

制订《信息处理用现代汉语常用词词表》 的原则与问题的讨论

梁南元 刘 源 沈旭昆 谭 强 杨铁鹰

《常用词表》制订组

【摘要】 本文讨论了《信息处理用现代汉语常用词表》(以下简称《常用词表》)的制订方法。提出按照《信息处理用现代汉语分词规范》,以定量原则为主,定性原则为辅的原则进行选词,《常用词表》首次提出选词函数的术语,并创造性地使用两个不同选词函数共同选词,使所选词条均匀分布性更好;以定量为原则的收词方法客观真实地反映了社会实际用词的规律,尽可能地避免了传统主观方法建立词典时的不足;采用联想的定性方法做为定量标准的补充,使《常用词表》中词条更加完整避免和减小了在词频统计中由于分类、选材、抽样、分词等引起的背景干扰。《常用词表》收词规范、收词频率高、覆盖率高,为“现代”各个时期、各个专业所通用。经验证,覆盖率在98.5%以上。

一、引 言

近些年来,随着我国汉语言信息处理研究的开展和深化,研究的重点逐渐从字的信息处理转移到词、短语、句子和篇章等的处理。语言中有意义的、可独立运用的最小单位是词,自然语言理解、机器翻译等都以词作为基本处理单位,汉字输入、汉字自动识别等也需有词的介入才能更好地提高其效率和性能。

制订《信息处理用现代汉语常用词词表》(以下简称《常用词表》)的目的就是按照《信息处理用现代汉语分词规范》(以下简称《分词规范》),根据定量原则为主,定性原则为辅的选词原则,为汉语信息处理提供一个常用词词表。它对汉字编码、汉字识别、语音识别、汉外翻译、汉语言理解等均有非常重要的意义。[34]把现代汉语常用词词表的建立列为六大问题之一。

《常用词表》首次提出选词函数的术语,并创造性地使用两个不同选词函数共同选词,使所选词条均匀分布性更好;以定量为原则的收词方法客观真实地反映了社会实际用词的规律,避免了传统主观方法建立词典时受专家文化素养、专业学科、社会地位、个人爱好和用

本文1991年5月13日收到

词习惯等影响的不足；采用联想的定性方法做为定量标准的补充，使《常用词表》中词条更加完整，也避免和减小了在词频统计中由于分类、选材、抽样、分词等引起的背景干扰。使用双选词函数和以联想为辅助的方法建立常用词表在我国尚属首创，在国外也尚未发现。

《常用词表》具有如下特性：

1. 为“现代”各个时期所通用；
2. 为各个专业所通用，增加各个专业学科的基础术语，就可形成各专业学科的信息处理电子词表；
3. 为人们经常使用；
4. 收词频度高，覆盖率高；经验证，覆盖率在98.5%以上。
5. 收词量适当，一级常用词表收词 6994 条，二级常用词表收词 27970 条，共 34964 条，此外有单字词条 3522 条。
6. 收词规范，词条均符合《信息处理用现代汉语分词规范》（报批稿）。

1987年以北京航空航天大学为主签订了“信息处理用现代汉语常用标准词库的研究”的合同。合同规定在“七·五”期间制订《常用词表》。签订合同后，成立了由北京航空航天大学 and 燕山公司系统部等单位组成的制订组开始了制订工作。在三年的时间里，制订组经过讨论研制计划、收集词条、讨论分词规范、确定选词标准、对词进行逐条筛选等大量工作，现在终于制订了该词表。

《常用词表》的制订过程如下：

1. 确定制订《常用词表》的原则
 - a. 明确该词表的用途、目的及服务对象。
 - b. 确定该词表应具备的性质。
2. 确定词表的构成和要达到的性能指标
 - a. 分析现代汉语言词汇系统的构成及其特点。
 - b. 确定词表的构成。
3. 建立原始词表
 - a. 明确收词原则。
 - b. 确定词表词源。
 - c. 建立原始词表。
4. 确定选词标准，构造初表
 - a. 分析评价现有选词标准。
 - b. 建立新的选词标准。
 - c. 建立初表一稿，制订组内部评审。
 - d. 建立初表二稿，制订组内部评审。
 - e. 建立初表三稿，制订组内部评审。
 - f. 建立初表四稿，制订组内部评审。
 - g. 形成征求意见稿。
5. 评审征求意见稿
6. 验证征求意见稿

