教育与学习 4 卷 11 期
ISSN: 2705-0408(Print); 2705-0416(Online)

语料库辅助下的英语词汇自主学习研究

李秋一

西安外国语大学英文学院 陕西 西安 710000

摘　要：作为语言学习的基石，英语词汇一直是英语学习的重点之一，同时也是英语学习者面临的难点。语料库是指大量搜集起来的结构化语言数据。在之前的众多学者的研究中已经证实基于语料库的词汇学习的有效性，但在现实生活中使用该方法的个人学习者并不是很多。本文基于之前学者的研究，探究语料库直接使用在词汇学习中存在的问题，并就这些问题提出具体的解决方案以此来促进语料库在词汇自主学习上的进一步运用。

关键词：语料库；英语词汇学习；数据驱动学习

Research on independent learning of English vocabulary with the assistance of corpus

Qiuyi Li

School of English, Xi'an University of Foreign Chinese Shaanxi Xi'an 710000

Abstract: As the cornerstone of language learning,English vocabulary has always been one of the priorities of English learning,but also a difficulty faced by English learners.Corpus refers to a large collection of structured language data.The effectiveness of corpus-based vocabulary learning has been demonstrated in previous studies by numerous scholars,but not many individual learners use the method in real life.Based on the research of previous scholars,this paper explores the problems existing in the direct use of the corpus in vocabulary learning,and puts forward specific solutions to these problems to promote the further application of the corpus in autonomous vocabulary learning.

Keywords: Corpus;English vocabulary learning;Data-driven learning

引言

词汇是语言的核心，在二语学习中的重要性不言而喻。词汇和语言使用熟练程度存在着较高的正向关系，主要体现在和语言的听、说、读、写能力的紧密相连。词汇知识的掌握越多，语言使用的熟练程度往往越高。词汇的学习对于学习者来说，往往也是比较困难的过程。而基于语料库的数据驱动型学习能够为词汇学习带来新的生机，学习者借助语料库巨大的数据容量和丰富真实的语料来更好地学习英语二语词汇并从中提升不同层次的词汇能力。

1 语料库概述

1.1 语料库语言学

语料库语言学是基于语料库展开的有关语言的研究。语料库是指大量存储在电子设备中可被机器读取的真实语料。是语料库语言学研究中重要的基础和材料。储存在语料库中的语料有代表性、真实性以及计算机可读性等特点。计算机技术的发展为语料库语言学的发展起到了重要的促进作用。一方面，计算机大容量储存技术的发展使语料库的规模逐渐扩大，种类也越来越多。另一方面，电子分析工具以及语料标记信息的发展大大提升了语料分析的精准度和效率。在语料处理方面，研究者一般通过处理工具（例如AntConc，Wordsmith，SketchEngine，CQPweb等）来分析语料的频率（frequency）、关键词（keywords）、搭配（collocation）、共现（concordance）等信息。除了作为一门学科，语料库语言学更多的是作为一种方法被其他学科所使用，几乎所有的语言学研究（认知、词汇、语用、文体学、社会语言学、语言与教学）都可以基于语料库来进行。语料库语言学对于词汇的教学和学习帮助也很大，在之前的大量实证研究中，除了对语料库在词汇学习上可行性进行了论证，还证实了基于语料库的词汇学习和教学的有效性。对于这一点在后文会进行详细说明。

1.2 语料库与词汇学习

在词汇教学方面，语料库通过直接使用和间接使用两个途径来实现。间接使用是指语料库给教学大纲、词汇测试以及教学材料的设计编写提供启示和参考。例如，在教学的过程应该关注语言中最常见的词形，以及这些词形在使用中构建出的最核心的模式组合。这时就可以通过语料库比较频率来制定相关的教学计划和大纲。

直接使用指的是数据驱动学习（DDL），即学习者自身参与到语料库的分析当中来。约翰斯（Johns，1991）年提出数据驱动学习的概念（Data Driven Learning），

