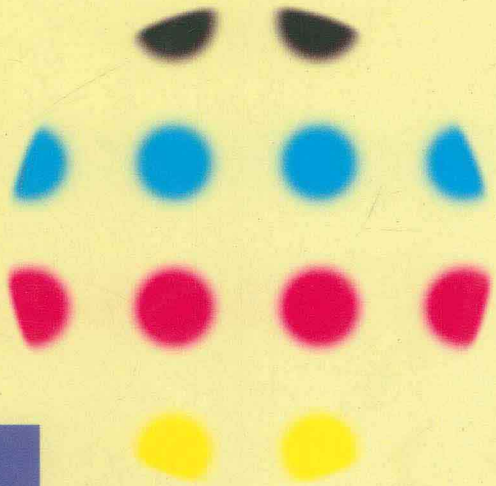
全国高协组织教材研究与编写委员会审定

SEMANTIC ATTAINMENT IN INTERLANGUAGE —A CORPUS-BASED STUDY OF VERB USES THROUGH CHINESE L2 LEARNERS' WRITTEN ENGLISH TEXTS

从中介语看语义获得

—基于语料库的中国学生英语书面语动词使用情况研究

马跃 著



中国科学文化出版社
香港教科文出版有限公司

本书由全国高协组织教育发展中心、香港教科文出版有限公司资助出版
全国高协组织教材研究与编写委员会审定

从中介语看语义获得

马跃 著

中国科学文化出版社
香港教科文出版有限公司
2001 · 香港

从中介语看语义获得

马跃 著

出版发行：中国科学文化出版社
香港教科文出版有限公司
排　　版：新天地文印中心
印　　刷：盛源印务有限公司
开　　本：16K
印　　张：12.6
字　　数：255 千字
版　　次：2001 年 8 月第 1 版
书　　号：ISBN 962－8467－48－4/G · 13
定　　价：25.00 元

作 者 简 介

马跃，男，1959 年生于山东济南。1982 年获山东师范大学文学学士学位（英语语言文学），1987 年获广州外国语学院文学硕士学位（语言学及应用语言学），1990 年获加拿大圣玛丽大学教育硕士学位（课程与教学论），2001 年获广东外语外贸大学文学博士学位（外国语言学及应用语言学）。现任暨南大学华文学院副院长、副教授。长期从事第二语言教学与研究工作，主要研究领域为应用语言学。

序 言

桂 诗 春

在英语教育史中曾经出现过一个所谓词汇（控制）运动。在上一个世纪的早期，H. Palmer，M.West，E. Thorndike，L. Faucett 等英语教育家是这场运动的推动者。Palmer 认为在最常用的英语词中，头 1000 个词的使用频率占了所有书面语的 85%，第二个 1000 词占了 7%，第三个 1000 词占了 3%。合共为 95%。H.Bongers 对这个估计做了不少论证，表明 Palmer 的说法是有根据的，例如 Shaw 的 The Doctor's Dilemma 为 96.1%，Bennett 的 The Card 为 95.25%。West 根据他在印度孟加拉的双语调查设计了一套改善印度学习英语儿童的阅读能力的教材，叫做《新法读本》（The New Method Readers）。读本的主要特点是控制词汇，尽量使用常用词，以培养儿童的阅读兴趣。由此开了简写本的先河。在 West 的建议下，1934 年和 1935 年相继在纽约和伦敦开 Carnegie 会议，上面所说的几个人都参加了，特别是 Thorndike，作为词频统计专家而担当顾问。1936 年发表了 Carnegie《英语通用词表》（General Service List of English Words），约有 2000 个标题词。当时还没有计算机，词汇统计工作都是人工做的，耗费了很多人力。

但是，绝大多数的常用词都是多义词，词表必须包括义项，才能起作用。West 又进一步对词表的 2000 个标题词的意义进行统计，用百分比来表示各种意义的比例。这更是一项十分繁重的工作，从 1939 年开始，到 1953 年才结束问世。这份通用词表成为每一个英语教师的案头手册。

50 年代中叶以后，由于转换生成语法的出现，词汇教学在英语教学中的地位有所削弱。但是最近 20 年来它又复苏了，词汇能力到底是学生语言能力的一个不能忽略的组成部分。而且由于计算机的普及和它的功能的不断增大，语料库的建立再不是一件难事。词汇的频率统计更趋于精确和科学化。

