Advances in Social Science, Education and Humanities Research, volume 615
Proceedings of the 2021 4th International Conference on Humanities Education and Social Sciences (ICHESS 2021)

Study on Language Deep Learning by DDL under the Circumstances of A.I.
--A Comparative Analysis of Chinese and Korean English Learners' Corpus

HOU Jinrong[1,a]

[1]*School of Foreign Languages, Heze University, Heze, Shandong, China*
[a]*helenhouhou@126.com*

ABSTRACT
Artificial intelligence (AI) improves the progress of educational concepts, accelerates the educational reformation, promotes the development of educational technology, changes the teaching and learning methods. In the field of second language acquisition, data-driven learning contributes to the language deep learning because it focus on the process of learning and pays attention to learners' intuition and reflection. Based on Chinese and Korean learners' written corpus, this study applies the three-step procedure of "identify, classify and generalize" to analyze the similarities and differences among Chinese, Korean learners and native speakers on the use of high-frequency words, key words and collocation. It aims to draw attention to learners' cognitive process and improve learners' language normality.

***Keywords**: Artificial intelligence, data-driven learning, deep learning*

人工智能环境下数据驱动促进语言深度学习的方式研究
--基于中韩学习者语料库的对比分析

侯晋荣[1, a]

[1]*菏泽学院外国语学院，菏泽，山东，中国*
[a]*helenhouhou@126.com*

摘要
人工智能改进教育教学理念，加快教育改革的步伐，推动教育技术的创新，促进教学方式和学习方法的变革。在外语教学领域，数据驱动学习关注学习者的学习过程、语言意识的构建并促进学习的反思从而促进语言的深度学习。本研究基于中韩学习者英语作文语料库，通过识别、分类和概括三个步骤对比分析中韩英语学习者与英语母语者在高频词汇使用、关键词和词汇搭配方面的异同，促使学习者关注认知过程，提高语言使用规范。

***关键词：**人工智能，数据驱动，深度学习*

1. 前言

《中国教育现代化 2035》报告指出人工智能已成为教育领域的发展方向。人工智能的快速发展加快教育改革的步伐，推动教育技术的进步，改进教育教学理念，促进教学方式和学习方法的变革。在外语教学领域，人工智能可提供客观反馈，促进老师教和学生学的积极性。语言活动是一个复杂的认知过程，它推动了大脑的综合反应、协调和操作。在此活动中，信息通过三个方式传递：语言输入、语言处理和语言输出。语言处理是学习过程的核心部分，大脑中的信息加工驱动着语言的认知过程。基于人工智能的深度学习关注学习者的学习过程、语言意识的构建并促进学习的反思。

人工智能研究的是人与机器的互动。在语言学习领域，人工智能的发展促进语料库的建设和应用，使学者研究大量真实语料成为现实。语料库有关的研究分为两种：基于语料库的研究和语料库驱动的研究[1]。基于语料库的研究将语料库作为研究语言和语言学的方法，旨在验证、反驳或完善已经存在的理论或假设，分析语言的结构，观测语言的使用。语料库驱动的研究认为语料库本身应该是关于语言假设的唯一

来源，人们分析语料库呈现的数据得出新的观点或理论。

2. 数据驱动学习研究

20 世纪 60 年代起，语料库广泛应用于词典编撰、话语分析、翻译等领域。语料库可以将文本转化为计算机可读语言，同时处理大量信息并跟踪上下文，为语言研究开辟了道路[2]。Tim Johns 提出数据驱动学习（DDL）的概念，认为“每个学生都是福尔摩斯”，可直接从语言数据发现事实并进行学习[3]。1994 年，英国兰卡斯特大学举办第一届教育与语料库论坛，学者开始研究语料库技术在教学中的应用。Leech 认为数据驱动的学习方法应该是语料库在教育中应用的核心，因为语言学习势必要研究语言的使用,教师应引导学生发现语言事实，指导学习过程[4]。

语料库语言学的研究方法应用于二语学习课堂，将语料分析变成一个教学工具，以提高学习者对语言使用的认识和敏感性，同时增强语言学习策略[5]。Boulton and Cobb 对 64 名二语学习者进行实证研究，发现数据驱动学习适用于中等及中等以上水平的外语学习者的通用英语和学术英语的学习。语料库的索引呈现对于学习者的词汇和词汇语法的学习有很大帮助[6]。Lee 通过对 29 位二语学习者的观察发现数据驱动学习可促进中等熟练水平学习者深入理解所学知识[7]。