作者简介：
李秋一（1997.10-），男，汉族，湖北恩施人，硕士研究生，西安外国语大学英文学院，研究方向：外国语言学及应用语言学

学习者参与自身语言的分析并且关注特定的语言特征。DDL通过鼓励学习者自主发现自然语境中语言的规则来提升学习者的主动性。直接接触语料的过程让学习者沉浸在自然语料当中，能够让他们自己去验证关于语言特征的假说、激发学习者自己学习的意识。对于教师本身来说，从中也能受益很多，语料库的支撑能够增强老师的自信，并且提升语言意识。

另外，学习者语料库（learner corpora）的研究也为二语习得的研究提供了实证数据的支持。因为语言习得涉及到学习者的心理变化过程，对这一过程的观察比较难实现，而基于学习者语料库的错误分析为研究者提供了一个观察窗口，让研究者能够观察语言学习的心理过程。研究的结果也为外语教学提供了教学启示。本文侧重点在学习者自主直接使用语料库进行词汇学习的方式，从其词汇能力的层次出发，分析学者在词汇习得的过程的不同阶段的词汇能力。

2 语料库与词汇学习

2.1 词汇能力

在最基本层次的理解上，词汇能力被概括成词性和意义的简单联系。但是许多学者不认同这一点，他们指出词汇能力的构建是一个复杂的过程，受到各方面的影响。其中影响最深的表述是内申（Nation）所提出的成分方法，他将词汇能力看作是从几个角度和多个方面共同构建的结果。所以他从形式、意义和使用三个角度，以及输入输出两个方面来概括词汇能力。另外一个比较重要的词汇能力概括是从词汇的深度和广度来区分词汇能力。安德森和弗里博迪（Anderson&Freebody，1981）首次从这两个维度来解释词汇能力：词汇的广度（或者容量）指学习者所知道的词汇数量，词汇深度指多大程度上了解这些词。米拉（Meara，2010）则从心理语言学的角度来解释词汇能力，他坚持从更加整体的角度去看词汇能力，将词汇能力看作是组建学习者心理词典的相互连结网络。

在词汇能力评价中使用较多的是亨里克森（Henriksen，1999）的框架，他从以下三个主要方面来区分词汇知识：①精确知识，即能够理解目标词项的定义，②深入的知识，即能够陈述指称含义（例如同义词，反义词）以及目标词汇的句法特征（例如搭配模式）项目，以及③生产性使用能力，即能够使用目标词汇项。下面将从这三个层次来看语料库在词汇学习中的具体作用。

2.2 语料库在词汇学习中具体作用

精确知识方面，语料库能够帮助学习者建立起词与其基本意义的联系。不仅能够将词语放在原文语境中呈现，方便学习者将具体意义放在原文中进行理解。还能通过平行语料库使目标语和学习者的母语相对应。语料库中单词出现的频率反映到现实中所对应的是单词的使用频率。也就是说越经常使用的词语，在语料库中出现的频率也就越高。学习者可以选取最常出现的单词和意义进行优先学习，优化学习效率。

在深入知识方面，语料库能帮助学习者理解多义词以及区分同义词。越经常使用的词语，更有可能含有多种意思，语料库是大量储存的自然语言集合，是探究词语在不同的真实语境下构建不同意思的有力工具，通过对词汇所在语境的探索，学习者能够清楚的区别多义词的具体含义和使用方法。在同义词方面，因为词语不是单独存在的，他们在和其他词语使用时会共同构建词汇关系，有时由不同词汇构建的这一关系所表达的意思是一样的。学习者在不能区分这类同义词的时候，可以借助语料库方法。因为在语料库研究中，就算是同义词他们之间区别也能够清楚地看出来，意义相同的单词或者词汇关系会在不同的语域或者交流模式上存在差异。