在我国，培养学生的词汇能力始终是英语教育中的一个受到广泛注意的焦点。从教学大纲、教材、教学方法到测试，都少不了词表。但是多数词表的制定仍然是经验性的。而且更重要的是，我们对学生所掌握的词表和我们所制定的词表有些什么距离，缺乏科学的研究。中国学生英语语料库的建立使我们有可能了解学生使用词汇的实际情况。马跃博士的《从中介语看语义习得——基于语料库的中国学生英语书面语动词使用情况研究》就是根据中国学生英语语料库来观察他们词汇能力的一项具有重大现实意义的研究。

这项研究的特点是把中国学生使用英语动词的义项比例和 West 的通用词表的比例加以比较，从而了解我们所制定的词表和学生所使用的词表的差别。对词义进行标注，是语料库语言学的一个难点。目前还没有自动标注的程序，必须根据它所出现的上下文来进行人工标注。这项研究没有绕过这个难点，而且选择了通用词表中的动词，再根据中国学生英语语料库的实际使用情况进行检索和标注，从而得到义项的比例，以资比较。从这个方面来说，这项研究是具有开创性意义的。

但是到目前为止，这项研究的价值仅在于建立一个模型，还有许多工作要做。如果作者能够继续努力，把通用词表所有标题词的义项都进行标注和统计，和 West 的统计并列在一起进行比较，那就能够为我国英语教学，特别是制定英语教学词表做出一件奠基性的工作。我们翘首以待。

《从中介语看语义获得》

（基于语料库的中国学生英语书面语动词使用情况研究）

SEMANTIC ATTAINMENT IN INTERLANGUAGE _ A CORPUS-BASED STUDY OF VERB USES THROUGH CHINESE L2 LEARNERS' WRITTEN ENGLISH TEXTS

内容提要

在第二语言习得（SLA）研究中，研究人员经常从第二语言学习者的中介语入手考察二语学习者的成绩和存在的问题。因此，中介语也就成了人们在评价学习者二语获得的成绩时经常要研究的对象。随着语料库语言学的发展，我们现在可以利用电脑对二语学习者留下的大量的书面语料来进行此类研究。

往常对于学生第二语言习得情况的评估大多是通过错误分析进行的。人们对于中介语中的错误进行分析，以便决定一个学生掌握所学外语的情况，或者是对于学生学习外语的实际情况做出评估。相形之下，本文所采用的基于语料库的研究却是通过对于学生正确使用所掌握的二语来表达自己的意思的了解，特别是如何使用英语动词来表达自己意思的情况，考察学生语义获得的状况。

语料库语言学所取得的令人振奋的发展使教师们在案头上进行前所未有的大规模的经验性研究成为可能。因而本文所进行的研究也想为语言教师们提供一个范例，以展示孤军奋战的语言教师们也可以受益于当今现代科学技术所提供的机遇，利用电脑来从事第二语言习得研究。我们根据二语学生所提供的大量的书面语料，对学生运用英语动词的侧面加以观察，以评估学生语义获得的情况。这些语料取自于一个叫做 *CLEC* 的语料库（即 *C*hinese *L*earners' *E*nglish *C*orpus，中国英语学生语料库）。 这是我的导师桂诗春教授牵头建立的国内首个较大型

的外语中介语语料库，是他所领导的一个大型的“九五”人文社科基金研究项目的一部分。我们从该语料库中提取语料进行研究，并与 Michael West 在 1953 年发表的经典著作 *A General Service List of English Words*（英语通用词表）中的语义频率数据加以对照分析。后者是一部非常独特的著作，虽然其中的数据是根据电脑时代之前所建立的一个语料库中的材料所得出的，但它所提供的英语词汇的语义频率信息是空前的，在某种程度上可以说迄今为止仍然未有第二部著作能够为读者提供如此详细的语义频率信息。 另外，我们还将其与其他英语本族语语料库中所提供的可比信息进行对照分析。这些分析对比的目的是想考察把英语作为第一语言和第二语言的使用者在英语动词词义的使用上究竟是否存在显著性的差异。这种差异是否关乎到诸如英语本族语的使用者读起第二语言学生的书面语来会感到“洋腔洋调”之类的问题。然而，由于学习年限所限，在当前的研究中我们仅限于调查和比较动词的不同义项的使用。这儿所做的一切仅仅是为今后同类的研究做出一点铺垫或者说是提供一种研究的模式。我们的最终目的是要对 West 的通用词表中所涉及的所有词的语义信息进行分析。对于二语语料所进行的调查得出的数据将会与 West (1953) 中对于英语本族语语料调查所得的语义频率信息并列在一起供人们参考利用。

