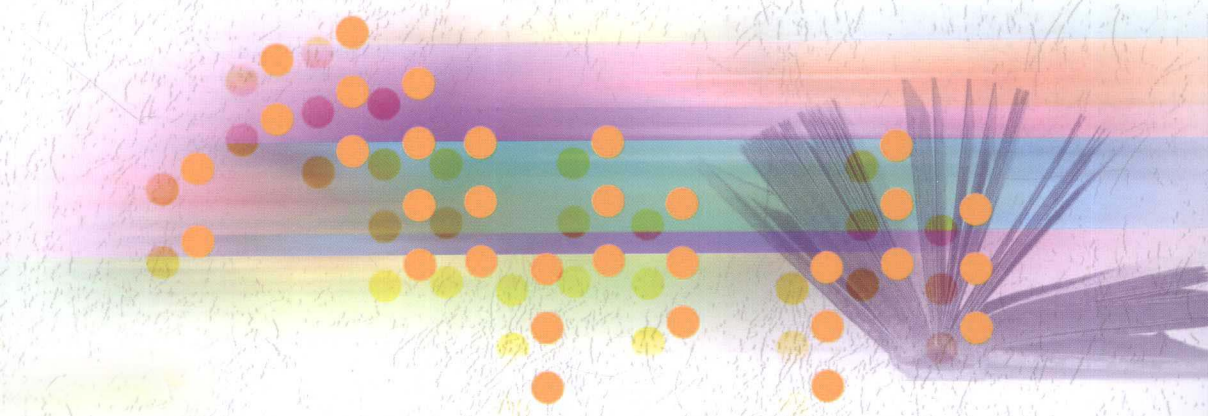
汉语理解处理中的动态词研究

唐兴全 ◎ 著



科学出版社

汉语理解处理中的动态词研究

唐兴全 著

科学出版社

北京

内 容 简 介

动态词的组合识别和语义认定是全局理解的基础。动态词研究将帮助提高句类分析技术的水平，对正处于句处理阶段的中文信息处理有重要意义，对人用或机用的各类词典的编纂也具有指导作用。本书在 HNC (hierarchical network of concepts，概念层次网络)理论概念基元符号体系和句类体系的基础上，重点研究了动态词的性质、组合模式以及描述方法；探究了动态词内部的构成方式以及动态词对句类分析的影响；介绍了字小专家研究的重要价值；并研究了字小专家“得”的用法。

本书可供语言学及应用语言学等相关专业人员参考。

图书在版编目(CIP)数据

汉语理解处理中的动态词研究 /唐兴全著. —北京：科学出版社，2012.1

ISBN 978-7-03-033188-5

I. ①汉… II. ①唐… III. ①汉语－词语－研究 IV. ①H136

中国版本图书馆CIP数据核字(2011)第277510号

责任编辑：胡 凯 龚 勋 / 责任校对：赵桂芬

责任印制：赵 博 / 封面设计：王 浩

科学出版社出版

北京东黄城根北街 16 号

邮政编码：100717

http://www.sciencep.com

源海印刷有限责任公司 印刷

科学出版社发行 各地新华书店经销

*

2012 年 1 月第 一 版 开本：B5 (720 × 1000)

2012 年 1 月第一次印刷 印张：12 3/4

字数：310 000

定价：48.00 元

(如有印装质量问题，我社负责调换)

序

我是书文的热心翻阅者或浏览者，逐步形成了一种积习。即使是需要我写序的新书，也是如此对待。但这一次，这个积习终于被打破了。那是由于翻阅到唐兴全博士新著《汉语理解处理中的动态词研究》中的一段话，拷贝如下。

