规模真实文本所需的机用词典,已成为计算语言学研究的一个重要课题.而语义词典的收词范围是我们在建造机用词典的过程中首先要解决的一个问题.以现有的汉语词典的收词为参考来确定机用词典的收词标准是一个重要的途径,为此我们对现有的几种常用汉语词典进行了统计分析,对不同的词进行了分类讨论.

目前为计算语言学界经常参考研究的三部词典:《现代汉语词典》(以下简《现汉》),《现代汉语通用词典》(以下简称《现通》),《同义词词林》(以下简称《词林》).其中,《现汉》是词典,它的编写目的是推广普通话,促进汉语规范化并在汉语教学方面起到它

① 国家自然科学基金资助项目 69575011

应有的作用。在收词方面除一般语汇外，也收了一些常见的方言词语，以及某些习见的专门术语。《现通》是一部字典，它根据字的义项进行编排，所收词语均作为例证按义项分系于各个字头之下，少数不能或不易分析字义的复音词则单独简释词义。收词以现代汉语为主，酌收少量常见古汉语词和方言词。《词林》是为创作和翻译工作者提供的一本从词义查词的工具书。它按照“以词义为主，兼顾词类，并充分注意题材的集中”这样的原则，并参考其它语义分类体系，建立了一套分类体系。将语义分为大、中、小三个层次。可见，《词林》是一部类义词典，主要选收现代汉语语词，也酌收了一些常见的方言词与古语词；此外，还有部分词素，词组，成语，俗语等。

我们在建立各部词典对应的机内词表时，参考的印刷版本分别为：《现汉》1996年7月修订第三版，《现通》1987年版，《词林》1983年10月第一版。在基本不影响统计分析结果的前提下，三部词典中含有国标一、二级6763个汉字以外的字的词将不被包含在我们统计用机内词表中。由于这三部词典所收语词中，除词以外，均收入了部分词组、成语、俗语等，所以本文在对这三部词典的收词进行统计分析过程中，对所有被这三部词典所收的词组、成语、俗语以及部分词素(主要收在《词林》中)，一般不作概念性区分，在不影响分析结果的前提下，只简单称为词或语词。

本文首先根据各词典所收词的同现将词汇划分为不同的集合，进而对各集合中的词进行了分类分析，在第三节中给出了词典的同现比值定义并进行了相关的定量计算及分析。这些统计分析结果对于我们制定的机用词典的收词规范起了重要作用，对于其它电子词典的收词及印刷版词典的修订也可起到一定的参考价值。另外，这里使用的一些统计分析方法对于电子词典或印刷词典的收词评价也有参考作用。

2 词集的划分与词集的分类分析

由于我们所分析的三部词典所收词不统一，有些词被三部词典同时收录，也有些词同时出现在两部词典中，而不被第三部词典所收；另外一些词则只被其中一个词典收录，而不出现在另外两部词典中。这样，我们根据各词典所收词的同现情况，将词汇划分为不同的词集，并对不同的词集进行了讨论，由于本文篇幅所限，这里只简要叙述其中两个词集的分类讨论的情况。

2.1 词集的划分

为以下叙述方便，本文约定分别用XH, XT, XL表示《现汉》，《现通》，《词林》对应的词集。它们两两相交形成七个词集。如图－1所示：

图－1中A_i (i = 1, .., 7)分别表示XH, XT, CL相交形成的不同的同现集合。(在附录中给出它们的部分词)。不难看出，它们可以分别表示：

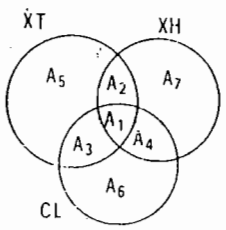

图-1 词集划分示意图

$A_1 = XH \cap XT \cap CL$；　$A_2 = XT \cap XH - A_1$；　$A_3 = XT \cap CL - A_1$；

$A_4 = XH \cap CL - A_1$；　$A_5 = XT - XH - CL$；　$A_6 = CL - XH - XT$；

$A_7 = XH - XT - CL$

我们对不同的词集分别作了逐词分析，根据各词集的特点对词进行了分类讨论。下面仅介绍了其中的部分词集的分类分析情况。

2．2 词集 A_2 的分类分析:

从图－1中可以看出，A_2 中的词同时存在于 XH，XT 中，但不为 CL 所收，A_2 中共有 8300 余词，对这些词进行逐条分析后，大致可分为以下几类:

一. 专用词。这类词约占 A_2 总数的 40 %，其中有些是领域性较强的专业术语，如：咖啡碱、多糖、分贝等；另一些则是较为常用的事物的名称，如：货机、纺车、飞碟等。这些词不被收入 CL，主要原因是无同义现象。

二. 无同义现象的通用词。这类词约占 A_2 总数的 27 %，其中含有部分成语，如：防不胜防、整装待发等，在这类词中多数是无明显的同义现象的词；如：反坐、结仇、凡是、拔罐子、开斋等，其中也有些词在现有体系下不易归类。

