英汉对照
计算语言学
词语汇编

English-Chinese Lexicon
of
Computational Linguistics

俞士汶　　朱学锋
E. Kaske　　冯志伟

北京大学出版社

英汉对照
计算语言学词语汇编

English-Chinese Lexicon of Computational Linguistics

俞士汶　　朱学锋
E. Kaske　冯志伟

北京大学出版社
·北　京·

内 容 简 介

本书收录了从80年代至90年代初国内外出版的相关论著中精选出的计算语言学专用词语，共计5415条。正文按英语词语排序，正文后除附有汉语拼音索引外，还增加了汉语拼音逆序索引和英语轮排索引，这两种索引是目前英汉对照词典中所罕见的，它们为读者查找、整理，以及研究所需要的术语提供了极大的方便。本书收集的词语对计算语言学的覆盖面广，英汉对译准确可靠，且多种索引皆是由计算机自动生成的，这为计算机技术在词典编纂中的应用提供了实例。所以本书既是从事计算机科学、语言学、认知科学和信息管理等学科领域中教师、学生和科技人员的宝贵资料，也是从事辞书编纂、术语开发、文献资料管理等人员的重要参考资料。

书　　名：英汉对照计算语言学词语汇编
著作责任者：俞士汶等四人
责 任 编 辑：邱淑清
标 准 书 号：ISBN 7-301-02883-0/TP·261
出　版　者：北京大学出版社
地　　址：北京市海淀区中关村北京大学校内　100871
电　　话：出版部 62752015　发行部 62559712　编辑部 62752032
排　印　者：北京大学印刷厂
发　行　者：北京大学出版社
经　销　者：新华书店
版 本 记 录：787×1092毫米　32开本　14.125印张　616千字
1996年8月第一版　1996年8月第一次印刷
定　　价：35.00元

序

电子计算机于40年代问世之后不久，人们很快就想到利用计算机进行机器翻译研究，这是数字计算机在非数值领域应用的最早尝试。随着语言信息处理研究的发展，逐步形成了计算语言学这门新兴的学科。计算语言学已成为自然语言人机接口、机器翻译、情报检索、语言文字计量研究等应用技术的理论基础。

我国虽然早在50年代就进行过机器翻译的试验，但是关于计算语言学的比较系统的、规模较大的研究是到了80年代才开始的。1986年中国中文信息学会所属的计算语言学专业委员会成立，有力地推动了我国计算语言学研究的发展。同年，在著名语言学家朱德熙先生的倡导下，北京大学也成立了我国的第一个计算语言学研究所。从那时起，北大的学者在这个领域中辛勤劳作，取得了丰硕的成果。北大计算语言学研究所已成为我国计算语言学研究的重要基地之一。

本书的中国作者都在若干语言工程项目中承担了任务，有着丰富的知识与经验。尤其难能可贵的是，他们在完成繁重的工程任务的同时，始终注意提高自身的素质，因为他们认识到，计算语言学是文理结合的交叉学科，不同基础的研究人员在进入这个领域之后，必须注意知识更新和知识结构的调整。本书的编纂就是他们在

这方面的努力的例证。本书的德国作者 E. Kaske 小姐在北大进修期间，除了完成学业，还对本书的编纂作出了重要的贡献。另外，本书的编纂也得到了日本朋友的帮助。

我很荣幸地于1992年主持了本书的基础研究成果“英、日、德、汉对照计算语言学词语库”的技术鉴定。鉴定委员会的专家们高度评价了该项研究的学术水平和实用价值。鉴定会之后，作者们又对词语库进行了补充与仔细的推敲，并将词语库交给一些单位使用，听取意见。1995年该项成果获得了北京大学科技成果三等奖。现在，以该项成果为基础，北京大学出版社出版两本《计算语言学词语汇编》，一本是英汉对照的版本，另一本是英、日、德、汉4种语言对照的版本。在当前尚找不到一本计算语言学词典的情况下，这两本词语汇编的出版顺应了学科发展的需要，在学科建设及其应用技术的开发中将长期发挥作用。收录的5000多条词语学科针对性强，广泛地覆盖了自然语言处理、机器翻译等相关领域的概念和用语，英、日、德、汉4种语言的对译准确可靠。

作者与编辑对这两本词典的编排颇具匠心。两本汇编都以英语词语为主，从英语可以很容易地查到相应的汉语、日语和德语。两本汇编都有汉语拼音索引。4 种语言对照的版本中自然还有日语索引和德语索引。英汉对照的版本又增添了英语轮排索引和汉语拼音逆序索引，