动态词的辨识是汉语理解处理中的必需模块。动态词在汉语中占有很高的比例，一个信息处理用的词表，不管规模多大，都不可能穷尽所有的词，因此，动态词是语句分析的重大障碍。动态词研究将帮助提高句类分析系统的效率和准确性，对于句处理阶段的中文信息处理有重要意义。动态词的组合识别服务于语义块感知和句类假设、语义块构成分析等句类分析的主要环节，对最终解决汉语的计算机理解问题具有主要意义。动态词组合模式具有较强能产性和规则性，动态词的识别与语义认定应该建立在对动态词组合模式详尽描写的基础上。如果我们搞清楚了能产的组合模式，就可以根据这些模式来处理在言语中随时可能出现的由这些模式所造成的动态词的语义。因此，应根据动态词组合模式的不同特点，制定不同的识别策略(见该书第 3.3 节)。

这段话很有一点康德先生所提倡的理性法官气势，它推动着我往下查看其落实情况。于是，我细读了该书的第四章。结论是：作者对汉语动态词组合模式的三大类划分(“以概念类别为纲”、“以概念组合结构为纲”和“特殊组合模式”)是高瞻远瞩的，第一大类的四小类划分是明智的；第二大类的八小类划分是精巧的；第三大类的三小类划分是周到的。大部分论述符合康德先生要求于理性法官的透彻性和齐备性标准。这使我非常惊异，如同在茫茫大漠里遇到了一片绿洲。

于是，我接着细读了该书的第五章，一路惊喜。因为这十多年来出现了关于语言本体的熙攘浪潮，但是，那语言本体之“本”究竟何在？那熙攘浪潮对这个根本问题实质上是采取回避态度的。现在，一本敢于正视这一根本问题的著作终于出现了，这是语言信息处理领域的一件大事！所以，我原本打算向作者写点热烈祝贺的话。

但是，最后我还是决定引而不发。为什么？因为仅仅立足于语义思考，毕竟还不能达到一流语言理性法官的高度。动态词识别的终极解决方案必须立足于语境分析，即语言理解基因与动态记忆的激活。词语的语义分析是语言理解处理的核心环节么？人类语言脑的语言理解处理过程存在独立的词语处理和句处理阶段

么？作者是完全有潜力完成这一重大思考的，从而在第五章与第三章之间建立起更好的呼应，期待作者再接再厉。

然而，即使是当前的版本，它已经是同类著作中最优秀的了，对此，我没有任何疑义。

黄曾阳

2011 年 6 月 9 日

于北京阑珊庐

前　言

汉语理解处理需要在概念层面进行，概念层面关注的是语言成分连用后的内容效应，而非形式效应。汉语理解处理系统在将待处理语串与词库匹配后，会剩余一些孤零零的单字。由于大多数汉字的用法复杂，义项众多，因此这些汉字在语串中的意义、地位和用法都需要进一步确定才能确保达到对整个语串的理解。其中一部分单字需要与其邻接成分临时、动态组合在一起整体构成一个语义单位来理解。这个语义单位就是动态词。动态词是在汉语理解处理过程中，根据语义理解需要，单字跟与其邻接的两个或多个字词按照一定组合模式临时组合而成的语义单位。动态词可以是一个多层嵌套组合的复杂结构，本书所研究的动态词限定在组合层次在两层以内，长度不超过五个字的范围内。

本书对动态词的研究、分析是以 HNC 理论为指导进行的。HNC 理论是一个关于自然语言理解(natural language understanding，NLU)处理的理论体系，由中国科学院声学研究所黄曾阳先生创立。

本书的研究素材主要来源于动态词标注语料。本书以 HNC 理论概念基元符号体系和句类体系的相关理论为基础，重点研究了动态词的性质、组合模式以及描述方法，探究了动态词内部的构成方式以及动态词对句类分析的影响。全书除绪论和结语外，共五章，主要根据动态词现象描述、动态词组合模式研究、服务于动态词识别处理的规则描述与字库建设三部曲行文。