2018 年，中国正式实施《教育信息化 2.0 行动计划》，我国的外语教育进入人工智能时代。人工智能在外语教学中的应用主要表现在智能语音测评、自动批改、分析阅读、学情分析和教育机器人等领域。人工智能时代的外语学习从传统的行为主义转向认知主义和建构主义。外语学习不仅仅强调外部环境对学习者的促进作用，进行语言的反复操练和刺激，而是关注语言外部刺激和学习者内在习得过程相互作用的过程。学习过程中，应树立学生的主体地位，观察学生的学习过程，关注学生知识体系的建构，促进学生形成高阶思维进行深度学习。

不难看出，国内外学者对于人工智能时代的外语教学和深度学习极为关切，对学习方式、学习评价和学习效果等方面进行探索，但较少关注学习过程，即如何使用数据驱动的方式促进语言的深度学习。本研究基于中韩大学生英语作文语料库，分析中韩学生在写作中与英语母语者的差异，聚焦二语学习者的学习过程，分析人工智能和语料库信息技术对于外语学习者深度学习的促进作用。

3. 研究设计

3.1. 研究问题

目前大多数数据驱动学习研究都集中在母语语料库上。如何应用于课堂教学的学习者语料库研究相对较少。学习者通过对比分析学习者语料库和母语语料库可以使他们注意到英语学习过程中的典型错误，并发展识别母语和非母语之间差异的能力。通过观察学习者的典型错误，学生可以有意识重建自己的语言。Hong 提出数据驱动的外语学习主要有识别、分类和概括三个步骤[8]。识别指分析语料数据，找出外语学习者和母语使用者的差异。分类指对典型差异进行深入分析，通过上下文观察语言结构的搭配。概括指构建外语学习者的习得意识，改进语言的熟悉度。本研究尝试回答以下三个问题：

（1）中国英语学习者、韩国英语学习者与母语使用者的高频词汇使用有何异同？

（2）中国英语学习者、韩国英语学习者与母语使用者的关键词使用有何异同？

（3）中国英语学习者、韩国英语学习者与母语使用者的词串使用有何异同？

3.2. 研究语料

本研究构建学习者语料库，设定十亿词英国国家语料库为参照库。学习者语料库包括中国学习者和韩国学习者两个子库。中国学习者语料库库容 234 万词，韩国学习者语料库库容 110 万词。参照库英国国家语料库库容 15 亿词。

3.3. 研究工具

使用 Wordsmith 8.0 的 Wordlist，Keyness 和 Concordance 工具分别分析学习者语料库和母语语料库的高频词、词串和语境。Wordsmith 是一款多功能语料库检索软件，Wordlist 可按照频次或字母顺序得出词汇列表。Keyness 对比分析学习者语料库和参照库的词汇列表，得出高频和低频词汇，从而进行主题分析。研究者亦可设置词串长度，提取高频词串。Concordance 具备语境共现功能，呈现关键词和高频词串的上下文。

3.4. 研究步骤

研究前需收集中国学习者和韩国学习者英语作文语料并进行清洁，分类存放到学习者语料库文件夹并下载英语国家语料库。第一步，识别。使用 Wordsmith 的 Wordlist 制作中国学习者语料库、韩国学习者语料库和英国国家语料库的词表。第二步，分类。使用 Keyness 提取各个语料库的高频词汇和词串，并进行分类整理，观察外语学习者和母语使用者的词汇和词串使用差异。第三步，概括。使用 Concordance 进行语境共现，结合上下文考察中韩英语学习者和母语使用者的词串搭配，分析写作内容和写作特点，描述和归纳中韩二语学习者英语作文的词汇使用特点和写作风格，参照母语使用者提出改进意见和建议。

4. 结果与讨论

4.1. 宏观数据分析

标准化类/形符比（STTR）和句长最能体现文本特征，标准化类/形符比可体现不同长度的文本中词汇量的多少，而句长可以体现篇章的难易程度。英语国家语料库（BNC）标准化类/形符比最高（42.66%），说明英语母语使用者用词丰富。中国学习者语料库（CEFLC 34.6%）和韩国学习者语料库（KEFLC 33.8%）标准化类/形符比远远低于英语国家语料库，一方面是因为受作文主题的影响词汇局限于作文主题相关词汇，另一方面说明学习者在词汇使用有待丰富和提升。中韩学习者语料库类标准化类/形符比的差别体现出不同国家英语学习者的差异。中国学习者语料库标准化类/形符比稍高于韩国学习者语料库，说明中国外语教学中比较重视词汇教学。

表 1 学习者语料库与英语国家语料库宏观特征

语料	CEFLC	KEFLC	BNC
库容	234 万词	110 万词	15 亿词
形符	387216 词	185378 词	97860872 词
类符	9804 词	9328 词	512588 词
STTR	34.6%	33.8%	42.66%
平均句子长度	18	16	21

