

从大规模语料库看 双语词典释义的适用性和客观性原则

解建和 谭家祥

(南通医学院外文教研室)

肖建平

(中国人民银行南通分行计算机室)

摘要: 自建300万字英-汉双语语料库, 利用英汉双向检索系统, 对医学英语文献中50个常用词组进行检索。结果表明, 这些词组在句子中的汉语等值译义数目, 远远超过一般英汉双语词典所给出的释义数量。提示, 这些引伸的词义, 应该引起英汉双语词典编纂家们的注意。检索结果为语料库帮助词书编辑的功能, 提供了实验证据。表明, 大型语料库对双语词典编辑具有实用价值。支持“把主要以语言工作者的语感为依据的传统词典学编辑方法同主要以机储语料的大量实例为基础的计算机词典学编辑方法结合起来”的论点。

关键词: 双语语料库 检索系统 释义原则 适用性 客观性

Significance of a Large English-Chinese Corpus to the Principle of Application and Objectiveness in Word-Definition in a Bilingual Dictionary

Xie Jianhe Tan Jiexiang

(English Department, Nantong Medical College)

Xiao Jianping

(Computer Section, Nantong Branch, People's Bank of China)

Abstract

A 3-million English-Chinese corpus was established and a relative bilingual searching system was developed. With the help of the searching system, 50 frequently used phrases in English medical literature were examined. The results of the searching have shown that the number of the Chinese equivalents of these phrases in the examples picked out from our corpora is far greater than that given in ordinary E-C dictionaries on sale, which should draw E-C lexicographers' attention to the derivative meanings of words and phrases. The search-

ching results have provided a certain experimental evidence for the assisting function of a corpus in lexicography and shown the practical value of a bilingual corpus in bilingual lexicography. In addition, the results support the argument that the conventional lexicographical approach mainly based on linguists' language sense perception should be combined with the computer lexicographical approach mainly based on an abundance of illustrative examples from computer-stored corpora.

**Key words: bilingual corpus, searching system,
principle of word definition, application, objectiveness**

双语词典的释义原则包括“准确性”、“适用性”、“简明性”、“客观性”和“规范性”。^[1]这已是多数词书编纂者的共识。本文仅就利用真实文本(已有的知识成果)语料库,对双语词典释义的适用性和客观性原则意义作一初步探讨。

原苏联语言学家B.Γ.Γ AK 院士,在“词典学发展的某些规律性”一文中指出:“编纂词典时,必须运用两种方法:借鉴已出版的词典和根据一个词在上下文和语言环境中的实际使用来分析该词的词义。”Γ AK 院士特别强调:“在词典编写中,第二种方法极为重要。如果词书的编者毫无批判地抄录前人的成果,那么,他的词典便会落后于语言的实际,反映不出词语在现代语言中的真正作用。”^[2]

根据词的实际使用来分析该词的词义,就需要积累大量的语言材料。但是,冰冻三尺,非一日之寒。大量语言素材的收集,并非个人在短时间内所能做到的。然而利用计算机语料库,情况就不同了。基于语料库的语言研究已经成为计算语言学的重点。^[3]

1 国内外语料库开发研究简况

六十年代,英美各国开始研制建立语料库。比较著名的有:

1. 1964, 美国布朗(Brown)大学:当代标准编辑美国英语语料库—美语;100万词;有语法和句法分析标记

2. 1970-1980, 伯明翰(Birmingham)大学语料集—(主要)书面英语;2千万词

3. 1980, 英国兰开斯特(Lancaster)大学/IBM口语英语语料库—正式口语英语;有语法和句法标记及节奏韵律标记

近年来开发的超大型语料库,有:

1. 跨国英语语料库—来自15个讲英语国家的语料库,每库100万词;有语法标记

2. 英国国家语料库—英语;1亿词

3. 伯明翰英语库—2亿词

4. 朗文(Longman)-兰开斯特语料库—书面英语;三千万词

5. 美国计算语言学协会计算语言学语料库—(以美语为基础);1亿词

英国兰开斯特大学著名语言学家 Geoffrey Leech 在 1991 年提出了语料库语言学的理论,并总结了迄今为止语料库的四大功能:^[4]