这两种索引为读者从不同角度查找多语言对照的计算语言学词语提供了方便。这样，两种版本相互补足，既能满足读者的不同需要，又充分照顾了读者的利益。这两本词语汇编包括各种索引的编写都是在计算机辅助下进行的，这也为计算机技术在词典编纂中的应用积累了经验，提供了实例。

衷心希望《计算语言学词语汇编》的出版能在计算语言学的研究、普及与应用中发挥作用，也希望作者们在此基础上继续努力。需要做的工作很多，如果近期内能有一本带详细释义的《计算语言学词典》出版，那自然是计算语言学界的一大幸事。

陆汝为

1995年2月20日于北京

编写说明

1. 研究背景与编写过程

计算语言学在科学、技术、经济、文化各个领域中正发挥越来越大的作用。从事计算语言学研究及其应用技术开发的人也越来越多。由于计算语言学是研究自然语言处理的，其研究工作的一个重要特点是必须面对多语言现象。像机器翻译系统至少要处理两种以上不同的语言。任何学科的发展都离不开国际间的交流（包括专业术语及其同一性的确认与协调）。从事计算语言学研究的中国人需要阅读外文资料自不待言，特别困难的是常常需要知道两种以上外语的知识。例如读一篇关于日英机器翻译的英语文献，文献中不可避免地会出现日语句子以及有关日语语言学的概念与术语。中国人研究计算语言学，总要面向汉语，因此需要建立汉语信息处理特有的术语，并要努力使这些术语以及这些术语所表达的概念、理论与方法得到国际间学者与产业界的理解与认同。计算语言学又是文理结合的新兴学科，在我国尚无计算语言学专业、硕士点、博士点的情况下，进入这个领域的研究者往往是或者只具有计算机科学的基础，或者只具有语言学的基础，而学际间的交流与融合也是十分重要的。

基于上面的认识，作者俞士汶、朱学锋从1986年起在阅读文献资料的同时就注意收集不同语言对照的计算语言学词语，并在计算机上建立了英、日、汉对照的计算语言学词语库。日积月累，到1992年初已更新了9个版本（将全库内容打印出来进行校对，算作一个版本，零星的联机增补和删改不算在内），词语总数达5000余条。可以说，多语种对照的计算语言学词语库已初具规模。

德国访问学者Elisabeth Kaske自1992年3月起加入了词语库

的开发工作。她不仅仔细地校对了整个词语库，特别是英语部分；而且在3个多月的期间内，将大约90%的词语译成了德语，从而使词语库扩充为英、日、德、汉4种语言互相对照的词语库。1992年7月北京大学组织国内专家对这项自选研究课题的成果进行了评审与鉴定。以中国工程院院士和中国中文信息学会理事长陈力为教授为主任、以中国国家语言文字工作委员会副主任傅永和教授为副主任的鉴定委员会高度评价了这项成果的意义与价值。

鉴定会之后，俞士汶、朱学锋不仅根据专家们的意见再次进行了润色加工，而且请日本朋友富士通公司的大冈智雅先生和菅野芩先生对日语词语特别是日语词语的假名音序索引进行了校对。但是，E. Kaske 小姐回德国后受条件限制，剩下10%左右的德语词语的增补工作进展迟缓。幸好，1993年4月冯志伟回国后立即承担了增补德语词语及审校全库的任务。1994年 E. Kaske 利用暑期访华的机会对德语部分又作了一次校对。至此，英、日、德、汉对照的计算语言学词语库的开发工作告一段落。

现在出版的《英汉对照计算语言学词语汇编》与《英·日·德·汉对照计算语言学词语汇编》两本词典就是在上述词语库的基础上编纂的。

2. 词语汇编的概况

2.1 大多数是术语

术语是表述事物的概念的，特别是用来表述各个专业领域中的概念。通常认为术语的中心词应该是名词。按照这种理解，本汇编所收的词语大部分是术语。本汇编也收录了一些计算语言学领域中需要使用的、多少有些特殊含义的动词与形容词，如

activate	v.	激活
parse	v.	句法剖析

negative	adj.	否定的
synchronic	adj.	共时的

现在对术语的认识也在发展，除了体词性的术语外，也存在谓词性的术语。按照这种广义的理解，上面这些动词与形容词也可看作是术语。本汇编中，有的词语较长，不太像是一个术语，如

abstraction of relation between variables
analysis by synthesis method

因此，为慎重起见，作者将本词典命名为《计算语言学词语汇编》。

2.2 词语的来源

本汇编所收的词语主要有以下3个来源：

(1) 美国斯坦福大学语言与信息研究中心选编的《80年代自然语言处理文献目录》。这本书包含1980年至1987年间问世的1764篇论文或著作的目录，书后附有轮排的关键词索引。

