Multidimensional Analysis Tagger of Mandarin Chinese

The Multidimensional Analysis Tagger of Mandarin Chinese (MulDi Chinese) adapts Biber's (1988) analyses of English register variation to Mandarin Chinese. MulDi Chinese describe dimensions of register variation in Chinese, assessing the degree of orality, literacy, narration, explicit evaluation, abstractness, concreteness, classicality, and modernity. The programme tags 60 linguistic features based on ICTCLAS (H.-P. Zhang, Yu, Xiong, & Liu, 2003) and word lists in Chinese linguistics research. It generates scores along 4 dimensions of register variation. It will also plot the variation of the input text(s) against 15 registers in an upsampled Brown family (Francis & Kučera, 1964, 1971, 1979) ToRCH2014 corpus (J. Xu, Chen, Song, & Liu, 2017).

1 Referencing the Tagger

To cite the tagger, please use the following:

> Liu, N. 2019. Multidimensional Analysis Tagger of Mandarin Chinese. Available at: https://github.com/Nannan-Liu/Multidimensional-Analysis-Tagger-of-Mandarin-Chinese.

MulDi Chinese is based on the ICTCLAS, and it is advised to reference ICTCLAS when MulDi Chinese is used. Please refer to https://dl.acm.org/citation.cfm?id=1119280.

2 Requirements

MulDi Chinese requires Python to run (https://www.python.org/). The Python packages needed are:

- Python wrapper of ICTCLAS, i.e., PyNLPIR (https://pypi.org/project/PyNLPIR/)
- Pandas (https://pandas.pydata.org/)

- Categorised plaintext corpus reader from NLTK (Bird, Loper, & Klein, 2009) (`from nltk.corpus import CategorizedPlaintextCorpusReader`)
- Standard scaler from scikit learn (Pedregosa et al., 2011) (`from sklearn.preprocessing import StandardScaler`)
- NumPy (https://numpy.org/)

3 List of Variables

This section describes the linguistic features used in MulDi Chinese in alphabetic order of feature names. The abbreviations are consistent with those in the English tagger (Nini, 2018, pp. 17–31). Note that all occurrences are standardised by the length of input text.

3.1 Abstract nouns

MulDi Chinese counts occurrences of words in the following list (Fang, 2019): 社会, 问题, 生活, 经济, 关系, 作用, 中国, 现在, 情况, 时候, 人民, 活动, 方面, 科学, 条件, 思想, 过程, 影响, 方法, 要求, 技术, 事, 时间, 世界, 教育, 社会主义, 组织, 地方, 文化, 运动, 历史, 地区, 物质, 形式, 政治, 自然, 东西, 结构, 现象, 理论, 工业, 人类, 精神, 结果, 时期, 意义, 语言, 内容, 计划, 水平, 产品, 基础, 环境, 特点, 能力, 知识, 经验, 实际, 性质, 政府, 作品, 目的, 规律, 力量, 办法, 心理, 原则, 商品, 实践, 行为, 矛盾, 原因, 因素, 地位, 方向, 资本主义, 程度, 政策, 范围, 法律, 声音, 时代, 质量, 阶段, 方式, 人物, 速度, 自由, 价值, 困难, 中心, 事情, 事物, 对象, 现代, 事业, 利益, 材料, 内部, 音乐, 形象, 国际, 温度, 年代, 观点, 战争, 阶级, 希望, 家庭, 空气, 身体, 本身, 感情, 身上, 生命, 效果, 思维, 一部分, 意见, 标准, 无产阶级, 会议, 信息, 功能, 态度, 概念, 高度, 手段, 基础上, 理想, 说话, 化学, 措施, 目标, 帝国主义, 生物, 新闻, 行动, 民主, 资源, 物体, 资料, 意识, 观念, 道德, 实际上, 位置, 道路, 本质, 军事, 商业, 集体, 体系, 祖国, 机关, 意思, 机会, 习惯, 宗教, 领域, 机构, 国民经济, 形态, 哲学, 比例, 马克思主义, 类型, 成果, 脸上, 情绪, 能量, 成分, 健康, 成绩, 文艺, 空间, 品种, 主义, 主体, 规模, 形势, 方针, 意志, 责任, 队伍, 原理, 颜色, 项目, 委员会, 情感, 重点, 整体, 生产资料, 工程, 战略, 消息, 事件, 情形, 行政, 科技, 交通, 数学, 营养, 成本, 专业, 财政, 食物, 路线, 权力, 利润, 大部分, 元素.

3.2 Adverbial marker *di* 地

The tagger counts the occurrences of the part-of-speech (pos) tag ‘particle 地’ followed by standardisation.

3.3 Adverbs (RB)

MulDi Chinese counts occurrences of all words tagged as 'adverb'. Below is an example.

> 而 对于 在 流行性 传染病 蔓延 过程 中 受到 经济 损失 的 企业 和 个人 尚 [adverb] 无 类似 基金 的 设立 。(ToRCH2014_B27_SEG)

3.4 Adversative conjunctions

MulDi Chinese counts occurrences of words in the following list: 但, 但是, 可, 可是, 不过, 然而, 倒是, 然, 只是 (Jin & Jin, 2001)

3.5 Amplifiers (AMP)

MulDi Chinese counts occurrences of words in the list below followed by standardisation.