在生产使用方面，语料库能帮助学习者正确合适地使用词汇。对于一般词语的用法和搭配，可以使用语料库搭配检索，通过不同的算法（MI，MI2 等）来搜索目标词汇的搭配。比如说，begin和start都可以和to搭配使用，那他们的区别在哪儿？通过语料库搜索对比就可以发现答案，begin和to搭配使用更加常见，并且多在小说文体中使用。而start虽然可以和to搭配，但实际使用中较少，其在小说和学术文体上使用差异也不明显。隐喻意义是词语使用的非文学方式，隐喻是通过比较来解释事物的。习语指的是一些受到语言和文化的影响的固定词组，他们整体所表达的意思不是其简单的字面意思相加。对于隐喻和习语的使用，语料库能够作为识别隐喻用法的真实样本，能为学习者提供具体语境来展示他们的用法和意义。而在语域变体方面，语料库也能够探究其差异。因为语体变化关注的是语言在不同因素作用下的使用变化（交流情况、正式度、年龄、行业等），而现代语料库取样范围很广，并且对这些语料从各个方面都进行了标注。例如英国国家语料库对语料库中语料的性别、年龄、社会阶层、地区、教育背景等多做了详细标注。

2.3 语料库辅助下词汇学习存在的问题

首先，基于语料库的词汇学习方法的实施需要电脑和语料库软件的支持，这使得基于语料库的词汇学习场景较为固定，只能在部分有条件的学校实施。学习者课下自己实施起来难度大一点，从而导致该方法没有在个人学习者

中普及，也有教师尝试过纸质语料库索引的方式，通过打印目标词汇的索引列表让学习者直接使用，这样一来可以直接观察打印的文本，但这方法又过于费时间并且灵活性差。其次，语料库储存的语料是母语者的真实对话和文本使用，对于低水平的学习者来说，这些语料可能过于复杂，他们无法自己处理这些数据，从而导致无法从数据中推断出目标词汇的准确含义和使用方法。再加上大部分学习者运用语料库能力的缺乏，也会导致学习的效果大打折扣。最后，过去的数据驱动教学主要从事教学和教师角度出发，没有从学习者本身需求出发，致使学习者的需求没有得到满足。所以需要从这三方面入手对语料库辅助下的词汇学习改进。

3 解决措施

3.1 做好语料预筛选

语料库方法对低阶和高阶学习者来说效果不是很好。这是因为通过语料库软件检索出来的语境和语料是随机呈现或者按照文档顺序出现的。在句子的难度和理解层面检索的结果并没有做到分类或预处理。这就导致了所呈现出的语料对于一部分学习者来说过于困难，无法从当前语境中推断出词语的意义。又或者所呈现的句子结构过于简单以至于文本信息过少而无法推断出目标词汇的意义。因此在原有语料库结果上应对语料进行筛选。换句话说就是为学习者提供合适的输入材料。

维果茨基（Vygotsky，1978）提出了最近发展区（ZPD）的概念，他将其定义为“由学习者自身独立解决问题所确定的实际发展水平与通过在成人指导下或与更有能力的同伴合作下解决问题确定的潜在发展水平之间的距离”。在该概念下，应注意区分两个关键的发展水平：实际的和潜在的，而处于这两个发展水平之间的距离就是最近发展区。当这一理论应用于语言学习领域时，人们通常会利用它来选择适合语言学习者的输入，使输入比学习者当前的水平稍微困难一点，从而让学习者获得更好的学习效果。词汇覆盖（lexical coverage）是指一段文本中已知单词所占比例。施密特（Schmidt，2011）发现，已知词汇的百分比与阅读理解之间存在线性关系，这表明知道的单词越多，学习者就更容易理解目标语篇。

语料的筛选需要借鉴最近发展区和词汇覆盖的概念。在做语料筛选的过程中，首先对学习者的词汇能力做好测试，建立学习者专属的已知词汇集合。当学习者搜索目标词汇时，将语料库中的包含目标词汇的语料按照已知词汇比例从高到低进行排序展现给学习者。而这一过程的实现，可以借助python中相关的自然语言处理工具进行。

3.2 语料库手机交互程序设计

前面提到的另外一个问题是语料库学习方法的实施需要借助电脑和语料库软件来进行，因此局限性可能较大。虽然近些年来科技的发展使电子计算机已经十分普及。但由于便携性的问题还是使语料库的使用比较繁琐面对这一问题，则需要研究者开发相关的手机端程序来方便学习者快速学习，智能手机的普及率与便携性都高出电脑很多。

