多语种单语平衡语料库建设手册（1.0版）

许家金、周顾盈 编制（2021年5月5日）

该语料库共建项目由北京外国语大学中国外语与教育研究中心许家金发起，并负责总体建设方案的制订。

**建设目标**

* 2‒3年建成至少5个语种百万词级的单语平衡语料库。
* 语料库命名方式：XxBrown（Xx为语种缩率代码，Brown代表语料库所依据的取样方案）。目前已启动的多语种单语平衡库有ArBrown（阿拉伯语）、DeBrown（德语）、EsBrown（西班牙语）等。

**取样方案**

* 严格对照布朗语料库（Brown Corpus）的收集方案。有关布朗语料库的介绍可见

<http://www.helsinki.fi/varieng/CoRD/corpora/BROWN/basic.html>

* 布朗语料库详细文本来源见

<http://korpus.uib.no/icame/manuals/BROWN/BCMLOS.HTM>

* 目前北外语料库团队建成的2个100万词次的英语布朗语料库、2个100万词次的汉语布朗语料库和1个260万字词的英汉平行语料库，可从以下链接下载：

<http://corpus.bfsu.edu.cn/Crown_CLOB.zip>

* <http://corpus.bfsu.edu.cn/ToRCH2009.zip>
* <http://corpus.bfsu.edu.cn/ToRCH2014.zip>
* <http://corpus.bfsu.edu.cn/Yiyan_Corpus.zip>
* 语料类型及相应篇数

|  |  |  |  |
| --- | --- | --- | --- |
| **体裁**  **类型** | **子体裁代码** | **子体裁**  **类型说明** | **文本**  **数量（篇）** |
| **新闻** | A | 新闻报道 | 44 |
| B | 社论 | 27 |
| C | 报刊评论 | 17 |
| **通用** | D | 宗教 | 17 |
| E | 日常技艺及消遣爱好 | 36 |
| F | 通俗读物 | 48 |
| G | 传记、回忆录等 | 75 |
| H | 政府或机构公文及文宣 | 30 |
| **学术** | J | 学术 | 80 |
| **小说** | K | 一般小说 | 50 |
| L | 侦探小说 | 12 |
| M | 科幻小说 | 12 |
| N | 历险悬疑小说 | 13 |
| P | 言情小说 | 30 |
| R | 幽默 | 9 |
| **合计** |  |  | **500** |

* **总体原则**：在语料收集时，应注重保证文本来源的广泛多样（例如，多家报纸；或小说的前中后各取部分内容）。
* **发表时间**：优先收集新闻语料（近1-2年的新闻文本），之后收集通用/学术/小说三类文本（收最近10年内的文本）。
* **地区分布**：法语只收法国本土的法语文本，不收法国以外其他国家和地区的文本；俄语只收俄罗斯本土的俄语文本；西语目前只收西班牙出版的文本；阿拉伯语应综合考虑各国人口和影响力。其他语种的语料地域分布问题可再单独商讨。
* **文本来源：**各国新闻网站、期刊数据库、在线电子书（https://z-lib.org、http://gen.lib.rus.ec等网站）、国外出版社试读/预览文档、谷歌图书预览等。

**语料加工**

* **文本长度**：每篇文本2000词（+/-50单词），可用Word或EditPad Pro通过正则表达式统计词数。
* **文本格式**：存储为txt格式，UTF-8编码。
* **短文接长**：单篇文本若字数不够2000（如新闻类文本单篇字数常常不够2000），可用多篇同类型文本拼合，单独保存，在文件名后加上A/B/C…以示区分，如DE09A01A/ DE09A01B…同一本小说截取前中后三个部分拼接而成一个2000词的文本。
* **文字识别**：ABBYY FineReader可以把扫描图片转成文字。
* **文本命名**：语言（DE）+分类字母+第几篇+是否拼接（如是，用ABC表示），如DEA01A.txt。
* **出版信息**：收语料的同时整合语料元信息（metadata），填入Excel表格。可参考北外团队建成语料库的随附Excel表。

**补充说明**

1. 新闻类（收**近1-2年**新闻文本，建库之初优先收新闻体裁）：

注意：录入txt时从正文开始，不要前言或摘要；新闻类语料单篇文本字数往往不够2000，常常需要拼接。

* 1. Press reportage 新闻报道
* 主题覆盖面广：新闻报道还可以再根据主题细分，没有特别严格要求说一定要有哪些主题，覆盖面尽量广泛，可参考各国各新闻网站上的分类自行安排。
* 出处尽量来源广泛：全国性报纸、地区小报、新闻门户网站上的；周报、日报…最好都能涉及。
  1. Press editorials 社论
* 新闻社论，包括社评、记者评述、读者来信、专栏文章等，尽可能覆盖广泛。
  1. Press reviews 报刊评论
* 新闻评论，是指发表在报刊上的，针对某一个文学作品、电影、音乐等的评论，尽可能覆盖广泛。

除新闻外其他三个类别的都收**最近10年**的。

1. 通用类

注意：正文开始前可能还有其他文字内容，如献词、前言、作者解释、开场白等，只保留文学当中认为是书的一部分的绪论、开篇，但也不全收，大概按照2/10的比例收入语料。

1. Religious writing 宗教

* 跟宗教主题有关的内容，尽量杂一点，不要偏学术的（比如发在期刊上的专业论文），要那种闲暇时看着玩的。

1. Instructional writing 日常技艺及消遣爱好

* 经验、指导性的内容，传授非常具体的技能（比如怎么搭帐篷、如何做菜等）

1. Popular lore 通俗读物

* 通俗社会生活/热门知识/大众读物，和E的区别在于，F更加侧重日常生活、家长里短、娱乐八卦、心灵鸡汤等，看起来不费脑子的消遣读物。

1. Biographies and essays 传记、回忆录等

* 只要传记和回忆录。

1. Reports and official documents 政府或机构公文及文宣

* 报告和官方文件，比如布朗语料库里收的Government Documents/ Foundation Reports/ Industry Reports/ College Catalog/ Industry House organ。

1. 学术类
2. Learned（又称academic或scholarly）

* 目标语料是学术期刊里的论文或者学术专著。
* 覆盖面要尽可能广，比如在学科上应尽可能涵盖文理等多个学科（如，Natural sciences/ Medicine/ Mathematics/ Social and behavioral sciences/ Political science, law, education/ Humanities/ Technology and engineering）。
* 从标题到摘要到关键词都要，参考文献不要，英文摘要不要，只留目标语言。
* 处理插图和表格：删掉，并在删掉的地方留下<table>，<figure>等说明。
* 处理页眉/页脚/脚注/尾注：如果是pdf文章，页眉页脚不用（页眉：小标题；页脚：页码）。注释性的文字（脚注/尾注）可以保留，放到文章末尾比较合适，人工把脚注加工成尾注。
* 处理理工科论文里的复杂公式：删，留下<equation>做说明。
* 处理文科论文里的其他语言引用：引用古代的文字/翻译的文字，删，留下<quote>作说明，因为研究目标是当代语言，不能反映当代语言情况的都删。如果一篇论文中引用古代文字或外文的比重偏高，删，留下<quote>作说明；如果只有一两句引用的可以保留不删。找语料的时候尽量不要找带大量引文的，如果出现大量引用，需要补足2000词。

1. 小说类

注意：正文开始前可能还有其他文字内容，如献词、前言、作者解释、开场白等，只保留文学当中认为是书的一部分的绪言或开篇，但也不全收，大概按照2/10的比例收入语料。

1. General fiction一般小说

排除了明显标签的、除了下面的那几种分类外的、不是特别严肃的小说，就都可以算入这一分类。

1. Mystery and detective fiction 侦探小说
2. Science fiction 科幻小说
3. Adventure fiction历险悬疑小说

* 中文语料库在此分类下的语料包括进了“武侠小说”，与此类比，德国文学中骑着龙去冒险的玄幻冒险类也可以作为此分类下的目标语料。包含要素：个人英雄主义、善恶斗争、门派对抗。

1. Romantic fiction言情小说
2. Humor幽默小说，包括反讽/讽刺（有幽默效果的那种）文学, 喜剧……

\*小说的分类可以参照亚马逊图书或者自行找到各国针对图书的网站中的分类。

**🗐🗐🗐🗐🗐🗐🗐**

第一期“七库全书”多语种语料库项目参建人员名单

1. ArBrown：毛浚语（北京外国语大学阿拉伯学院）
2. DeBrown：周顾盈（北京外国语大学德语学院）
3. EsBrown：赵冲（北京外国语大学西葡语学院）
4. FrBrown：贾芳（天津外国语大学欧语学院法语系）
5. JpBrown：周彤（北京外国语大学日语学院）
6. KrBrown：徐俊（大连外国语大学韩国语学院）
7. RuBrown：陶源（东南大学外国语学院）