- a. 选择验证方法。
 - b. 验证征求意见稿。
 - c. 分析验证结果。
7. 修改征求意见稿, 形成鉴定稿

二、指导思想

2.1 符合《分词规范》

《常用词表》中的词条经严格审查, 均符合《信息处理用现代汉语分词规范》(报批稿)。因而《常用词表》收取的都是分词单位。在不引起混淆的场合, 我们也称它们为词。为方便信息处理, 实际上《常用词表》中可以包含一些常用固定词组[35]。在初表一稿、二稿中都选取了一些高频词组, 例如“一种, 一个”。经仔细分析, 这些词组分开后均可由计算机方便地分析和处理, 并不会产生较大的影响, 故在第三稿和以后各稿中一律未收。

2.2 以定量为主, 以定性为辅

目前在编纂词典中有以下收词标准:

1. 主观标准

主观标准是作者根据自己的学识、经验和兴趣, 主观判断一个词是否重要、是否有用。从三千年前古代巴比伦楔形文字里保存的最早词表, 到世界上如今名目繁多的词典和词表, 包括我国目前出版的辞书, 绝大多数都是这样收词的。主观标准在编纂辞书曾起过、也正在起着很大的作用。由于专家判断力的权威性, 这样的词表一般较易为公众所接受。它依靠一个人或多人的经验完成, 不需要进行大量的统计工作。但是这种方法的收词受到个人文化素养、学科专业、社会地位、生活爱好习惯的局限和影响, 使收词有较大的个体随机性, 缺乏统一的客观标准。此外客观上人们记忆有限且容易迅速更新, 这就更增添了词表的局限性。

2. 频率标准

频率标准指按词的使用频率决定其重要与否。频率越高, 词越重要。这种方法是由德国人克定首先在编纂《德语词频率词典》中使用的。

3. 分布率标准

加拿大学者范德·贝克在1935年首次采用分布率标准。他认为, 一个词如果有5位专家各使用一次, 比1位专家使用10次的词更为重要, 他以一个词在取样中出现的篇数多少来实施这个标准。在此之后, 分布率进一步发展为不同学科(语体)、不同时期等的分布率。

4. 联想性标准

法国学者米诞阿最早提出“易联想性”的概念。

以上第二、第三两种可以统称为统计法。统计法建立在定量的基础上, 它所依据的结果不以人们意志为转移, 收词标准统一, 且客观和严格。当然词汇的各项统计也额外的增加了工作量。目前许多学科都在由定性研究向定量研究发展。由于计算机的出现使统计法得到了飞速发展, 使统计结果可重复利用, 所选词易于复查, 可不断完善, 因而统计法比主观法有更大的优势, 特别是对那些收词量大, 取舍词差别甚微的情况。

由于在统计过程中存在语料的分类、选材、抽样的背景干扰及汉语分词的失误, 因而统计

数据也存在一定的偏差,这时使用联想性方法可以弥补和校正统计法的不足,使词表更完善。

2.3 常用性

《常用词表》是一个供各专业学科使用的通用性的词表,它包含了汉要民人民现代常用的词汇。有些专业学科的术语,例如“电子、计算机”等由于为整个社会所普遍使用,因而也收录在《常用词表》中。在《常用词表》的选词函数中,常用词的均匀性由十类学科分布率来度量。

常用性反映词表具有很高的覆盖率,而覆盖率反映了词表所收词的动态特性,不能选取过多的低频词。

2.4 稳定性

稳定性用于衡量词在现代中国不同时期的使用情况。《常用词表》所选词是稳定的,在选词函数中,稳定性由现代各时期的分布率来度量。

三、词 条 来 源

《常用词表》所依据的统计数据来自北航等11个单位完成的“现代汉语词频统计”[27][29],辅之词频统计结果[30]、[31]、[32]、[33]。

3.1 “现代汉语词频统计”简介

该项词频统计开始于1981年,完成于1986年,它首先按照表1的计划,从1919年至1949年、1950年至1966年、1957至1967、1977至1982年共四个时期,每个时期又分为自然科学和社会科学共十类学科中选取母体3亿汉字,然后从中抽样2千余万汉字,由计算机分词后分时期分学科进行词频统计。这次词频统计具有选材抽样分布合理、背景干扰小、分词标准一致、统计精度较高等优点。直至现在为止,它仍是国内外规模最大的汉语通用词频统计。制订组根据这次词频统计的数据,构造了两个选词函数,作为定量选词的基础。

