国家哲学社会科学“九五”规划项目

中国学习者英语语料库

Chinese Learner English Corpus

桂诗春　杨惠中　编著

外教社

上海外语教育出版社

图书在版编目(CIP)数据

中国学习者英语语料库/桂诗春,杨惠中著. - 上海:
上海外语教育出版社,2002
ISBN 7-81080-531-2

Ⅰ. 中… Ⅱ. ①桂… ②杨… Ⅲ. 计算机应用-英语-语言教学-研究 Ⅳ. H319

中国版本图书馆 CIP 数据核字(2002)第 050515 号

出版发行：上海外语教育出版社
（上海外国语大学内） 邮编 200083
电 话 021-65425300（总机），35051812（发行部）
电子邮箱 bookinfo@sflep.com.cn
网 址：http://www.sflep.com.cn http://www.sflep.com
责任编辑：李法敏

印 刷：上海江杨印刷厂
经 销：新华书店上海发行所
开 本：787×1092 1/16 印张 45 字数 1119 千字
版 次：2003 年 1 月第 1 版 2003 年 1 月第 1 次印刷
印 数：2 300 册

书 号：ISBN 7-81080-531-2 / H · 199
定 价：76.00 元

前　言

“基于语料库的中国学习者英语失误分析”是国家社科基金“九五”规划的一个项目，由桂诗春、杨惠中负责。按照原定的计划，这个项目由两个部分组成：一是建立一个100万词的中国英语学习者的书面语语料库；二是根据这个语料库对中国学习者的英语失误进行分析。这两部分分别体现为《中国学习者英语语料库》和《中国学习者英语失误分析》两本书；它们互相补充，结成姐妹篇。

现在呈献在读者面前的是《中国学习者英语语料库》，首先需要说明的是：为什么只收集100万词的语料？按目前的技术条件，建立几百万词、甚至上千万词的语料库并非难事。但是我们的语料库必须对言语失误按照统一的言语失误表进行标注，而失误标注又颇为耗费人力和时间。由少数人来做，易于统一，但需假以时日；由较多的人来做，可以加快进度，但却不易统一。如果增加语料，工作就更为繁复，所以我们定在100万词，以便于操作，同时又可以积累经验，以利于语料库以后的扩充。所以我们对语料库的理解是它是一个不断扩充的工程，只要有可能，我们的语料库将来还可以在规模、层次、标注等方面继续扩充和完善。

语料库的出版也不是易事。BROWN语料库的篇幅是424页，Carroll的 *Word Frequency Book* 的篇幅是856页；而我们的语料库还增加了关于言语失误的各种统计资料，篇幅更不会小。好在现代电子技术发达，我们不但有可能把一些次要的表格、甚至连整个语料库本身也可以收录在光盘里。在目前向读者提供的光盘里，我们除了各种表格外，还提供了整个语料库和各个分体语料库，并且有一个简单的检索程序，让读者可以根据教学需要查阅全体（或不同类型的）学习者的语料和言语失误，进行研究。读者也可以使用一些通用的语料库工具，做更多的分析。因为我们的语料库都用统一的.txt格式储存。语料库中所使用的各种统计分析手段都是目前语料库研究中通用的手段，绝大部分都是利用目前较为成熟的程序。在正文里我们也做了一些初步的解释，但不可能太详尽。好在中国的图书市场上现在也有几本关于语料库的著作，如Biber等人的 *Corpus Linguistics*，Kennedy的 *An Introduction to Corpus Linguistics*，读者可以参考。

参加本项目的同志除桂诗春、杨惠中外，还有杨达复、何安平、李文中、濮建忠、卫乃兴、雷秀云、周海中、常新萍、廖海青、常晨光、王哲、戴凡、王初明、许罗迈、曾用强、李金辉、杜金榜、魏新华、贾冠杰、田宝堂、罗颖、王龙吟等，他们都是上海、广州、河南等地的高等院校教师。他们在编制本语料库的工作中付出了大量的劳动，有的负责制订编制英语失误分类表、有的负责标注、有的负责编写程序。在整个语料库的标注完成后，又由桂诗春、杨达复、何安平三人进行标注的统一。语料库中的中学和大学英语专业的语料采集自北京、上海、广州、河南等地的中学和高等院校，大学英语的语料由CET考试委员会提供，清华大学外国语学院也提供了一些大学英语六级的语料素材。目前这个语料库虽然也还比较粗糙，在许多地方也有待完善，但是它总算是我们的阶段性成果的一个标志。我们热诚地希望海内外的读者和同行提出批评和建议。另外，上海外语教育出版社对《中国学习者英语语料库》的出版给予了十分可贵的支持，责任编辑李法敏同志在该书的策划、编辑、校对等方面十分认真负责，在此我们表示衷心的感谢。

桂诗春
杨惠中
2002年9月

目　录

1 导　言

中国学习者英语语料库(Chinese Learner English Corpus,CLEC)是国家社科基金"九五"规划项目"基于语料库的中国学习者英语失误分析"(Corpus-based Analysis of Chinese Learner English,CBACLE)的一个重要组成部分。本书所载的是 CLEC 的各种统计资料和列表;对中国学习者英语错误的各种分析另收集在《中国学习者英语失误分析》一书里。两书为姐妹篇,供读者互相引证。

在某种意义上说,语料库语言学是一种研究方法,而这种研究方法是借助计算机来实现的,故 Leech(1998a)主张把语料库语言学(corpus linguistics)说成是计算机语料库语言学(computer corpus linguistics)。随着计算机的普及和现代技术(高速的中央处理器、精密的扫描仪和字母识别程序、大容量内存和硬盘,等等)的发展,这种研究方法在最近 20～30 年间有了很大的发展。McEnery 和 Wilson(1996)对使用语料库方法来进行语言学研究的发展归纳如表 1.1。根据英国 Lancaster 大学 Taylor, Leech 和 Fligelstone 等人在 1989 年的统计,英语的机读语料库当时已有 36 种,非英语的有 18 种。Hofland 等人 (1999)更把 18 个大型的英语语料库制成 ICAME(International Computer Archive of Modern English)英语语料库光盘(第二版),公诸于世。

表 1.1　语料库研究方法的发展

时期	研究数目
1965	10
1966－1970	20
1971－1975	30
1976－1980	80
1981－1985	160
1986－1991	320

语料库方法可以广泛地应用在语言学的各个领域(句法学、语音学、语义学、语用学、社会语言学、心理语言学、应用语言学,等等)。语料库和语言教学有密切的关系,它成为 1994 年 ICAME 年会的主题,1997 年由 Wichman 等人将该年会论文编辑为《教学与语言语料库》(1997)。根据 Leech(1997)的说法,语料库运用到教学可以是直接的(如对学习者讲授语料库语言学、教他们使用语料库、利用语料库进行教学, 等等),也可以是间接的(如编辑词典、编写教材、语言测试,等等)。Leech 还提出建立专门用途英语语料库、母语和二语语言发展的语料库、双语和多语语料库,以进一步探索语料库对教学的作用。二语语言发展的语料库也可称为学习者语料库(以后均略为 LC,learner corpus)。Granger (1998)所编著的《计算机上的学习者英语》收录了 15 篇关于 LC 的论文,体现了语料库语言学研究者近年来探索在语言教学中使用语料库的各种努力。LC 还可分为有标注(tagged)和无标注(untagged)两种,而有标注的 LC 还可以从不同的角度进行标注。从语法角度的标注叫做语法标注(grammatical tagging),主要是对词类(parts of speech, POS)标注;现在已经可以根据概率的原则,用计算机来进行自动化处理,准确率最高达 95%～97%。另一个角度从

学习者的言语失误[①] 来标注，叫做失误标注(error tagging)。它需要由人工进行，难度大而工作繁重，所以尽管有一些人在做这方面的努力，到目前为止，还未有一个对言语失误进行标注的 LC 正式问世。我们所建立的 100 万词的 CLEC 组织了一批教师对言语失误进行标注，体现了一种很有意义的尝试。它现在已经放在因特网上供教师试用，希望能获得反馈，以作进一步的改进；我们更希望教师们利用语料库所提供的信息，对中国学习者的英语特点和英语失误进行探索，产生更多的研究成果，推进我国的英语教学。

根据 Leech(1998b)的说法，建立 LC 的目的是：

- 比较 LC 和以目标语为母语的语料库(以后均略为 ECNS，English corpus of native speakers)，看有哪些语言特征是超用的(overused)或少用的(underused)？
- 学习者的母语在多大程度上影响了他们使用目标语的行为？
- 学习者的目标语在哪些方面达到或未达到目标语说话人的言语行为？
- 学习者有哪些主要方面(按照频数)未能符合目标语说话人的言语行为而需要特别的帮助？

这意味着我们需要从两个方面来分析学习者语料：一个方面是对比分析 LC 和 ECNS 的异同，我们选择了美国英语的 BROWN 语料库和英国英语的 LOB 语料库，因为这两个语料库的数目都是 100 万个词左右，而且我们的学习者有的学美国英语，有的学英国英语。但是这两个语料库反映的是 20 世纪 60～70 年代英语的使用情况，所以我们也尽可能使用 FROWN 和 FLOB 来进行对比。这两个语料库是德国 Freiburg 大学根据 BROWN 和 LOB 两个语料库的采样方案收集 90 年代美国和英国英语语料建成的语料库。另一个方面是分析 LC 的言语失误，这是他们言语行为偏离目标语说话人的主要方面。应该说明的是学习者的语料偏离 ECNS 有许多方面，例如语言风格、文化色彩和母语影响等等，我们目下还未能对它们进行标注。因为对它们的标注有争议，而且带有较强的主观成分；而我们参与标注的人较多，难以统一。但我们的语料库一旦公诸于世，研究者就可以根据自己的需要对失误进行再分类和再标注，以便根据特定需要作更深入的研究。

---

① 失误是失检(mistakes)和错误(errors)的合称。一般人把语言运用(performance)中的误差叫做失检，这些误差是学习者可以自行检查出来，并作更正的；而错误则是语言能力(competence)中的误差，学习者不能自行更正。但是这种区分是从解释误差的角度提出来的，学习者语料库仅能提供言语误差，至于它们是失检，还是错误，则无法说明。例如一个词拼写错了，是因为学习者已经懂得它的正确拼写法，但在使用中不小心拼错了，还是因为学习者根本不懂其正确拼写法，需要研究者根据具体情况来解释，语料库是无能为力的。所以我们把这两种情况笼统称为失误。

2 CLEC 的建立

2.1 样本的选定

表 2.1 CLEC 语料分布

类型	词次
ST2	208088
ST3	209043
ST4	212855
ST5	214510
ST6	226106
总计	1070602

LC 和 ECNS 最主要的不同是学习者本身是有差异的，他们的语言发展居于不同的阶段，所以样本必须来自不同发展阶段的学习者，而制订 ECNS 抽样方案则考虑文体类型(genre)而不是语言能力。学习者的写作能力只是停留在"一般的"英语(例如我们不能期望我们的学习者去写小说、社论、科技文章，而这些类型是一般 ECNS都有的)。从整体上看，我们所建立的 LC 基本上是同质的(homogeneous)，都是中国的英语学习者；从分体上看却是异质的(heterogeneous)，他们处于不同的发展阶段。我们定为 5 个阶段，如表 2.1[①]：

a) 中学阶段，主要是高中生，因为初中生还没有写作课。代号为 ST2。
b) 大学英语 4 级，大学 1～2 年级非英语专业学习者，多数学习者将参加 CET4 级考试。代号为 ST3。
c) 大学英语 6 级，大学 3～4 年级非英语专业学习者，多数学习者将参加 CET6 级考试。代号为 ST4。
d) 英语专业 1～2 年级学习者。代号为 ST5。
e) 英语专业 3～4 年级学习者。代号为 ST6。

整个语料库的语料有 100 万词，每一类型的学习者的语料为 20 万词。为了避免学习者在考试时往往采取回避策略(strategy of avoidance)，避免写一些没有把握的东西，因此语料采样应不仅来自考试的试卷，还应来自课内外的作业。前者称为试卷作文，后者称为自由作文[②]。

由于采样和录入的困难，目前的语料库严格来说是书面英语的语料库。但是初级的英语

---

① 这是经过处理后的数字，未经处理的原始语料库为 1207879 词，整理原则见 3.11 词频排列表(按频数)。

② 我们在研究过程中发现，试卷作文和自由作文在语言运用方面有很大差异：试卷作文是在考试环境下的语言运用，不仅有时间和考试规则的限制，而且不允许考生查阅词典和参考书，此外还有考试焦虑因素的影响等等，因此试卷作文是一种非常态语言运用。目前 CLEC 中所收集的学习者语料，ST2、ST5 和 ST6 都是自由作文，而 ST3 和 ST4 主要是试卷作文。由于来源不同，目前 CLEC 中的数据仅适宜于做同类语料的比较，但不适合作纵向的比较，即不能用来说明中国学习者的语言发展过程。从长远来看，本项目将开展后续研究，把整个 CLEC 语料库分为两个子库：CLEC1 全部由自由作文构成，CLEC2 全部由试卷作文构成，这样才能作纵向的比较。关于试卷作文和自由作文的进一步讨论请参阅《中国学习者英语失误分析》一书中的有关文章。

学习者不会在语体上区别目标语，所以他们所写的往往就是他们要说的话。

2.2 样本的处理

样本的处理在 LC 里也是比较特殊的。一般的语料库可以通过扫描仪和光学字母识别程序来建立，十分方便。但是我们的样本都是学习者的手写文字，需要组织专人来输入，而且还要找人来校对，以免出错。我们的处理程序如下：

a）选好样本。

b）输入样本。

c）校对。

d）对言语失误进行标注。

e）对标注进行统一。统一最好由一个人进行，使标准得以统一。但语料太多，我们最后由三个人把失误类型分为三大部分来进行统一。

f）对语料库和言语失误进行统计分析。

g）建立语料库索引检索器(concordancer)。

h）将语料库索引检索器和整个语料库放在因特网上试用。

i）将语料库的统计数据公开发表。

2.3 言语失误分类表的制订

对语料库中的学习者的言语失误进行标注是本语料库最主要的特点。在制订言语失误分类表前，我们考虑了几条编制原则：

a）简单合理，易于系统操作。参与标注的人比较多，分类表过于繁复，就难于掌握。我们采取两级分类，第一级有 11 类：词形、动词短语、名词短语、代词、形容词短语、副词、介词短语、连词、词汇、搭配、句法。每一类里再用数目字细分。如[cc]为词语搭配不当，[cc1]表示名词和名词的搭配，[cc2]表示名词和动词的搭配，[cc3]表示动词和名词的搭配，等等。

b）分类表的类别要适中。过粗容易统一，但信息太少，不利于分析学习者的失误；过细难以统一，容易把同一种失误归到不同类别。目前我们采取的办法是对常见的失误从细，对少见的失误从粗。现在的分类表有 61 个失误码，是属于中等规模的分类表。

c）提供足够的失误信息(失误本身、失误类型和失误发生范围)。例如 In the past, people are [vp6, 4 -] kind to each other...，失误标注用方括号表示，放在失误内容之后。[vp6] 为 vp(动词)第 6 种(时态)失误；4 - 为失误发生的范围，- 表示失误的位置，4 表示失误前有 4 个词。要联系这 4 个词，才能判断 are 这个词用错了。

d）开放性。容许研究者根据需要对失误类型进行补充或进一步再分出细类。例如[sn8]为句子结构有缺陷，研究者可以对这种失误再分为若干细类来研究。这需要把 sn8 的失误全部检索出来，然后定出第三级的分类范畴，如 sn81，sn82 等等。

e）对语体误用或失误的来由暂不作标注，因为这需要标注者较多的主观判断，更难以

统一。

表 2.2　言语失误分类表(总数:61)[①]

词形		动词短语		名词短语		代词	
码	类型	码	类型	码	类型	码	类型
fm1	spelling	vp1	pattern	np1	pattern	pr1	reference
fm2	word building	vp2	set phrase	np2	set phrase	pr2	anticipatory it
fm3	capitalization	vp3	agreement	np3	agreement	pr3	agreement
		vp4	finite/non-finite	np4	case	pr4	case
		vp5	non-finite	np5	countability	pr5	wh-
		vp6	tense	np6	number	pr6	indefinite
		vp7	voice	np7	article		
		vp8	mood	np8	quantifiers		
		vp9	modal/auxiliary	np9	other determiners		

形容词短语		副词		介词短语		连词	
码	类型	码	类型	码	类型	码	类型
aj1	pattern	ad1	order	pp1	pattern	cj1	pattern
aj2	set phrase	ad2	modification	pp2	set phrase	cj2	set phrase
aj3	degree	ad3	degree				
aj4	-ed/-ing confusion						
aj5	predicative/attributive						

词汇		搭配		句法	
码	类型	码	类型	码	类型
wd1	order	cc1	noun/noun	sn1	run-on sentence
wd2	part of speech	cc2	noun/verb	sn2	sentence fragment
wd3	substitution	cc3	verb/noun	sn3	dangling modifier
wd4	absence	cc4	adj/noun	sn4	illogical comparison
wd5	redundancy	cc5	verb/adv	sn5	topic prominence
wd6	repetition	cc6	adv/adj	sn6	coordination
wd7	ambiguity			sn7	subordination
				sn8	structural deficiency
				sn9	punctuation

① 我们没有把对表 2.2 和表 2.3 的说明译成汉语,因为没有统一译法,勉强统一容易引起误解。

表 2.3 标注说明

码	分类	类型	说明
fm1	word	spelling	spelling, coinage, abbreviation, apostrophe
fm2	word	word building	derivation, inflection, compounding, plurality (noun), irregularity(verb), 3rd person singular form(verb), syllabification, hyphenation, word division or fusion
fm3	word	capitalization	lower initial letter for upper initial letter or vice versa
vp1	vb phr	pattern	error in transitivity(vi as vt or vice versa), transitive verb pattern/grammatical (cf *Oxford Advanced Learner's Dictionary of Current English* edited by A. S. Hornby)
vp2	vb phr	set phrase	phrasal verb and verbal phrase: error in form or use
vp3	vb phr	agreement	number agreement with its subject (noun or pronoun)
vp4	vb phr	finite/non-finite	finite verb for non-finite verb or vice versa
vp5	vb phr	non-finite	infinitive error: form and use; infinitive for participle or vice versa; -ed participle for -ing participle or vice versa
vp6	vb phr	tense	error in tense use within a sentence; the sequence of tenses between sentences
vp7	vb phr	voice	error in the use of voice: active for passive or vice versa
vp8	vb phr	mood	error in the use of mood: imperative, subjunctive; improper structure of conditional sentences
vp9	vb phr	modal/auxiliary	misuse of modal/auxiliary verbs; wrong form of modal verb (or auxiliary verb) and verb combination (eg tense form, voice form, etc)
np1	nn phr	pattern	error in combination with other words/grammatical
np2	nn phr	set phrase	omission or replacement of a fixed element that goes after a certain noun
np3	nn phr	agreement	number agreement of a noun with its determiner or a word that refers to it
np4	nn phr	case	possessive case error: form or use
np5	nn phr	countability	uncountable noun used as countable noun
np6	nn phr	number	countable noun used with no determiner or -s; a or -s with plural noun
np7	nn phr	article	a/an confusion or definite/indefinite confusion
np8	nn phr	quantifiers	misuse or confusion between many/much, (a) few/(a) little, some/any, etc
np9	nn phr	other determiners	misuse or confusion of demonstratives, wh- determiners, numerals, etc

续表

码	分类	类型	说明
pr1	pron	reference	incorrect/ambiguous pronoun reference/anaphoric
pr2	pron	anticipatory it	improper or wrong use of anticipatory it; it replaced by a demonstrative, etc
pr3	pron	agreement	number agreement with a noun it refers to
pr4	pron	case	case error of any personal pronoun
pr5	pron	wh-	misuse or confusion of interrogative, relative and conjunctive pronouns
pr6	pron	indefinite	misuse or confusion of indefinite pronouns such as all/both, few/little, some/any, either/neither, etc
aj1	adj	pattern	error in the combination with other words/grammatical
aj2	adj	set phrase	error in the idiomatic use of an adjectival phrase; omission or replacement of a fixed element that goes after a certain adjective
aj3	adj	degree	adjective degree error: form and use
aj4	adj	-ed/-ing confusion	-ed adjective for -ing adjective or vice versa
aj5	adj	predicative/attributive	predicative adjective used as attributive adjective
ad1	adv	order	improper adverb placement/wrong position
ad2	adv	modification	adjective modifier used as verb modifier; other kinds of confusion
ad3	adv	degree	adverb degree error: form and use
pp1	prep	pattern	unacceptable combination with other words/grammatical
pp2	prep	set phrase	error in the formation or use of an idiomatic prepositional phrase
cj1	conj	pattern	unacceptable combination with other words/grammatical
cj2	conj	set phrase	error in the formation or use of a phrase functioning as a conjunction
wd1	word	order	misplacement of any word other than an adverb
wd2	word	part of speech	error in part of speech: right root but wrong word class
wd3	word	substitution	error in word choice: right word class but wrong selection (any part of speech)
wd4	word	absence	omission of a word(any part of speech)
wd5	word	redundancy	oversuppliance of a word(any part of speech)
wd6	word	repetition	unnecessary repeating of a word

续表

码	分类	类型	说明
wd7	word	ambiguity	not clear word meaning/semantic
cc1	collocation	noun/noun	improper noun(phrase) and noun(phrase) combination/semantic
cc2	collocation	noun/verb	improper noun(phrase) and verb(phrase) combination/semantic
cc3	collocation	verb/noun	improper verb and noun(phrase) combination/semantic
cc4	collocation	adj/noun	improper adjective and noun(phrase) combination/semantic
cc5	collocation	verb/adv	improper verb and adverb (or ad/v) combination/semantic
cc6	collocation	adv/adj	improper adverb and adjective combination/semantic
sn1	sentence	run-on sentence	improper addition of clauses/fused sentence
sn2	sentence	sentence fragment	subordinate clause as a sentence; any phrase as a sentence
sn3	sentence	dangling modifier	illogical adverbial modification of a clause
sn4	sentence	illogical comparison	error in the comparison of words or phrases in a sentence which can not be compared
sn5	sentence	topic prominence	the co-occurrence of an initial noun phrase and its equivalent(usually a pronoun) in the same sentence
sn6	sentence	coordination	faulty parallelism of clauses (or words/phrases) in a sentence
sn7	sentence	subordination	faulty attachment of a subordinate clause to the main clause
sn8	sentence	structural deficiency	error in the grammatical construction of a sentence: improper splitting, pattern shifting, confusing structure, etc
sn9	sentence	punctuation	overuse, absence, choice, apostrophe, comma splice, etc

2.4 语料库的制作工具

语料库是在计算机上实现的一个数据库，必须使用合适的软件来进行加工。这方面的软件已有不少，如 WordCruncher, MicroConcord, Longman's Concordancer, Concordance, Concordancer, Lexa, TACT, Wordsmith, 等等。经过实验和比较，我们决定使用 TACT 和 Wordsmith，因为它们的功能比较强大，而且是免费软件或共享软件。但是我们有特殊的标注要求，而且这些软件大都不能处理汉语(我们的 LC 虽然是英语的，但偶尔也有汉字，影响了文件的处理)，故我们也编写了一些专门的软件，如 Corpfind (供标注用。有的同志还用 Word 的自动图文集的功能编制言语失误分类表，找到失误后，按鼠标键入码，效果也

很好)，Cbrowser(供检索用)，Cleantxt(供清除汉字符号用)，Paragraph(供清除转行符用)，Merge(供合并和统计词表用)，PosTagger(供做语法标注用)，Lemma(作词目归并用)，Wordlist(作改正拼写后归并词表用)。所有的这些软件都要求语料库的文件是纯文本(.txt)格式。另外我们觉得 Microsoft Office 的 Excel 制作表格的功能十分强大，我们所做的表格都是 Excel 的.xls 格式的，必须装有 Excel 才能打开。对这些表格我们不作进一步转换，以便用户在 Excel 状态下进行数据处理。如有需要，用户可以在 Excel 下把文件另存为别的格式。Excel 本身也能做一些统计和制图工作；在需要做进一步的统计分析和制图时，我们使用了 SPSS，Statistica 和 Harvard Chart。

TACT 和 Wordsmith 都可以对语料库作统计分析，并进行索引检索。但是 TACT 可以定出检索条件(如全部语料或某一类学习者的语料)来检索词语或失误，而 Wordsmith 有一个特殊的功能，叫做 keyness(关键词性)，可以把两个语料库的词语频数进行比较，找出比参照语料库超用或少用的词语。例如我们可以把 5 类学习者的词表与一个参照语料库的词表进行比较，看哪些词语是各类学习者多用或少用的。在光盘里，我们提供了这两个软件，要使用 Wordsmith 的全部功能，必须注册。

3 CLEC 的统计分析

3.1 统计列表

3.1.1 词频排列表(按频数)

词频排列表(Rank List)按频数把语料库的词型从高到低进行排列,例如 the 的出现频数最高,共有 61787 次,排在第一位。对词频也可以按字母顺序排列,叫做字母排列表(Alphabetical List)。这两个表的数据是一样的,只是排列次序不同。本书只提供按频数的词频排列表,在光盘中编号为 I。光盘还提供按字母排列的词频排列表,编号 II。为了把 CLEC 的词频排列表和别的 ECNS 的词频排列表进行比较,我们必须对 CLEC 的语料做一些筛选处理。语料中有许多汉语拼音的专有名词和我们加到语料库里的失误标注,还有许多拼写失误,例如 *abilitical, *abilitities, *abilitys, *abillities, *ablelity, *ablity, *abtilities 等等,都是 ability 和 abilities 的拼写失误的不同形式。如果我们把它们都作为词型算进词频排列表里来和 ECNS 的词频排列表比较,则中国学习者的词汇量显然含有水分。故我们在编制词频排列表时,把汉语拼音的专有名词和失误标注加以剔除,把拼写失误的都改正过来。经过处理后,原来语料库的词次(tokens,语料库所有单词出现的次数)从 1207879 减为 1070602,词型(types,语料库中所有拼写相同的连续词符串,如 do, does, did, doing, done 是 5 个词型)从 25562 减为 15313。但我们仅在编制词频排列表时做了改动,原始的语料并没有减少和改正,以保持原貌。在使用词语检索器进行其他统计时,仍按原来 1207879 个词计算,望读者留意。

一般语料库的词频排列表都要提供一些重要参数如频数(frequency)和分布率(dispersion)。AHI(American Heritage Intermediate Corpus)还提供 U 值(一个词在 1000000 词中的理论频数)和标准频数指数(SFI)。我们采取了 AHI 的几个参数来整理我们的词频排列表。具体的公式和它的含义见词频排列表前的说明。

3.1.2 拼写失误表

拼写失误表,在光盘中编号为 III。我们在编制词频排列表时,为了了解学习者所使用的词汇量,把他们的拼写失误改正了。但不同类型学习者的拼写失误对教学很有参考意义,故我们把词频排列表中改正了的拼写失误形式单独列出一个拼写失误表。拼写失误共有 10540 词次、5810 词型。拼写失误表先列出正确的拼写形式,然后列出各类学习者的失误形式。我们可以看到有些常用词是学习者容易拼写错的,如 knowledge(22 种),society(21 种),important(13 种),government(13 种),opinion(12 种),beautiful(12 种),because(11 种),industry(11 种),people(11 种),等等。

3.1.3 词目表

词目表,在光盘中编号为 IV。词频排列表所排列的词型来自原始语料库,所以 take,

took,taken,taking 都作为词型而统计。我们需要把这些不同形式的词型归并而成为词目(lemmas),这就是词目归并(lemmatization),目的是了解学习者实际使用了多少词。

在编制词目表时,我们以 1998 年 Yasumasa Someya 所编制的 E-lemma 表为依据,编成专门软件。在 E-lemma 里,代词、副词并没有归并。词目表仍按词频排列表所设定的参数来统计,可参考词频排列表前的说明。

经过词目归并后,词型大概减少 1/3 强,见表 3.1。

表 3.1 词目归并前后的词型对比

学习者类型	词目归并前	词目归并后
ST2	5844	3981
ST3	5343	3578
ST4	5481	3891
ST5	8459	5726
ST6	9978	6781
整个语料库	15313	9861

* 参见第 3 页脚注 2

3.1.4 词频分布表

词频分布表(Word Frequency Distribution),在光盘中编号为 V。它和词频排列表所提供的数据是一致的,但是排列方式不一,主要是从排列的序号看词频的分布。在书中,我们提供了整个 CLEC 的词频分布总表,但在光盘里则增加了各类学习者的分表(编号 VIII～XII)。

3.1.5 词目分布表

CLEC 词目分布表,在光盘中编号为 VI。FLOB 词目分布表,在光盘中编号为 VII。编制这两个表的目的是为了了解词目归并后词频分布的变化情况。

3.1.6 语法标注频数表

CLEC 语法标注表,在光盘中编号为 XIII。根据 LOB 的 Tagset 进行词类的自动标注,标注后再进行归类统计。因为各类学习者的语料不完全一样,故表中既提供原始的语法标注频数,又提供经标准化处理后的频数,两者可以进行比较。标准化处理的方法见表 4.2 前的说明。

S. Johansson 和 K. Hofland(1989)曾按上述 Tagset 的 14 大类比较了 LOB 和 BROWN 的频数,现增加 CLEC 的频数,以作比较。表 3.2 显示,几个语料库的语法标注的比例比较一致,用得最多的是名词类和动词类,其词汇密度(名词、动词、形容词、副词、数词等实义词所占的比例)亦很一致,在 58%～60%之间。

如果我们把这 14 类语法标注作图,就可以看到 CLEC 的限定词和介词用得少些,而代词

又用得多些，如图3.1。这可能反映了CLEC的特点，学习者作文中有很多与个人和社会生活有关的题材，故使用了较多的代词。至于限定词(特别是the，a(n))和介词用得少些，则可能是中国学习者受汉语影响，掌握得不好，有意或无意地少用。

表3.2 CLEC，LOB，BROWN语法标注比较

	ST2	ST3	ST4	ST5	ST6	CLEC	LOB	BROWN
名词	49074.9	48709.2	44910.5	50201.8	51256.2	244152.6	254992	272984
动词	39010.2	41381.6	42229	37203.2	37876.6	197700.6	179975	185393
限定词	22730.3	21773.4	22051.8	22743.5	23537	112836	125018	123321
介词	18145.7	21264.8	19462.1	21477.3	22851.3	103201.2	123440	122613
形容词	14758.4	15614.6	13731.5	15509.5	17297.7	76911.7	73546	72034
代词	24272.1	19258.3	23039.4	19574.1	13908.7	100052.6	71498	66879
副词	11437.2	11985.7	12113.6	12140.9	11297.7	58975.1	56083	53283
连词	9025.4	10222.4	10481.6	10354.6	11158.7	51242.7	55516	60328
数词	3078.2	1087.4	1844.6	1897	1354.9	9262.1	19126	20853
不定式	2906.5	3739.6	4422.3	3764.8	4044.2	18877.4	15837	15030
Wh-词	2193.1	2233.3	2511.8	2265	2595.5	11798.7	15718	14921
Not	1825.3	1831.2	2385.9	1725.4	2105.8	9873.6	7465	6979
There	744.2	715.5	578.3	631	551.3	3220.3	2794	2280
感叹词	238.8	54.3	98.9	114.8	43.3	550.1	1109	629
总计	199440.3	199871.3	199861.3	199602.9	199878.9	998654.7	1002117	1017527
实义词	117358.9	118778.5	114829.2	116952.4	119083.1	587002.1	583722	604547
比例	0.588441	0.594275	0.574544	0.585925	0.595776	0.587793	0.582489	0.594134

* 参见第3页脚注2

图3.1 CLEC，LOB，BROWN语法标注比较