1. 非常, 十分, 真的, 特别, 很, 最, 肯定(Wei, 2019)
2. 挺, 顶, 极, 极为, 极其, 极度, 万分, 格外, 分外, 更, 更加, 更为, 尤其, 太, 过于, 老, 怪, 相当, 颇, 颇为, 有点儿, 有些, 最为, 越发, 越加, 愈加, 稍, 稍微, 稍稍, 略, 略略, 略微, 比较, 较, 暴, 超, 恶, 怒, 巨, 粉, 奇 (L. Wu, 2006)
3. 很大, 相当, 完全, 显著, 总是, 根本 (G. Wu & Pan, 2010)
4. 真, 真的, 一定

Note that amplifiers and emphatics were merged in this list.

3.6 Auxiliary adjectives

MulDi Chinese counts the occurrences of the tag 'auxiliary adjective' (Liu, Niu, & Liu, 2012) followed by standardisation.

> 突然 [auxiliary adjective] 有点 怅然 还 清晰 [auxiliary adjective] 记得 第一 次 见 您 是 什么 时候 (ToRCH2014_F01_SEG)

3.7 Auxiliary verbs

MulDi Chinese counts the occurrences of the tag 'auxiliary verb'.

3.8 Average clause length (ACL)

Average clause length is a salient predictor of register variation in Chinese (Hou, Huang, & Liu, 2017; Hou, Huang, Ahrens, & Lee, 2019). Clause end markers are defined as comma ， , colon ： , semicolon ； , and all sentence end markers (see Section 3.10; General Administration of Quality Supervision, Inspection and Quarantine and Administration [2011]). MulDi Chinese counts the number of words (词; note the difference with the character 字) within the boundary of two clause end markers and then divides it by the total number of clauses in the input text.

3.9 Average word length (AWL)

MulDi Chinese sums up the total number of characters in a text and divides it by the total count of words therein (Cf. M. Wang, 2017; Z.-S. Zhang, 2017).

3.10 Average sentence length (ASL)

Sentence ends are defined to include the period 。, question mark ？, ellipses ……, exclamation mark ！, and em dash ——(General Administration of Quality Supervision, Inspection and Quarantine & Administration, 2011). The tagger counts the number of words within the boundary of two sentence end markers and then divides it by the total number of sentences in the input text.

3.11 Chinese person names

The tagger counts the occurrences of the tags 'personal name', 'Chinese given name', and 'Chinese surname' followed by standardisation.

3.12 Classifiers

MulDi Chinese counts the occurrences of all kinds of 'classifier' tags followed by standardisation.

> 参与 侦破 了 三四百 起 [classifier] 命案 (ToRCH2014_L01_SEG)

3.13 Classical grammatical words

MulDi Chinese counts occurrences of 所, 将, 之, 于, and 以 (Feng, 2006; Z.-S. Zhang, 2017) followed by standardisation.

3.14 Classical syntax

MulDi Chinese counts occurrences of words in the following list replicated from Feng (2006) followed by standardisation: 备受, 言必称, 并存, 不得而, 抑且, 不特, 不外乎, 且, 不外乎, 不相, 中不乏, 不啻, 称之为, 称之, 充其量, 出于, 处于, 不次于, 从属于, 从中, 得自于, 得力于, 予以, 给予, 加以, 深具, 之能事, 发轫于, 凡此, 大抵, 凡, 所能及, 所可比, 非但, 庶可, 之故, 工于, 苟, 顾, 广为, 果, 核以, 何其, 或可, 跻身, 跻于, 不日即, 藉, 之大成, 再加, 略加, 详加, 以俱来, 见胜, 见长, 兼, 渐次, 化, 混同于, 归之于, 推广到, 名之为, 引为, 矣, 较, 借以, 尽其, 略陈己见, 而言, 而论, 决定于, 之先河, 苦不能, 莫不是, 乃, 泥于, 偏于, 颇有, 岂不, 岂可, 乎, 哉, 起源于, 何况, 切于, 取信于, 如, 则, 若, 岂, 舍, 甚于, 时年, 时值, 使之, 有别于, 倍加, 所在, 示人以, 随致, 之所以, 所以然, 无所, 有所, 皆指, 所引致, 罕为, 鲜为, 多为, 唯, 尚未, 无一不, 无不能, 无从, 可见, 毋宁, 无宁, 务, 系于, 仅限于, 方能, 需, 须, 许之为, 一改, 一变, 与否, 业已, 不以为然, 为能, 为多, 为最, 以期, 不宜, 宜于, 异于, 益见, 抑或, 故, 之便, 应推, 着手, 着眼, 可证, 可知, 可见, 而成, 有不, 有所, 有待于, 有赖于, 有助于, 有进于, 之分, 之别, 多有, 囿于, 与之, 同/共, 同为, 欲, 必, 喻之, 曰, 之际, 已然, 在于, 则, 者, 即是, 皆是, 云者, 者有之, 首属, 首推, 莫过于, 之, 之于, 置身于, 转而, 自, 自况, 自命, 自诩, 自认, 自居, 自许, 以降, 足以.

3.15 Concessive conjunctions (CONC)

MulDi Chinese counts occurrences of words in the following list.

1. 纵然, 即使, 虽然, 虽说, 虽, 固然, 尽管 (Ling, 2007)
2. 就是 (C. N. Li & Thompson, 1989, p. 637)

3.16 Complement marker *de* 得

The tagger counts the occurrences of the tag ‘particle 得’ followed by standardisation.

