Multidimensional Analysis Tagger of Mandarin Chinese

The Multidimensional Analysis Tagger of Mandarin Chinese (MAT Chinese) is a program that extends Bibers functional analysis of English (1988). Its point of departure is to study register variation and communicative effect of texts. The program tags 54 linguistic features based on ICTCLAS (H.-P. Zhang, Yu, Xiong, & Liu, 2003) and word lists in Chinese linguistics research. It performs statistical analysis to indicate 5 dimensions of register variation. The program plots the variation of the input text or corpus as against 15 registers in an unsampled ToRCH2014 corpus. It also offers visualisation options using existing Python packages.

1 Referencing the Tagger

To reference the tagger, please use the following:

> Liu, N. 2019. Multidimensional Analysis Tagger of Mandarin Chinese. Available at: https://github.com/nnl93/Multidimensional-Analysis-Tagger-of-Mandarin-Chinese.

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

2 Requirements

This program requires Python to run (https://www.python.org/). The Python packages needed are NLTK (Bird, Loper, & Klein, 2009), Python wrapper of ICTCLAS – PyNLPIR (https://pypi.org/project/PyNLPIR/), and Factor Analyzer (https://factor-analyzer.readthedocs.io/en/latest/factor_analyzer.html).

3 List of Variables

This section describes the variables in alphabetic order. Next to the name of the variable is the acronym used in the tagger. The acronyms are consistent with those in MAT English (Nini, 2018, p. 17). An asterisk in the name indicates variables are sufficiently important to be included in final feature set.

3.1 Abstract nouns*

The tagger counts occurrences of words in the following list, then standardises the occurrences by the length of the input text.
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 (Fang, 07-Aug-2019)

3.2 Adjectives (JJ)

The tagger counts occurrences of tag ‘adjective’ minus those of tag ‘noun-adjective’ (a normalisation feature, see Section 3.62) and ‘auxiliary adjective’ (Section 3.7).

> 新 [adjective] 技术可以让许多开始变质的作品焕发出旧时光彩
> ToRCH2014_C01_SEG

3.3 Adverbs (RB)*

The tagger counts occurrences of all words tagged as 'adverb'.

> 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

The tagger counts occurrences of words in the following list, then standardises the occurrences by the length of the input text.
但, 但是, 可, 可是, 不过, 然而, 倒是, 然, 只是 (Jin & Jin, 2001)

3.5 Amplifiers (AMP)*

The tagger counts occurrences of words in the following list, then standardises the occurrences by the length of the input text.

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

N.B. Amplifiers and emphatics were merged in this list.

3.6 Attributive adjectives

Attributive adjectives are tagged as 'distinguishing word' by ICTCLAS. The tagger counts all occurrences of words tagged as 'distinguishing word' then standardises the occurrences by the length of the input text.

> 我要向我所有 [distinguishing word] 的读者表示谢意……见证了不计其数的非正常 [distinguishing word] 死亡 ToRCH2014_L01_SEG

3.7 Auxiliary adjectives*

The tagger counts all occurrences of words tagged as 'auxiliary adjective' (Liu, Niu, & Liu, 2012) then standardises the occurrences by the length of the input text.

> 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.

3.8 Average word length (AWL)*

The tagger sums up the total number of characters (字) and divides them by the total count of words (词) in the input text (Cf. M. Wang, 2017; Z.-S. Zhang, 2017).

3.9 Average sentence length (ASL)*

The training corpora consist of written registers, so the end of sentence is marked by punctuations, defined as Chinese period 。, question mark ？, ellipses ……, exclamation point ！, and em dash ——(Cf. Z.-S. Zhang, 2017, p. 55). The tagger counts the number of words (词) within the boundary of two sentence end markers, then divides it by the total number of sentences in the input text.

3.10 Average clause length (ASL)*

Following (Hou, Huang, & Liu, 2017; Hou, Huang, Ahrens, & Lee, 2019), average clause length is deemed a salient predictor of register variation in Chinese. Clause end markers are defined as comma ，, colon ：, semicolon ；, and all sentence end markers. The tagger counts the number of words (词) within the boundary of two clause end markers, then divides it by the total number of clauses in the input text.