绪论部分说明了动态词研究的具体内容和研究目标。动态词的组合识别和语义认定是全局理解的基础。动态词研究将帮助提高句类分析技术的水平，对正处于句处理阶段的中文信息处理有重要意义。动态词的研究对于人用或机用的各类词典的编纂也具有指导作用。绪论部分还介绍了语言学、中文信息处理、HNC 理论对动态词进行的相关研究。

第一章“研究背景和 HNC 理论简介”概要介绍了对自然语言理解与中文信息处理的认识，论述了中文信息处理的现状和难点等。同时，对 HNC 理论和语言理解技术进行了简要介绍。

第二章“现代汉语中的动态词”，给出了动态词的定义、界定标准及动态词标注工作，并探讨了汉语信息处理用词库收词的指导方针。动态词是在汉语理解处理过程中，根据语义理解需要，单字跟与其邻接的两个或多个字词按照一定组合模式临时组合而成的、需要经过辨认处理作为一个意义单位理解的语义单位。本章还说明了动态词与基本概念短语、缩略、重叠等形式的区别，并从语义、结构、

韵律、词长、频率等方面对动态词进行了界定。本书对动态词的研究素材来源于真实语料。本章还介绍了作为动态词分类标准的 HNC 理论的概念类别与概念组合结构，并对动态词标注工作作了说明，揭示了不同概念组合结构的动态词的分布情况。本章还探讨了信息处理用词库收词问题，指出汉语信息处理用词库的规模不是关键，词库应描述中文字、词的组合规律。

第三章“动态词的组合模式与识别策略”给出了动态词组合模式的定义，说明了动态词组合模式与构词法和造词法的区别，并将动态词组合模式分为形合模式和意合模式两种类型。形合模式是指在组合模式中有标志性的单字，以该字为核心，能形成一系列词内组合关系和组合整体效果(语义层面和词法层面)相同的词，而且标志性单字的左向或右向组合成分的替换率应达到一定的数量规模。意合模式是指模式中没有标志性单字但在组成成分的概念类别和概念内涵以及内部组合关系上有规则可循的组合模式。本章在将动态词组合模式二分的基础上，提出了动态词识别处理的规则描述与字库描述两种策略。

第四章“动态词组合模式分类研究”在动态词标注语料的基础上，分别以 HNC 理论的概念类别和概念组合结构为纲，深入分析、刻画了动态词的组合模式，并对部分组合模式进行了规则化的描写。以概念类别为纲，对表时、空、数、量四类基本概念的动态词进行了分析；以概念组合结构为纲，分析了不同概念组合结构下的动态词组合模式。其中，将联合式区分为对比性联合、对偶性联合、包含性联合、平行性联合和搭配型联合四类；从概念组合关系和句类的角度，深入细致地刻画了作用式动态词的构成。另外，本章中还对汉语活跃语素构词、待嵌格式、否定式组合等特殊形式的动态词进行了分析。

第五章“字知识库建设与动态词识别”讨论了服务于汉语理解处理的 HNC 单字知识库建设方法、建设内容。在字知识库建设实践的基础上，本章探讨了单字知识库的描述对象与中心内容，说明了单字组合知识描述的要点，并对字库对单字组合能力的描述方式进行了详细说明；介绍了字小专家研究的重要价值，并详细描写了“W 得”字小专家的组合知识。

结语部分对本书的内容进行了总结，并展望了动态词研究下一步的工作。

本书最后附录了 HNC 语义网络节点表、HNC 句类表示式、语句格式代码表等内容供读者参考。

目　录

绪　论

0.1　研究内容与研究目标

汉语理解处理需在概念层面进行。汉语理解处理系统接收一个待处理的语串，首先会逐一扫描并与词库相匹配，词库中存在的多字词将先行识别。我们先看下面几个句子的分词结果[以《现代汉语词典》(2005 年版)(以下简称《现汉》)词表为分词底表，词间用“/”标记]：