例如，由编号为449的篇名"Noun phrases in lexical functional grammar"自动生成的文中关键词索引项 KWIC 计有

functional grammar, /phrases in lexical	449
grammar, /phrases in lexical functional	449
lexical functional/, Noun phrases in	449
phrases in lexical functional/, Noun	449
Noun phrases in lexical functional/	449

这些 KWIC 为选取所要的词语提供了线索。作者依据这些 KWIC 确定了以下术语

functional grammar
grammar
lexical functional grammar
noun
noun phrase

phrase

(2) 日本80年代中后期出版的有关计算语言学与自然语言处理技术的十多本著作后面所附的索引。如1986年京都大学长尾真教授的《言語工学》书后就有约300个日语关键词的索引。这些关键词，如“格文法”、“概念依存文法”、“深層構造”、“文脈自由型文法”等就是适合要求的术语。但某些书后的像“日本語からの変換”就不适合需要。因此，对这些书后的索引也需要进行筛选或者作适当的加工。

(3) 中国机电部计算机与信息发展研究中心、中国中文信息学会学术委员会合编的《语言信息处理词典词条汇编》(1989年第一版，俞士汶和冯志伟参与了这项工作)。计算语言学词语库吸收了其中一部分源于汉语以及汉语信息处理的词条，像“轻声”、“汉语拼音方案”、“汉字编码方案”、“量词”、“自动切词”等。

除了上述3个主要来源外，作者还从所阅读的英语、日语、中文文献资料中零星收集了相当一部分词语。

在收录词语时，作者注意了学科的针对性，即主要收集自然语言处理、机器翻译等计算语言学研究领域的词语，而相关学科(如计算机科学、语言学、认知科学等)的一些术语若较多地在计算语言学论著中出现，也酌情收入。

初选的词语绝大部分是单语种的，只有很少一部分已有双语对照的基础。英、日、德、汉4个语种间的翻译和对照工作是由作者完成的。

如果希望进一步研究本汇编所收集的词语，请参阅书后所列文献[1]—[28]。

2.3 词语汇编的规模与构成

《英汉对照计算语言学词语汇编》与《英·日·德·汉对照计算

语言学词语汇编》均由正文、索引、缩略语表、参考文献构成。

正文以英语词语为主，共有5415条词语，每个词语前有编号。英语词语按英语字母顺序排列，且大小写字母混合编排，大小写字母相同时大写字母排在前面。以阿拉伯数字打头的词语排在最前面，希腊字母则排在最后面。

两种版本均可由正文中的英语词语查到对应的汉语词语。两种版本都有汉语索引。每条索引包括汉语拼音、汉语词语及对应的英语词语的编号。一条索引中可能并列两个以上的英语编号。

关于汉语拼音索引，说明如下。

(1) 索引首先按拼音字母顺序排列，如

baike cidian 百科词典 1508

baike quanshu 百科全书 1111，1508

(2) 声调在排序中起作用，即同一音节按一声、二声、三声、四声、轻声的顺序排列，不过在印刷时略去了声调符号，如

baohe fanchou 饱和范畴 4178

baogao 报告 4057

如果不考虑声调，baogao 应在 baohe fanchou 之前。由于“饱”发第三声，“报”发第四声，所以如上排列。

(3) 汉字在排序中也起作用，即发音相同的不同汉字应分别集中在一起，如

bao 包 3344

baohan 包含 2186

baorongshi daici 包容式代词 2187

baoyunju 包孕句 1490

baoci 褒词 724

baoyici 褒义词 724

如果只按照拼音排序，baoci 应排在 baohan 之前。显然这是不合适的。以上编排方法与《现代汉语词典》的原则是一致的。但发音相同的不同汉字的排列顺序则依据国家标准 GB2312。

(4) 词组型术语的汉语拼音按词切开，中间置以空格，但空格符在排序时不起作用。如

biaozhunhua 标准化 4603

biaozhun jiegou 标准结构 4601

如果空格符起作用，则 biaozhun jiegou 应排在 biaozhunhua 的前面。因为在 ASCII 编码体系中，空格符位于 h 之前。这个约定同《现代汉语新词词典》(于根元主编)是一致的。