3.11 Standard deviation of average clause length*

Furthering (Hou et al., 2017, 2019), the standard deviation of ACL is also regarded as a potential predictor. The standard deviation of ACL is obtained through Python built-in statistics package.

3.12 Be 是 (BE)*

The tagger counts all occurrences of 是 tagged as 'verb 是'.

> 还清晰 [auxiliary adjective] 记得第一次见您是 [verb 是] 什么时候 ToRCH2014_F01_SEG.

3.13 Book parentheses

The tagger counts all occurrences of the tuple (‘ ⟪’, ‘left parenthesis/bracket’), then standardises the occurrences by the length of the input text.

3.14 Causal connectives (CAUS)

The tagger counts occurrences of words in the following list, then standardises the occurrences by the length of the input text.

1. 因为, 固 (Xiao, He, & Yue, 2010)
2. 所以, 则, 从而, 故, 结果, 所以, 为此, 以至, 以至于, 因, 因此, 因而, 由于, 于是, 之所以, 致使 (Ni, 2008)

3.15 Classifiers*

The tagger counts all occurrences of words tagged as ‘classifier’ then standardises the occurrences by the length of the input text.

> 参与侦破三四百起 [classifier] 命案 ToRCH2014_L01_SEG

3.16 Classical function words*

The tagger counts occurrences of 所 tagged as ’particle 所’, 将 as ’adverb’, 将 as ’preposition’, 之 as ’particle 之’, 于 as ’preposition’ and 以 as ’preposition’ (Feng, 2006; Z.-S. Zhang, 2017), then standardises the occurrences by the length of the input text.

3.17 Classical syntax

The tagger counts occurrences of words in the following list replicated from (Feng, 2006), then standardises the occurrences by the length of the input text.
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

3.18 Concessive conjunctions (CONC)

The tagger counts occurrences of words in the following list, then standardises the occurrences by the length of the input text.

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

3.19 Conditional conjuncts* (COND)

The tagger counts occurrences of words in the following list, then standardises the occurrences by the length of the input text.
如果, 只有, 假如, 除非, 要是, 要不是, 只要, 假如, 倘若, 倘或, 设使, 设若, 如若, 若, and 的话 tagged as 'particle 的话', 的时候 tagged as ' 的', 'particle 的/底', ' 时候', 'noun'

3.20 Consecutive nouns*

The tagger counts all occurrences of two consecutive nouns, represented by regex <noun.*><noun.*>, then standardises the occurrences by the length of the input text. In the following example, four pairs of 'consecutive nouns' can be tagged.

> 各省 [noun] 、自治区 [noun]、直辖市 [noun] 及副省级城市 [noun]、
> 新疆生产 [noun-verb] 建设 [noun-verb] 兵团 [noun] 民 [noun morpheme]
> (宗)委(厅、局),各民族 [noun] 院校 [noun] ToRCH2014_H01_SEG

3.21 Consecutive verbs

The tagger counts all occurrences of two consecutive verbs, represented by regex <verb.*><verb.*>, then standardises the occurrences by the length of the input

text. In the following example, the underlined two words constitute 'consecutive verbs'.

> 抱 [verb] 紧了蜷 [verb] 起 [directional verb] 的双腿 ToRCH2014_M01_SEG

3.22 Dash

The tagger counts all occurrences of the tag 'dash', then standardises the occurrences by the length of the input text.

> 我写信给石头:[dash] 不管你有多大,走到哪……ToRCH2014_G01_SEG

3.23 Demonstrative pronoun* (DEMP)

The tagger finds words tagged as 'demonstrative pronoun', , then standardises the occurrences by the length of the input text.

> 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

3.24 Descriptive words*

Descriptive words are named 'status word' by ICTCLAS. The tagger counts all occurrences of words tagged as 'status word' then standardises the occurrences by the length of the input text.

> 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

3.25 Discourse particles (DPAR)

The tagger counts occurrences of words in the following list, then standardises the occurrences by the length of the input text.

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

3.26 Disposal marker *ba* 把

The tagger counts all occurrences of word '把' tagged as 'preposition 把' then standardises the occurrences by the length of the input text.