人们通常认为二语学生在使用外语上较本族语使用者逊色许多。他们不仅说起话来南腔北调，口音浓重，而且他们的外语书面语打眼就能看出是老外的作为。除去较为明显的语法和拼写错误是中介语的显著特征之外，还有其他东西能够使中介语具有明显的不同于本族语的区别性特征的吗？换句话说，除去错误，我们是否可以从中介语中二语学生对于外语的“正确”使用的过程中发现一些区别于本族语使用者的东西呢？ 既然“意义”是语言交际过程中的一个重要组成成分，我们认为针对二语学习者的“意义”而不仅是“形式”进行深入的了解一定会令我们得到一些有启发意义的东西。通过对语中介语中的这些侧面的深入了解，我们可能会给语言教师们提供一些有益的线索，以便让他们更好地决定在课程中教什么和不教什么，他们需要强调学生做什么等来正确引导学生的注意力。

毋庸讳言，“意义”是语言学研究和语言习得研究中历来最具争议的研究领域。为避免争议，我们完全可以通过一个设计严谨的心理语言学实验来研究理解和产生意义的过程，或者用其他手段和严密的控制来研究某个有限的语境中语言

单位的语义特征和特点。迄今为止通过大量的语料从义项的使用频率上来研究语义的做法是十分罕见的。 Michael West 的 General Service List of English Words 代表的就是这样一种难得的尝试。这一项研究是由一批“人类分析专家”根据一个早期的非机读语料库来进行的。McEnery and Wilson（1996）是这样来形容 West（1953）的，这是“一部词义频率词典，迄今为止还没有人能够超越它”。人们迄今对此类工作缺乏兴趣的原因明显是因为这样性质的研究需要庞大的人力组合、时间、精力和金钱才能完成。但是，由于计算机技术的进步，基于语料库的研究使人们利用机读语料来对巨大的数据库来进行系统分析已经成为可能，这种研究对于赤手空拳孤军奋战的语言教师来说曾经是连想都不敢想的，更不必说去尝试着做了。虽然全自动的词义分析时代离当今的普通语言研究者和教师们来说还相去甚远，但是现在人们在家中案头上利用电脑和从互联网上可以下载的软件来比较轻松地分析处理的数据量之大已经是空前的了。尽管我们还存在利用全自动的设备进行词义分析的困难，本研究课题所涉及的领域利用当今计算语言学所提供的现成技术已经足可以让我们放胆涉足我们关心的问题。不过单枪匹马的从一个有上百万词的语料库中去研究词的义项不会是一项轻松的工作，这仍然是一项单调、耗时和十分具有挑战性的尝试。当然，比起前计算机时代对于同样大小的非机读语料库来说，现在可是好多了，效率不知该提高了多少倍。

鉴于当前的研究是一项基于语料库的旨在发掘二语学习者中介语的状况的由数据驱动型的研究，而非为检验假设而通过严密控制因素的实验从与二语学习者进行直接交流的过程中来获取数据的实验性研究，在我们对于数据的分析中自然既涉及定性分析又涉及定量分析。我们要对学生的书面语料进行标注，以便于使用文本分析工具进行处理，我们还要把其中的部分数据抽出来再进行详细的分析。对于其中的有些数据我们要进行量化的统计分析，以便从二语学生的看上去是松散和孤立的英语动词的书面语的使用中推断出一些有规律性的东西。该研究本身所具有的归纳性的特征和缺乏与原作者当面交流的问题使得我们不可能通过进一步的有严格控制的实验程序和推断分析来发现其产生这些数据的过程，这是其遗憾之一。对数据分析结果的意义所做的评估和解读以及对于出现的问题和可能的原因的追朔均难以离开定性分析。大概这是对于非介入性的研究所能够做的最好的补救。在本研究的过程中，我们将致力于对于二语学习者可以观测到的行为的研究，而不是去猜测当语言处理发生时在大脑这个黑箱里到底发生了什

