THE NEW CHINESE CORPUS OF LITERARY TEXTS LITCHI

Mateja PETROVČIČ
University of Ljubljana, Slovenia
mateja.petrovcic@ff.uni-lj.si

Radovan GARABÍK
Ľ. Štúr Institute of Linguistics, Slovak Academy of Sciences, Bratislava, Slovakia
garabik@kassiopeia.juls.savba.sk

Ľuboš GAJDOŠ
Faculty of Arts of the Comenius University in Bratislava, Slovakia
lubos.gajdos@uniba.sk

Abstract

The aim of the article is to introduce the corpus of Chinese literary texts and to describe the process and design principles behind the corpus construction. The authors provide information regarding the reasoning behind the chosen structure and annotation of the corpus, and further discuss possibilities the corpus opens for linguistic research and language learning. The article provides several examples of how the corpus can be used at various levels of language research.

Keywords: Chinese; corpus linguistics; building and using corpora; literary texts; Litchi

Povzetek

Namen članka je predstaviti korpus kitajskih literarnih besedil ter opisati procese in principe njegove izdelave. Sledi utemeljitev za izbrano strukturo korpusa in obrazložitev uporabljenega sistema označevanja. V nadaljevanju prispevek predstavi možnosti uporabe korpusa za jezikoslovne raziskave in učenje oziroma poučevanje jezika. Različne zahtevnostne stopnje smo avtorji tudi ponazorili s številnimi primeri.

Ključne besede: kitajščina; korpusno jezikoslovje; izdelava in uporaba korpusov; literarna besedila; Litchi

Acta Linguistica Asiatica, 10(2), 2020.
ISSN: 2232-3317, http://revije.ff.uni-lj.si/ala/
DOI: 10.4312/ala.10.2.65-81

1 Introduction

Corpus linguistics is a well-established indispensable part of linguistic research in general. We can find the most prominent use of monolingual huge corpora both in scientific research or practical uses, notably in lexicography, language education, natural language processing, and as a valuable data source in machine learning and data mining / data science.

There is a lack of corpora of different text types and genres in general, although it is indisputable that texts have many different functions in social life and result in corresponding differences in form and substance (Council of Europe, 2011, p. 93). Awareness of these features are mentioned several times in the Common European framework of reference for languages, advising the users to consider with which text types the learner will need/be equipped/be required to deal receptively, productively, interactively, and in mediation (Council of Europe, 2011, p. 96). Therefore, specialized corpora are needed in L2 acquisition as complementary learning resources.

Among freely available Chinese corpora, the BCC corpus (http://bcc.blcu.edu.cn) is an exception in this respect, since it distinguishes the following subcategories of texts: literature *wénxué* 文学, press *bàokān* 报刊, multi-domain[1] *duōlǐngyù* 多领域, *Wēibó* 微博, science and technology *kējì* 科技, and ancient Chinese *gǔ Hànyǔ* 古汉语. However, even though the BCC corpus enables users to go beyond simple search, its functions are very limited compared to CQL supported corpora.[2] Therefore, the authors realized a creation of an extra corpus of Chinese literary texts (named *Litchi*) with advanced functionality would be a useful addition to freely available corpora of Chinese language.

Perhaps surprisingly, there are only a few academic institutions in Europe that build and use corpora of Chinese language, even though Chinese as a second language and Chinese exolinguistic research is steadily gaining in popularity. Comenius University in Bratislava (Slovakia) is one of the institutes working on Chinese language corpus linguistics and corpus creation. The first corpus (the web-corpus Hanku)[3] was created in 2016, followed by the corpus of legal Chinese in 2018.[4]

Although the usefulness of language corpora is indisputable, we nevertheless sometimes encounter questions of practical considerations. Why is there a strong need

[1] In the present version, this part is called *duōlǐngyù* 多领域 (multi-domain), and in the previous versions it was called *zōnghé* 综合 (comprehensive). As stated on the website, this section includes texts from the newspapers, literature, Weibo, science and technology. These contents are independent and do not intersect with other sections of the BCC corpus. The goal of this part was to build a "balanced" corpus.

[2] See retrievable examples *jiǎnsuǒshì shìlì* 检索式示例 (http://bcc.blcu.edu.cn/help) for details.

[3] See more in Gajdoš, Garabík and Benická (2016).

[4] See more at http://158.195.113.63/run.cgi/corp_info?corpname=zh-law.

for such corpora and how can they be effectively exploited? We firmly believe that it is indispensable to make further research on registers of modern Chinese to answer such questions. Sadly, the current utilization of available corpus linguistic resources in scientific research in this area still does not reach satisfactory levels in many respects.

2 Availability

The *Litchi* is available via the website of Comenius University,[5] using the NoSketch Engine web-based corpus manager (Rychlý, 2007; Kilgarriff et al., 2014). The process of the building has begun in autumn 2019, and the corpus structure is modelled after our previous implementation to keep user access compatible across different corpora. Main parameters of the corpus are summarized in the following table.

Table 1: Parameters of the corpus Litchi

Parameters	Status	Notes
Type	synchronous	literary texts from the Internet
Language of interface	Slovak, English, Chinese, others	explanatory notes in English
Size (May 2020)	92 613 119	in tokens
Tokenization	into words (*cí* 词)	automatic statistical word segmentation
POS annotation	yes	Penn Chinese Treebank tagset
Bibliographic annotation	yes	title, author’s name, alternative author's name(s), authors’ geographical origin, gender, date of birth indication
Style and genre annotation	no	
Phonetic annotation	yes	Hànyǔ pīnyīn: tones marked by diacritics; tones marked by numerals
Syntactic annotation	yes	Penn Chinese Treebank compatible dependency annotation
Statistic tools	yes	absolute frequency, relative frequency average reduced frequency
Save results directly from the interface	yes	in text or XML format

[5] Available at https://fphil.uniba.sk/katedry-a-odborne-pracoviska/katedra-vychodoazijskych-studii/cinsky-jazykovy-korpus/litchi/.

Parameters	Status	Notes
KWIC	yes	KWIC or sentence view
Collocations search	yes	many collocation measures
Advanced search options	yes	Boolean operators—conjunction, disjunction, negation; possibility to use regular expressions at the character, word, pinyin, and metadata level; full CQL etc.
Sorting by	yes	Multi level sorting hierarchy; left, right, node, references etc.
Availability	registration required	free to use for registered users, registration not restricted

3 Corpus compilation

The Litchi corpus is compiled of freely available literary works in Chinese published on the Internet. The source texts are stored in the GB18030 character encoding (as a national standard of the People's Republic of China) (Lunde, 2009, p. 105). However, all the subsequent processing and annotation are performed in the UTF-8 encoding,[6] to ensure maximum compatibility of processing tools, corpus manager and user access.

3.1 Cleanup

Text cleanup consists of removing unwanted characters, collapsing whitespace to a single ordinary space, replacing control characters with a space, and unifying line endings. Tokenization is performed by *ZPar* (Zhang & Clark, 2011), with tokens equal to Chinese words (*cí* 词), but tokens include also numbers, punctuation and other symbols (with the exception of white space and control characters).

Tones are written either using standard diacritics or, for users lacking the means to enter Hànyǔ pīnyīn diacritics, there is a possibility to use digits 1 to 5 (with the neutral tone having the number 5). The transcription into Hànyǔ pīnyīn was performed by the *xpinyin* package.[7]

[6] In practice, we can treat the GB18030 as an alternative ASCII-extending encoding of the Unicode character repertoire (on par with UTF-8).

[7] See more at https://lxneng.com/posts/70.

3.2 Corpus structure

Since the Litchi corpus was compiled with Chinese as foreign language instruction in mind, the annotation has been designed to facilitate queries by inexperienced Chinese speakers (e.g. students of the language).[8]

3.3 Positional attributes

Positional attributes describe token-level annotation – the basic unit of the corpus is a *token*, it usually corresponds to a word, but also punctuation characters and numerals are separate tokens. Given the specifics of written Chinese language, tokens in the Litchi corpus are equal to Chinese words (*cí* 词); tokenization (word segmentation) in Chinese is a nontrivial task and a certain amount of errors is to be expected.

Each token can be assigned several attributes, further describing or specifying the token, its grammatical or lexical features. Following positional attributes are used in the Litchi corpus: *word, lemma, tag, pinyin, npinyin, head, deprel*.

The fundamental attribute ***word*** is the basic unit of the corpus (token). It is in the original form of the word (*cí* 词) in the text as written in Hànzì. Example: 斯洛文尼亚 (Slovenia).

We repurposed the default attribute ***lemma*** to be an all-encompassing default query type. It is a combination of a word written in Hànzì, each individual character (*zì* 字) of a word in Hànzì, Hànyǔ pīnyīn transcription of a ***word***, using both diacritics and numerals to mark tones, a transcription with the tones omitted, as well as a union of transcriptions of individual characters (字) of the word. We aim for inclusiveness – if a user enters a single syllable (in either Hànzì or one of the two Hànyǔ pīnyīn transcriptions, or even in Hànyǔ pīnyīn without tones), the corpus manager will search for all the words containing the syllable. For example, the word *Sīluòwénníyǎ* 斯洛文尼亚 will be assigned the “lemma” luo|luo4|luò|ni|ni2|ní|si|si1|si1luo4wen2ni2ya4|siluowenniya|sī|sīluòwénníyà|wen|wen2|wén|ya|ya4|yà|亚|尼|文|斯|斯洛文尼亚|洛.

The attribute ***tag*** is the part of speech tag, a two or three uppercase ASCII character denoting the part of speech of the *word*. For example, the word 斯洛文尼亚 will be likely part-of-speech tagged as NR (i.e. Proper Noun).

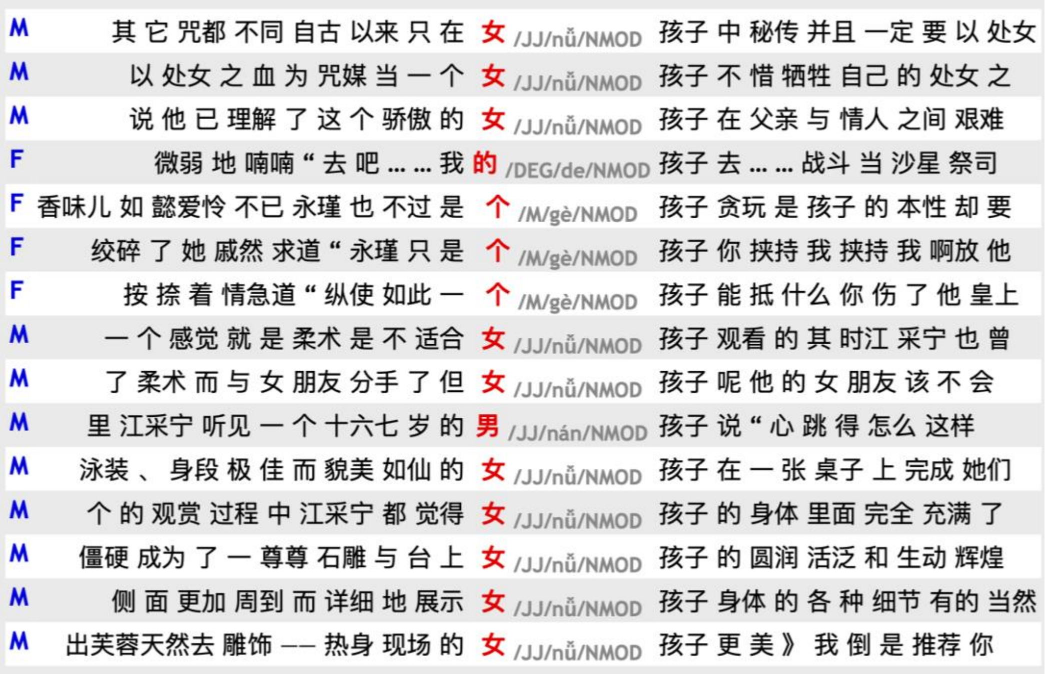
The ***pinyin*** attribute is the transcription of word using the Hànyǔ pīnyīn method, tones are indicated by diacritics. The transcription is in lowercase. For example, 斯洛文尼亚 will be transcribed as *sīluòwénníyà*. Characters with multiple readings will be assigned only the first (in some collation) reading.

[8] For details see Chapter 4.

The ***npinyin*** is again the Hànyǔ pīnyīn transcription, but this time the tones are indicated by numerals 1 to 5 (5 stands for the neutral tone). For example, 斯洛文尼亚 will be transcribed as *si1luo4wen2ni2ya4*.

Tokens in the current sentence are numbered (counted from zero) and the attribute ***head*** is the token number the current word is in relation to.

The attribute ***deprel*** marks the syntactical relation of the word (node) to the governing word (node).



M	其 它 咒都 不同 自古 以来 只 在	女 /JJ/nǚ/NMOD	孩子 中 秘传 并且 一定 要 以 处女
M	以 处女 之 血 为 咒媒 当 一 个	女 /JJ/nǚ/NMOD	孩子 不 惜 牺牲 自己 的 处女 之
M	说 他 已 理解 了 这 个 骄傲 的	女 /JJ/nǚ/NMOD	孩子 在 父亲 与 情人 之间 艰难
F	微弱 地 喃喃 “ 去 吧 … … 我	的 /DEG/de/NMOD	孩子 去 … … 战斗 当 沙星 祭司
F	香味儿 如 懿爱怜 不已 永瑾 也 不过 是	个 /M/gè/NMOD	孩子 贪玩 是 孩子 的 本性 却 要
F	绞碎 了 她 戚然 求道 “ 永瑾 只 是	个 /M/gè/NMOD	孩子 你 挟持 我 挟持 我 啊放 他
F	按 捺 着 情急道 “ 纵使 如此 一	个 /M/gè/NMOD	孩子 能 抵 什么 你 伤 了 他 皇上
M	一 个 感觉 就 是 柔术 是 不 适合	女 /JJ/nǚ/NMOD	孩子 观看 的 其 时江 采宁 也 曾
M	了 柔术 而 与 女 朋友 分手 了 但	女 /JJ/nǚ/NMOD	孩子 呢 他 的 女 朋友 该 不 会
M	里 江采宁 听见 一 个 十六七 岁 的	男 /JJ/nán/NMOD	孩子 说 “ 心 跳 得 怎么 这样
M	泳装 、 身段 极 佳 而 貌美 如仙 的	女 /JJ/nǚ/NMOD	孩子 在 一 张 桌子 上 完成 她们
M	个 的 观赏 过程 中 江采宁 都 觉得	女 /JJ/nǚ/NMOD	孩子 的 身体 里面 完全 充满 了
M	僵硬 成为 了 一 尊尊 石雕 与 台 上	女 /JJ/nǚ/NMOD	孩子 的 圆润 活泛 和 生动 辉煌
M	侧 面 更加 周到 而 详细 地 展示	女 /JJ/nǚ/NMOD	孩子 身体 的 各 种 细节 有的 当然
M	出芙蓉天然去 雕饰 —— 热身 现场 的	女 /JJ/nǚ/NMOD	孩子 更 美 》 我 倒 是 推荐 你

Figure 1: Example of the use of the attribute *deprel* (NMOD – functionally corresponds to an attributive); searching for nominal modifiers of the word *háizi* 孩子

3.4 Structures

The corpus possesses a hierarchical structure – the so-called *structures* describe information about grouping of tokens, or intra-token information. The corpus can be thus seen as a stream of tokens, interrupted by special marks denoting a start of a structure, end of a structure, or a structure between two tokens.

The Litchi corpus uses following structures (compatible with de facto standards in written language corpora):

<doc> stands for one document, which is a logically and conceptually separate standalone unit, typically a book, a short story etc. The structure contains several attributes, providing annotation of the document (metadata).

<p> marks paragraphs, units conveying a sort of coarse-grained segmentation of text; paragraphs are inferred from the structure of the text itself, without resorting to linguistic information

<s> marks sentences, segmented according to heuristic-statistical model of the *ZPar* segmentation.

The structure ***<g>***, often used in other corpora to mark that there was no whitespace between tokens is not used since spaces in written Chinese are mostly irrelevant (and not used).

3.5 Document annotation

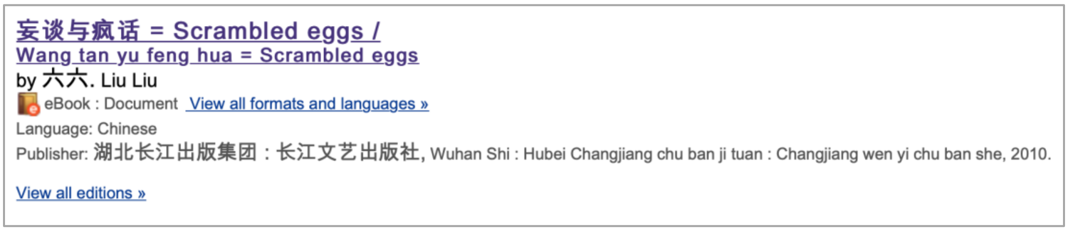
Each document has a certain set of metadata (document annotation) that are kept in the compiled corpus and can be queried or the results can be filtered by the metadata.

doc.title is the name of the document (e.g. book title), written in Hànzì. The Litchi corpus includes 1312 different literary works.

doc.author is the name of the author (pen name, if different from the real name), written in Hànzì.

doc.alter_name comprises alternative author names (either the real name or other pen names). If the author is of non-Chinese origin, this string includes the name (or multiple name variants) either in the original language or in a well-known transcription.

The motivation for this labeling was to maintain the original pair title-author. For example, according to the WorldCat, the work 妄谈与疯话 (Wàng tán yǔ fēnghuà) is written by author 六六 (Liu Liu) (see Figure 2).



妄谈与疯话 = Scrambled eggs /
Wang tan yu feng hua = Scrambled eggs
by 六六. Liu Liu
eBook : Document View all formats and languages »
Language: Chinese
Publisher: 湖北长江出版集团：长江文艺出版社, Wuhan Shi : Hubei Changjiang chu ban ji tuan : Changjiang wen yi chu ban she, 2010.
View all editions »

Figure 2: The results of the query “妄谈与疯话” in WorldCat

However, Liu Liu is a pen name of the author with the real name Zhang Xin 张辛, as provided elsewhere on the Internet, for example the Xiabook.com (https://www.shutxt.com/writer/61/).

Similarly, the author of the work entitled 尼尔斯骑鹅旅行记 (*Níěrsī qí é lǚxíngjì*) is 西尔玛·拉格洛夫 (Xī’ěrmǎ Lāgéluòfū) in the *doc.author* field, and Selma Lagerlöf in the *doc.alter_name* field.

All the bibliographic records in Litchi (including origin, gender, and age range) have been checked manually to verify and complement the meta data. As a result, 539 different authors are included in the corpus.[9] Some tasks were quite intriguing, for example the author "B·N·崔可夫" (N B Cuikefu), which stands for "Vasily Ivanovich Chuikov". In the original Chinese text, "B" is a letter of a Cyrillic script and stands for the letter "V" in Latin script. However, instead of "И" (e.g. letter "I" in Latin script), the mirror form "N" has been used. The Chinese version is therefore a mixture between "Vasily Ivanovich Chuikov" and "Василий Иванович Чуйков".

doc.authors_origin provides information on authors' origin in the geographical or linguistic sense (e.g. China, Korea, Japan, etc.) to enable the user to distinguish originally Chinese texts from the translated works into Chinese. This is a two-letter abbreviation of the region.

doc.gender is the gender of the author, we use the value self-described by the author or the gender the author is commonly considered to be of. This is not strictly a binary valued item – currently, there are three values present in the corpus annotation: M for male, F for female, N/A for unknown gender.

doc.born_in provides information related to the authors' age, in 15-year intervals. Authors born in the previous centuires have only the century of their birth recorded here (written in English, with the numeric part at the beginning).

The value "N/A" has been assigned to all the bibliographic data where no clear and straight expressions were available in the authors' online profiles. For example, if a person's brief presentation avoided the use of 3rd person personal pronouns and used neutral expressions, such as *bǐzhě* 笔者 (writer), *zuòzhě* 作者 (author), *qí* 其 (his/her), *běnrén* 本人 (I/me), it was not possible to assign a clear gender value. Similarly, if the author states to be born "in a small village in Yue nan" (生于粤南一小村), this does not necessarily mean "in the South Guangdong". Unlike the expressions Eastern/Northern/Western Guangdong (粤东/粤北/粤西), the notion *Yuè nán* 粤南 (lit. Southern Guangdong) doesn't seem to refer to any real geographical places.[10] If the description was further masked with blurred expressions such as "graduated from the BA studies at a certain university" (某名校本科毕业), this further justified the use of the "N/A" value.

The final proportion of known and unknown information for fields *doc.gender*, *doc.authors_origin* and *doc.born_in* is presented in Figue 3:

[9] The BCC corpus includes works from 469 different authors.

[10] See http://wap.yuexinet.net/view.php?aid=38

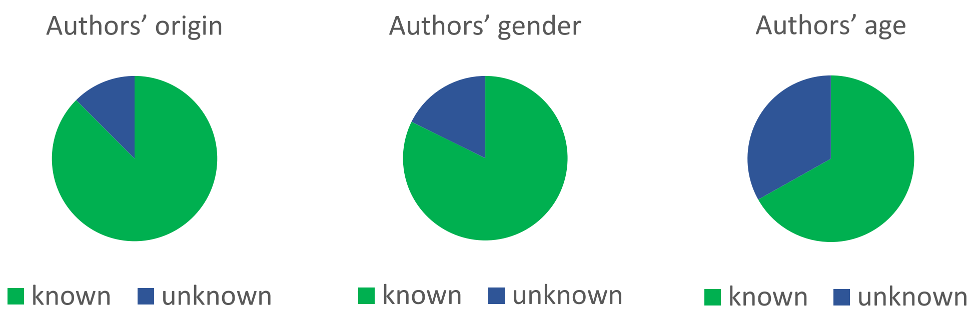


Figure 3: Proportion of known/unknown information in the authors' data

4 Usage in linguistics research

The corpus manager is a very powerful tool when in the hands of an experienced user. originally aimed at scientific research in linguistics and related fields, in the last decades of ever-increasing importance of corpus linguistics the usage of corpora converged to a subfield of descriptive linguistics with its own terminology, approaches, good practices and established rules. Nevertheless, the learning curve is not prohibitively steep, the corpus manager can even be used by completely casual users, if we prepare the corpus adequately and provide sane defaults.

For pedagogical reasons, we arbitrarily divide the corpus usage into these levels:

- basic
- advanced
- expert

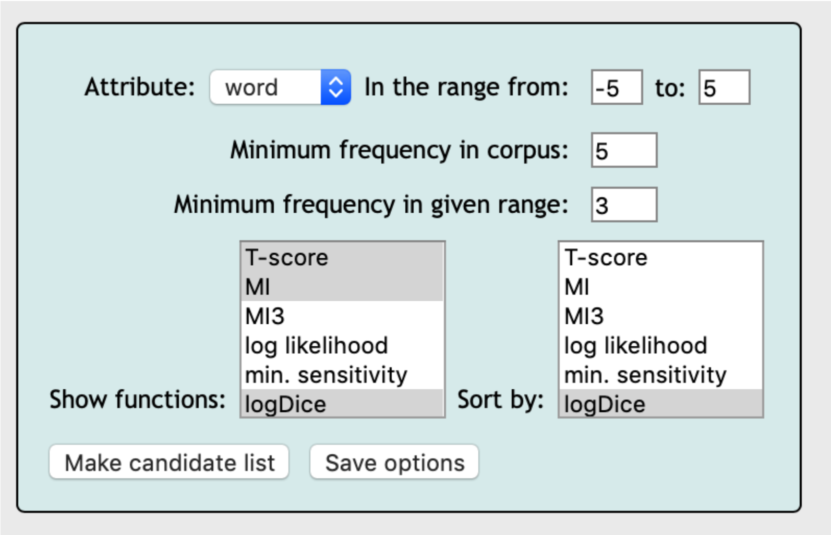
Needless to say, this division is based on our experience and the dividing lines between the levels are not strictly delineated.

4.1 Basic use

At the basic level a user may search for a word as KWIC (Key Word In Context). This is a very basic option when searching for concordance (context) of KWIC and it is very useful for students of foreign languages or translators. This usage usually does not require any additional instructions – users just type the word and get a readable list of occurrences. For Chinese language corpora, the situation is a bit complicated by the need to enter Hànzì characters. Although the plethora of input methods is a thing of the past and (in a non-professional setting) the prevalent input method is based on toneless Hànyǔ pīnyīn transcription, language model selecting (and ordering) the most probable Hànzì characters and the user picking up the appropriate character. While easy for native or fluent speakers, it can be challenging for students or less literate, less proficient non-native speakers. Also the specific tokenization matters – users have to be familiar with our chosen segmentation into words (词).

This is the basic motivation behind our *lemma* attribute – by default, the users can query the corpus by a single character (字) or a word (词); both of them can be written in Hànzì, in Hànyǔ pīnyīn with standard diacritics, in Hànyǔ pīnyīn with tones indicated by numbers, or in toneless Hànyǔ pīnyīn. Thus users with either technical obstacles preventing them typing Hànzì or diacritics, or users less proficient in written Chinese can still benefit from the corpus, by entering the search term in an intuitive way and still getting (a superset of) relevant results. This is obviously very important in teaching Chinese as foreign language.

In addition to searching for given words or characters, one of the nontrivial results we can obtain from huge corpora is the collocation analysis by various collocation measures. By default, the logDice measure is selected, empirically found to provide the best results for lexicographic purposes (and by extension, for almost any other purpose as well) (see Figure 4). The NoSketch Engine UI makes it very easy to search for collocation candidates in the corpus.



Attribute: word In the range from: -5 to: 5

Minimum frequency in corpus: 5

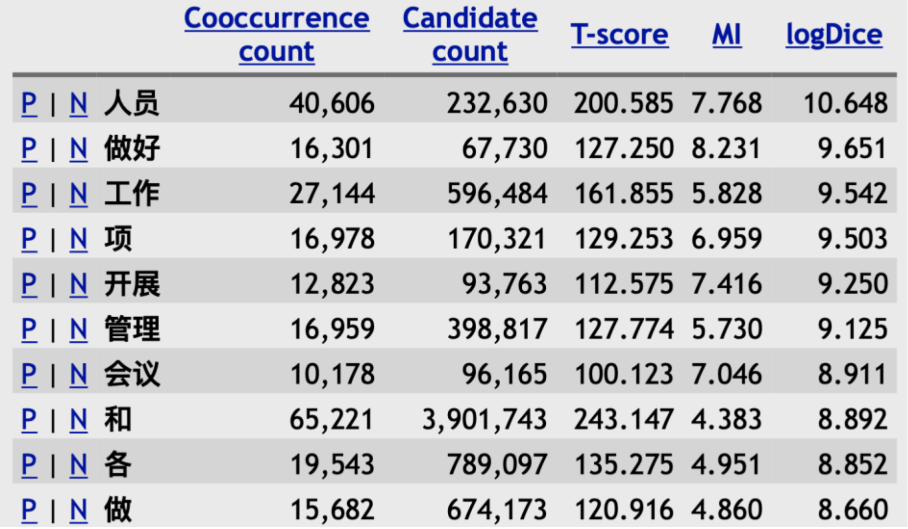
Minimum frequency in given range: 3

Show functions: T-score, MI, logDice

Sort by: logDice

Figure 4: The Collocations candidates parameters selection UI

The collocation candidates for e.g. the token *gōngzuò* 工作 (work) are presented in Figure 5.



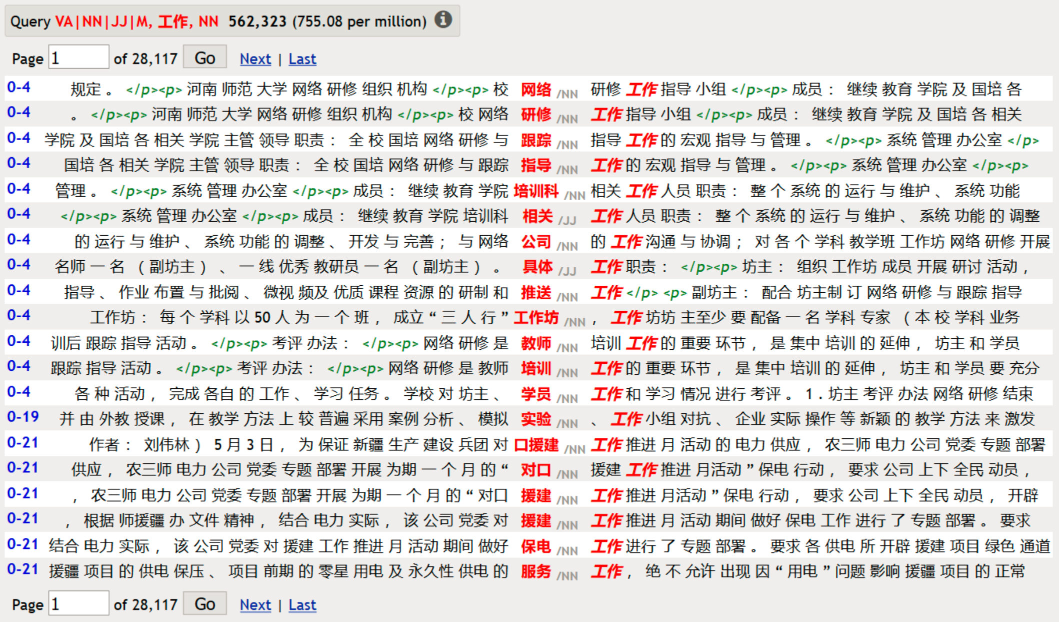
	Cooccurrence count	Candidate count	T-score	MI	logDice
P | N 人员	40,606	232,630	200.585	7.768	10.648
P | N 做好	16,301	67,730	127.250	8.231	9.651
P | N 工作	27,144	596,484	161.855	5.828	9.542
P | N 项	16,978	170,321	129.253	6.959	9.503
P | N 开展	12,823	93,763	112.575	7.416	9.250
P | N 管理	16,959	398,817	127.774	5.730	9.125
P | N 会议	10,178	96,165	100.123	7.046	8.911
P | N 和	65,221	3,901,743	243.147	4.383	8.892
P | N 各	19,543	789,097	135.275	4.951	8.852
P | N 做	15,682	674,173	120.916	4.860	8.660

Figure 5: Top 10 collocation candidates of a word *gōngzuò* 工作, word range -5 to 5 (i.e. up to five tokens in both directions)

Results of this simple query show that this word is frequently found in the phrases such as *gōngzuò rényuán* 工作人员 (staff member); it takes the verb *zuòhǎo* 做好 (to do/to finish), as in *zuòhǎo gōngzuò* 做好工作 (to do a job well), *zuòhǎo zìjǐ de gōngzuò* 做好自己的工作 (to do one's own work), *zuòhǎo yuǎnjiāo gōngzuò* 做好远教工作 (to do a distance teaching work); as a noun, it takes the classifier *xiàng* 项; it is expected to be used together with the conjunction hé 和 (with), e.g. to work with somebody, etc.

4.2 Advanced use

At the advanced level, it is possible to search for combinations of a few words conforming to a specified condition (e.g. usage of negation words (Gajdoš, 2019), concrete word order, part-of-speech tags, syntactic role, Boolean operators etc.) by using CQL expressions. In this example, we search for the most frequent attributives to the noun *gōngzuò* 工作 (work). The CQL query for this task would be (meet [tag="VA|NN|JJ|M"] 1:[word="工作" & tag="NN"]1 2). See Figure 6.



Query VA|NN|JJ|M, 工作, NN 562,323 (755.08 per million)

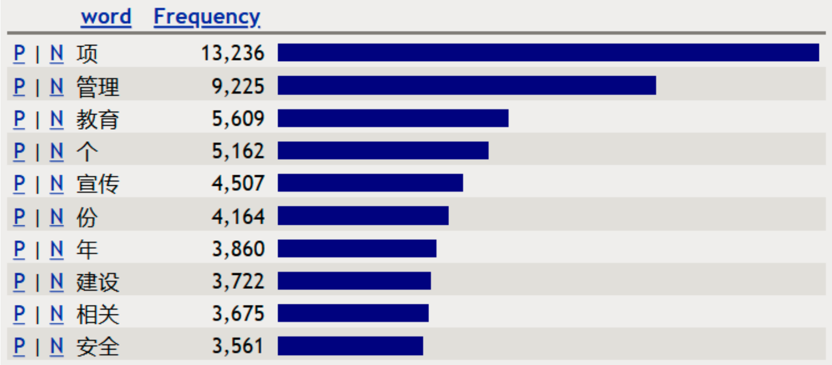
Page 1 of 28,117 Go Next | Last

0-4	规定 。 </p><p> 河南 师范 大学 网络 研修 组织 机构 </p><p> 校	网络 /NN	研修 工作 指导 小组 </p><p> 成员 ： 继续 教育 学院 及 国培 各
0-4	。 </p><p> 河南 师范 大学 网络 研修 组织 机构 </p><p> 校 网络	研修 /NN	工作 指导 小组 </p><p> 成员 ： 继续 教育 学院 及 国培 各 相关
0-4	学院 及 国培 各 相关 学院 主管 领导 职责 ： 全 校 国培 网络 研修 与	跟踪 /NN	指导 工作 的 宏观 指导 与 管理 。 </p><p> 系统 管理 办公室 </p>
0-4	国培 各 相关 学院 主管 领导 职责 ： 全 校 国培 网络 研修 与 跟踪	指导 /NN	工作 的 宏观 指导 与 管理 。 </p><p> 系统 管理 办公室 </p><p>
0-4	管理 。 </p><p> 系统 管理 办公室 </p><p> 成员 ： 继续 教育 学院	培训科 /NN	相关 工作 人员 职责 ： 整 个 系统 的 运行 与 维护 、 系统 功能
0-4	</p><p> 系统 管理 办公室 </p><p> 成员 ： 继续 教育 学院 培训科	相关 /JJ	工作 人员 职责 ： 整 个 系统 的 运行 与 维护 、 系统 功能 的 调整
0-4	的 运行 与 维护 、 系统 功能 的 调整 、 开发 与 完善 ； 与 网络	公司 /NN	的 工作 沟通 与 协调 ； 对 各 个 学科 教学班 工作坊 网络 研修 开展
0-4	名师 一 名 （ 副坊主 ） 、 一 线 优秀 教研员 一 名 （ 副坊主 ） 。	具体 /JJ	工作 职责 ： </p><p> 坊主 ： 组织 工作坊 成员 开展 研讨 活动 ，
0-4	指导 、 作业 布置 与 批阅 、 微视 频及 优质 课程 资源 的 研制 和	推送 /NN	工作 </p> <p> 副坊主 ： 配合 坊主制 订 网络 研修 与 跟踪 指导
0-4	工作坊 ： 每 个 学科 以 50 人 为 一 个 班 ， 成立 “ 三 人 行 ”	工作坊 /NN	， 工作 坊坊 主至少 要 配备 一 名 学科 专家 （ 本 校 学科 业务
0-4	训后 跟踪 指导 活动 。 </p><p> 考评 办法 ： </p><p> 网络 研修 是	教师 /NN	培训 工作 的 重要 环节 ， 是 集中 培训 的 延伸 ， 坊主 和 学员
0-4	跟踪 指导 活动 。 </p><p> 考评 办法 ： </p><p> 网络 研修 是 教师	培训 /NN	工作 的 重要 环节 ， 是 集中 培训 的 延伸 ， 坊主 和 学员 要 充分
0-4	各 种 活动 ， 完成 各自 的 工作 、 学习 任务 。 学校 对 坊主 、	学员 /NN	工作 和 学习 情况 进行 考评 。 1 . 坊主 考评 办法 网络 研修 结束
0-19	并 由 外教 授课 ， 在 教学 方法 上 较 普遍 采用 案例 分析 、 模拟	实验 /NN	、 工作 小组 对抗 、 企业 实际 操作 等 新颖 的 教学 方法 来 激发
0-21	作者 ： 刘伟林 ） 5 月 3 日 ， 为 保证 新疆 生产 建设 兵团 对	口援建 /NN	工作 推进 月 活动 的 电力 供应 ， 农三师 电力 公司 党委 专题 部署
0-21	供应 ， 农三师 电力 公司 党委 专题 部署 开展 为期 一 个 月 的 “	对口 /NN	援建 工作 推进 月活动 ” 保电 行动 ， 要求 公司 上下 全民 动员 ，
0-21	， 农三师 电力 公司 党委 专题 部署 开展 为期 一 个 月 的 “ 对口	援建 /NN	工作 推进 月活动 ” 保电 行动 ， 要求 公司 上下 全民 动员 ， 开辟
0-21	， 根据 师援疆 办 文件 精神 ， 结合 电力 实际 ， 该 公司 党委 对	援建 /NN	工作 推进 月 活动 期间 做好 保电 工作 进行 了 专题 部署 。 要求
0-21	结合 电力 实际 ， 该 公司 党委 对 援建 工作 推进 月 活动 期间 做好	保电 /NN	工作 进行 了 专题 部署 。 要求 各 供电 所 开辟 援建 项目 绿色 通道
0-21	援疆 项目 的 供电 保压 、 项目 前期 的 零星 用电 及 永久性 供电 的	服务 /NN	工作 ， 绝 不 允许 出现 因 “ 用电 ” 问题 影响 援疆 项目 的 正常

Page 1 of 28,117 Go Next | Last

Figure 6: The results of the query
(meet [tag="VA|NN|JJ|M"] 1:[word="工作" & tag="NN"]1 2)

In the next step, results of the previous query may be ordered by Node forms in the Frequency menu, to get a list of the most frequent attributives (Figure 7).



	word	Frequency
P | N	项	13,236
P | N	管理	9,225
P | N	教育	5,609
P | N	个	5,162
P | N	宣传	4,507
P | N	份	4,164
P | N	年	3,860
P | N	建设	3,722
P | N	相关	3,675
P | N	安全	3,561

Figure 7: Top 10 most frequent attributives of the noun *gōngzuò* 工作 (work)

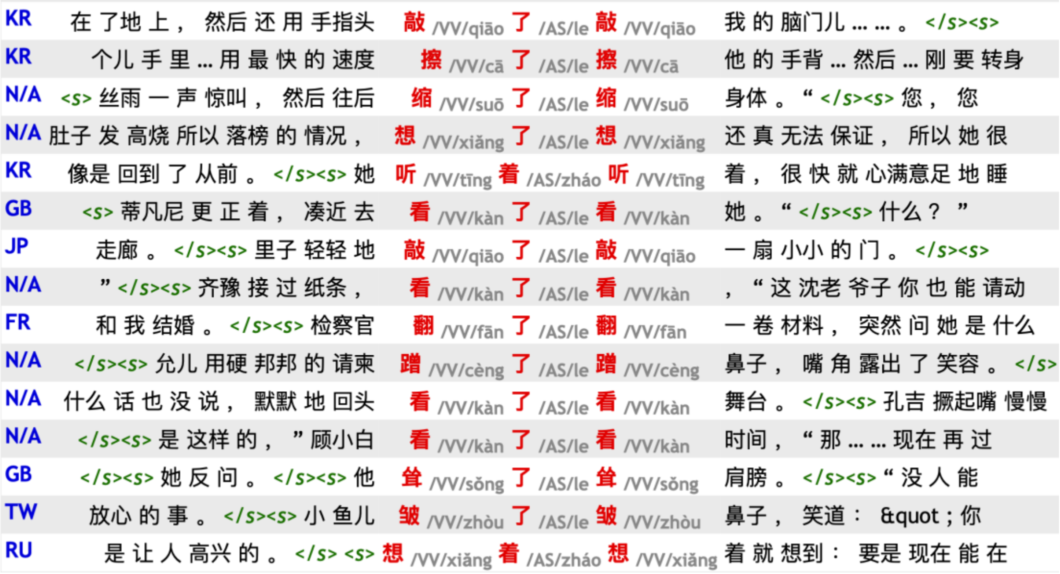
Results reveal that the most frequent noun phrases with the head noun *gōngzuò* 工作 (work) include *guǎnlǐ gōngzuò* 管理工作 (management work), *jiàoyù gōngzuò* 教育工作 (educational work), *xuānchuán gōngzuò* 宣传工作 (promotional work), *jiànshè gōngzuò* 建设工作 (construction work), and others. Its most frequent measure words are *xiàng* 项, *gè* 个 or *fèn* 份, etc.

In our opinion, this level of usage is suitable for most cases – language pedagogy as well as linguistics research.

4.3 Expert use

The expert level is an extension of the previous one and it is often used in linguistics research. The corpus manager offers an arbitrary combination of POS tags, word order, context filters (e.g. MEET, WITHIN), conditions for bibliographic annotation etc. For example, with bibliographic annotation in the Litchi corpus, it is possible to search for a concrete grammatical phenomenon in the works of one author (doc.author) or in the works of all female authors (doc.gender). The following figure demonstrates the possibility of conditions combination (search only in texts by authors *not* from Mainland China; find all "regular" verbs (VV) with an "aspect" marker (AS) followed again by the same verb).

CQL query:
(1:[tag="VV"] 2:[tag="AS"] 3:[tag="VV"] within <doc authors_origin!="CN"/>) & 1.word=3.word

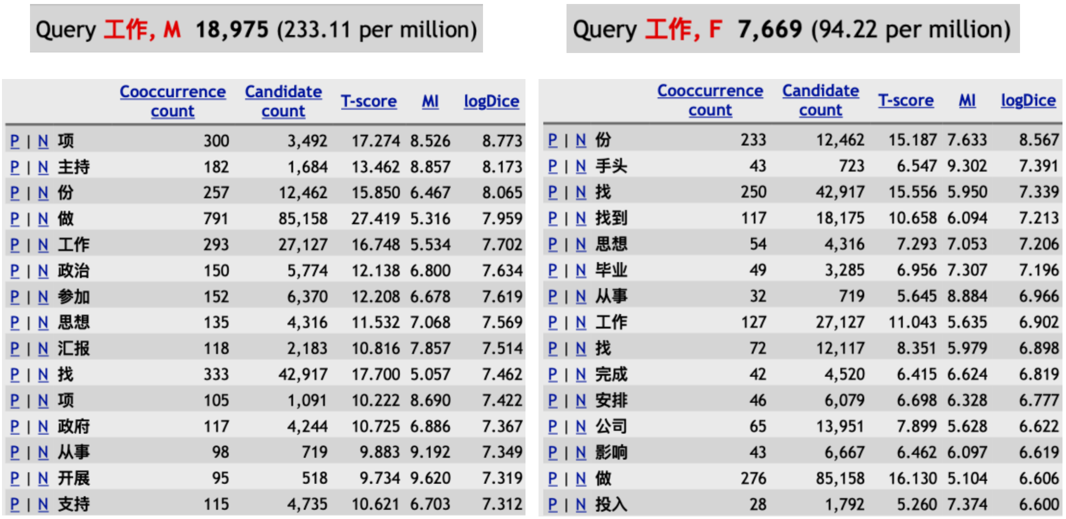


KR	在 了地 上， 然后 还 用 手指头	敲 /VV/qiāo 了 /AS/le 敲 /VV/qiāo	我 的 脑门儿 … … 。 </s><s>
KR	个儿 手 里 … 用 最 快 的 速度	擦 /VV/cā 了 /AS/le 擦 /VV/cā	他 的 手背 … 然后 … 刚 要 转身
N/A	<s> 丝雨 一 声 惊叫， 然后 往后	缩 /VV/suō 了 /AS/le 缩 /VV/suō	身体 。 " </s><s> 您， 您
N/A	肚子 发 高烧 所以 落榜 的 情况，	想 /VV/xiǎng 了 /AS/le 想 /VV/xiǎng	还 真 无法 保证， 所以 她 很
KR	像是 回到 了 从前 。 </s><s> 她	听 /VV/tīng 着 /AS/zháo 听 /VV/tīng	着， 很 快 就 心满意足 地 睡
GB	<s> 蒂凡尼 更 正 着， 凑近 去	看 /VV/kàn 了 /AS/le 看 /VV/kàn	她 。 " </s><s> 什么？ "
JP	走廊 。 </s><s> 里子 轻轻 地	敲 /VV/qiāo 了 /AS/le 敲 /VV/qiāo	一 扇 小小 的 门 。 </s><s>
N/A	" </s><s> 齐豫 接 过 纸条，	看 /VV/kàn 了 /AS/le 看 /VV/kàn	， " 这 沈老 爷子 你 也 能 请动
FR	和 我 结婚 。 </s><s> 检察官	翻 /VV/fān 了 /AS/le 翻 /VV/fān	一 卷 材料， 突然 问 她 是 什么
N/A	</s><s> 允儿 用硬 邦邦 的 请柬	蹭 /VV/cèng 了 /AS/le 蹭 /VV/cèng	鼻子， 嘴 角 露出 了 笑容 。 </s>
N/A	什么 话 也 没 说， 默默 地 回头	看 /VV/kàn 了 /AS/le 看 /VV/kàn	舞台 。 </s><s> 孔吉 撅起嘴 慢慢
N/A	</s><s> 是 这样 的， " 顾小白	看 /VV/kàn 了 /AS/le 看 /VV/kàn	时间， " 那 … … 现在 再 过
GB	</s><s> 她 反 问 。 </s><s> 他	耸 /VV/sǒng 了 /AS/le 耸 /VV/sǒng	肩膀 。 </s><s> " 没 人 能
TW	放心 的 事 。 </s><s> 小 鱼儿	皱 /VV/zhòu 了 /AS/le 皱 /VV/zhòu	鼻子， 笑道： &quot; 你
RU	是 让 人 高兴 的 。 </s> <s>	想 /VV/xiǎng 着 /AS/zháo 想 /VV/xiǎng	着 就 想到： 要是 现在 能 在