这次词频统计的词条来自:

1. 现代汉语词典;
2. 辞海;
3. 汉英词典;
4. 汉法词典;
5. 汉日词典;
6. 标准汉英词典;
7. 常用字构词词典;
8. 列车时刻表(中的全国火车站名);
9. 行政区划表;
10. 国际著名人名录;
11. 现代汉语八百词;
12. 中国地名手册;
13. 世界地名手册;

14. 汉语拼音词汇;
15. 汉语小词典;
16. 成语词典;
17. 汉英小词典;
18. 常用词语用词典;
19. 常用汉字音形教学手册;
20. 外国哲学社会科学人名录;
21. 当代国际人物词典;
22. 世界报刊通讯社、电台译名手册;
23. 现代汉语词表。

表1 “现代汉语词频统计”选材抽样明细表

时 期	社会科学 (S)						自然科学 (N)					
	1 文体生活	2 历史哲学	3 政治经济	4 新闻报导	5 文学艺术	6 社会科学总和	1 建筑运输	2 农林牧渔	3 电子轻工	4 重工工业	5 基础科学	6 自然科学总和
第一时期1919--1949	0.54	1.08	1.61	1.62	2.15	7						
第二时期1950--1966	1.08	2.16	3.22	3.24	4.3	14	1.0	2.0	2.0	2.0	5.0	12
第三时期1967--1976	0.756	1.512	2.254	2.268	3.01	9.8	0.5	1.0	1.0	1.0	2.5	6
第四时期1977--1981	3.024	6.048	9.016	9.072	12.04	39.2	1.0	2.0	2.0	2.0	5.0	12
时期总和1919--1981	5.4	10.8	16.1	16.2	21.5	70	2.5	5.0	5.0	5.0	12.5	30

3.2 添加词

“现代汉语词频统计”共统计出 77482 个词的频度，以此为基础表 1。为完善基础表 1，进行了如下工作：

1. 添加《现代汉语新词词典》相异词条及课题组人员联想 词 条 1391 条，形成基础表 2。
2. 添加《现代汉语词典》中频度为零的词条 16655 条，形成基础表 3。
3. 添加北京语言学院的词频统计结果 [30] 中的相异词条 1202 条，形成初表一稿，共有词 条 96730。

初表一稿中有部分词条无词频数据，部分词条只有参考词频数据（[30] 的数据），为统一起见，这些词条以下一律称之为联想词条。

3.3 加工处理

1. 初表一稿包括了众多的专有名词，为统一处理起见，经审查去掉了专 有 名 词 8000 条。
2. 规范化处理，挑出不符合《分词规范》的词条。
3. 对剩下的频度为零的 11980 个词，按主观标准根据常用程度分级，第一次分五级，第一、二级较常用，共 9000 余词条。第二次把第一级的四千余词条进一步分为三级，全部零

频度词分为六级，其中第一级作为一级常用词词表的补充词条，第二级作为二级常用词词表的补充词条，一、二级分别为 2674 和 2451 条。

4. 联想。在整个工作过程中，制订组成员不断进行了词条联想。联想方式包括主题联想、同类联想、反义联想、同结构联想及异结构联想等。

四、词表的构成

《常用词表》由一级常用词表（6994条）、二级常用词表（27970条）和单字词表（3522条）组成；另外还有12个专有名词附表做为可选词表由用户根据不同领域和应用自行选择。词表中的每个词条有序号、词条、频度、两个选词函数值、来源等信息，以方便使用者的使用。

4.1 一级常用词词表

一级词表有词条 6994 条，其中包括

1. 由选词函数 Z 和 T 共同选出的一级词表部分 6046 条。
2. 只被选词函数 Z 选出的一级词表部分（经联想处理）。399 条。
3. 只被选词函数 T 选出的一级词表部分（经联想处理）。284 条。
4. 联想词中的一级词 128 条。
5. 由二级常用词词表中联想提高到一级的词 137 条。

4.2 二级常用词词表

二级词表有词条 27970 条，其中包括：

1. 由一级常用词词表下调到二级的部分 2765 条。
2. Z 函数和 T 函数同时选中的二级词 22657 条。
3. Z 函数单独选中的二级词（经联想处理）30 条。
4. T 函数单独选中的二级词（经联想处理）1383 条。
5. 联想词中的二级词 1735 条。