三. 有同义现象、且可以归类的词。这类词约占 A_2 总数的 33 %。例如："度日"《现汉》释义为：过日子。而"过日子"在《词林》中作为一个词收入，义类代码为"Hj01"，属于大类 H(活动)下属的中类 Hj(生活)中的一个小类。这类词中还含有一小部分词是书面上的文言词语(《现汉》释义中标有<书>)。例如，冀望：<书>希望。居积：<书>积累。在词林中"希望"义类码为 Gb04，"积累"义类码为 He13。

由于《词林》的分类体系及其编码形式具有简洁明了、可操作性强等特点，许多计算语言学及情报检索的研究工作者均直接或间接地以此为基础资源。在使用过程中，对于《词林》收词不全的问题都深有感触，以上对 A_2 词集的分类分析对于将一些词归入《词林》分类体系，标以适当的义类代码，将有一定的辅助参考作用。

2．3 词集 A_3 的分类分析

从图－1中可以看出，A_3 中的词是同时为《现通》，《词林》所收，而不为《现汉》所收，A_3 中共有近 3200 词，对其中的词进行逐词分析可以大致分为以下几类:

一. 专用词。这类词在 A_3 中占 16 %，其中有专业术语，如焦比、酸性、硫磺泉；也有一些是较为常见的事物名或某种人为的名称，如：海鸥、洋菜、马球、龙须草、惊风等。

二. 在《现汉》中未被解释，但作为例词出现的词。这类词占到 A_3 的 17 %。其中大多数可表示为两个或多个词义的组合。如解释"博学"一词时用"博学多才"作为例词；解释"不堪"时，用"不堪一击"为例词。而"博学多才"、"不堪一击"不单独收词解释。

三. 未作例词出现，但可拆成不同词或词素的组合。这类词在 A_3 中占到 46 %。如:

船工、付清、盖章、现钞、已婚等。这类中也有少数词是含有前缀或后缀的词，如：阿爸、阿爹、站长，市长等。

四. 不易拆开的词。这类词占到 A_3 的 19 %。这类词使用中结合紧密，大多数不能拆成词或词的组合，且多数为二字、四字词组。如：少顷、有鬼，东张西望、不折不扣、飞针走线、坐失良机等。

五. 在 A_3 中有一小部分词(约占 2 %)虽未被《现汉》所收，但它与其它词的组合却被《现汉》收入解释，由于它们的特殊性，将其单独分为一类。如：《现汉》中收有“花拳绣腿、无轨电车、整装待发、慢慢悠悠、结发夫妻”；但“花拳、无轨、整装、慢慢、结发”未收入。

3 对三部词典收词的定量分析

3．1 三部词典收词的相近性

在对不同词典的收词进行分析比较的过程中，我们不难发现，任何两部词典的收词都是既有相同的词，又有许多不同的词。如图－2所示，词集 N 为两词表 A、B 的交集，而 M_1、M_2 则映应两个词表的差异。我们认为 M_1、M_2、N 三个词集综合起来应该能够反映词表 A、B 的相近程度。为了能反映词表间的两两相近程度，我们给出以下定义：

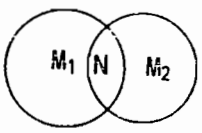

图-2 词汇同现示意图

设 A、B 为任意两词集，则称 $R(A, B) = A \cap B / A \cup B$ 为词集 A 与词集 B 的相近性。由此定义可以看出函数 R(A，B)反映的是 A 与 B 的交集的词的数量与其并集中词的数量的比例。显然，R(A，B)满足以下性质：

(1) $R(A, B) = R(B, A)$

(2) 当 $A = B$ 时，$R(A, B) = 1$

(3) 当 $A \cap B = \Phi$ 时，$R(A, B) = 0$

所以，对于任意两个词集 A，B，对应的 R(A，B)值能基本反映 A，B 间的相近程度。

我们对 XH，XT，CL 两两组合计算结果，如下：

$R(XH, XT) = 0.4792$

$R(XH, CL) = 0.4782$

$R(XT, CL) = 0.4325$

这些数据首先反映出任意两个词表的相近性均在 0.4－0.5 之间，也就是说。任意两词表的交集词数占其并集词数的不足 50%，而另外 50%以上和词，是两部词典收词有分歧的词。

显然这三部词典收词的相近程度有如下关系：

$R(XH, XT) > R(XH, CL) > R(XT, CL)$

3．2 三部词典收词的相异性

我们从图－1看出，A_5，A_6，A_7 中的词分别为 XT，CL，XH 单独收录的词，它们在各自词典中所占比例反映这些词典中较“偏僻”词所占的比例。我们用 S 表示这个比值，计算结果如下：

$$S(XH) = A_7 / XH = 22.97\ \%$$

$$S(XT) = A_5 / XT = 22.59\ \%$$

$$S(CL) = A_6 / CL = 25.38\ \%$$

以上数据说明每部词典中至少有 20% 以上的词不被其它词典所收录。而且《词林》中的“偏僻”词所占比例明显大于《现汉》与《现通》。

4　机用语义词典分级收词的初步设想

从以上统计分析的结果反映出词集 A_1 A_7 各有其特点，这些对于我们制定机用语义词典的收词标准有很大的参考价值。由于各词集所含词的规范性及对信息处理的重要性存在明显差异，我们初步提出机用语义词典中词的分级处理的设想。