3.17 Conditional conjuncts (COND)

MulDi Chinese counts the occurrences of words in the following list followed by standardisation: 如果, 只有, 假如, 除非, 要是, 要不是, 只要, 倘若, 倘或, 设使, 设若, 如若, 若 (Yu, 2007), 的话, and 的时候 (C. N. Li & Thompson, 1989, p. 663).

3.18 Demonstrative pronoun (DEMP)

MulDi Chinese finds all kinds of tags containing 'demonstrative pronoun'.

> 其中 [demonstrative pronoun], 线性 谐振 子 作为 动力 系统 中 的 基础性 模型 , 不同 形式 的 激励 噪声 对 其 [demonstrative pronoun] 共振 行为 影响 显著。(ToRCH2014_J01_SEG)

3.19 Descriptive words

Descriptive words are named 'status word' by ICTCLAS. The tagger counts the occurrences of this tag followed by standardisation.

> 坐 在 桌前 的 女孩子 已经 可以 用 面色 惨白 [status word] 来 形容 了 (ToRCH2014_K01_SEG)

3.20 Discourse particles (DPAR)

MulDi Chinese counts occurrences of words in the following list in conjunction with the tags 'particle 的话' and 'particle 来讲/来说/而言/说来'.

1. 我跟你说, 你知道吗, 我告诉你, 我跟你讲, 你知道(Wei, 2019)
2. 不好意思, 就这样, 无所谓, 没问题, 不得了, 不用说, 不怎么, 不怎么样, 对了, 好了, 你看, 罢了, 话说回来, 不要说, 要说, 算了, 就是了, 不像话, 不要紧, 没事儿, 再说吧, 巴不得, 怪不得, 就得了, 得了, 你说呢, 说真的, 没劲, 没什么, 有的是, 怎么搞的, 话是这么说, 说不好, 说了算, 要我说, 一句话, 本来嘛, 别看, 够朋友, 说白了(Ji & Liu, 2015)
3. 总的来说, 总而言之, 只不过, 这样子, 想不到

3.21 Disyllabic negation

The tagger counts occurrences of 没有 (C. N. Li & Thompson, 1989, p. 415).

3.22 Disyllabic words

The tagger counts occurrences of words in the following list reproduced from Feng (2006): 安定, 安装, 办理, 保持, 保留, 保卫, 保障, 报道, 暴露, 爆发, 被迫, 必然, 必修, 必要, 避免, 编制, 变动, 变革, 辩论, 表达, 表示, 表演, 并肩, 补习, 不断, 不时, 不住, 布置, 采取, 采用, 参考, 测量, 测试, 测验, 颤动, 抄写, 陈列, 成立, 成为, 承担, 承认, 持枪, 充分, 充满, 充实, 仇恨, 出版, 处于, 处处, 传播, 传

达, 创立, 次要, 匆忙, 从容, 从事, 促进, 摧毁, 达成, 达到, 打扫, 大力, 大有, 担任, 导致, 到达, 等待, 等候, 奠定, 雕刻, 调查, 动员, 独自, 端正, 锻炼, 夺取, 发表, 发动, 发挥, 发射, 发生, 发行, 发扬, 发展, 反抗, 防守, 防御, 防止, 防治, 非法, 废除, 粉碎, 丰富, 封锁, 符合, 负担, 负责, 复述, 复习, 复印, 复杂, 复制, 富有, 改编, 改革, 改进, 改良, 改善, 改正, 干涉, 敢于, 高大, 高度, 高速, 格外, 给以, 更加, 公开, 公然, 巩固, 贡献, 共同, 构成, 购买, 观测, 观察, 观看, 贯彻, 灌溉, 光临, 规划, 合成, 合法, 宏伟, 缓和, 缓缓, 回答, 汇报, 混淆, 活跃, 获得, 基本, 集合, 集中, 极为, 即将, 计划, 记载, 继承, 加工, 加紧, 加速, 加以, 驾驶, 歼灭, 坚定, 减轻, 检验, 简直, 建立, 建造, 建筑, 交换, 交流, 结束, 竭力, 解决, 解释, 紧急, 紧密, 谨慎, 进军, 进攻, 进入, 进行, 尽力, 禁止, 精彩, 进过, 经历, 经受, 经营, 竞争, 竟然, 纠正, 举办, 举行, 具备, 具体, 具有, 开办, 开动, 开发, 开明, 开辟, 开枪, 开设, 开展, 抗议, 克服, 刻苦, 空前, 扩大, 来自, 滥用, 朗读, 力求, 力争, 连接, 列举, 流传, 垄断, 笼罩, 轮流, 掠夺, 满腔, 盲目, 猛烈, 猛然, 梦想, 勉强, 面临, 明明, 明确, 难以, 扭转, 拍摄, 排列, 攀登, 炮打, 赔偿, 评价, 评论, 赔偿, 评价, 评论, 破坏, 普遍, 普及, 起源, 签订, 强调, 抢夺, 切实, 侵略, 侵入, 轻易, 取得, 全部, 全面, 燃烧, 热爱, 忍受, 仍旧, 日益, 如同, 散布, 丧失, 设法, 设立, 实施, 实现, 实行, 实验, 适合, 试验, 收集, 收缩, 树立, 束缚, 思考, 思念, 思索, 丝毫, 四处, 饲养, 损害, 损坏, 损失, 缩短, 缩小, 贪图, 谈论, 探索, 逃避, 提倡, 提供, 提前, 体现, 调节, 调整, 停止, 统一, 突破, 推迟, 推动, 推进, 脱离, 歪曲, 完善, 万分, 万万, 危害, 违背, 违反, 维持, 维护, 围绕, 伟大, 位于, 污染, 无比, 无法, 无穷, 无限, 武装, 吸取, 袭击, 喜爱, 显示, 限制, 陷入, 相互, 详细, 响应, 享受, 象征, 消除, 消耗, 小心, 写作, 辛勤, 修改, 修正, 修筑, 选择, 严格, 严禁, 严厉, 严密, 严肃, 研制, 延长, 掩盖, 养成, 一经, 依法, 依旧, 依然, 抑制, 应用, 永远, 踊跃, 游览, 予以, 遇到, 预防, 预习, 阅读, 运用, 再三, 遭到, 遭受, 遭遇, 增加, 增进, 增强, 占领, 占有, 战胜, 掌握, 照例, 镇压, 征服, 征求, 争夺, 争论, 整顿, 证明, 直到, 执行, 制定, 制订, 制造, 治疗, 中断, 重大, 专心, 转入, 转移, 装备, 装饰, 追求, 自学, 综合, 总结, 阻止, 钻研, 遵守, 左右.