4.3 单字词条 3522 条

4.4 专有名词附表

专有名词包括如下内容：

1. 中国地名部分，包括：
 - 32个省、市、自治区全名，简称。
 - 地区级行政区名；
 - 县级行政区名；
 - 名山大川，著名旅游胜地；
 - 重要建筑，地点等。
2. 世界地名部分，包括：

- 世界各国国家全名、简称、首都名；
- 著名城市；
- 名山大川，各大洋、海，著名旅游胜地；
- 重要建筑、地点等。
- 3. 中国各民族名，民族语言名（部分少数民族使用汉语，无自己的民族语言）。
- 4. 外国各大民族名，民族语言名。
- 5. 中外名著名，各类文学作品名，刊物名。
- 6. 中外著名人物名。
- 7. 中外组织，机构名，包括新闻社、出版社、电台、电视台等，中国行政机构及团体名。
- 8. 节日、年号、二十四节气名。
- 9. 知名产品名，如柯达、可口可乐、飞鸽等。
- 10. 其它专有名词。

专有名词的选取不应全部依据定量标准。由于某类事件的影响及抽样选材背景干扰，某个人名或地名等有时频度出其的高。对于上述1、2、3、4、7、8的选取应主要依据分级标准。为了方便不同的应用，在《常用词表》后增加了如下专用词表：

1. 世界地名，包括世界上200个国家的全名、简称及首都名。
2. 中国省、市、自治区名及省会名共59个。
3. 中国地区、县名（包括地区级市及县级市）2368个。
4. 中国56个民族名。
5. 中国52个主要山峰名。
6. 中国4个主要高原名，5个主要盆地名，3个主要平原名。
7. 中国45个主要江河名。
8. 中国13个主要湖泊名。
9. 中国4个主要近海名，3个海峡名。
10. 中国18个主要岛屿名。
11. 中国17个主要关隘、山口名。
12. 中国8个主要宗教名。

以上专用词表在使用时，除31之外需要按照《分词规范》将专名与通名切分开。

24节气、重大节日、六种外国主要货币名称、主要外语均按同类联想方式列入二级常用词表。

其它专有名词因使用甚不稳定，没有选录。

4.5 选词函数

根据“现代汉语词频统计”的选材、抽样和统计数据，在进行大量分析统计后，构造了两个选词函数，使选出的词条更均匀，更符合社会常用词的实际分布。

$$Z(w) = Z_t(w) \times \log_2(Pd(w)) \times T(w) \times J(w) \times k(w) \times D(w)$$

$$T(w) = Z_t(w) \times \log_2(Pd(w)) \times S(w) \times N(w) \times K(w)$$

$$T(w) = Z_t \times (10 \log_2 Pd + stn) + k$$

其中： $Z_t(w)$ ：词w覆盖的子样本数。“词频统计”语料分为4个时期10类学科，故子

样本数最多为40。

$Pd(W)$: 词 W 的出现次数。

$S(W)$: 词 W 覆盖的社会科学类子类数, 此处其值分布在 0 到 5 之间。

$N(W)$: 词 W 覆盖的自然科学类子类数, 此处其值分布在 0 到 5 之间。

$$T(W) = \begin{cases} 3 & \text{如果 } S(W) + N(W) \geq 9 \\ 2 & \text{如果 } 6 < S(W) + N(W) \leq 8 \\ 1.5 & \text{如果 } 5 \geq S(W) + N(W) \geq 4 \\ 1 & \text{如果 } S(W) + N(W) \leq 4 \end{cases}$$

$$K(W) = \begin{cases} 1.4 & L(W) \geq 4 \\ 1.2 & L(W) = 3 \\ 1 & L(W) = 2 \end{cases}$$

这里表示词 W 包含的汉字个数。

$$J(W) = \begin{cases} 3 & 25 \leq It(W) \\ 2.5 & 20 \leq It(W) < 25 \\ 2 & 15 \leq It(W) < 20 \\ 1.5 & 10 \leq It(w) < 15 \\ 1 & 1 \leq It(W) < 10 \end{cases}$$

$$D(W) = \begin{cases} 3 & [s(w) + n(w) > 2][zt(w)/st(n) > 2] \\ 1 & \text{其它情况} \end{cases}$$

五、几个问题的讨论

5.1 单字词

单字词的频率占词频的50%以上, 因而单字词应当在《常用词表》中占据重要的地位, 由于以下原因, 在《常用词表》征求意见稿中没有列入单字词:

1. 由于汉语用字及用词的灵活性, 任一个单字在特殊情况下都可以单独出现, 因而一个实用的词表应当包含基本的汉字集, 而不管它们是否为词。例如, “澡、讯”有许多语言学家认为它们不是词, 但是在下列例句中它们可单独出现, 如不将其收入词表, 将难以进行后续处理:

这个澡真热乎。

新华社3日讯。

2. [40]认为, 把语素全部列入词典, 在标注信息时给予词、离合词、粘合词、词头词尾、其它单音节语素以不同的标志是一种好的办法。在信息处理时, 这种方法仍不失为一种好的处理方式。

单字词和单字语素的划分是一个有较多争论的问题, 目前尚无一致的划分结论。我们在以前的工作中曾列出一个单字词词表[26]。为方便使用, 我们在《常用词表》正式稿中增加了一个有 3500 余词条的单字词条表。

5.2 收词数量

1. 汉语词与字不同。汉字是一个闭合的系统,其数量是有限的。汉语的词则是一个开放的集合,其数量可以说是接近无穷的。2500个常用字覆盖率可达99.63%;3500个常用字的覆盖率则达99.97%[42];GB2312-80中有汉字6763个,覆盖率在99.99%以上[43]。但从表二可以看出,频度最高的前六万词覆盖率只能达到99.8188%。截止覆盖率为90%的近8000词中,低频词的频率已不到0.002%。在9500条词之后,增加1000条词仅能扩大不到0.8%的覆盖率。把覆盖率截止到92.6994%,此时词条数为10500,去掉单字词后的词数为6900。[30]的统计结果高频词比[27]要集中一此,前9000高频词覆盖率为95.84%。

表2 汉语词覆盖率变化表[27]

词种数	覆盖率	增加词	增加覆盖率
500	53.3116		
1500	70.2003	1000	16.8887
3500	81.7676	2000	11.5673
5500	86.8745	2000	5.1069
7500	89.9086	2000	3.0341
10500	92.6994	3000	2.791
15000	95.1050	4500	2.1056
49065	99.0000	14065	3.8950
60000	99.9188	10835	0.8188
77482	100	17482	0.0812

表3 出现词种数表(频率大于等于0.001%) [27]

音节数		一		二		三		四		五		一~五		二~五
词种	数/累频	CK	LP	CK	LP	CK	LP	CK	LP	CK	LP	CK	LP	词种数
子类														
N * *		2464	56.77	6693	36.69	985	2.23	422	0.57	67	0.09	10631	96.35	8167
S * *		2975	54.52	8264	37.72	669	1.49	437	0.88	53	0.09	12398	94.70	9423
* * 1		2696	54.63	7920	39.99	522	0.01	230	0.0026	23	0.002	11664	94.63	8695
* * 2		2891	54.85	8210	37.90	854	1.58	418	0.74	50	0.06	12423	95.13	9532
* * 3		2727	56.84	8103	35.43	1036	2.32	546	0.11	88	0.15	12500	94.85	9773
* * 4		3069	53.14	8733	38.74	678	1.72	450	0.70	62	0.08	12992	94.38	9923
* * *		3081	55.25	8572	36.56	901	1.73	431	0.72	55	0.08	13040	94.34	9959