么。在本项研究中，为保证样本的代表性，我们依据在 West 和英国国家语料（BNC）中均为常用动词这样一个标准，共从 CLEC 中抽取了 683 个类的动词进行标注分析。可最终发现仅有 380 个动词能够符合我们进行深入比较分析的条件，仅占最初抽样的 56%。其主要原因是：一些动词在 CLEC 中的频率太低，没有比较意义，或者这些词在 West 中频率太高，原书没有给出具体的义项频率数字，因而无法进行有意义的比较。我们把从 CLEC 中抽出的样本按照高中低三个组进行了划分，分别对其中所包含的动词做了标注和统计。我们把按照动词义项的数目与 West 的常用词表中所提供的语义频率进行了比较，并根据英美几个著名的语料库中所提供的一些信息进行了间接的对照比较，得出了一些有用的数据。我们还依据 Biber 等人（1999）对于 Longman 语料库中的动词所做的义域分类对我们的动词做了类似的处理，并与我们能够了解到的 Longman 语料库的一些信息进行了分析与对比，发现了一些很有趣的事情。我们把一些重要的数据和一些有参考价值的数据表作为附录收在本文的最后。

在本研究中我们做了三个关于我们可能会有什么样的发现的基本假设或者说是预测。首先，我们认为第一语言和第二语言的数据在词义选择方面是随二语学习者的习得或学习水平高低呈正相关关系。我们相信在第一语言和第二语言的语料中相似性和差异并存。其次，我们相信通过聚类分析我们可以发现二语学习者语义获得的程度与本族语人群的语义选择相比存在有明显的规律性。我们会自然预期水平越高的学生与本族语的人群在语义选择上就越接近，反之，差异就增大。再者，由于词语的搭配可被视作是表现动词在使用过程中是否广泛的一个程度指标，我们期待对于词语搭配的研究可以解释本族语的语料会比第二语言的语料涉及更多的词语搭配类别。有趣的是，我们的研究发现均证实了上述三项预料所期待的答案。分析的结果表明在动词词义选择方面第一语言和第二语言的语料中的确相似和差异并存。第二语言学习者的语义获得与本族语的数据存在明显的有规律性的相关，与学生对目标语的接触程度呈正相关关系。虽然词语搭配研究似乎显示第一语言和第二语言使用者在动词词义的范围或广度选择上存在着显著的差别，我们还是发现第二语言学习者在语义获得方面稍好于句法获得。我们从本研究的发现中受到了鼓舞，今后会继续通过此类基于语料库的研究方法去更为深入地探讨第二语言学习者的中介语中所隐藏的奥秘。

ACKNOWLEDGMENTS

The author wishes to express sincere appreciation to Professor Gui Shichun, my supervisor, for his constant support and encouragement in the preparation of this dissertation. It was Professor Gui, forerunner of applied linguistics in China, who first opened my eyes to the promising new land of the corpus-based applied linguistic studies and provided me with a walking stick for obtaining my primary education in corpus linguistics from the Internet. His immanent eagerness and interest in pursuing the truth and meeting the ever-emergent challenges push me into the probe in the new realm of SLA studies that I have never alone had the courage to tread.

Special thanks are due to all the other faculty members of the Institute of Linguistics and Applied Linguistics at GDUFS, who have joint the effort by Professor Gui in training all their PhD candidates in ways few domestic institutions have ever attempted to ensure the quality of education that bears the seal of GDUFS. Both GDUFS and Jinan University deserve my thanks for providing me with the opportunity of in-service training and for testing me trustfully and friendly by adding unwanted administrative responsibilities to me during the course of my PhD candidacy. Thanks are also due to the members of the CLEC project for their kind permission for my access to the fruit of their candid and painstaking efforts.

I would also like to thank my parents, both aged over 80, who have tolerated my absence from visiting them during the entire course of prolonged study and dissertation writing. Equally I would like to thank my wife and daughter for my unintentional negligence of my duties at times as a husband and father.

