**DiSCUSS现代汉语平衡口语语料库**

（该说明文档由孙铭辰撰写，许家金审定）

DiSCUSS库引用方式：孙铭辰，2022，DiSCUSS现代汉语平衡口语语料库的创建，《语料库语言学》（2）。

DiSCUSS现代汉语平衡口语语料库（Diversified Spoken Chinese Uttered in Social Settings）是按照国际英语语料库ICE模式（Greenbaum & Nelson 1996）创建的百万词级平衡现代汉语口语语料库，由北京外国语大学中国外语与教育研究中心、人工智能与人类语言重点实验室主持建设，许家金教授统筹设计。

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 文本类别 | | | | 文本名称 | 文本数量 | 文本词数 |
| DiSCUSS  语料库 | 对话 | 私人 | 当面交谈 | S1A-001至SlA-090 | 90 | 296,694 |
| 电话交谈 | S1A-091至SlA-100 | 10 | 33,106 |
|  | | | | |
| 公开 | 课堂教学 | S1B-001至SlB-020 | 20 | 67,108 |
| 媒体讨论 | S1B-021至SlB-040 | 20 | 64,915 |
| 媒体采访 | S1B-041至SlB-050 | 10 | 33,320 |
| 赛场辩论 | S1B-051至SlB-060 | 10 | 31,944 |
| 法庭质证 | S1B-061至SlB-070 | 10 | 35,153 |
| 商业交易 | S1B-071至SlB-080 | 10 | 34,618 |
|  | | | | |
| 独白 | 脱稿 | 自发评论 | S2A-001至S2A-020 | 20 | 67,788 |
| 无稿演讲 | S2A-021至S2A-050 | 30 | 106,089 |
| 演示介绍 | S2A-051至S2A-060 | 10 | 33,703 |
| 法庭陈述 | S2A-061至S2A-070 | 10 | 35,178 |
|  | | | | |
| 念稿 | 媒体新闻 | S2B-001至S2B-020 | 20 | 66,634 |
| 媒体讲话 | S2B-021至S2B-040 | 20 | 62,887 |
| 有稿演讲 | S2B-041至S2B-050 | 10 | 33,401 |
|  |  |  |  |  | 300 | 1,002,538 |

**表1.** DiSCUSS现代汉语平衡口语语料库构成情况

DiSCUSS语料库按照ICE口语部分的建库模式创建，将ICE口语部分的60万词既定规模扩大到百万词级。由表3所示，DiSCUSS共包含300文本，1,002,538词（正则表达式：[\u4e00-\u9fa5a-zａ-ｚＡ-ＺA-Z0-9０-９\.%％]+），分为对话、独白两大类，私人、公开、脱稿、念稿四小类，并可细分为15个子类。各子类代表不同口语语境，与DiSCUSS语料库名称中的“Social Settings”相呼应，语料库详细情况见表1。DiSCUSS语料库的“对话—私人”部分（S1A001至SlA-100）、“对话—公开—商业交易”部分（S1B-071至 SlB-080）的110个文本由许家金、康卉、苏杭、李银美、朱周晔、钱一华、刘朝霞、孙铭辰等收集并转写，其余190个文本取自网络。DiSCUSS语料库建设过程中得到了董通、陈哲、刘芳芳、王波、王义娜、王彦，马博森、权立宏，陆军等老师、同学们的坚定支持，在大家共同努力之下，DiSCUSS语料库得以问世。

DiSCUSS语料库分为三个版本：纯文本版、分词版和分词标注版，以UTF-8编码的txt纯文本格式储存。文本基于正则表达式经过了系统清洗，并使用Jieba（https://github.com/fxsjy/jieba）组建进行了分词与标注处理。

口语语料库的建设一直以来都是困扰学界的难题。书面语语料库建设的优先级往往高于口语语料库，因为后者收集转写语音的难度更大、成本更高、速度更慢（Love et al. 2017）。Burnard（2002）更是认为建设100万词口语语料库的工作量是建设同样规模新闻报纸类语料库的10倍。相较于国外，国内现代汉语口语语料库的研制兴起于20世纪90年代前后，虽起步较晚，但奋起直追，且势头愈加迅猛，其中比较有代表性的平衡或专门口语语料库有：“北京口语调查”项目（宋孝才 1987）、CALLHOME语料库（Alexandra & Zipperlen 1996）、北京地区现场即席话语语料库（顾曰国 2002）、中国传媒大学有声媒体文本语料库（http://ling.cuc.edu.cn/RawPub/）、城市青少年汉语口语语料库（许家金 2009）、BCC语料库口语子库（荀恩东等 2016）、CCL语料库口语子库（詹卫东等 2019）等。

相较于以往汉语口语语料库的研制，DiSCUSS语料库具有三个显著特点。

1. 代表性与平衡性。DiSCUSS语料库的建设严格按照国际标准，采用ICE口语部分的采样框架。国际英语语料库（International Corpus of English，ICE）项目始于1988年，由伦敦大学学院的 Sidney Greenbaum（1988）发起，下辖近30支语料库建设队伍，旨在基于相同的ICE采样框架，建设代表世界各国家、地区英语变体的多个可比语料库。ICE项目的口语采样框架继承了伦敦大学学院“英语用法调查”项目（Quirk et al. 1972）的宝贵经验，并进一步改进创新，具有很高的效度、信度和认可度。采用该框架，能够使DiSCUSS语料库较好保证代表性与平衡性的同时，收集多种语境下的口语语料，充分体现DiSCUSS语料库名称中“Diversified”的特点。
2. 可比性。DiSCUSS语料库的精简版（56万词）将构成国际可比语料库汉语分库ICC-CN的口语部分。国际可比语料库（International Comparable Corpus，ICC, https://korpus.cz/icc）项目是采用相同ICE采样框架建设的多语种可比语料库，以同ICE项目的各英语变体、及ICC项目内各语言之间形成多语种对比研究。DiSCUSS语料库精简版，即ICC-CN口语子库，将在未来与ICE家族内近30种英语变体语料库、ICC家族内10余种多语种语料库构成可比语料库，从而开展对比研究。
3. 开源性。用户可全文免费下载DiSCUSS语料库的纯文本版、分词版和分词标注版三个版本。

**获取方式**

感兴趣的读者可访问“北外语料库语言学”网站（http://corpus.bfsu.edu.cn），在http://corpus.bfsu.edu.cn/info/1070/1335.htm页面全文下载DiSCUSS现代汉语平衡口语语料库，或登陆北外CQPweb平台（http://114.251.154.212/cqp/，账号：test；密码：test）在线检索使用。

**参考文献**

ALEXANDRA C., ZIPPERLEN, G. CALLHOME. Mandarin Chinese Speech LDC96S34 [DB/OL]. Philadelphia: Linguistic Data Consortium, University of Pennsylvania, 1996.

BURNARD L. Where did we go wrong? a retrospective look at the British National Corpus [C]//KETTEMANN B, MARKUS G. Teaching and learning by doing corpus analysis. Amsterdam: Rodopi, 2002:51-71.

GREENBAUM S. Proposal for an International Corpus of English [J]. World Englishes, 1988, 7:315.

GREENBAUM S, NELSON G. The international corpus of English (ICE) project [J]. World Englishes, 1996, 15:3-15.

LOVE R, DEMBRY C, HARDIE A, BREZINA V, MCENERY T. The Spoken BNC2014: designing and building a spoken corpus of everyday conversations [J]. International Journal of Corpus Linguistics, 2017(3):319-344.

QUIRK R, GREENBAUM S, LEECH G, SVARTVIK J. The Grammar of Contemporary English [M]. London: Longman, 1972.

顾曰国. 北京地区现场即席话语语料库的取样与代表性问题[C]//中国社会科学院，法国国家科研中心. 全球化与21世纪首届“中法学术论坛”论文集. 北京：社会科学文献出版社，2002：484-500.

宋孝才. 谈“北京口语调查”[J]. 世界汉语教学，1987(02)：25-29.

许家金. 青少年汉语口语中话语标记的话语功能研究[M]. 北京：外语教学与研究出版社，2009.

荀恩东，饶高琦，肖晓悦，臧娇娇. 大数据背景下BCC语料库的研制[J]. 语料库语言学，2016(01)：93-109.

詹卫东，郭锐，常宝宝，谌贻荣，陈龙. 北京大学CCL语料库的研制[J]. 语料库语言学，2019(01)：71-86.