①上/周/艺术/界/的/同行/们/相/聚/一起/共/谋/发展。

②今年/是/中/日/建交/第/三/十/五/年/。

③他/把/东西/搬/出/了/老/张/的/房间/。

④他/昨/晚/在/李/老师/家/里/大/闹/特/闹/。

⑤其间/有/一/个/十一/二/岁/的/少年/，/项/带/银/圈/，/手/捏/一/柄/钢/叉/，/向/一/匹/獒/尽力/的/刺/去/。

可见，经与词库匹配后，语串[①]中会剩下一些孤零零的单字。由于大多数汉字的用法复杂，义项众多，所以这些汉字在语串中的意义、地位和用法都需要进一步确定才能确保达到整个语串的理解。HNC 理论[②]将这些分词碎片的处理称为“孤魂”处理。孤魂的出现表明统计模型不是对人脑自然语言感知模式的恰当模拟，要消除孤魂，就需要计算机掌握自然语言概念联想脉络的恰当激活模式。“孤魂”处理包括“合”与“分”两类，“分”是指那些在语句中独立使用，不必与其他成分组合的单字，如例句②中的“是”。从概念分析的角度看，语言信息处理研究的重点在于如何从小的结构颗粒所表达的概念意义组合成更大的合成结构所表达的概念意义，这种组合分析是关键之关键。对“孤魂”处理而言，最重要的是单字的组合处理，上述例句分词后切分出来的单字大多需要进一步组合才能实现意义的完整。例如，五个例句中的单字还应该进一步组合为：

上周　　艺术界　同行们　相聚　　共谋　　中日　　第三十五年

搬出　　老张　　昨晚　　李老师　家里　　大闹特闹

---

① 语串是句子的下一层次概念。当语句内包含以逗号、分号为分割标记的若干片段时，这些片段称为语串；当语句内不含语串标记时，该语段等于一个语串。语串基本上相当于一般意义的小句。

② HNC 理论是面向整个自然语言理解的理论框架，由中国科学院声学研究所黄曾阳研究员创立，本书第一章将对 HNC 理论简要介绍。以下简称 HNC。

一个　十一二岁　项带　银圈　手捏　一柄　钢叉　一匹　刺去　……

这些都是在汉语理解处理过程中需要动态、临时组合在一起的。类似的组合还有很多，它们不见于一般词典中。信息处理用词库即使规模很大，也只能收录其中很少的一部分。我们将这类组合称为动态组合词，简称动态词。必须强调的是，这里的**动态词并不对词和短语严格区分，仅仅是对动态组合单位的命名。动态词是在汉语理解处理过程中，根据语义理解需要，单字跟与其邻接的两个或多个字词按照一定组合模式临时组合而成的、需要经过辨认处理作为一个意义单位理解的语义单位**。动态词的组合识别和语义认定是全局理解的基础，而动态词的构成描述则是基础的基础。

本书在 HNC 概念基元符号体系和句类体系的基础上，重点研究动态词的性质、组合模式以及描述方法，探究动态词内部的构成方式以及动态词对句类分析的影响，以期为计算机的处理提供帮助。动态词可以是一个多层嵌套组合的复杂结构，本书所研究的动态词限定在组合层次在两层以内，长度不超过五个字的范围内，如：

车门　猪肉　蓝天　大桥　改建　重修　唱歌
打昏　搞好　揭开　成批　逐年　数百　可贵
啤酒厂　上班族　准教师　环境保护署

动态词也可根据词性的不同分为动态动词、动态名词、动态形容词等。

语言学中曾有动态词的提法，但所指语言现象并不一样。有人把“着、了、过、在”称为动态词(邢志群　2003)，认为它们表示动作的行为或状态。很多人在写文章过程中，也使用了动态词的名称，比如说“珍稀动物是个动态词，会随着时间发生变化”，这里的动态词是指词义随着时间变化而变化的词；再如“‘堆起、挺立、擎起’等动态词极优美地体现了诗的语言的张力，而‘浓郁、淡漠、沉着、静静’则是整首诗的外在色彩”[①]，这里的动态词是指词义描述的是动作行为，而非静止的属性、状态等。以上动态词的说法与我们所说的动态词并不是一码事。