Table of Contents

List of Figures x
List of Tables xi
Acknowledgement xii
1. Introduction 1
1.1. Purpose of the present study 1
1.2. Raison d'être 2
1.3. Overview of the present study and major terms defined 4
2. Relevant Theoretical and Practical Issues in Perspectives 6
2.1. IL and relevant issues in SLA 6
2.1.1. Interlanguage 8
2.1.2. Error analysis 10
2.1.3. Linguistic Universals and L1 Transfer 13
2.2. Corpus linguistics and its applications 15
2.2.1. Corpus and computational linguistics 16
2.2.2. Applied corpus linguistics and the Internet 19
2.2.3. Word sense disambiguation 25
2.2.4. WordNet, Semcor and FrameNet 30
2.2.5. Cobuild and TACTweb: online aids to corpora studies 33
2.3. The role of lexis in L2 teaching and learning 35
2.3.1. A historical perspective on vocabulary teaching and research through vocabulary selection 35
2.3.2. West's A General Service List of English Words 39
2.4. Verb classifications 43
3. Research Design and Methodological Issues 47
3.1. Research Design 48
3.1.1. Rationale 48
3.1.1.1. Why a contrastive study between West's list and Chinese learners' performance? 49
3.1.1.2. Why choosing verbs as a model construction interface? 51
3.1.1.3. Similarities and dissimilarities 52
3.1.2. Hypotheses and predictions 53
3.1.2.1. Hypothesis one: Correlations in L1 and L2 data in sense selection are positively corresponded to the level of L2 acquisition and learning 53

3.1.2.2. Hypothesis two: Clustering in L1 and L2 data in sense selection is patterned according to learner levels in L2 acquisition and learning 55
3.1.2.3. Hypothesis three: There should be more collocates in L1 data than in L2 data 56
3.1.3. Data 57
3.1.3.1. Groupings of CLEC 58
3.1.3.2. Groupings of verbs 59
3.1.3.2.1. A complete list of the chosen verbs 59
3.1.3.2.2. Grouping of the verbs according to their meanings 63
3.2. Methodological issues 73
3.2.1. Tagging of meanings 74
3.2.2. Comparison of semantic frequency 80
3.2.3. Studies of collocates 82
3.2.4. Identification of similarities and dissimilarities by statistical means 85
3.3. The corpora involved 86
3.3.1. CLEC 86
3.3.2. LOB 87
3.3.3. BNC 88
3.3.4. Longman corpus 88
3.4. Textual analysis tools 89
3.4.1. MicroConcord 90
3.4.2. *TACT* 91
3.4.3. WordSmith Tools 92
4. Findings and Discussion 94
4.1. Quantitative Analytical Procedures 94
4.1.1. Comparison with West's *Service List* 94
4.1.2. Comparison against Longman Corpus 98
4.1.3. Comparison against LOB 104
4.2. Qualitative procedures: results from case by case examinations 110
4.3. Keyword studies 116
4.4. Findings in summary 120
4.5. Related Issues in Perspectives 122
4.5.1. The vocabulary lists in curriculum guides examined 123
4.5.2. Proposal for newer vocabulary guides 126
5. Conclusion 129
5.1. Summary 129

5.1.1. Conclusions regarding Hypothesis One: Patterns hidden in the seemingly different L1 and L2 verb uses 129
5.1.2. Conclusions regarding Hypothesis Two: L2 learners' repertoire of verbs comparable to L1 speakers' 130
5.1.3. Conclusions regarding Hypothesis Three: Analysis of verbs sense selection patterns point to tendency of better semantic than syntactic attainment for L2 learners 131
5.2. Implications for language teaching and learning 133
5.3. Recommendations for further studies 134
Glossary 137
Bibliography 144
Appendix A: Verbs by Semantic Domains from West and the CLEC 151
Appendix B: Verb lists from BNC & West's Service List 158
Appendix C: Percentage curves for top 10 verbs 168
Appendix D: Comparison of Verb Sense Uses in Percentages between 380 verbs in both West (1953) and the CLEC 169
Appendix E: Key Verb Uses Compared in CLECMA against BROWN and CLECMA against LOB 182