对于 A_1 中的词，由于它们具备《词林》的义类信息，且具有《现汉》、《现通》的释义、义项的信息，所以将被全部收入语义词典，并作适当处理，给出尽可能的语义信息，以满足信息处理的需要。另外对于词集 A_2，A_3，A_4 中的词将根据分类分析的结果，依据一定的原则将其中一部分词收入语义词典。其中，对于 A_3 中的被收入语义词典的词作《词林》义类与《现通》义项间的意义对齐，对 A_4 中被收入语义词典的词作《词林》义类与《现汉》义项间的意义对齐，这两项工作可以参考已有的研究成果。而对于 A_2 中被收入语义词典的词，需要标注《词林》义类代码，这一工作困难大，我们已作的自动标注算法的研究，效果较为满意，但仍有部分词需要人工干预，甚至手工机助标注。由于词集 A_5、A_6、A_7 中的词多为古汉语用词、地名、专业术语等，所以依据一定的收词原则只能对其中一小部分词收入语义词典。

在我们对三部汉语词典的收词分析过程中，我们越来越意识到对于编纂任何一部汉语印刷版词典或建造一部机用词典来说，收词都是一个不容忽视的问题。众所周知，对于汉语来说，给“词”下定义是一件很困难的事，也很不容易完全统一认识。那么，我们是不是应该在词的外延或具体词的个体上多作些讨论。不仅不同词典对词的认识上不统一，同一部词典中认识也不完全一致，或者缺乏严格标准。更为复杂的是对于成语、专用词、俗语、词组的收录上。具体讲，有这样一些问题值得探讨：关于成语，哪些成语该收，哪些成语不收，标准怎样定。关于词组，哪些接合紧密更接近于词应该收，哪些则能拆开来讲。对于专科词，尤其是一些常见事物的名称，一些常见疾病的名称，一些较为普及的科技领域中的常用词，收到什么程度。以上问题对于任何信息处理用的机用词典也是应当认真考虑的。

参考文献

[1] 中国社会科学院语言研究所词典编辑室编，现代汉语词典(修订本)，商务印书馆，1996

[2] 梅家驹等，同义词词林，上海辞书出版社，1983

[3] 傅兴岭主编，现代汉语通用字典，外语教学与研究出版社，1987

[4] 杨尔弘、黄昌宁、张津，利用机读资源建造机用词典，ICCC'94 国际会议论文集

附录：《现汉》、《现通》、《词林》三部词典收词词集的划分与部分例词

A_1 词集(同时被三部词典所收的词)

阿斗　白口　带子　背静　鞭笞　病号　不一　仓库　产妇　尘埃　痴子　出借　垂本　解送　精巢　就医　竣工　可谓　裤衩　困干　黎民　两栖　留洋　轮廓　毛虫　米蛋　惟我独尊　乌云

A_2 词集(同时出现于《现汉》、《现通》，不被《词林》所收)

啊哟　傲然　百万　包围圈　北宋　闭门造车　辨证　并发　不锈钢　惨变　查抄　滚筒海防　好像　黑压压　呼哨　画院　黄蜡　汇流　机具　挤兑　甲板　键槽　铰链　窃据穷匮　拳师　韧带

A_3 词集(同时出现于《现通》、《词林》，不被《现汉》所收)

阿爸　霸权主义　半推半就　本年　彬彬有礼　不合时宜　不知所以　仓卒　查帐　怀乡　基本功　记忆犹新　甲烷　交角　介绍人　浸剂酒器　剧饮　慷慨解囊　昆弟　郎君省长　时不再来　市合

A_4 词集(同时出现于《现汉》、《词林》，不被《现通》所收)

八仙桌　保护色　笨　扁桃体　不能自已　惭　缠绕茎　澈底　赤脚医生　解囊　晶状体　旧国军事基地　拷贝　口头语　拉亏空　老婆子　楼阁　埋汰　煤炭　磨刃　拿主意　蔫不唧　盆浴

A_5 词集(只出现于《现通》中)

鸣啼　跋前顾后　百徒　薄扇　本着　播种机　栽纸机　侧重　城隅　坚韧不拔　酱瓜浸软　酒槽　撄捕　看头　溃溢　老少皆宜　礼尚　两侧　水战　松脐　锁边　摊鸡蛋　套筒　条痕　铜叶

A_6 词集(只出现于《词林》中)

八荒　百页　报罢　笔记小说　兵刃　不成气候　不清楚　操券　撑肠拄腹　旱秧田　合订本　烘干　忽忽不乐　画诺　黄钻　火车绒　急进派　家君　艰难竭蹶　凝睇　女中魁首　朋故　平方丈

A_7 词集(只出现于《现汉》中)

成龙配套　赤膊上阵　出阵　垂挂　存蓄　倒背如流　冬至　骄躁　解码　经贸　具体劳动　老娘们儿　例禁　灵境　漏失　麦季　玫瑰　弥合　逆反应　数理逻辑　说开　两边倒　贪杯　摩托艇