我们把需要收入信息处理用词库的词称为应登录词，或称词库词。那么哪些语言成分有必要作为词库词呢？总的来说，模式识别度、使用频度是衡量一个词能否进入词库的主要标准。我们将对词库收词的指导方针和收词原则作简要分析。

以往的现代汉语词汇研究主要是对现代汉语词汇进行抽样式定性描写，所涉及的词条是有限的，更重要的是，对词的结构、意义等方面的研究所针对的对象往往就是规范型汉语词典中已收录的词，所作调查研究也是基于有限的词典条目。

① 世甫. 寂寞沙洲，文化苦旅. 网址：http://book.youren.com/story/95840.html. 2006 年 9 月 4 日访问。

对汉语信息处理而言，研究真实语料中鲜活的、动态的语言组合成分，谋求实现计算机对动态词的识别与语义认定有更大的研究意义和应用价值。

分类是科学研究的基础，动态词有多种组合模式。本书将对动态词的组合模式进行研究，重点关注模式中前后成分之间的语义关系，组合规则及组合结果。对动态动词来讲，应该关注两点：一是两个字词应该合并起来充当同一个特征语义块[①]的构成部分，而不能割裂开；二是两个字词合并起来之后的句类是否与合并之前相同，换言之，句子的语义框架是否发生变化。对动态名词而言，应该关注名词内部成分间是何种关系，名词的语义如何判定。

动态词的识别主要是语句的局部构成分析中的组合过程，但其语义的认定有时则依赖于全局语义分析的结果。本书主要讨论可以通过局部动态组合来认定语义的动态词，重点关注动态词组合模式的规则研究以及面向信息处理的单字知识库在动态词识别处理中的作用。如动态名词“碧山、碧水、碧宵、碧天、碧树、碧草”等应该可以通过将“碧”与后面相邻单字名词或名语素捆绑来处理，这就需要在字库中对“碧”这一组合能力进行详细描述。局部动态组合识别的正确性有时影响到全局语义分析的结果，如类似“播音器、扫描仪、割草机”其中含有动词的动态词如果不能正确识别出来，将对整个句子的语义类型的正确判定造成严重干扰，系统可能会误将动态词中所含的动词作为全局的特征语义块进行句类假设，从而造成分析错误，降低系统的效率。

正确分析动态词的组合模式，形成汉语理解处理系统可以应用的判定规则，提高系统分析的效率和准确率，是我们的研究目标。

0.2　研究意义与价值

语义研究是当前和以后中文信息处理研究的热点和难点，也是中文信息处理从字处理、词处理到句处理阶段突破的瓶颈所在，突破了这一瓶颈，中文信息处理将会迎来新的发展契机。而理解处理过程中的词义认定是句子语义理解的基础和前提。因此，在词库已经对词库词的意义和用法进行了定性描述的基础上，对动态词的组合识别与语义认定进行研究就成了词汇处理中的重要课题，也是从字词处理向句处理阶段过渡绕不过去的一道关口。

由于汉语本身“字义基元化、词义组合化”(黄曾阳　1988)的特点，汉语合成词具有强大而灵活的能产性，动态词在汉语文本中占有很高的比例。一个信息处理用的词表，不管规模多大，都不可能穷尽所有的可能组合。因此，动态词是语句分析的重大障碍。动态词组合模式具有较强能产性和规则性，如果我们对动

① 特征语义块是 HNC 术语之一，大体相当于中心动词。第一章中将有介绍。

态词组合模式进行了详尽的描写，就可以根据这些模式来处理在言语中随时可能出现的由这些模式所造成的组合的语义。在构建一定容量词表的基础上，让计算机贮存一些动态词组合模式规则，可以更好地解决动态词词义理解问题。

新事物层出不穷，对各种新事物的命名的识别是中文信息处理遇到的困难之一。而对于实体的命名往往是依据一定的组合模式进行的，摸清了这些组合模式将有助于提高对命名实体的识别水平。

动态词研究将帮助提高句子语义分析技术的水平，对正处于句处理阶段的中文信息处理有重要意义。动态词的组合识别将服务于语义块感知和句类假设、语义块构成分析等句类分析的重要环节。如果动态词属于特征语义块核心部分，则会影响语义块感知和句类假设[①]。比如 0.1 节中所举动态词“搬出”，“搬”作特征语义块时的句类为基本作用句 XJ[②]，含有 A、B 两个广义对象语义块，而该例句却含有“他”、“东西”、“房间”三个广义对象语义块，这是由于“搬”与“出”的组合作特征语义块造成的。如果计算机已经将“搬”“出”优先组合作为特征语义块，并进行句类假设，则会得到正确的句类检验。如果动态词属于广义对象语义块的核心部分，则会影响句类检验。如果动态词属于语义块的说明部分，则会对语义块构成分析产生影响。动态词是一种依据规则自动组合而成的全新的意义单位，计算机需要在这一单位基础上进行语义块内部的构成分析，获得语义块内部的构成。因此，动态词的组合识别研究对于提高句类分析系统的效率和准确性有重要意义，动态词的辨识就成了汉语理解处理中必需的模块。

动态词的研究对于人用或机用的各类词典的编纂也具有指导作用。各类词典首先应该全面收录那些内部不可分析或内部构成模式不具有能产性的词汇性成分。对于那些能产性较强的组合模式所造成的形式，可以根据词典规模和词典的适用对象适当收录。人用词典因为人的类推和概括能力较强，就可以不收或少收动态词。作为一部规范的人用词典，所收词语条目是否齐全固然极为重要，但是更为重要的是所收词语类别是否齐全。对于某些结构，《现汉》仅收录其中部分组合作为代表，而没有全部收录。如对于“从×”结构（“从”意为采取某种方针或态度，×代表单字形容词），《现汉》收录了“从缓”“从宽”“从权”“从严”“从简”，而没有收录同样常用的“从轻”“从重”“从速”等。而机用词典由于计算机缺乏人的类推和概括能力，却具有较大存储能力，对动态词就可以适当多收，部分能产性弱的组合模式造成的动态词可以全部收录，如“星期”与数字“一”到“六”的组合一共只有六个，可以全部收入词典中。

---

① 句类、语义块、语义块感知及句类假设、句类检验等都是 HNC 理论中的术语，有特定的含义，第一章中将有说明。

② 各句类及其子类的名称、代码和表示式见本书附录二。

0.3 研究方法

以有限驭无限，是训诂学和 HNC 理论在研究方法上的共同之处。训诂学力求以有限的语言现象描述、阐释更多的其他同类语言现象。HNC 则提出“概念无限而概念基元有限，语句无限而句类有限，语境无限而语境单元有限”的“三无限三有限”观点。动态词是一个开放的系统，动态词数量无限，但动态词是一种相邻两字或多字的临时的规律性的组合，其表达的方式和手段又是有限的，因此需要我们用有限的组合模式去概括无限的动态词出现情况。

论文将在材料的获取上采用语料提取与演绎相结合的方法。首先制定动态词标注规范，并依据规范对真实文本中的动态词进行标注。标注的目的是获取动态词的真实素材，这是动态词分类、构成模式、识别处理等研究的基础。对一定规模不同领域的真实语料中的词汇使用情况尤其是动态词使用情况的调查，一方面，可以使所研究的对象更贴近于词汇使用的真实情况，另一方面也可以全面有效地概括动态词的种种情况，为汉语信息处理铺路。但同时，再大规模的动态词标注语料也只是言语的有限片断，仍有一些语言现象无法从标注语料中直接获得，因此我们根据内省增加了一些语料中未出现但却在日常生活中出现的动态词条。对于以汉语为母语者，通过演绎、内省获得对语言现象的判断在语言研究中是必不可少的。