List of figures

Number *Page*

1. Psycholinguistic sources of errors 11
2. The role of the L1 in L2 communication and learning 14
3. Corpus-aided language teaching 22
4. Levels of WSD Algorithms 29
5. Distribution of the CLEC verbs by semantic domains 72
6. Search the verb 'balance' with an asterisk 76
7. Search the verb 'balance' with different inflectional forms 76
8. Summary screen for the verb 'balance' 77
9. Concordance of the verb 'balance' 77
10. Larger context for the verb 'balance' 78
11. Sense marking for the verb 'balance' 78
12. Part of collocates for verb 'do' from the LOB 84
13. Part of collocates for verb 'do' from the CLEC 84
14. Data before correction and data after correction 96
15. Cluster analysis of the West's and CLEC data 97
16. Percentages of West's verbs among the top 3000 in Longman 99
17. Distribution of semantic domains in LSWE & CLECMA compared 100
18. Distribution across registers in LSWE corpus 101
19. Frequencies of the top twelve lexical verbs in LSWE Corpus 102
20. Frequencies of the twelve verbs from the sub-corpus of the CLEC 102
21. Top 12 most common (lexical) verbs in the sub-corpus of the CLEC 103
22. Top twelve most common lexical verbs distributed by registers 103
23. Plot of collocates from 10 verbs from the CLEC against the LOB 109
24. Percentage curves for the verb DO 114
25. Vocabulary for middle schools mapped against West (1953) 124

List of Tables

Numbers *Page*

1. Areas of studies in SLA 7
2. Issues of IL Studies in Comparison 9
3. Growth of Corpus Linguistics 16
4. Correlation coefficients and the degree of data correspondence 54
5. Number of valid cases from raw data count after data trimming 60
6. A complete list of the verbs chosen for the current investigation 63
7. Verbs and their occurrences by semantic domains in the sub-corpus of the CLEC 72
8. Correlations between BNC, BROWN and West verb lists 82
9. Linear correlation between West and the CLEC data 95
10. Correlation on a nonparametric model 95
11. Number of collocates for ten selected verbs in both CLEC and LOB 104
12. Collocates (types) after removal of less relevant features 105
13. Collocates of the verb 'do' from the CLEC compared against the LOB ones 107
14. Statistical comparison between the collocate types from the CLEC and the LOB 108
15. Similarity matrix for CLEC and LOB verbs 109
16. Frequencies for the upper 10 verbs in both West and the CLEC 111
17. Frequencies for the middle 10 verbs in both West and the CLEC 111
18. Frequencies for the lower 10 verbs in both West and the CLEC 112
19. Basic information of the three corpora compared 118
20. Keyword figures between the three corpora compared 119

Chapter 1

INTRODUCTION

1.1. Purpose of the present study

In second language acquisition (SLA) studies, interlanguage (IL) is often the area that researchers will look at in assessing the learners' attainment of aspects of the L2. One common way of determining the learners' conditions in L2 acquisition and learning is through *error analysis* (EA); in which learners' interlanguage errors are addressed in order to decide how well a learner knows the language or how a learner actually learns a language. In contrast, in the present study we intend to address the issues through detection of how well the learners can *USE* the language by looking directly at their ability in expressing meanings in the L2, in particular how well they can use verbs to express themselves in English.

We approach the problem by looking at a large body of written texts produced by different levels of Chinese learners of English, which we have abstracted from *the CLEC* (see 3.3.1 below for details), viz. *C*hinese *L*earners' *E*nglish *C*orpus, a project developed under the supervision of Professor Gui Shichun. Typologically speaking, the present study would naturally fall within the field of corpus linguistics; more exactly it is an area within the field that deserves the label *corpus-based applied linguistic studies*, which examines the more practical and applied aspects of corpus linguistics rather than involving itself in the very technical area of the basic theory development and arguments in corpus construction and related issues. It does, however, intend to set up a model for assessing L2 students' linguistic abilities or at least to explore the L2 acquisition and learning issues through a way made possible by computers that formerly would loom too time-consuming and definitely unattainable by an individual researcher because of the huge manpower mobilization involved in such researches. The exciting development in corpus linguistics has made it possible for language teachers to undertake unprecedented empirical researches on desktops. Thus it is also our hope to demonstrate the feasibility of how individual language teachers can cash in on the opportunities that new science and technology have brought about in this age of SLA study aided by computers. Sampson (1992:181-200) points out: