iWriteBaby学习者语料库的创建及应用

许家金

2019.3.23

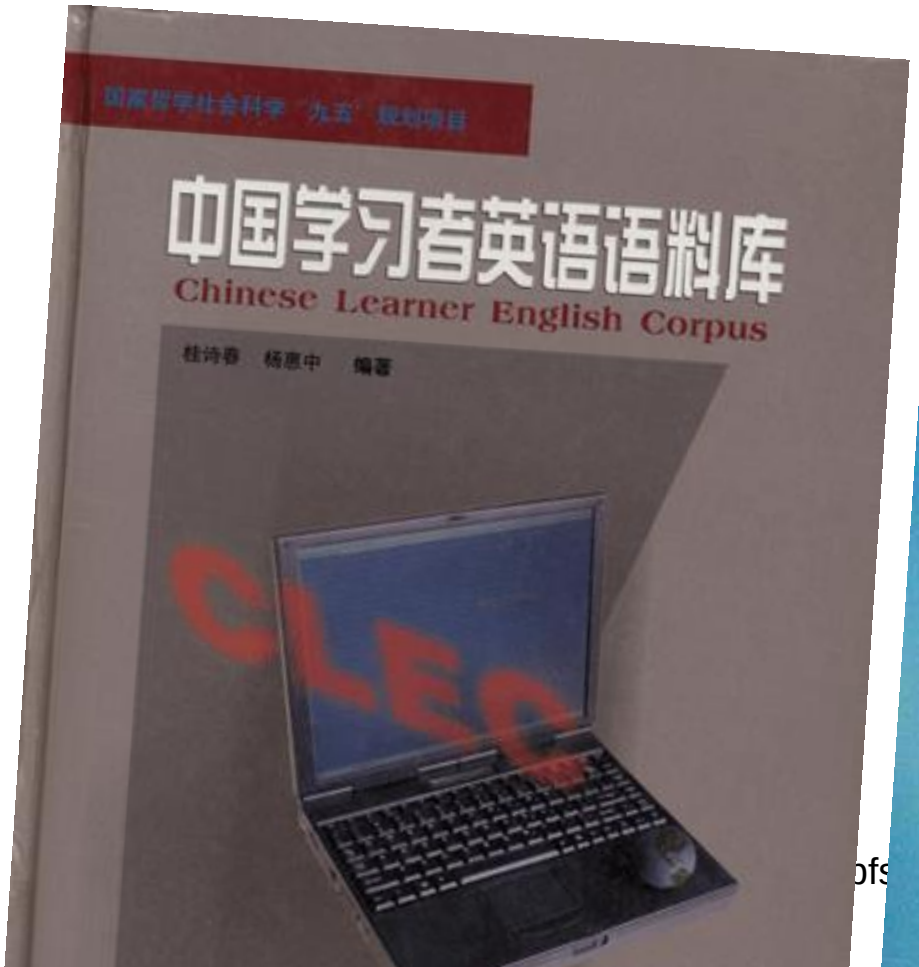
iWriteBaby学习者语料库

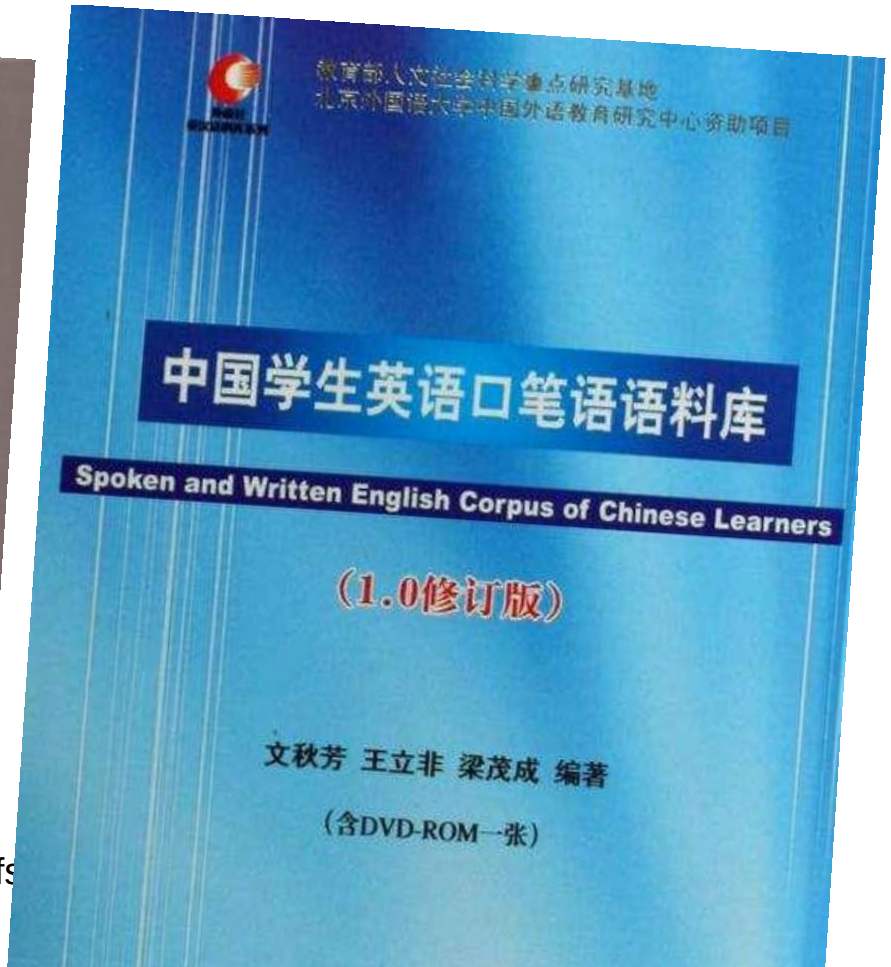
- iWrite Corpus（iWrite总库）上亿词次大数据语料库
- iWriteBaby，精选自iWrite总库的800万词次子库