注:CK表示词种数;LP表示累计频率;N**表示自然科学样本;S**表示社会科学类样本; **1、**2、**3、**4表示四个时期;***则表示总结结果。

表4 科目状态数分布表

词种 时期	状态 数	1	2	3	4	5	6	7	8	9	10
一	T_1	1277	682	458	513	1400					
	T_2	1450	835	611	537	472	362	367	344	409	875
	T_3	1356	723	508	423	356	259	257	235	267	621
	T_4	1163	925	738	573	549	477	432	441	512	1339
	T	972	817	678	614	538	522	507	500	590	1873
二	T_1	10989	4269	2104	1349	1141					
	T_2	12966	6259	3990	3005	1860	1310	970	823	712	1435
	T_3	11205	4734	2891	1759	1205	759	544	484	638	21
	T_4	12596	7523	5188	3950	2970	2003	1503	1280	1223	2141
	T	11743	8083	6085	4995	3963	2738	2066	1703	1717	3688
三	T_1	1355	277	87	25	39					
	T_2	2775	1245	605	257	180	87	57	49	37	47
	T_3	2773	786	378	190	97	51	27	21	15	13
	T_4	4664	1855	930	546	298	171	110	80	79	63
	T	5478	2362	1301	891	487	235	187	156	115	135
四	T_1	1187	202	38	13	8					
	T_2	2849	1182	557	307	124	46	27	23	24	12
	T_3	2416	781	349	122	59	23	20	10	7	1
	T_4	3929	1758	1009	532	273	143	67	70	43	33
	T	4303	2160	1364	922	472	234	131	90	81	79
五	T_1	95	8	3	0	0					
	T_2	404	145	49	32	11	0	1	0	1	0
	T_3	319	76	27	7	2	3	0	1	1	0
	T_4	704	202	78	34	9	5	3	2	3	2
	T	836	326	129	67	29	13	3	3	5	2
六	T_1	36	5	1	0	0					
	T_2	152	70	29	18	7	3	1	1	1	0
	T_3	162	36	11	3	4	2	0	2	0	0
	T_4	342	108	42	16	5	3	1	3	2	2
	T	393	152	61	35	12	10	2	5	3	2
七	T_1	10	0	0	0	0					
	T_2	47	17	6	1	4	1	1	0	0	1
	T_3	47	7	3	1	0	0	0	0	0	0
	T_4	103	36	13	4	0	1	1	0	1	0
	T	118	57	16	7	3	2	1	0	1	1

2. 各专业学科和各时期用词。根据[27]的结果, 尽管在各时期, 科目的选材总量不同, 但各时期总的用词情况变化不大; 自然科学则比社会科学集中一些。但主要差别在低频词上, 对于频率大于等于万分之一的词条, 累计频率都在95%左右, 用词数在9000左右, 详情见表三。

对于频率大于等于0.00005%的词条, 累计频率在99%, 各时期和学科用词数大多在34000左右。

3. 各专业学科的分布状态。从表4可以看出, 在十类科目中专业学科状态数为9.10时的词条段为8293 (不包括单字词为6000); 在**2、**3、**4、***中对于专业学科状态数大于6的词数分别为6218 (不包括单字词为5825)、3168 (不包括单字词为2644, 第三时期为文革时期, 抽样字数较少)、9435 (不包括单字词为9037)、13665 (不包括单字词为13407)。

4. 综上所述, 《常用词表》以收词34000条左右较合适, 其中一级常用词词表以7000条左右较合适。根据实际选词函数的运算及联想, 最后一级常用词选定6994条, 二级常用词27970条。

六、词表验证

验证是制订《常用词表》的最后一项工作, 制订组直接引用新华社词频统计结果对《常用词表》进行了验证工作。

6.1 验证材料

用于《常用词表》验证的新闻语料词频统计底表含词147955条, 词出现的总频度为7455171次。详细情况见附表7。

附表7 新华社词频统计分布表

音节数	1	2	3	4	5	6	7	8
词数	8173	64204	30297	32405	6476	3478	1387	1535
频度	3523270	2980898	471797	393630	51306	26292	7896	142

6.2 验证结果

《常用词表》中所收均为多音节词, 验证时删去了词频统计结果中的单音节词及其频度。统计的直接结果如下

6.3 结果分析

制订组对验证的初步结果进行了分析, 发现影响覆盖率的主要原因有两种:

1. 收入了大量不属于《常用词表》收词范围的词, 象人名、地名等。如蔡炎书出现20次, 梁希30次, 李鹏3505次。

	新华社词频统计
一级词数	6582
一级词覆盖率	48.84%
二级词数	23842
二级词覆盖率	9.54
总覆盖率	58.38%

2. 收入大量不符合分词规范的词。如裁军谈判出现 46 次、裁军问题 41 次、裁军协议 24 次、裁军谈判会议 70 次、财产损失 58 次、财务管理 22 次等。

针对上述情况,制订组对词条进行了规范化,得出最终结果如下:

一级词覆盖率	89.15%
二级词覆盖率	10.15%
总覆盖率	99.3%

参 考 文 献

- [1] 制订组, 信息处理用现代汉语分词规范, 第一稿至第(报批稿)七稿, 1988年至1990年。
- [2] 制订组, 《信息处理用现代汉语分词规范》编制说明, 1990年9月。
- [3] 中国社科院语言研究所词典编辑室, 现代汉语词典, 1983年, 商务印书馆。
- [4] 编辑委员会, 辞海, 上海辞书出版社, 1982年。
- [5] 编辑委员会, 《辞海》增补本, 上海辞书出版社, 1982年。
- [6] 编写组, 汉英词典, 商务印书馆, 1979年。
- [7] 汉法词典。
- [8] 汉月词典。
- [9] 标准汉英词典。
- [10] 常用字构词词典。
- [11] 列车时刻表。
- [12] 行政区划表。
- [13] 国际著名人名录。
- [14] 现代汉语八百词。
- [15] 中国地名手册。
- [16] 世界地名手册。
- [17] 汉语拼音词汇。
- [18] 汉语小词典。
- [19] 成语词典。
- [20] 汉英小词典。
- [21] 常用词语三用词典。
- [22] 常用汉字音形教学手册。
- [23] 外国哲学社会科学人名录。
- [24] 当代国际人物词典。

(下转第25页)

的话语模型还有待进一步扩充和完善;

(3) 篇章小结结构的模型,这是用 TR 链 结构的主题推进顺序来保证话语段的前后连贯。

参 考 文 献

- [1] 李东、黄昌宁:功能合一语法在汉语生成中的应用,《1990年国际中文与东方语言计算机处理学术会议论文集》,1990年4月,长沙。
- [2] 李东、黄昌宁:基于合一的汉语生成,《论文集:中国人工智能90》,1990年7月,长春。
- [3] 王福祥:《汉语话语语言学初探》,商务印书馆,1989年,北京。
- [4] 沈开木:《句段分析(超句体的探索)》,语文出版社,1987年,北京。
- [5] McKeown, K. R., Discourse Strategies for Generating Natural Language Text, Artificial Intelligence, Vol. 27, No. 4, Oct.-Dec., 1985.
- [6] Paris, C. L., Combining Discourse Strategies to Generate Descriptions to Users along a Naive/Expert Spectrum, Proceedings of IJCAI 87, vol. 2, 1987.

Chinese Discourse Generation Based on Planning

Huang Changning, and Li Dong

(Department of Computer Science and Technology,
Tsinghua University, Beijing 100084)

Abstract

Discourses, also called sentence group or super-sentence, is a set of sentences which are coherent on meaning and structure. Discourse generation is a planning procedure, it has to solve such problems as "what to say?" and "how to say?" within given situation of communication. This paper introduces a Robotic Discourse Generation System (RDGS)* based on planning. The planning procedure is implemented by a Discourse Model for deep planning and a Theme Promotion Pattern for surface planning.

(上接第37页)

- [25] 世界报刊通讯社、电台译名手册。
- [26] 现代汉语词表。
- [27] 研制组,现代汉语词频、字频统计鉴定材料,1986年5月。
- [28] 北航研制组,梁南元执笔,现代汉语自动分词系统PC-CWSS技术报告,1990年6月。
- [29] 刘源、梁南元等,现代汉语常用词频词典,宇航出版社,1990年。
- [30] 王还、常宝儒等,现代汉语频率词典,北京语言学院出版社,1986年。
- [31] 北航、新华社等,新闻语料词频统计,1990年。
- [32] 北航、国家信息中心,三百万字经济法频统计,1989年。
- [33] 词频统计,人民大学等。
- [34] 柳维专,我国汉字信息处理技术发展现状及今后的任务,计算机世界,1983年3月15日。
- [35] 张国防,陶沙,信息处理用中文词库系统的理论和实践, ICCIP'87, 1987年。
- [36] 张拱贵,常用词表编制中的若干问题,辞书研究,1989年第6期。
- [37] 张社英,王德进,关于建立《信息处理用现代汉语通用词表》的若干问题, ISSCIP'89, 1989年3月。
- [38] 郑林曦等,普通语三千常用词表,文字改革出版社,1959年。
- [39] 现代汉语新词词典。
- [40] 郝恩美,语词词典如何处理语素问题,辞书研究,1989年第1期。
- [41] 吕叔湘,汉语语法分析问题,商务印书馆,1979年。
- [42] 国家语委汉字处,现代汉语常用字表,语文出版社,1988年。
- [43] 华北计算技术研究所,《信息交换用汉字编码字符集基本集》编制说明,1980年。
- [44] 信息处理用现代汉语五千词表。
- [45] 许欣文等,工程实用词语库的设计学实现。
- [46] 张社英, BH-CWD词表、BH-86软件系统及汉语的几个统计规律,北京航空航天大学硕士论文,1986年。
- [47] 谭强,制订《信息处理用现代汉语常用词表》的理论和方法,北京航空航天大学硕士论文,1991年。