英语国家语料库平均句长 21 个单词，中国学习者语料库和韩国学习者语料库平均句长分别为 18 和 16，说明英语母语者书面语呈现出句子较长的总体特征，篇章相对难度较高。较之英语母语者，中国学习者使用句子稍短，篇章稍易，韩国学习者平均句子最短，篇章最易。

4.2. 高频词汇对比分析

大部分高频词汇是定冠词、介词、连词和情态动词等语法词汇。此外，出现较多的还有代词。研究发现排名前 10 的中国学习者语料库和韩国学习者语料库高频词汇完全一致，仅出现频次不同。这说明，不同国家的学习者呈现出二语习得和产出的较大共性。之后的高频词汇呈现出不同，说明不同国家的学习者之间亦存在差异。本研究重点分析不同国家的学习者之间英语使用的差异及与母语使用者的不同。

研究发现，学习者语料库和英国国家语料库出现频次排名前 4 位的词为 the, to, of, and。其中，学习者语料库 to 使用频次（中国 3.35%，韩国 3.29%）比英国国家语料库高（2.61%），and 使用频次（中国 2.32%，韩国 2.57%）比英国国家语料库低（2.64%）。学习者语料库出现频次高于英国国家语料库的还有 in, is, that, it, for。学习者语料库出现频次低于英国国家语料库的有 a, on, with。在人称代词方面，二语学习者和英语母语者均较多使用 it。不同的是，中国学习者使用 they 较多，韩国学习者使用 they 和 I 较多，而英语母语者使用 I 较多。这说明中国学习者倾向于对他者进行描述，而英语母语者和韩国学习者关注自我的观点和情感态度的表达。

4.3. 关键词对比分析

关键词是指文本中使用频率高，用以标明主题、风格和句法特征的词汇。关键词既可以是反映文本内容的实意词，也可以是传递细微信息的语法词。以英国国家语料库为参照库，分别提取中国学习者语料库和韩国学习者语料库中的关键词。P<0.001 说明提取的学习者语料库中的词汇与英国国家语料库有显著差异（表 2）。

研究发现，不同国家的学习者作文中使用的主题词汇各不相同。中国学习者作文中出现 credit（6.32%），smoking（3.35%），students（3.14%），cyber（2.32%），restaurants（2.29%），recycling（1.93%），abortion（1.82%），card（1.38%），banning（1.18%），cafes（0.97%）频次较高，说明写作的主体是学生，主题围绕信用卡的使用和禁止吸烟等社会关切的内容，地点涉及网吧和餐厅等公共场所。

表 2 中国学习者语料库和韩国学习者语料库关键词

CEFLC		KEFLC	
credit	6.32%	students	4.22%
smoking	3.35%	university	3.29%
students	3.14%	smoking	2.71%
cyber	2.32%	Korea	2.57%
restaurants	2.29%	people	2.12%
recycling	1.93%	communication	2.04%
abortion	1.82%	organs	1.89%
card	1.38%	smokers	1.21%
banning	1.18%	can	1.09%
cafes	0.97%	school	1.06%

韩国学习者使用 students（4.22%），university（3.29%），smoking（2.71%），Korea（2.57%），people（2.12%），communication（2.04%），organs（1.89%），smoker（1.21%），can（1.09%），school（1.06%）较多，说明与中国学习者一样，他们写作的主体是学生，但是由于作文任务的不同，主题词存在较大差异。韩国学习者写作主题围绕禁止吸烟、面对面交流、人体器官的再利用、教育等问题进行。

中国学习者和韩国学习者共同围绕的禁止吸烟作文中，中国学习者使用明确的指令 smoking banning，而韩国学习者使用 prevent，prohibit，stop 等搭配，但是频次不够突出。韩国学习者主要围绕吸烟导致的健康问题展开叙述，而中国学习者则态度明确，公众场合禁止吸烟。两国学习者在公共事务和管理管理方面的态度存在一定差异，中国学习者更注重社会的约定。

4.4. 词汇搭配和类连接对比分析

虽然语言具有任意性的特征，理论上讲人们可以

自由选择使用语言。但大部分我们接触的语言是规约的[9]。Sinclair 认为语言遵循成语规则（the idiom principle），母语使用者使用大量的约定俗成的短语，他们的语言自然流畅。数据驱动的方法为英语学习者提供机会接触真实语境，其最终目的是在教学中通过提取和学习真实语料中的词汇搭配达到类似母语的语言熟练度[10]。Erman 和 Warren 把母语者倾向于使用两个或者两个以上的单词组成的短语称之为语言预制（prefab），也有语言学家称之为语块、词簇或者词束。此类语言的机制是语言学习中的核心环节[11]。