corpus.bfsu.edu.cn

CLEC、SWECCL等

- 纸笔写作年代，前辈学者的开创性工作

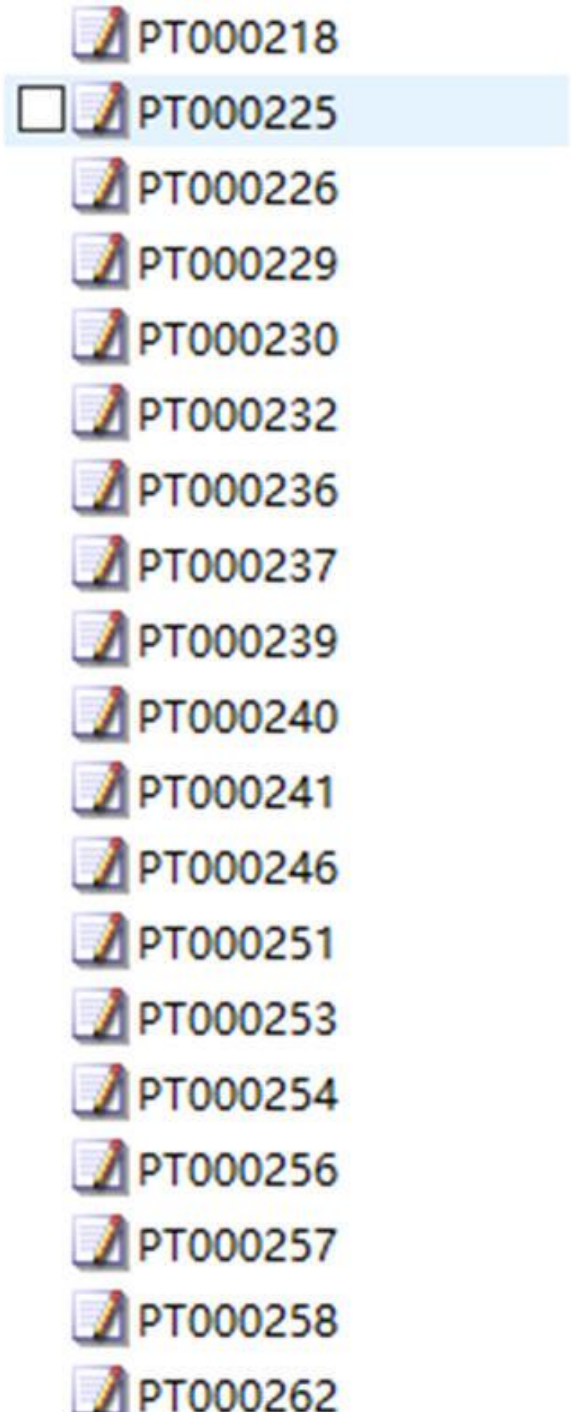




前辈学者

- 桂诗春、杨惠中
- 文秋芳、何安平、卫乃兴、李文中、濮建忠、杨达复、王立非、梁茂成

纸笔写作 → 在线写作


PT000218
PT000225
PT000226
PT000229
PT000230
PT000232
PT000236
PT000237
PT000239
PT000240
PT000241
PT000246
PT000251
PT000253
PT000254
PT000256
PT000257
PT000258
PT000262
PT000314
PT000316
PT000320
PT000322
PT000323
PT000325
PT000326
PT000327
PT000328
PT000331
PT000332
PT000333
PT000334
PT000336
PT000338
PT000340
PT000342
PT000343
PT000348
PT000400
PT000401
PT000402
PT000405
PT000408
PT000411
PT000412
PT000414
PT000416
PT000418
PT000420
PT000424
PT000426
PT000427
PT000429
PT000430
PT000432
PT000433
PT000434