目前已经在web端已经有很多在线语料库，不需要本地下载通过互联网就能够直接访问。例如，美国杨百翰大学马克·戴维斯（Mark Davies）在 2018 年推出面向语言学习者的iWeb语料库检索平台（可在手机上使用）。它通过大数据链接了 2000 万个网页，9.5 万个网站，拥有 140 亿词次的语料，可据此自建任何话题的语料库，可检索最常用的 6 万个英语词，不仅呈现其索引行、词性、定义、同义词、词丛，还能即时链接到与之相关的网页、话题、发音、图像、视频乃至 100 多种语言的译文（何安平，2019）。这为手机程序的实现提供了显示基础，如果能够调用在线语料库的资源，加上检索和相对应的个性化语料库预处理，就能够达到很好的效果。也就是说在学习者和语料库之间添加辅助学习软件，将学习者的输入转换成语料库工具可执行的命令（例如正则表达式），然后将结果筛选后通过不同的形式呈现给学习者。工作原理如下：学习者输入（发出指令）、软件解释命令并传递给语料库工具、对储存在服务器中的数据进行匹配检索、按照用户词汇覆盖降序排序、输出结果。

交互程序的实现，一方面可以增加基于语料库词汇学习的应用场景，是语料库辅助学习不再局限于固定的教学场景中，在课下只需要借助智能手机就能够进行词汇的学习。另一方面来说，交互程序的实现，对初学者比较友好，降低了语料库使用的门槛。初学者的学习成本大大降低，有利于语料库辅助学习的传播和发展。

3.3 关注学习者需求

学习者的动机和需求紧密相连。当前的语料库的语料容量巨大，大型的语料库的容量已经达到 10 亿次。语料库的种类也是多样的，有专门语料库、通用语料库、多语种语料库、平行语料库、学习者语料库、历史语料库以及监控语料库等待。如今的学习者缺少的不是学习资源，他们缺少的和个人学习需求相匹配的资源。

从词汇能力角度来看需求的话，可以将学习者的需求划分为三类。第一种学习者只要求将单词的形式和简单意思相对应即可；第二种学习者希望在此基础上了解一个单

ISSN: 2705-0408(Print); 2705-0416(Online)

词的词义扩展，即在不同语境下所表现出的不同意义，或者不同单词所表达的相同意义；而第三类学习者则希望掌握词汇的使用，希望自己的表达符合本族人的表达习惯，在具体的环境下选取合适的词汇表达。从用途角度来看，根据学习者的目标应用场景，其需求又可以进行不同的划分。例如、口语词汇、学术英语词汇、商务英语词汇等等。

之前的基于语料库的词汇学习，主要是从教学角度出发，而很少从学习者的角度出发。也很少考虑到学习者的个人需求。在后续的学习中，可以根据学习者自身的实际情况，在开始使用语料库词汇学习之间进行需求调查，根据学习者具体的学习需求匹配相对应的语料库作为数据驱动学习的样本。

4 结语

本文从语料库与词汇学习的关系入手，阐述了语料库辅助下的词汇学习的好处，总结了之前在此方面存在的一些问题，并对这些问题提出了相对应的解决措施。在如今互联网和大数据快速发展的时代，相信语料库能够在不远的将来能更好地帮助二语学习者解决在词汇学习上的一系列问题。

参考文献：

[1] Biber, D.S, Conrad&R. Reppen.(2000). Corpus Linguistics. Beijing, China: Foreign Language Teaching and Research Press.

[2] Schmitt, N. (2010). Researching Vocabulary: A research manual. Basingstoke: Palgrave Macmillan.

[3] Szudarski, P. (2018). Corpus Linguistics for Vocabulary: A Guide for Research .London: Routledge.

[4] 何安平.何安平谈语料库与语言教学[J].语料库语言学,2019:13-22.