词汇共现考察单词使用的语境。词汇的搭配观测节点词和其他实意词之间的关系，类连接分析节点词和其他语法词汇之间的关系。本研究通过对比分析学习者语料库和英语国家语料库的语境关键词的词汇搭配和类连接来呈现语言特征。使用 Wordsmith 的聚类分析分别提取中国学习者和韩国学习者使用频率最高的 3 词高频词串（表 3）。通过对比发现，in order to 和 according to the 是中韩两国学习者使用频率较高的词串。本文以 in order to 为例，分析中韩学习者和母语使用者词串使用和搭配的差异。

表 3 中国学习者语料库和韩国学习者语料库词串

CEFLC		KEFLC	
in order to	255	face to face	262
according to the	209	to face communication	188
according to a	142	we have to	102
to go to	126	face communication is	72
would like to	124	to make a	54
to solve the	113	we need to	53
to manage their	107	they want to	52
due to the	105	in order to	51
for them to	104	they have to	50
to repay the	97	according to the	49

英国国家语料库中 in order to 的高频使用短语为 in order to get, in order to be, in order to make，中国学习者语料库中出现较多的是 in order to solve（例 2）韩国学习者语料库中未发现与 in order to 高频搭配使用的动词。

Example

(1) Bradley flew to Hodges First Army headquarters in order to get the liberation started. (BNC)

(2) You can meet the minimum criteria in order to be considered by the Admissions Committee. (BNC)

(3) Therefore the government is proposing to legalize the soccer betting in order to solve the problem.(CEFLC)

(4) In order to prevent these related crimes and unethical issues, the sale of human organs shouldn't be legalized.(KEFLC)

In order to 的核心功能是表达动作的目的,英语国家语料库中，in order to 的典型用法为 in order to get/be+被动语态（例 1，例 2）。中韩学习者在此短语使用中只关注了语义而非语态。

5. 结论

基于语料库的数据驱动学习对师生均有益处。对老师而言，数据分析可观测英语母语使用者的典型用法，提高语言的敏感性，改进课程设计。学习者通过数据对比分析，可察觉语言使用中的错误用法并加以改正。

本研究通过讨论数据驱动的语料库在教学中的应用提升学习者的语言意识，改进语言学习的效果。数据驱动的学习使学习者通过识别、分类和概括观察并提升语言的熟练度。研究发现，中韩英语学习者与英语母语者在高频词汇使用方面存在较大差异， 中韩学习者使用 to, that, for, they, it 较多，使用 and, with, I, you 较少。数据驱动的学习促使学习者关注认知过程，提高自主学习能力。英语国家语料库中频繁使用的词串如 in order to+被动语态在中韩学习者语料库中均不多见，学习者可有意识增强典型词串的使用，使语言表达更加地道。

项目基金

山东省教育科学十三五规划年度课题《人工智能促进语言深度学习的方式研究》阶段性成果（课题编号：2020WBYB005）

REFERENCES

[1] McEnery, T., Hardie, A. (2012) Corpus Linguistics: Method, theory and practice. Cambridge University Press, Cambridge.

[2] Biber, D. (2020) Corpus Linguistics. Cambridge University Press, Cambridge.

[3] Johns, T. (1997) Contexts: The Background, Development, and Trialling of a Concordance-based CALL Program. In: Wichmann, A., Fligelstone, S., McEnery, T., Knowles, G. (Eds.) Routledege, London. pp. 100–115.

[4] Leech, G. (1997) Teaching and language corpora: a convergence. In: Wichmann, A., Fligelstone, S., McEnery, T., Knowles, G. (Eds.) Routledge, London. pp. 1–23.

[5] Pérez-Paredes, P. (2019) A systematic review of the uses and spread of corpora and data-driven learning in CALL research during 2011–2015. Computer Assisted Language Learning,43:20-31.

[6] Boulton, A., Cobb, T. (2017) Corpus use in language learning: A meta-analysis. Language Learning, 67:348–393.

[7] Lee, H., Warschauer, M., Lee, J. H. (2019) The effects of corpus use on second language vocabulary learning: A multilevel meta-analysis. Applied Linguistics, 40:721–753.

[8] Hong, Shin-Chul. (2011). A critical role for learners from the perspective of datadriven learning. English Language Teaching, 23:1-26.

[9] Siyanova-Chanturia, A., Martinez, R. (2015). The idiom principle revisited. Applied Linguistics, 36: 549-569.

[10] Sinclair, John (1991). Corpus, concordance and collocation. Oxford University Press, Oxford.

[11] Erman, B., Warren, B. (2000). The idiom principle and the open choice principle. Text, 20: 29-62..