1. 用于语言研究：词汇学，句法学，语义学和语用学，文章结构，会话分析，体裁风格
2. 用于出版词书和语法书：用法指南，分类词典，频率词典，搭配词典，电子词典
3. 用于教学：词汇、语法索引，教材索引，便览练习，专业文库(口语)(科技、医学、法律等)，频率研究及词汇、语法分级，儿童语言语料库，儿童教学文本语料，儿童语言索引，双语及多语种语料库(供翻译研究及词汇、语法对照分析用)
4. 用于言语和语言技术：结合多媒体技术制作各种声、像、文字软件

英国伯明翰(Birmingham)大学的《Collins Cobuild Dictionary》就是通过两千万词的英语语料库选编而成的。朗文公司利用语料库，编写成各种英语教科书及各种词书。

中科院文学研究所拥有千万词汉语语料库。台湾中央研究院也拥有近千万词的古代、近代、现代汉语语料库。清华大学计算机系拥有5000万字真实语料，并用来对现代汉语述语动词机器词典，进行了研究。^[5]

但是，在机器翻译的研究中尚未见到国内学者利用双语语料库的报道。^[6]

2 我们开发的语料库

我们成功地进行了英语医学文献百万字计算机词汇调查，编成了《医用英语基础词汇》，利用计算机技术自动生成《中文索引》，^[7]成功地进行了《朗文英汉双解活用词典》自动转换生成汉英词典的实验研究，^[8]并对汉英词典生成过程中的释义原则作了探索，^[9]建立了300万字的英-汉双语语料库，进行了英汉-汉英文字处理计算机技术的实验研究，确立了英汉-汉英文字检索系统的实验模型，利用此实验模型对语料库运用于英语教学及词典编辑等多项功能，作了初步应用研究。

我们的研究拟建立超大型语料库，开发任意检索系统，供各层次的英语学习者、英语教学与科研工作者及外事工作人员实际使用。

本文仅就大型语料库对双语词典释义适用性和客观性原则的意义，结合实例进行探讨。

我们所建立的语料库，除系统语料库外，还附有文本语料库、口语语料库和外贸经济词汇语料库等。语料选自：1.多部优质英汉词典 2.英汉双语译文文献的篇章、段落和句子。按一定格式录入计算机，利用 FoxBase+ 建立检索系统：“EC-CE Searching System (EC-CESS)”(英汉-汉英检索系统)。(检索系统的模块结构略)

3 语料库对适用性原则的意义

一. 发掘出新的义项

我们利用此检索系统对医用英语中常用的五十个词组，进行了检索，发现，真实文本语料提供的词义义项，多于市售词典所给出的义项，而且是医用英语中独特的词义。

以“associated with”和“be responsible for”两个词组为例。

(1) “A is associated with B”

先在系统语料库中检索，后利用文本语料库检索。选择“英-中”检索项，键入“associate-”(为“associate”的模糊检索)，屏幕出现如下例句：

1. We naturally associate the name of Darwin with the doctrine of evolution.

我们很自然地将达尔文的名字与进化论联系在一起。

2. Cigarette smoking has been associated with breast as well as lung cancer.

吸烟不仅与肺癌有关，也与乳癌有关。

3. However, prolonged exposures to cold or hot environments are usually associated with both pain and tissue injury.

然而，长时间接触冷、热环境常常会引起疼痛和组织损伤。

4. In 1836, it was pointed out that paralysis of the right side of the body was often associated with loss of speech, while patients whose left side was paralysed could usually talk normally.

1836年，有人指出，人体右半部瘫痪常常伴有失语，而左半部瘫痪的病人，常常仍能正常讲话。

上述4个例句中，1-2来自我们的核心语料库。源语料词典对“associated with”这个词组的释义均为：1.使联合，加入，参加(with) 2.联想，想起(with)

只有《英汉技术词典》和《常用医学英语短语》除上述词义外，还列有：涉及，伴随，伴有，产生，和…并发，与…有关，在…同时。

3-4来自我们的文本语料库。从例句译文来看，除了：1.使联合，加入，参加(with) 2.联想，想起(with)外还有：

1. A与B有关，A参与B 2. A为B所引起/产生 3. A引起B 4. A伴随B，A伴有B

(2) “A is responsible for B”

检索方法同(1)。经检索，在“responsible”单词下，出现如下例句：

1. The pilot of the plane is responsible for the passengers' safety.

飞机驾驶员应为旅客安全负责。

2. Mechanical defects were responsible for the accident.

这次事故是由于机械故障而引起的。

3. The main substance responsible for O₂ carriage is haemoglobin.

被用来运输O₂的主要物质是血红蛋白。

4. The enzymatic activity responsible for thrombin formation evolves during the interaction of a number of constituents of the blood.