BigCLEC、 BigSWECCL

- 大数据时代
- 向前辈致敬
- 站在前辈肩膀上，将学习者语料库、中介语研究往前推进一小步

iWrite语料库

- 1亿词次—∞
- 我们的一个小目标：iWriteBaby语料库
- 800万词次：目前我国已公开的最大规模学习者语料库
- 原计划1千万词次，经清理（雷同作文、不符合学生水平的作文约200万词）

iWriteBaby1.0beta概况

- 800万词次
- 5万多篇作文
- 69所高校（985/211大学与普通高校1:10）
- 23个省市自治区、48个不同的城市
- 154个学科专业
- 超过9000个不同的作文题

PT029363

- THE PLSEMCO NUIENCI KMC;LWMTjmmco mdpewmp okf3wf lfepkf m o, ermw fmewrmmv o, verer f, wer[kgv bmoem o, vre, mcoe l, ceorcpe, ceprmpem,, f, [pe, cercAkmcokw3 l, f-w, fpf,, c[ec,,, coemc, coemoe m, ore, fpero, eowfmomwf, ferov, o v, moervero m, vmvreo f, reofpoef, cow-ifwe vi f, omgmv mvcklwpvmierm cmemir mvroeigri moiefk coeme mfiwefmiowemk miewofj km qeoifj fjweij fmpwf m mf

全文无空格

- StressNowadays, thegofrommorningtillnight. Anditishardtoslowdown. Differentpeoplehavedifferentstress. FORtheworkingpeople, theymaybebusyinwork.forthestudents, theyofenbusyinstudy.Forourparents,theyoften earntheirliving.Inmyopinion,stressisgoodforus.

无奈、无助，还有那么点调皮

- wo zhen de shen me dou bu zhi dao. ying yu ken ding yao gua, yao shi bu gua qing jiao wo du shen. wo ying yu zhe me cha de re zen me hui xie zuo wen ne. Bu zhi dao xie shen me, wan quan bu zhi dao. ji kao zuo wen you ren kan me; yao shi you ren kan jiu gan ga le. zhen de hen gam ga hao ba. wo ye bu xiang zhe me xie de, ke shi wo shi zhen de bu hui a.

不同之处

- 比以往更真实：错得很离谱
- 在线写作学生们找到了真我
- 语言错误更真实，作文内容更走心

- Just as an old sayings goes live to old, learn to old.

以往语料库不会出现

- The biggest dream of my life is when I graduated, I can finding a job that have nice paid. So I can support my parents well. At the same time I hope that I can have a girlfriend. She needn't look nice, just I love she and she love me. As the time going, We will get married and we will have our children. I must study hard, because study can made me find a good job. When I find a good job, then I can find a girlfriend. Because when I have job, I can support her and our children.

iWriteBaby语料库的应用

- 目标：通过在线写作系统中产生的学生大数据语料，进一步分析汇总，发现典型问题，有的放矢，最终提升整体英语写作水平
- 库学同源、库研同步、库教同理
- 语料为本，取之学生，用之学生

大数据语料库在路上

- iWrite corpus, iWriteBaby 1.0, iWriteMini
- 专题库：错误库、同题库、追踪库、版本库
- 教案库

iWriteBaby研究初探

- 错误分布统计
- 中国学生作文的开篇模式
- Old/Chinese saying
- 句法复杂度研究

中国学生作文的开篇模式

- With
- As
- More
- There
- In

发布方式

- 语料云
  http://www.corpuscloud.cn
- BFSU CQPweb语料库平台
  http://111.200.194.212/cqp/
- 初期发布方式：在线检索
  （iWriteBaby1.0beta）
- 最终发布方式：全文提供

已知问题

- 作文话题仍不足
- 语料来源还应更广泛
- 单词黏连
- 漏网之鱼

致谢

- 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
- 梁茂成教授、徐一洁总经理、贾云龙先生
- 张晶、田夏春、牛琪雯、申晓蕾、邹露潇、任胜雷、徐维华等

欢迎指正