为了考察某类组合模式的能产性，我们借助了《现汉》的收词情况。不同的词典所收条目及收录标准都会有所差异，同一词典的收词也不可能完全符合人们的心理词库或称词感。但规范性的词典收词基本能反映人们的词感。《现汉》收词有其大概的选择标准，一般那些使用不具有普遍性和稳定性的词不会被收录，内部结构透明的词的收录也有所限制。因此，我们以《现汉》收词条目作为词库收录词的范围，考察某一组合模式与词库的关联，进而认定该组合模式是否是动态词的组合模式。

在研究中，我们将采用定性与定量结合、理论导向与材料导向并行的研究方法，依据动态词出现的实际情况进行归类定性描述，将 HNC 理论中有关概念类别、概念组合结构的观点运用到对实际的动态词标注材料的分析中去，同时，也会吸收一些有益的词法理论。

中文信息处理尤其需要从全局俯瞰局部的研究方法。我们认为，中文信息处理长期停留于词处理阶段的现状正是由于究其一点未及全局的研究方法的局限造成的。因此，对动态词的分析应从语句分析全局的角度，从服务于句类分析的目的出发，比如，对动态动词，应考虑它对句类的影响。

0.4 相关研究现状

语言学界、中文信息处理学界和 HNC 学界均对动态词或其相关的内容给予了关注。

0.4.1 语言学及中文信息处理学界对动态词相关问题的研究

应该说，语言学界和中文信息处理学界对动态词缺少专门的研究，对组合模式的研究主要是以词典已收录词为对象进行。和动态词相关的是如何确定词与非词，何为词汇词与语法词，中文信息处理如何认定词，构词研究如何进行等问题。

一、词的定义

词的定义问题一直是困扰汉语言学界的有不少争议的问题，所谓“语言中能独立使用的、最小的意义单位”的定义当放到中文信息处理中时，显得很不具有可操作性。同时，在语言系统的各个子系统中，对词汇系统的研究是公认的语言研究中比较薄弱的环节。这在很大程度上是因为词汇系统内成员众多，情况复杂。

由于看问题的角度不同，对于词的定义曾有“韵律词”、“词汇词”、“语法词”、“形式词”、“理论词”等说法。中外语言学家在词的定义问题上曾经提出了许多见解，可谓百家争鸣。例如，《中学教学语法系统提要》定义为“词是最小的能够自由运用的语言单位”；台湾国立编译馆《国语教学指引》(1988)中认为“词是一个字或两个字以上具有独立意义的”；“把一定的声音去表示某一定的概念” (许世瑛 1976)；语言中最小的意义单位(王力 1984；吕叔湘 1979)；说话的时候表示思想中一个观念的语词……语言中间一个一个观念的表示(黎锦熙 1978)；最小能够填进某功能框架里的空位单位(赵元任 1980)；“词是句中最小的能够独立运用的语言单位”(黄伯荣，廖序东 2002)；词是“意义单纯的语言单位”(刘叔新 1984)； “最小的或基本的造句单位”(高名凯，石安石 1963；张志公 1982)；由一定的语音形式和意义内容构成的一种语言单位，它的声音和意义是相互联系的、统一的(罗肇锦 1990)。在国外，布隆菲尔德认为词是“能够单独成句的单位”(Bloomfield 1933)。

上述定义基本可以区分成从词的意义上、从句子的组织上、从音义结合上三种定义方式或者定义角度。总括来说，其定义包含了一个完整的意义、自由运用和最小造句单位三种含义，在构成上则包含了字数与音节的限制。但在词库工程上，上述定义都很难作为严格的界定某一语言单位是否为词的标准：某一语言形式是不是一个意义单位？词缀算不算独立运用？自由运用是否过多依据使用者的自我判断？