(5) 在汉语拼音索引中以阿拉伯数字、英语字母和希腊字母(拼音中所含英语字母和希腊字母皆用斜体排印)开头的一些词语放在最后。

英汉对照版还包含有汉语拼音逆序索引和英语轮排索引。这两种索引目前还是罕见的，但对读者检索、整理词语都很有用。本书的这两种索引都是由计算机自动生成的。作者和编辑都相信，本书提供的这两种索引对中文辞书的编排方式将产生积极的影响。

生成汉语拼音逆序索引的步骤是首先将每个词语所含汉字的次序完全倒过来，再按照汉语的音节(包括声调)、汉字的次序排列(各项规定与正序索引相同)，不过印刷时各个拼音、词语仍保持正常顺序。如以字母 a 开头的汉语拼音逆序索引包括以下条目

zhang'ai 障碍 341

fang'an 方案 4183

bianma fang'an 编码方案 688

hanzi bianma fang'an 汉字编码方案 570

hanyu pinyin fang'an 汉语拼音方案 617，4184

在正序索引中，"方案"排在"障碍"的前面，这里倒了过来。这种索引的意义在于可以帮助读者实现模糊检索。当读者只记得某

个词语的最后一个字或若干个字时，利用逆序索引可以迅速找到所要的词语。由于上面示例的最后两个字都是“方案”，这就启示人们注意到这种索引对研究词语分类也有帮助。假设读者要求了解本书收入了多少最后3个字是“语言学”的词语，如果只有正序索引，只好从头到尾搜索一遍。现在有了逆序索引，这些词语都集中在一起，读者一目了然。

在汉语拼音逆序索引中，以数字、英语字母和希腊字母（拼音中所含英语字母和希腊字母皆用斜体排印）开头的词语都分散开了。

正序索引与逆序索引的进一步发展便是轮排索引。不过，从读者的实际需要出发，本书只对英语词语提供了轮排索引。构成5415个词语的所有英语单词都单独立项，按字母顺序排列，成为轮排索引的入口点，在每个立项的单词之下，列入了所有包括该单词的术语，而不论该单词在术语的什么位置上。如果一个单词本身也是术语，则后面附有正文的编号，如

adjunct 82
adverbial ～ 89
sentence ～ 4364

如某个单词不是术语，则后面是空白，如

about
reasoning ～ change 3925

显然，轮排索引为读者依据任何位置的一个单词查找有关的词组型英语术语提供了最大的便利。

英、日、德、汉对照版不包含英语轮排索引和汉语拼音逆序索引，但增加了日语索引和德语索引。日语索引按照标注日语词语读音的假名的五十音图顺序排列，平假名与片假名混排，两者相同时，平假名排在前面。德语索引按照德语字母顺序排列。4个特

殊字母的排列位置是：ä 排在 ae 的位置，ö 排在 oe 的位置，ü 排在 ue 的位置，β 排在 ss 的后面。

两种版本都有缩略语表。共收录了219个与本学科关系密切的缩写词，同时给出了相应的原文注释（绝大多数是英语，也有几个源自德语、法语或拉丁语）。

2.4 多语种对照的特点

各种语言之间的对照，大致可分为以下几种情况。

（1）各语种间的术语是一一对应的。如

computational linguistics	計算言語学	Computerlinguistik	计算语言学
machine language	機械言語	Maschinensprache	机器语言
verb group	動詞群	Verbgruppe	动词词组

（2）一个语种的词语对应其他语种两个以上的词语。如对应“language acquisition”，汉语有两个词语：“语言习得，语言获取”，这两个汉语词语虽然不同，但意思并无差别。这种现象在日语中更为普遍，如“logic”既常译为“論理”，也常用外来语“ロジック”。当然，也有一个英语词对应语义截然不同的两个汉语词的情况，如“case structure”对应“格结构；分情况结构”。这样的两个汉语词语之间用分号隔开。

两个以上英语词对应一个汉语的情况也是存在的。如对应汉语的“语法理论”，英语可用“grammar theory”，也可用“grammatical theory”。

（3）不同语种的术语之间的对应关系有复杂的多对多的关系。如

grammar	语法，文法
syntax	句法，语法

反过来

句法	syntax
语法	grammar，syntax

英语中“grammar”与“syntax”所指的概念不同，但汉语中的“语法”与“句法”就常常混用。

(4) 汉语和日语都使用汉字，但汉字相同(简化字与对应的繁体字仍算是相同的汉字)的汉语词语与日语词语所指的概念可能并不相同，这一点在理解与运用时需要特别留意。如日语中的“国語”经常指的是“Japanese”，而汉语中的“国语”则应译成英语的“Chinese”或“Mandarin”，尽管两种语言中的“国语”也都有“国家正式语言”或“民族语言”的含义。又如，汉语的“助词”与日语的“助詞”同属语法上的“功能词”，若不深入释义，很难区分它们。若参照英语，汉语的“助词”对应“auxiliary word”，而日语的“助詞”对应“postpositional word”，从英语可以看出它们之间还是有区别的。

(5) 英语与德语同属日尔曼语族，它们比较接近，有相当一部分词语特别是专业词语的拼法与语义都是相同的(可能有个别字母发生变化，如在英语中读[k]的 c 在德语中都换成了 k)。不过正因为它们接近，更应注意它们之间的差别。有时词形相同的词，在两种语言中意思完全不一样，如德语的 Generation 与英语的 generation 的词形完全一样，但德语的 Generation 就没有“生成”的意思。

从以上分析，可以了解，本汇编虽是面向翻译的，但因为有多种语言对照，多多少少也揭示了每个词语所指称的概念，即参照另外的语种，可以较为确切地了解该语种的词语的含义。

3. 关于若干具体问题的说明

3.1 英语的不同词形

(1) 意思相同的词语若有两种以上不同的结构形式，通常只取一种，如从“case of grammar”，“grammar case”，“grammatical

case"中只选用"grammar case"。

(2) 英语名词的修饰成分常有几种表达形式，如"inflected language"，" inflecting language"，"inflexional language"，本汇编中也只收集了一种。又例如"synchronic"，"synchronical"，"synchronous"都是形容词，本汇编也只选用一种。

(3) 英语拼写法中连字符"-"的使用比较随意，本汇编对同一个词原则上采用统一形式。based 与前面的相关的词一般用"-"隔开，如 rule-based parser；而 generalization based on explanation 中的 based，因其先与后面介词词组结合再修饰前面的名词，故不能在它与前面的词之间加连字符。但有些前缀，如 multi 之后是否用连字符，则取决于原始语料，如"multi-language processor"，"multilingual text processing"。

(4) 英语中某些名词有特殊的复数形态，例如：

corpus—corpora，medium—media

本汇编也是依据原始语料决定取舍。

3.2 日语的不同词形

(1) 对应同一英语词语，不同的日语词语尽量收集齐全，以反映日语文献用语的实际情况。

(2) 对应同一英语的不同日语之间用逗号隔开，用片假名表记的外来语置于平假名与汉字混用的表记之前。

(3) 日语外来语的书写不规范，如"パ-ザ"与"パ-サ"都是指的 parser。又如"インタ-フェ-ス"与"インタ-フェス"都是由 interface 转写来的。外来语不同单词间的分隔符(即圆点)用得也比较随意，如"マン・マシン・インタ-フェス"中间有圆点符，而"ニ-モニックコ-ディング"中间又没有圆点符。本汇编做了一些一致化的工作，但原始语料的痕迹还是到处可见的。

(4) 日语汉字的使用以及汉字后面所附的"送假名"用得也比较随意，如"分かち書き"与"分ち書き"，"読み取り"与"読取り"都是同一个词。本汇编未特意做一致化工作，目的也是为了反映真实文本的状况，有利于增强使用者(包括以后使用本汇编的计算机系统)辨认同形词的能力。

(5) 日语的同一单词可能有两种以上不同的读音，如"依存"可以读"いぞん"，也可读"いそん"。只要不影响意义，本汇编只采用其中一种。

3.3 关于德语译名

(1) 德语译语原则上应该以德语文献资料中使用的为根据，但由于在计算机科学与语言学等领域，德国人常常直接使用原语言(主要是英语)的术语，德语术语规范化程度不够高，因此作者在选择译名时颇费斟酌。

(2) 对于某个词语，若德语文献(包括有关词典)中只使用英语(外来语)，作者也就在德语项目中复制这个英语词语，随后再附上德语，作为注释。如

英：story grammar

德：story grammar　　Geschichtengrammatik

但与此类词语相关的其他词语，若找不到文献的依据，则作者自行译成德语。如

英：story tree　　德：Geschichtenbaum

(3) 对于单词常保留多个译名，对于词组一般只选留其中的一个。如

discourse　　Diskurs, Gespräch, Rede, Sprachgebranch, gesprochene Sprache

discourse analysis　　Diskursanalyse

discourse context　　Diskurs Kontext