3.23 Disyllabic prepositions (BPIN)

The tagger counts the occurrences of the following words: 按照, 本着, 按着, 朝着, 趁着, 出于, 待到, 对于, 根据, 关于, 基于, 鉴于, 借着, 经过, 靠着, 冒着, 面对, 面临, 凭借, 顺着, 随着, 通过, 为了, 围绕, 向着, 沿着, 依据 tagged as 'preposition'. The list is reproduced from Fang (2018).

3.24 Disyllabic verbs

The tagger counts occurrences of words tagged as any types of verbs that have a length of two.

3.25 Downtoners (DWNT)

The tagger counts occurrences of words in the following list (X. Lu, 2004): 一点, 有点, 有点儿, 稍, 稍微, 有些.

3.26 Emotion words

The tagger counts occurrences of words in the following list reproduced from X. Xu and Tao (n.d.): 烦恼, 不幸, 痛苦, 苦, 快乐, 忍, 喜, 乐, 称心, 痛快, 得意, 欣慰, 高兴, 愉悦, 欣喜, 欢欣, 可意, 乐, 可心, 欢畅, 开心, 康乐, 欢快, 快慰, 欢, 舒畅, 快乐, 快活, 欢乐, 畅快, 舒心, 舒坦, 欢娱, 如意, 喜悦, 顺心, 欢悦, 舒服, 爽心, 晓畅, 松快, 幸福, 惊喜, 欢愉, 称意, 得志, 情愿, 愿意, 欢喜, 振奋, 乐意, 留神, 乐于, 爱, 关怀, 偏爱, 珍爱, 珍惜, 神往, 痴迷, 喜爱, 器重, 娇宠, 溺爱, 珍视, 喜欢, 动心, 挂牵, 赞赏, 爱好, 满意, 羡慕, 赏识, 热爱, 钟爱, 眷恋, 关注, 赞同, 喜欢, 想, 挂心, 挂念, 惦念, 挂虑, 怀念, 关切, 关心, 惦念, 牵挂, 怜悯, 同情, 吝惜, 可惜, 怜惜, 感谢, 感激, 在乎, 操心, 愁, 闷, 苦, 哀怨, 悲恸, 悲痛, 哀伤, 惨痛, 沉重, 感伤, 悲壮, 酸辛, 伤心, 辛酸, 悲哀, 哀痛, 沉痛, 痛心, 悲凉, 悲凄, 伤感, 悲切, 哀戚, 悲伤, 心酸, 悲怆, 无奈, 苍凉, 不好过, 抑郁, 慌, 吓人, 畏怯, 紧张, 惶恐, 慌张, 惊骇, 恐慌, 慌乱, 心虚, 惊慌, 惶惑, 惊惶, 惊惧, 惊恐, 恐惧, 心慌, 害怕, 怕, 畏惧, 发慌, 发憷, 敬, 推崇, 尊敬, 拥护, 倚重, 崇尚, 尊崇, 敬仰, 敬佩, 尊重, 敬慕, 佩服, 景仰, 敬重, 景慕, 崇敬, 瞧得起, 崇奉, 钦佩, 崇拜, 孝敬, 激动, 来劲, 炽烈, 炽热, 冲动, 狂热, 激昂, 激动, 高亢, 亢奋, 带劲, 高涨, 高昂, 投入, 兴奋, 疯狂, 狂乱, 感动, 羞, 疚, 羞涩, 羞怯, 羞惭, 负疚, 窘, 窘促, 不过意, 惭愧, 不好意思, 害羞, 害臊, 困窘, 抱歉, 抱愧, 对不起, 羞愧, 对不住, 烦, 烦躁, 烦燥, 烦, 熬心, 糟心, 烦乱, 烦心, 烦人, 烦恼, 烦杂, 腻烦, 厌倦, 厌烦, 讨厌, 头疼, 急, 浮躁, 焦虑, 焦渴, 焦急, 焦躁, 焦炙, 心浮, 心焦, 揪心, 心急, 心切, 着急, 不安, 傲, 自傲, 骄横, 骄慢, 骄矜, 骄傲, 自负, 自信, 自豪, 自满, 自大, 狂, 炫耀, 吃惊, 诧异, 吃惊, 惊疑, 愕然, 惊讶, 惊奇, 骇怪, 骇异, 惊诧, 惊愕, 震惊, 奇怪, 怒, 愤怒, 忿恨, 激愤, 生气, 愤懑, 愤慨, 忿怒, 悲愤, 窝火, 暴怒, 不平, 火, 失望, 失望, 绝望, 灰心, 丧气, 低落, 心寒, 沮丧, 消沉, 颓丧, 颓唐, 低沉, 不满, 安心, 安宁, 闲雅, 逍遥, 闲适, 怡和, 沉静, 放松, 安心, 宽心, 自在, 放心, 恨, 恶, 看不惯, 痛恨, 厌恶, 恼恨, 反对, 捣乱, 怨恨, 憎恶, 歧视, 敌视, 愤恨, 嫉, 妒嫉, 妒忌, 嫉妒, 嫉恨, 眼红, 忌恨, 忌妒, 蔑视, 蔑视, 瞧不起, 怠慢, 轻蔑, 鄙夷, 鄙薄, 鄙视, 悔, 背悔, 后悔, 懊恼, 懊悔, 悔恨, 懊丧, 委屈, 委屈, 冤, 冤枉, 无辜, 谅, 体谅, 理解, 了解, 体贴, 信任, 信赖, 相信, 信服, 疑, 过敏, 怀疑, 疑心, 疑惑, 其他, 缠绵, 自卑, 自爱, 反感, 感慨, 动摇, 消魂, 痒痒, 为难, 解恨, 迟疑, 多情, 充实, 寂寞, 遗憾, 神情, 慧黠, 狡黠, 安详, 仓皇, 阴冷, 阴沉, 犹豫, 好, 坏, 棒, 一般, 差, 得当, 标准.