3.27 Disposal marker *jiang* 将

The tagger counts all occurrences of word '将' tagged as 'preposition' then standardises the occurrences by the length of the input text.

3.28 Disyllabic negation*

The tagger counts occurrences of 没有 tagged as 'adverb' and as 'verb' (C. N. Li & Thompson, 1989, p. 415).

3.29 Disyllabic words*

The tagger counts occurrences of words in the following list replicated from (Feng, 2006), then standardises the occurrences by the length of the input text. 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

3.30 Disyllabic prepositions (BPIN)*

The tagger counts occurrences of these 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 taken from (Fang, 2018).

3.31 Disyllabic verbs

The tagger counts occurrences of words tagged as any type of verb, represented by regex <*verb>, that have a length of two.

3.32 Downtoners (DWT)*

The tagger counts occurrences of words in the following list (Lu, 2004), then standardises the occurrences by the length of the input text.
一点, 有点, 有点儿, 稍, 稍微, 有些

3.33 Emotion words*

The tagger counts occurrences of words in the following list, then standardises the occurrences by the length of the input text. The list is replicated from (Xu & 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

3.34 English words*

The tagger counts all occurrences of English words, then standardises the occurrences by the length of the input text. This is done by removing all Chinese words and punctuations represented by regex [□-□□-龟一-龠々〇丨-夕十-卅□丄-鼇一-□豈-鶴侮-頻並-龗，。；：“”【】、！（）()——《》=？"].

3.35 Exclamations*

The tagger counts all occurrences of the tag ‘exclamation mark’, then standardises the occurrences by the length of the input text.

3.36 Existential 有 * (EX)

The tagger counts occurrences of 有 tagged as 'verb 有'.

3.37 Experiential 过

The tagger counts all occurrences of word '过' tagged as 'particle 过' then standardises the occurrences by the length of the input text.

3.38 First-person pronouns* (FPP1)

我, 我们

3.39 Hedges* (HDG)

The tagger counts occurrences of words in the following list (?, ?), then standardises the occurrences by the length of the input text.
可能, 可以, 也许, 较少, 一些, 多个, 多为, 基本, 主要, 类似, 不少

3.40 Honorific titles*

The tagger counts occurrences of words in the following list, then standardises the occurrences by the length of the input text.
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 (L. Wang, 2014)

3.41 HSK I vocabulary*

150 words, replicated from (Hanban, 2012)

3.42 HSK III vocabulary*

600 words (450 in operationalization, level I words and duplicates removed), replicated from (Hanban, 2012)

3.43 HSK IV vocabulary*

5000 words (4400 in final list, level I, level III and duplicate words removed), replicated from (Hanban, 2012)

3.44 Imperfect aspect markers*

The tagger counts all occurrences of the word '着' tagged as 'particle 着', the word '在' tagged as 'preposition', '正在' tagged as 'adverb', '起来' tagged as 'directional verb' and '下去' as 'directional verb', then standardises the occurrences by the length of the input text.

3.45 Indefinite pronouns (INPR)*

The tagger counts occurrences of words in the following list, then standardises the occurrences by the length of the input text.
任何, 谁, 大家, 某, 有人, 有个, 什么

3.46 Interrogative pronouns*

The tagger counts words tagged as 'interrogative pronoun' minus those tagged as 'predicate interrogative pronoun'.

3.47 Intransitive verbs

The tagger counts all occurrences of words tagged as 'intransitive verb' then standardises the occurrences by the length of the input text.

3.48 Knowledge modals

可以, 会, 可能 (Peng & Zhu, 2017)

3.49 Lexical density*

The tagger counts occurrences of any type of verbs (*verb), nouns (*noun), adjectives (*adjective), numerals (*numeral), adverbs (*adverb) and pronouns (*pro-

noun), and divides the occurrences by the length of the input text, then multiplies the result by 1000.

3.50 *men* 们 suffix

The tagger counts occurrences of 们 tagged as 'suffix'.

3.51 Modal particles*

The tagger counts all occurrences of words tagged as 'modal particle' and 'interjection' then standardises the occurrences by the length of the input text.

3.52 Modest titles

The tagger counts occurrences of words in the following list, then standardises the occurrences by the length of the input text.
学生, 兄弟, 小弟, 弟, 妹, 儿子, 女儿 (L. Wang, 2014)

3.53 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'.

3.54 Modifying marker *di* 地 *

The tagger counts all occurrences of the word '地' tagged as 'particle 的/底' then standardises the occurrences by the length of the input text.

3.55 Modifying marker *dedi* 的

The tagger counts all occurrences of the word '的' tagged as 'particle 的/底' then standardises the occurrences by the length of the input text.

3.56 Modifying marker *de* 得

* The tagger counts all occurrences of word '得' tagged as 'particle 得' then standardises the occurrences by the length of the input text.

3.57 Monosyllabic negation*

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

3.58 Monosyllabic verbs*

The tagger counts occurrences of words tagged as any type of verb '*verb' that have a length of one.

3.59 Motivation modals

The tagger counts occurrences of words in the following list, then standardises the occurrences by the length of the input text.

1. 能, 敢, 肯, 要 (Peng & Zhu, 2017)
2. 愿意, 能够 (C. N. Li & Thompson, 1989, pp. 182-183)

3.60 Necessity modals

The tagger counts occurrences of words in the following list, then standardises the occurrences by the length of the input text.

1. 能, 应该, 要, 必须 (?, ?)
2. 必得, 应当, 该 (C. N. Li & Thompson, 1989, pp. 182-183)

3.61 Nouns* (NN)

The tagger counts occurrences of tag 'noun', 'noun morpheme' and 'proper noun', minus those of tag 'noun-adjective' (nominalisation), 'noun-verb' (nominalisation), 'pronoun', 'noun of locality'.

3.62 Nominalisation (NOMZ)*

The tagger counts occurrences of tags 'noun-adjective', 'noun-verb', and any 'verb' followed by 的 ('的', 'particle'), then standardises the occurrences by the length of the input text.

3.63 Onomatopoeia

The tagger counts all occurrences of words tagged as 'onomatopoeia' then standardises the occurrences by the length of the input text.

3.64 Or (OR)

或, 或者, 或是, 还是

3.65 Other conjunctions

The tagger finds all conjunctions except for those tagged as 'coordinating conjunction', and 因为, 固, 所以, 则, 从而, 故, 结果, 所以, 为此, 以至, 以至于, 因, 因此, 因而, 由于, 于是, 之所以, 致使, 纵然, 即使, 虽然, 虽说, 固然, 尽管, 就是, 纵然, 即使, 虽然, 虽说, 尽管, 就是, 如果, 只有, 假如, 除非, 要是, 要不是, 只要, 假如, 倘若, 倘或, 设使, 设若, 如若, 若 tagged as 'conjunctions'.

3.66 Other personal pronouns*

The tagger counts all occurrences of words tagged as 'personal pronoun', minus counts of 我, 你, 她, 他, 它 (plurals are automatically included).

3.67 Passives (BYPA)

被 tagged as 'preposition 被', 受到 as 'preposition 受到', 遭受 as 'verb', 受 as 'verb', 遭 as 'verb', 挨 as 'verb', 加以 as 'performative verb' (Xiao, McEnery, & Qian, 2006; Yang & Cheng, 2016)

3.68 Perfect aspect markers (PEAS)*

The tagger counts all occurrences of the word '了' tagged as 'particle 了/喽', the word '过' tagged as 'particle 过', then standardises the occurrences by the length of the input text.

3.69 Performative verbs

The tagger counts all occurrences of words tagged as 'performative verb' then standardises the occurrences by the length of the input text.

3.70 Person names

The tagger counts all occurrences of the tags 'personal name' plus 'Chinese', minus by those of the tag 'transcribed personal name', then standardises the occurrences by the length of the input text.

3.71 Place (PLACE)

The tagger counts occurrences of words tagged as 'noun of locality', 'locative word', and 'toponym'.

3.72 Prepositions (PIN)

The tagger counts occurrences of tag 'preposition' minus those of disyllabic prepositions.

3.73 Private verbs (PRIV)*

The tagger counts the occurrences of the following words tagged as 'verb': 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 (Chen, 2009; Q. Li, 2016; ?, ?)

3.74 Phrasal coordination (PHC)*

和, 以及, 而, 与, 并, 以至, 及, 并且, 而且, 不但, 而且

3.75 Public verbs (PUBV)*

The tagger counts occurrences of the following words tagged as 'verb'.

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

3.76 Purpose connectives

为了, 起见, 为的, 为是, 为着, 以便, 借以, 以求, 以示, 以表, 以便, 以求, 省得, 免得, 以免, 以防 (?, ?)

3.77 Questions*

The tagger counts all occurrences of the tuple ('？', 'question mark'), then standardises the occurrences by the length of the input text.

3.78 Quotes

The tagger counts all occurrences of the tag 'quotation mark' divided by 2, then standardises the occurrences by the length of the input text.

3.79 Second-person honorific pronouns*

您, 您们

3.80 Second-person pronouns* (SPP2)

你, 你们

3.81 seem/appear (SMP)*

The tagger counts occurrences of words in the following list, then standardises the occurrences by the length of the input text.
好像, 好象, 貌似, 似乎

3.82 Simile*

The tagger counts all occurrences of the word ‘仿佛’ tagged as ‘adverb’, ‘宛若’ tagged as ‘verb’, ‘如’ tagged as ‘verb’, all words tagged as ‘particle 一样/一般/似的/般’, word ‘像’ tagged as ‘verb’ and ‘preposition’, then standardises the occurrences by the length of the input text.

3.83 Standardized type-token ratio (STTR)

3.84 Third-person pronouns* (TPP3)

The tagger counts occurrences of words in the following list, then standardises the occurrences by the length of the input text.
她, 他, 他们, 她们, 它, 它们

3.85 Time words (TIME)

The tagger counts occurrences of tag ‘time word’.

3.86 Wh- words (WH)*

The tagger counts occurrences of tag ‘predicate interrogative pronoun’.

3.87 Unique words*

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.

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. (07-Aug-2019). *Chouxiang Mingci Cibiao Jianban [An abridged list of abstract nouns]*. Personnel Correspondence.

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.

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

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.

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, 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). Shanghai Normal University, Shanghai, China.

Ni, C. (2008). *A Study of Causal Conjunctives in Modern Chinese* 现代汉语因果连词研究 (MA). Central China Normal University, Wuhan, China.

Nini, A. (2018, October). *Multidimensional Analysis Tagger (v. 1.3) – Manual.*

Peng, J., & Zhu, X. (2017). Yingyu muyu xuesheng de hanyu qingtai dongci xide [English speakers' acquisition of Chinese modals]. *Chinese as a Second Language Research*, *6*(1), 149–174. doi: 10.1515/caslar-2017-0007

Wang, L. (2014). *A Study of Honourific and Modest Titles in Chinese* 汉语尊称、谦称研究 (MA). Xi'an International Studies University 西安外国语大学, Xi'an.

Wang, M. (2017). An Analysis of the Linguistic Features and the Teaching Implications of Master Dissertations of Thai Students Majored in Chinese International Education 汉语国际教育专业泰国来华留学生硕士论文语言特征分析及教学启示. *Overseas Chinese Education*(10), 1384–1394. doi: 10.14095/j.cnki.oce.2017.10.009

Wei, Z. (2019). Study on Discourse Involvement Devices Used by Chinese Native Speakers and Nonnative Speakers 汉语本族语者和学习者互动交际时的话语参与行为研究. *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]* (MA). Jinan University, Guangzhou, China.

Xiao, R., He, L., & Yue, M. (2010). In pursuit of the third code: Using the ZJU corpus of translational Chinese in translation studies. In *Using corpora in contrastive and translation studies* (pp. 182–214). Newcastle upon Tyne:

Cambridge Scholars Publishing.

Xiao, R., McEnery, T., & Qian, Y. (2006). Passive constructions in English and Chinese. *Languages in Contrast*, *6*(1), 109–149. doi: 10.1075/lic.6.1.05xia

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, 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.

Yang, X., & Cheng, L. (2016). A Corpus-based Contrastive Study of Translationese 英汉翻译不同语域下被动标记形式及语义韵变化中的 "transiationese". *Chinese Translators Journal*(6), 5–12.

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