iWriteBaby中国学习者英语语料库的创建*

北京外国语大学 许家金

提要：我国大学生已广泛使用在线写作系统。这给英语写作带来深刻变革，也滋生出海量作文语料。本文介绍的800万词iWriteBaby中国学习者英语语料库采集自iWrite在线写作与评阅系统。该语料库是目前我国已公开的最大规模英语学习者语料库。其语料来源广泛、话题多样。本文还简介了该库的在线检索平台使用方法。

关键词：iWriteBaby中国学习者英语语料库、在线写作、语料库建设

1. 背景

从国际范围来看，我国学习者语料库研究起步较早。20世纪90年代初，北京语言学院（北京语言大学前身）便开始了外国留学生汉语中介语语料库的建设与研究（储诚志、陈小荷1993）。国际上的学习者语料库研究也大约在1993年左右逐步开展起来（Granger 2015：8）。在1996年前后桂诗春、杨惠中两位先行者开始了“中国学习者英语语料库”（Chinese Learners' English Corpus，CLEC）的研制。CLEC语料库于2003年出版发行。其后，各类英语学习者语料库如雨后春笋。其中代表性的成果包括：“中国学生英语口笔语语料库”（Spoken and Written English Corpus of Chinese Learners，SWECCL）、“中国大学生英汉汉英口笔译语料库”（Parallel Corpus of Chinese EFL Learners，PACCEL）、“大学英语学习者口语英语语料库”（College Learners' Spoken English Corpus，COLSEC）。详见徐秀玲、许家金（2017：63-64）的相关综述。

近年，我国大学生的英语写作方式发生了显著变化。在各大高校，学生们的日常写作，乃至测验和考试中的英语作文部分往往都是在线完成。由此产生的海量学生作文构成学习者英语语料库的宝贵素材。本文介绍的“iWriteBaby中国学习者英语语料库”（以下简称“iWriteBaby语料库”）便是在这一背景下诞生的。

* 本文系2017年度教育部人文社会科学重点研究基地重大项目“服务国家战略的外国语言与外语教育创新研究”子课题“大数据视野下的外语及外语学习研究”（编号：17JJD740003）阶段性成果。

2. iWriteBaby语料库建设概况

iWriteBaby语料库中的数据来自iWrite英语写作评阅引擎所产生的英语作文，我们将后者称为iWrite语料库，即iWrite总库，截至2019年4月其总规模达1.85亿词次。该语料库将不断扩容，动态增长。我们从iWrite总库中精选出800多万词，建成了iWriteBaby语料库，与学界共享。我们希望这些取之于学生的作文，最终可以用之于学生，对改进学生英语写作有所助益。

iWriteBaby语料库从2016年筹建，历经三年左右时间，于2019年3月23日在第四届全国高等学校外语教育改革与发展高端论坛期间正式发布。

iWriteBaby语料库起初设计规模为1,000万词次，意在将先前百万词级学习者语料库扩展至千万词级。在建设实践中我们发现，这1,000万词语料中仍然存在大量不合规文本。其中包含有与库中其他作文部分或完全雷同的作文，明显超出中国学习者英语水平的作文，含大量汉字的作文，随意敲入的任意字符串组成的文本，甚至全文以汉语拼音写成的作文，等等。如此删去约170多万词。

这次发布的iWriteBaby语料库为iWriteBaby 1.0版。其中包含学习者英语作文52,855篇，计8,299,066词次（单词定义为[a-zA-Z0-9-]+）。库中作文来自全国69所高校（其中重点大学与普通高校比例约为1: 10）。它们来自全国23个省市自治区，48个不同的城市。这些学生分布在154个不同的学科专业。入库的作文题目超过1,000个。

iWriteBaby语料库由北京外国语大学许家金总体设计，并完成相关的语料整理校对工作。语料库建设的全过程得到北京外研在线数字科技有限公司、汇智明德（北京）教育科技有限公司的资金和技术支持。语料库的整体设计得到梁茂成教授的指导。

3. iWriteBaby语料库在线检索平台

目前的单机版语料库软件已很难处理800万词规模的iWriteBaby语料库。因此，我们将该语料库部署在“语料云”在线平台（http://www.corpuscloud.cn）。该云平台可以实现WordSmith、AntConc、BFSU PowerConc等单机版语料库工具的相应功能，例如词表、索引分析、搭配等。语料云是在大数据时代BFSU PowerConc的网络实现（许家金、贾云龙 2013；许家金、吴良平 2014），强于分析大规模语料库数据。

在语料云平台注册账号后，可免费访问iWriteBaby语料库。登录后，用户需在页首导航栏找到“设置”，并在“显示设置”中勾选iWriteBaby 1.0 beta。这样便

可在首页“1. 语料库”栏中选择iWriteBaby 1.0 beta语料库进行检索分析。

3.1 词表功能

通过语料云的“工具”菜单找到“词表生成”，就可以创建iWriteBaby语料库的词频表。图1中显示的是iWriteBaby中最常用的词汇。在词表结果中显示的库容量为8,293,751词，与前文我们提供的总词数略有差别。这与该系统与我们的单词定义不同有关。若使用该云平台，则库容信息及其他相应频数都应统一以系统提供的数据为准。

语料云平台还允许我们生成2到5词的多词词表（或词块列表）。除词表外，该平台还可生成词性（串）列表等。

序号	检索项	总频次	子库数量	重要性	All in One
	库容量	8293751	1		8293751
	SUM	8285684	1		8285684
	COUNT	103650	1		103650
1	the	372381	1	372381.00	372381
2	to	278474	1	278474.00	278474
3	and	197588	1	197588.00	197588
4	of	176071	1	176071.00	176071

图1 词表功能

正如英语本族语者常用词一样，位列词表顶端的词汇基本都是功能词。iWriteBaby语料库词表也有十分相近的趋势，其高频词依次是the、to、and、of、is、a、in、we、it、can、I、you、more、that、are、for、people。其中people位列第17位，是排位最高的实义词。

3.2 检索功能

中国英语学习者为何使用people如此频繁？我们可以通过语料云的“检索”功能进一步加以了解。

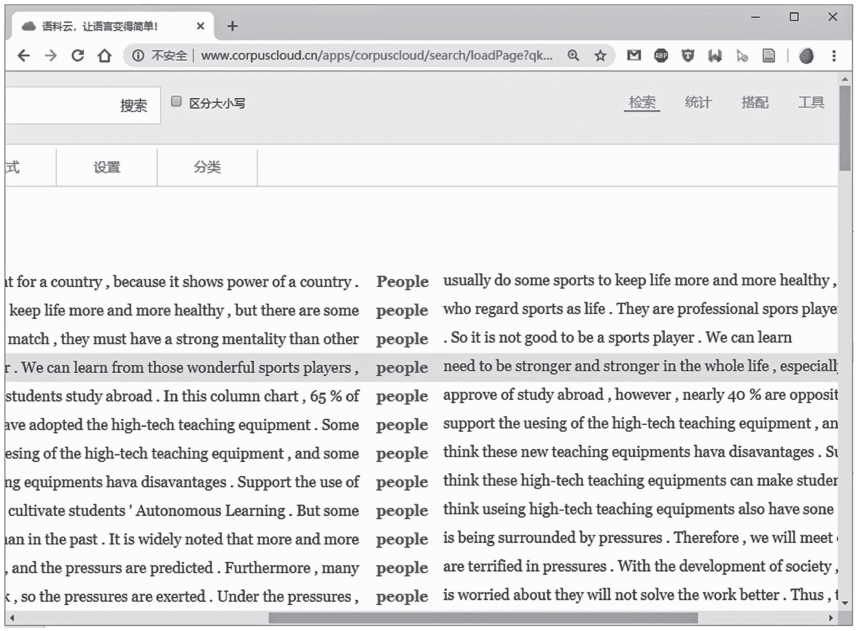
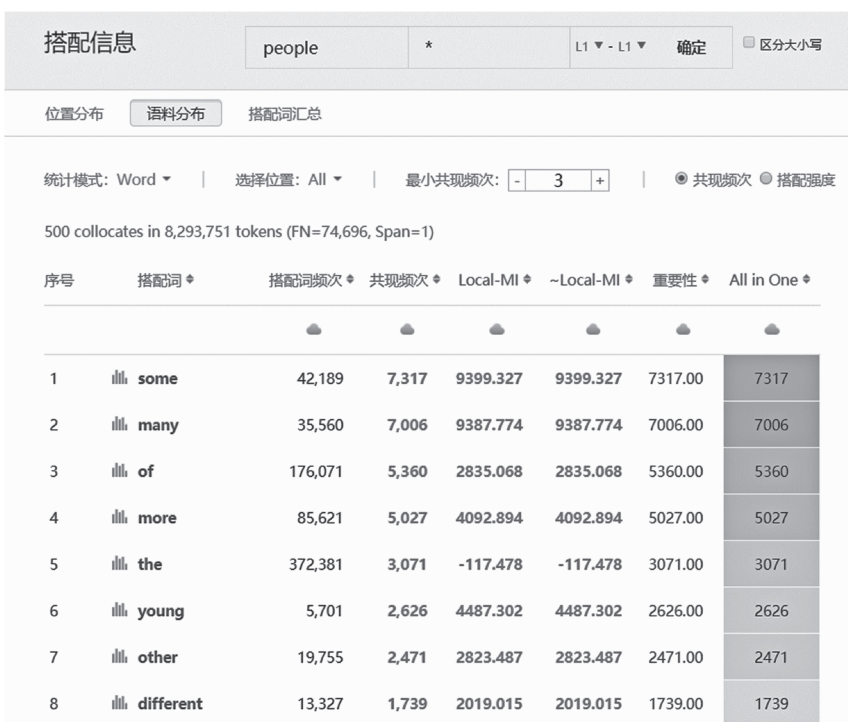


图2　检察功能

3.3　搭配功能

我们可通过搭配分析更多地了解people使用语境的典型搭配概率分布情况。



序号	搭配词	搭配词频次	共现频次	Local-MI	~Local-MI	重要性	All in One
1	some	42,189	7,317	9399.327	9399.327	7317.00	7317
2	many	35,560	7,006	9387.774	9387.774	7006.00	7006
3	of	176,071	5,360	2835.068	2835.068	5360.00	5360
4	more	85,621	5,027	4092.894	4092.894	5027.00	5027
5	the	372,381	3,071	-117.478	-117.478	3071.00	3071
6	young	5,701	2,626	4487.302	4487.302	2626.00	2626
7	other	19,755	2,471	2823.487	2823.487	2471.00	2471
8	different	13,327	1,739	2019.015	2019.015	1739.00	1739

图3　搭配功能

图3中的典型搭配词位于people的左边紧邻位置，构成people的限定语或修饰语。其完整形式为some people、many people、(a lot) of people、(more and) more people、other people、different people等。从上述典型搭配词及原文语境，我们认为people以及people构成的短语，在中国学习者英语中发挥着代词的作用。这也解释了为何people位列虚词聚集的词表顶端。People的这一中介语表现在汉英翻译中也存在（许家金 2016：18-19）。可以说，相当一部分的people用法是冗余的，而这些用法极有可能来自汉语母语的影响。例如，some people对应汉语的“有人”（相当于英语的someone、somebody）；other people对应汉语的“别人，其他人”（相当于英语的others、they）。

另外，iWriteBaby检索界面还提供子库分组功能。例如，用户可按大学类型（普通大学、重点大学）、性别（男生、女生）、任务类型（班级测试、课下写作、学校考试）分别检索分析结果，进而进行对比研究。

4. 结语

根据对iWriteBaby语料库的分析，我们针对中国英语学习者出现的典型英语表达错误，还创建了相关的教学案例库（http://ucreate.unipus.cn），用于改进英语写作。iWriteBaby语料库项目是在CLEC等开创性语料库项目基础上，意在将基于语料库的中介语研究向前推进一步。上述介绍的语料库建设工作只是我们的初期目标，后续还会利用iWrite写作平台尝试开发同题作文库、学习者作文追踪库等更多的教学研究资源。

参考文献

Granger, S. 2015. Contrastive interlanguage analysis: A reappraisal [J]. *International Journal of Learner Corpus Research* 1(1): 7-24.

储诚志、陈小荷，1993，建立“汉语中介语语料库系统”的基本设想 [J],《世界汉语教学》（3）：199-205。

许家金，2016，基于可比语料库的英语译文词义泛化研究 [J],《中国翻译》（2）：16-21。

许家金、贾云龙，2013，基于R-gram的语料库分析软件PowerConc的设计与开发 [J],《外语电化教学》（1）：57-62。

许家金、吴良平，2014，基于网络的第四代语料库分析工具CQPweb及应用实例 [J],《外语电化教学》（5）：10-15。

徐秀玲、许家金，2017，我国外语教学中的语料库应用40年 [J],《中国外语教育》（4）：62-68。

通讯地址： 100089 北京市北京外国语大学中国外语与教育研究中心

The construction of iWriteBaby Chinese learners' English corpus

XU Jiajin (105)

The widespread adoption of online English writing systems in Chinese universities has yielded enormous amount of English essays, in addition to its significant impact on English writing *per se*. The iWriteBaby Chinese learners' English corpus of eight million words was extracted from the iWrite English Writing Evaluation System. The corpus has been the largest publicly available Chinese learners' English corpus to date, which is characterized by its writer representativeness and topic diversity. The method of using the iWriteBaby online query system is also briefly demonstrated.