导致凝血酶产生的酶活动，是在血中许多成分相互作用时进行的。

上述4个例句中，1-2来自我们的核心语料库。源语料词典都列出其本义：对…负责，对…有责任。但大部分双语词典都未列出另一重要词义：由于…，是…的原因，尽管有些词典所引用的例句中已经出现这样的词义。

只有《英汉技术词典》列出较详细的词义：是(造成)…的(主要)原因，决定着，造成，导致；引起。《英语惯用法词典》专条示例说明：

The cold weather is responsible for his absence.

他因为天冷而缺席。

并加注强调，该句主语并不是“人”。这种用法主要通行在美国。(我们的文本语料库表明，

此用法亦流行于英国。)

[日]《英语科学论文用语辞典》专列条目强调说,在科学论文中,该词常用来指人以外的事物,意思不是“负责任的”,而表示“原因”等意义。并示例:

The enzyme is responsible for the conversion of food into energy.

食物之所以转化成能量是由于酶的作用。

3-4为我们的文本语料库中检索出的例句。译文中“responsible for”的词义有:

1. A负责B, A的功能/用途是B, A为B所用, A被用来(做)B, A对B起作用
2. A引起/导致/决定着B, A是引起B的原因
3. A构成B

检索结果表明:《英汉技术词典》和《常用医学英语短语》注意到科技文献中短语词组在科技文献中的特殊含义;检索结果同时也指出,这些词组还有数个引伸的词义,反映语言被使用的实际情况。应该引起英汉双语词书编纂家们的注意。我们的实验结果与清华大学陈群秀、黄昌宁等利用2000万字真实语料,对现代汉语述语动词的研究中,对“走”的句型实例分析结果一致。^[5]我们的实验结果与清华大学常新功、夏莹等利用字字同理(二元)统计所得概率矩阵,研究汉语字间关系,发现现行汉语词典未收入的词频很高的新词条的结果相吻合。^[10]

与上述两例类似的词组还有许多。如:

accompanied by, account for, amount to, appear to (do), attribute to, characterized by, contribute to, deal with, exposed to, implicated in, involved in, relieve A of B, result in, result from, serve to (do), available for, available to, characteristic of, continuous with, in the absence of, in the presence of, with reference to 等等,不下上百个。

象上面的例子,我们的文本语料库中,包含“associated with”的有157句,包含“responsible for”的有129句,而且都出自经典权威著作和期刊,如:《The Cecil Text-book of Medicine》,《Textbook of Medical Physiology》,《Science》,《Nature》,《The Lancet》等。如此众多的典型例子足以证明,一般词典对这些词组的释义,是远远不能满足读者需要的,应该增补引伸的意义。

二. 集众家之长,相辅相成,互补欠缺,为读者提供尽可能完整多样的对应释义以“abandon vt”为例:

《远东》: 1. 放弃, 弃绝(a ship, a car)

2. 舍弃, 弃守; 背弃, 离弃(one's home, one's friends, a city)

3. 放纵, 放肆(Do not ~ yourself to despair. 不要失望; 勿自暴自弃。

The sailors ~ themselves to drinking. 水手们恣意饮酒。)

4. 委付

《英华》: 1. 扔弃(地位等), 离弃(家园), 断绝(念头等), 戒除(恶习等)

2. 遗弃(妻子), 放弃(权利等) 3. 投保(货物)

(~ oneself to 恣意, 沉迷(酒, 色); ~ oneself to pleasures 恣意享乐;

~ oneself to despair 悲观失望; ~ oneself to emotion 感情用事)

《新英汉》: 1. 抛弃, 放弃 (~ smoking 戒烟; ~ medicine for literature 弃医学文;

~ an attempt 放弃尝试; ~ sb to the mercy of 听任某人由...摆布)

2. 离弃(家园、船只、飞机等) 3. 遗弃(妻、女等)

4. ~ oneself to sth 沉湎于某事; ~ oneself to despair 陷入绝望

《Collins》: 1. ~ (a baby, a car)[遗弃(婴儿, 汽车)]

2. ~ (a habit, hope)[戒除(恶习), 放弃(希望)]

3. ~ (office)[放弃(职位)] 4. ~ (a job, a game)[放弃(工作), (比赛)弃权]

5. ~ oneself to emotion [感情用事] 6. ~ (property) [将(财产)投保]

几部词书对“abandon vt”的释义,大同小异。然而,毕竟各有自身的特点,或多了义项,或多了释例。但大型语料库能做到集各家之长于一体,无一遗漏,将学术界、翻译界对词语研究的成果给以充分利用,使其发挥最大效能。这是任何手工操作很难做到的。

英语词语搭配构成中国人学英语的难点。英语教育专家们,如 Palmer 早在五十年代就注意到这个问题,并且有多部专著出版,在许多教科书中已纳入此类内容。因此,一般的固定词组,在一般词书中能查到。如: make sense, make progress, make a difference, do sb a favour, do wonders, have a fever, have a pleasant time, take a chance, take notice of 等。唯有松散的开放性词组(自由词组),如“形容词+名词”构成的搭配,研究者们认为是无限的,因而,尚未引起人们足够的注意。但是大型语料库能很方便地为读者提供大量语料,供选择使用。如, idea: 经检索,就会出现:

a clever idea 聪明的主意

a crazy idea 糊涂的想法

a useful idea 有用的主意

a lateral idea <喻>偏离正题的观点

a fanciful idea 不切实际的思想

a sensible idea 明智的想法

a striking idea 惊人的想法

an original idea 独树一帜的观点

a good/bright idea 好主意

an ingenious idea 独创性的见解

an improbable idea 不大可能的想法

an incredible idea 不可相信的主意

an intriguing idea 有诱惑力的想法

an interesting idea 一个有趣的念头

an inconsequential idea 一个无关紧要的想法

What a good idea. Let's do it. 多好的建议! 让我们来干吧!

.....

大型语料库对词书编纂中释义的适用性原则的贡献绝不止此。仅此几项,足以说明,大型语料库能使词书在最大限度上满足读者的需要。

4 语料库对客观性原则的意义

大型语料库对词条释义是否全面、广泛及其精确程度,具有检验、扩展、补充、深化的重大意义。从而反映语言实际运用的真实情况。

有一些历来不被认为是词组或句型的表达法,都不能进入双语词典。而实际上,英汉两种语言中都有相应的固定的表达形式。通过文本语料库的检索,含有这类句型的句子,大量出现。有的有上百条之多。如:

It is (well) known that 众所周知... It is shown that 已经表明...

It is reported that 据报道…

It is said that 据说…

It has been pointed out that 已经指出…

It is clear/evident/obvious that 很清楚/明显…

find/believe/think it possible/necessary to (do) 觉得/认为有可能/必要(做)…

There is no/There is evidence that 没有/有证据表明…

象这样具有相应的等值表达形式，而且用得如此广泛的语言材料，应该收进双语词典，为读者提供方便。

所谓客观性，主要是指辞书应该反映客观存在的语言事实，并非单纯只指排除政治性倾向或特定历史背景的痕迹，或妄加褒贬，或将目的语的社会文化背景强加进源语言词汇。

比较典型的要算“might/would otherwise (do)”(原来会…)了。在我们的语料库中，有12句，已经形成固定结构。但几乎无一本市售英汉词典收录。从检索例句中列出两例：

1. Any bacteria which enter the lymph current and might otherwise be carried into the blood stream are removed.

直译：进入淋巴流及原来会被带入血流的细菌都被清除。

译文：细菌进入淋巴后，即被清除，如果没有淋巴，细菌就会被带入血流。

2. Inbalance among nutrients can render inadequate an intake which would otherwise be ample.

各种营养之间的不平衡会使原来足够的摄入量变得不足。

市售英汉词典，对“otherwise”的释义：“另外，别样，不同地，用别的方法，在其它方面，此外；要不然，在不同的情况下，否则”等均不适用。究其原因，简单地说，就是缺少大量书证，无法使一些词语意义的内涵全貌，暴露出来。而实际使用中已经存在的语言现象，应该经过编纂者的分类、整理，使其全面、准确地反映出来。只有大型语料库才能满足这一要求。

5 结束语

荷兰语言学家阿尔-卡西姆在《词典编写和评价的标准》一文中，关于“语义材料和例证的标准”，有如下的概括：

1. 主要词义
2. 合适的等值翻译
3. 每个词义下是否都有一个例证
4. 例证是否代表条目语使用者的文化水平
5. 例证是否全部译成使用者母语
6. 例证是否言简意赅，说明词的用法^[11]

词典编纂专家历来主张：语言是人们使用的结果。使用的综合记录，便是词典。使用语言的过程形成了词典的内容。^[12]

他们反复强调：词典编纂者的工作就是，运用用过的语言实例为基础，解释有记载的事实，收集大量说明词的特点的例证，研究其意义，研究、编排、评价和解释语言事实。^[12]

他们认为，任何一个词多义和辩证，善于在上下文中表达新的意义色彩。词典需要的不是词的解释，而是词的实用的译文，适用于活的外语上下文中。^[13]

语言以使用为基础。任何东西被相当多的人普遍地使用，就会成为标准。^[12]

我们正是在“初次接触语言实际和教授语言的任务”^[2]时，不断吸收新的语言材料，丰富语料卡片^[14]的过程中审视词义的。为了方便读者，根据词在上下文语言环境中实际使用，分析词义的工作是词典编纂者不可缺少、也不能懈怠的常规工作。而大型语料库对释义选择、深化和扩展，具有重大意义。我们的实验证实了 Leech 关于语料库功能的论述的正确性和实践性，为语料库语言学提供了部分实验证据。并以比较详实的语料表明，大型语料库对双语词典编辑的实用价值。支持把主要以语言工作者的语感为依据的传统词典学编辑方法同主要以机储语料的大量实例为基础的计算词典学编辑方法结合起来的论点。^[5]

当然，词义的收录与取舍，也不是任意的，也并非可以“望文生义”或“随文撰义”。一方面，以大量的实际语言材料为根据，一方面要看读者的需要。

如果利用大型英汉-汉英文字语料库对基础词汇逐条核查，则可能使现行英汉双语词典的释义更加完善。这将是一件对亿万读者千古不朽的大好事。利用大型语料库这样编成的词典将有助于机助翻译的选词，为其提供尽可能完善的优质支持。（另文探讨）

参 考 文 献

- [1] 钱厚生，双语词典的释义原则与问题，《辞书研究》，1993.3：50
- [2] [苏]B.Γ.Γ A K，词典学发展的某些规律性，《词典学论文选译》，商务印书馆，1981：85-102
- [3] 周明、黄昌宁，具有自学能力的语料库句法标注工具CSTT，《计算语言学研究与应用》，北京语言学院出版社，1993：155
- [4] Geoffrey Leech，The State of the Art in Corpus Linguistics，In English Corpus Linguistics，Edited by Karin Aijmer and Bengt Altenberg，London，Longman，1991
- [5] 陈群秀、黄昌宁，现代汉语述语动词机器词典研究初探，《计算语言学研究与应用》，北京语言学院出版社，1993：231
- [6] 陈力为，《计算语言学研究与应用》前言，北京语言学院出版社，1993
- [7] 邵循道，《医生必备英语单词8000-医用英语基础词汇》，湖南科技出版社，1992.8
- [8] 解建和等，英汉词典自动转换成汉英词典的计算机实验研究，《现代外语》，1995.1：1-6
- [9] 解建和等，计算机转换成汉英词典的探讨，《辞书研究》，1994.6：28-37
- [10] 常新功、夏莹，用语料库语言学知识指导文本识别研究，《计算语言学研究与应用》，北京语言学院出版社，1993：92
- [11] [荷]阿尔-卡西姆，词典编写和评价的标准，《词典学论文选译》，商务印书馆，1981：162-166
- [12] [美]Chalton Laird，Language and the Dictionary，《Webster's New World Dic. of the American Language》，2nd College Ed.，The World Publishing Company，1972
- [13] [苏]Γ.В.谢尔巴，俄法词典再版前言，《词典学论文选译》，商务印书馆，1981：167-177
- [14] [苏]B.Γ.索洛科列托夫，苏联俄语词典编纂学传统，《词典学论文选译》，商务印书馆，1981：185-187