3.27 Exclamations

The tagger counts the occurrences of the tag ‘exclamation mark’.

3.28 Existential *yǒu* 有 (EX)

The tagger counts occurrences of the tag ‘verb 有’.

3.29 First-person pronouns (FPP)

The tagger counts occurrences of 我, 我们 followed by standardisation.

3.30 Hedges (HDG)

The tagger counts occurrences of words in the following list (G. Wu & Pan, 2010): 可能, 可以, 也许, 较少, 一些, 多个, 多为, 基本, 主要, 类似, 不少.

3.31 Honorifics

The tagger counts occurrences of words in the following list (L. Wang, 2014): 千金, 相公, 姑姥爷, 伯伯, 伯父, 伯母, 大伯, 大哥, 大姐, 大妈, 大爷, 大嫂, 嫂夫人, 大婶儿, 大叔, 大姨, 哥, 姐, 大娘, 妈妈, 奶奶, 爷爷, 姨, 老伯, 老兄, 老爹, 老大爷, 老爷爷, 老太太, 老奶奶, 老大娘, 老板, 老公, 老婆婆, 老前辈, 老人家, 老师, 老师傅, 老寿星, 老太爷, 老翁, 老爷子, 老丈, 老总, 大驾, 夫人, 高徒, 高足, 官人, 贵客, 贵人, 嘉宾, 列位, 男士, 女士, 女主人, 前辈, 台驾, 太太, 先生, 贤契, 贤人, 贤士, 先哲, 小姐, 学长, 爷, 诸位, 足下, 师傅, 师母, 师娘, 人士, 长老, 禅师, 船老大, 大师, 大师傅, 大王, 恩师, 法师, 法王, 佛爷, 夫子, 父母官, 国父, 麾下, 教授, 武师, 千岁, 孺人, 圣母, 圣人, 师父, 王尊, 至尊, 座, 少奶奶, 少爷, 金枝玉叶, 工程师, 高级工程师, 经济师, 讲师, 教授, 副教授, 教师, 老师, 国家主席, 国家总理, 部长, 厅长, 市长, 局长, 科长, 校长, 烈士, 先烈, 先哲, 荣誉军人, 陛下, 殿下, 阁下, 阿公, 阿婆, 大人, 公, 公公, 娘子, 婆婆, 丈人, 师长, 义士, 勇士, 志士, 壮士, 学生, 兄弟, 小弟, 弟, 妹, 儿子, 女儿.

3.32 HSK Level I vocabulary

150 words, reproduced from Hanban (2012)

3.33 HSK Level III vocabulary

600 words (450 in operationalisation, Level I words and duplicates removed), reproduced from Hanban (2012)

3.34 Imperfect aspect markers

The tagger counts the occurrences of the words 着, 在, 正在, 起来 and 下去 (McEnery & Xiao, 2010, p. 12).

3.35 Indefinite pronouns (INPR)

The tagger counts occurrences of words in the following list: 任何, 谁, 大家, 某, 有人, 有个, 什么.

3.36 Intransitive verbs

The tagger counts the occurrences of the tag 'intransitive verb'.

3.37 Lexical density

The tagger counts occurrences of any open-class type of verbs (verb), nouns (noun), adjectives (adjective), and adverbs (adverb) (Jurafsky & Martin, 2019, pp. 144–145) followed by standardisation.

3.38 Modal particles

The tagger counts the occurrences of the tags 'modal particle' and 'interjection'.

3.39 Modals

The tagger counts the occurrences of the following modal verbs and auxiliaries: 乐意, 会, 允许, 可以, 可能, 喜欢, 将, 应当, 应该, 得, 必得, 必要, 必须, 情愿, 想, 想要, 愿意, 敢, 爱, 肯, 能, 能够, 要, 该, 需要 (Chappell & Peyraube, 2016).

3.40 Modifying adverbs

The tagger counts the occurrences of the following words tagged as 'adverb': 也, 都, 又, 才, 就, 就是, 倒是, 越来越, 一边, 再, 甚至, 却, 原本, 只, 毕竟, 仍然, 反正, 刚, 常常, 已经, 就要, and 连 tagged as 'particle 连', 等 tagged as 'particle 等/等等/云云'.

3.41 Monosyllabic negation

The tagger counts occurrences of 别, 不, and 没 (C. N. Li & Thompson, 1989, p. 415).

3.42 Monosyllabic verbs

The tagger counts occurrences of any types of verbs that have a length of one.

3.43 Nominalisation (NOMZ)

The tagger counts occurrences of tags ‘noun-adjective’, ‘noun-verb’ (Z.-S. Zhang, 2017, pp. 39–40), and any types of verbs followed by the ‘particle 的’ (C. N. Li & Thompson, 1989, pp. 575–576).

3.44 Onomatopoeia

The tagger counts the occurrences of the tag ‘onomatopoeia’.

3.45 Second-person pronouns (SPP)

The tagger counts the occurrences of the following words: 你, 你们, 您, 您们.

3.46 Third-person pronouns (TPP)

The tagger counts occurrences of words in the following list: 她, 他, 他们, 她们, 它, 它们.

3.47 Other personal pronouns

The tagger counts the occurrences of the tag ‘personal pronoun’, minus the results of FPPs (see 3.29), SPPs (see 3.45), and TPPs (see 3.46).

3.48 Perfect aspect markers (PEAS)

The tagger counts the occurrences of the tags ‘particle 了/喽’ and ‘particle 过’ (McEnery & Xiao, 2010, p. 11).

3.49 Phrasal coordination (PHC)

This tag was assigned for any ‘coordinating conjunction’ that is preceded and followed by the same tag and such a tag is one of the adverb, adjective, noun, or verb tags.

3.50 Prepositions (PIN)

The tagger counts occurrences of tags ‘preposition’, ‘preposition 把’ and ‘preposition 被’.

3.51 Private verbs (PRIV)

The tagger counts the occurrences of the following words: 三思, 三省, 主张, 了解, 亲信, 以为, 企图, 会意, 伤心, 估, 估摸, 估算, 估计, 估量, 低估, 体会, 体味, 信, 信任, 信赖, 修省, 假定, 假想, 允许, 关心, 关怀, 内省, 决定, 决心, 决意, 决断, 决计, 准备, 准许, 凝思, 凝想, 凭信, 分晓, 切记, 划算, 判断, 原谅, 参悟, 反对, 反思, 反省, 发现, 发觉, 吃准, 合计, 合谋, 同情, 同意, 否认, 听信, 听到, 听见, 哭, 喜欢, 喜爱, 回味, 回忆, 回念, 回想, 回溯, 回顾, 图谋, 图, 坚信, 多疑, 失望, 失身, 妄图, 妄断, 宠信, 害怕, 察觉, 寻思, 尊敬, 尊重, 小心, 希望, 平静, 幻想, 当做, 彻悟, 得知, 忆, 忖度, 忖量, 忘, 忘却, 忘怀, 忘掉, 忘记, 快乐, 念, 忽略, 忽视, 怀念, 怀想, 怀疑, 怕, 思忖, 思想, 思索, 思维, 思考, 思虑, 思量, 恨, 悟, 悬想, 情知, 惊恐, 想, 想像, 想来, 想见, 想象, 愉快, 意会, 意想, 意料, 意识, 感到, 感动, 感受, 感悟, 感想, 感激, 感觉, 感觉, 感谢, 愤怒, 愿意, 懂, 懂得, 打算, 承想, 承认, 担心, 拥护, 捉摸, 掂掇, 掂量, 掌握, 推度, 推想, 推敲, 推断, 推测, 推理, 推算, 推见, 措意, 揆度, 揣度, 揣想, 揣摩, 揣摸, 揣测, 支持, 放心, 料想, 料, 斟酌, 断定, 明了, 明察, 明晓, 明白, 明知, 明确, 晓得, 权衡, 梦想, 欢迎, 欣赏, 武断, 死记, 沉思, 注意, 洞察, 洞彻, 洞悉, 洞晓, 洞达, 测度, 浮想, 淡忘, 深信, 深思, 深省, 深醒, 清楚, 清楚, 满意, 满足, 激动, 热爱, 熟悉, 熟知, 熟虑, 爱, 爱好, 牢记, 犯疑, 狂想, 狐疑, 猛醒, 猜, 猜度, 猜忌, 猜想, 猜测, 猜疑, 玄想, 理会, 理解, 琢磨, 生气, 生疑, 畅想, 留心, 留神, 疏忽, 疑, 疑心, 疑猜, 疑虑, 疼, 盘算, 相信, 盼望, 省察, 省悟, 看, 看到, 看见, 看透, 着想, 知, 知悉, 知晓, 知道, 确信, 确定, 确认, 空想, 立意, 笃信, 笑, 答应, 策划, 筹划, 筹算, 筹谋, 算, 算计, 粗估, 约摸, 置疑, 考虑, 考量, 联想, 腹诽, 臆度, 臆想, 臆断, 臆测, 自信, 自省, 蒙, 蓄念, 蓄谋, 衡量, 裁度, 要求, 观察, 觉察, 觉得, 觉悟, 觉醒, 警惕, 警觉, 计划, 计算, 计较, 认为, 认可, 认同, 认定, 认得, 认知, 认识, 讨厌, 记, 记取, 记得, 记忆, 设想, 识, 试图, 试想, 详悉, 误会, 误解, 谋划, 谋算, 谋虑, 赞同, 赞成, 走神儿, 起疑, 轻信, 轻视, 迷信, 迷信, 追忆, 追怀, 追思, 追想, 通彻, 通晓, 通, 遐想, 遗忘, 遥想, 酌情, 酌量, 醒, 醒悟, 重视, 铭记, 阴谋, 顾全, 顾及, 预卜, 预想, 预感, 预料, 预期, 预测, 预知, 预见, 预计, 预谋, 领会, 领悟, 领略, 高估, 高兴, 默认 (A. Lu & Zhang, 2007; Chen, 2009; Q. Li, 2016).

3.52 Preposition 被 *bèi*

The tagger counts the occurrences of the tag 'preposition 被' followed by standardisation.

3.53 Public verbs (PUBV)

The tagger counts occurrences of the following words.

1. 表示, 称, 道, 说, 讲, 质疑, 认为, 坦言 (Xin, 2013)
2. 指出, 告诉, 呼吁, 解释 (G. Wu & Pan, 2010)
3. 问 and 建议

3.54 Questions

The tagger counts the occurrences of the tag 'question mark'.

3.55 seem/appear (SMP)

The tagger counts occurrences of words in the following list and followed by standardisation: 好像, 好象, 貌似, 似乎.

3.56 *shì* 是 (be)

The tagger counts the occurrences of the tag 'verb 是'.

> 还 清晰 [auxiliary adjective] 记得 第一 次 见 您 是 [verb 是] 什么 时候 (ToRCH2014_F01_SEG)

3.57 Simile

The tagger counts the occurrences of the tag 'particle 一样/一般/似的/般' and the words 仿佛, 宛若, 如 and 像.

3.58 Total other nouns excluding nominalisation (NN)

The tagger counts occurrences of the tags 'organization/group name', 'noun', 'noun morpheme', 'noun phrase', 'other proper noun', 'proper noun', and 'noun of locality'.

3.59 WH-words (WH)

The tagger counts occurrences of tags containing 'interrogative pronoun'.

3.60 Unique words ratio

Unique words are words that only appear once in a text.

References

Biber, D. (1988). *Variation across speech and writing*. Cambridge: Cambridge University Press.

Bird, S., Loper, E., & Klein, E. (2009). *Natural Language Processing with Python*. Newton, MA: O'Reilly Media Inc.

Chappell, H., & Peyraube, A. (2016). Modality and mood in Sinitic. In J. Nuyts & J. van der Auwera (Eds.), *The Oxford Handbook of Modality and Mood* (pp. 296–329). Oxford: Oxford University Press.

Chen, Z. (2009). "zhidao" and "mingbai" lei dongci yu yiwen xingshi [The Verbs"zhi dao" and "ming bai" and interrogatives]. *Hanyu Xuexi*(4), 27–37.

Fang, Q. (2018, August). Shumian yu "shuangyin jieci + NP chou" de guimo xingshi yu laiyuan ["Two-Syllable Preposition + Abstract Noun" Collocation and Its Written Style and Evolution]. *Hanyu xuexi*(4), 29–38.

Fang, Q. (2019, August). *Chouxiang Mingci Cibiao Jianban [An abridged list of abstract nouns].* Personal Correspondence.

Feng, S. (2006). *Hanyu Shumin Yongyu Huibian [Expressions of Written Chinese]*. Beijing: Beijing Yuyan Daxue Chubanshe.

Francis, W. N., & Kučera, H. (1964, 1971, 1979). *The Standard Corpus of Present-Day Edited American English (the Brown Corpus)* [University of Helsinki]. http://www.helsinki.fi/varieng/CoRD/corpora/BROWN/basic.html.

General Administration of Quality Supervision, Inspection and Quarantine, & Administration, S. (2011). *Zhonghua renmin gonghe guo guojia biaozhun GB 15834-2011 biaodian fuhao yongfa [General rules for punctuation].* PRC General Administration of Quality Supervision, Inspection and Quarantine and Standardization Administration of China.

Hanban. (2012). *Xin hanyu shuiping kaoshi cihui [New Hanyu Shuiping Kaoshi (HSK) vocabulary].* Hanban/Confucius Institute Headquarters.

Hou, R., Huang, C.-R., Ahrens, K., & Lee, Y.-M. S. (2019, February). Linguistic characteristics of Chinese register based on the Menzerath—Altmann law and text clustering. *Digital Scholarship in the Humanities*. doi: 10.1093/llc/fqz005

Hou, R., Huang, C.-R., & Liu, H. (2017, March). A study on Chinese register characteristics based on regression analysis and text clustering. *Corpus Linguistics and Linguistic Theory*, *15*(1), 1–37. doi: 10.1515/cllt-2016-0062

Ji, C., & Liu, F. (2015). Luxue sheng hanyu shumian yu zhong de kouyu hua

qingxiang yanjiu [Colloquial Trends in Chinese Compositions of Foreign Students]. *Yuyan jiaoxue yu yanjiu*(01).

Jin, Y., & Jin, C. (2001). Xiandai hanyu zhuanzhe lianci zu de tong yi yanjiu [A Comparative Study of the Synonym Group of Adversative Conjunctions in Morden Chinese]. *Hanyu xuexi*(2), 34–40.

Jurafsky, D., & Martin, J. H. (2019). *Speech and Language Processing An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition* (Third Edition draft ed.).

Li, C. N., & Thompson, S. A. (1989). *Mandarin Chinese: A Functional Reference Grammar*. Berkeley, CA: University of California Press.

Li, Q. (2016). *Xinli dongci dapei ji qi zai duiwai hanyu jiaoxue zhong de yingyong yanjiu [A Study on Psychological Verb Collocation and Application in Teaching Chinese as a Foreign Language]* (Unpublished doctoral dissertation). Hunan University, Changsha, China.

Ling, Y. (2007). *Rangbu lianci yanbian ji yufa gongneng yanjiu li shuo [A Case Study of the Evolution and Grammatical Function of Concessive Subordinators]* (Unpublished doctoral dissertation). Zhejiang University, Hangzhou, China.

Liu, B., Niu, Y., & Liu, H. (2012). Jiyu yicun jufa biaozhu shu ku de hanyu yuti chayi yanjiu [Word Class, Syntactic Function and Style: A Comparative Study Based on Annotated Corpora]. *Yuyan wenzi yingyong*(4), 132–142.

Lu, A., & Zhang, J. (2007). Hanyu xinli dongci de zuzhi he fenlei yanjiu [A Study of the Classification of Chinese Mental Verbs]. *Huanan shifan daxue xuebao (shehui kexue ban)*(1), 117-123+160.

Lu, X. (2004). *Xiandai hanyu "X dian", "X xie" yanjiu – Jianlun fuci "shaowei" yu "youdian" de qubie [A Study of "X dian" and "X xie" in Modern Chinese and Adverbs "shao wei" and "you dian"]* (MA Thesis). Shanghai Normal University, Shanghai, China.

McEnery, T., & Xiao, R. (2010). *Corpus-Based Contrastive Studies of English and Chinese*. London: Routledge.

Nini, A. (2018, October). *Multidimensional Analysis Tagger (v. 1.3.1) – Manual.*

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Wang, L. (2014). *Hanyu qiancheng zuncheng yanjiu [A Study of Honorific and Modest Titles in Chinese]* (MA Thesis). Xi'an International Studies University, Xi'an.

Wang, M. (2017). Hanyu guoji jiaoyu zhuanye taiguo lai hua liuxue sheng shuoshi

lunwen yuyan tezheng fenxi ji jiaoxue qishi [An Analysis of the Linguistic Features and the Teaching Implications of Master Dissertations of Thai Students Majored in Chinese International Education]. *Overseas Chinese Education*(10), 1384–1394. doi: 10.14095/j.cnki.oce.2017.10.009

Wei, Z. (2019). Hanyu benzu yu zhe he xuexi zhe hudong jiaoji shi de huayu canyu xingwei yanjiu [Study on Discourse Involvement Devices Used by Chinese Native Speakers and Nonnative Speakers]. *Overseas Chinese Education*, *1*, 95–102. doi: 10.14095/j.cnki.oce.2019.01.012

Wu, G., & Pan, C. (2010). Hanyu xueshu lunwen zhong zuozhe lichang biaoji yu yanjiu [Authorial Stance Markers in Chinese Research Articles]. *Yuyan jiaoxue yu yanjiu*(3), 91–96.

Wu, L. (2006). *Xiandai hanyu chengdu fuci zuhe yanjiu [A Study on the Combination of Degree Adverb in Mandarin Chinese]* (Unpublished master's thesis). Jinan University, Guangzhou, China.

Xin, B. (2013). Zhongwen baozhi xinwen biaoti zhong de zhuanshu yanyu xia [Indirect speech in Chinese newspaper titles II]. *Dangdai xiuci xue*(6), 20–25. doi: 10.16027/j.cnki.cn31-2043/h.2013.06.001

Xu, J., Chen, Z., Song, R., & Liu, Y. (2017, August). *ToRCH2014 Corpus Launch.* Foreign Language Education Research Informed by Corpora (FLERIC), Beijing Foreign Studies University.

Xu, X., & Tao, J. (n.d.). *Hanyu qinggan xitong zhong qinggan huafen de yanjiu [A Study into the Classification of Emotions in Chinese Emotion System].* Chinese Academy of Sciences Human Machine Speech Interaction Group.

Yu, N. (2007). Xiandai hanyu jiashe he tiaojian lianci yanjiu zongshu [A Literature Review of Hypothetical and Conditional Connectives in Modern Chinese]. *Xiandai yuwen*(7), 14–15.

Zhang, H.-P., Yu, H.-K., Xiong, D.-Y., & Liu, Q. (2003). HHMM-based Chinese Lexical Analyzer ICTCLAS. In *Proceedings of the Second SIGHAN Workshop on Chinese Language Processing - Volume 17* (pp. 184–187). Sapporo, Japan: Association for Computational Linguistics. doi: 10.3115/1119250.1119280

Zhang, Z.-S. (2017). *Dimensions of Variation in Written Chinese*. London: Routledge. doi: 10.4324/9781315673141