Figure 8: The combination of conditions in CQL

Or, to continue with an example using *gōngzuò* 工作 (work), it can be observed, that male authors tend to write more about work than female authors. Moreover, their focus seems to be on different aspects of work, as roughly indicated in the data. The most frequent collocation candidates in men works are *zhǔchí gōngzuò* 主持工作 (take charge of the work), *zhèngzhì gōngzuò* 政治工作 (political work) or *cānjiā gōngzuò* 参加工作 (participate in work); whereas the most frequent collocation candidates in works of female authors include *shǒutóu gōngzuò* 手头工作 (work at hand), *zhǎo gōngzuò* 找工作 (to look for a job), *zhǎodào gōngzuò* 找到工作 (to find a job).[11]

[11] For more relevant results, a thorough research should be conducted.



Query 工作, M **18,975** (233.11 per million)

	Cooccurrence count	Candidate count	T-score	MI	logDice
P | N 项	300	3,492	17.274	8.526	8.773
P | N 主持	182	1,684	13.462	8.857	8.173
P | N 份	257	12,462	15.850	6.467	8.065
P | N 做	791	85,158	27.419	5.316	7.959
P | N 工作	293	27,127	16.748	5.534	7.702
P | N 政治	150	5,774	12.138	6.800	7.634
P | N 参加	152	6,370	12.208	6.678	7.619
P | N 思想	135	4,316	11.532	7.068	7.569
P | N 汇报	118	2,183	10.816	7.857	7.514
P | N 找	333	42,917	17.700	5.057	7.462
P | N 项	105	1,091	10.222	8.690	7.422
P | N 政府	117	4,244	10.725	6.886	7.367
P | N 从事	98	719	9.883	9.192	7.349
P | N 开展	95	518	9.734	9.620	7.319
P | N 支持	115	4,735	10.621	6.703	7.312

Query 工作, F **7,669** (94.22 per million)

	Cooccurrence count	Candidate count	T-score	MI	logDice
P | N 份	233	12,462	15.187	7.633	8.567
P | N 手头	43	723	6.547	9.302	7.391
P | N 找	250	42,917	15.556	5.950	7.339
P | N 找到	117	18,175	10.658	6.094	7.213
P | N 思想	54	4,316	7.293	7.053	7.206
P | N 毕业	49	3,285	6.956	7.307	7.196
P | N 从事	32	719	5.645	8.884	6.966
P | N 工作	127	27,127	11.043	5.635	6.902
P | N 找	72	12,117	8.351	5.979	6.898
P | N 完成	42	4,520	6.415	6.624	6.819
P | N 安排	46	6,079	6.698	6.328	6.777
P | N 公司	65	13,951	7.899	5.628	6.622
P | N 影响	43	6,667	6.462	6.097	6.619
P | N 做	276	85,158	16.130	5.104	6.606
P | N 投入	28	1,792	5.260	7.374	6.600

Figure 9: Comparison of frequency and collocation candidates
for the word *gōngzuò* 工作 (work) in relation to authors' gender

Data also show that there are more works written by men, but this does not influence the relative frequency of the selected word.

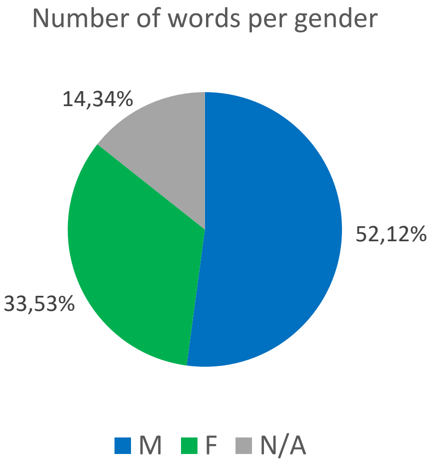


Figure 10: Distribution of the gender annotation value,
as a percentage of the number of words (词) in the corpus

Following two tables demonstrate the use of the corpora to identify keywords that are more relevant in one corpus, as compared to the second (reference) corpus, using the Simple maths method (Kilgarriff 2009) – the words with their relative frequency much higher in one corpus. We focus on rare words.

Table 2: Comparison of most relevant keywords in the *zh-law* corpus, as compared against *zh-lit* as the reference corpus

	zh-law		zh-lit		
word	Freq	Freq/mill	Freq	Freq/mill	Score
行政	33949	4712.4	197	2.4	1378.1
国务院	11302	1568.8	27	0.3	1178.8
下列	9974	1384.5	15	0.2	1169.9
应当	57301	7953.8	530	6.5	1059.1
条例	6761	938.5	7	0.1	865.1
规定	50465	7004.9	583	7.2	858.3
生产	16099	2234.6	161	2.0	750.7
认证	5486	761.5	3	0.0	735.4
直辖市	5970	828.7	12	0.1	723.1
申请人	5643	783.3	11	0.1	690.9
法规	8449	1172.8	59	0.7	680.5
共和国	8772	1217.6	71	0.9	650.9
注册	8703	1208.0	83	1.0	598.6
自治区	6212	862.3	43	0.5	564.9
受理	4655	646.1	14	0.2	552.2

Table 3: Comparison of most relevant keywords in the *zh-lit* corpus, as compared against *zh-law* as the reference corpus

	zh-lit		zh-law		
word	Freq	Freq/mill	Freq	Freq/mill	Score
想	200,276	2460.4	4	0.6	1582.7
那	266,923	3279.2	16	2.2	1018.4
什么	219,236	2693.4	12	1.7	1010.8
你	720,026	8845.7	65	9.0	882.7
吧	80,800	992.6	1	0.1	872.5
那	149,508	1836.7	16	2.2	570.6
那么	48,539	596.3	2	0.3	467.5
却	124,025	1523.7	18	2.5	435.8
里	208,813	2565.3	39	5.4	400.1
句	41,029	504.0	2	0.3	395.3
呢	48,497	595.8	4	0.6	383.7
为什么	39,129	480.7	2	0.3	377.0
拿	32,804	403.0	1	0.1	354.8
东西	29,531	362.8	1	0.1	319.5
现在	69,307	851.5	13	1.8	304.0

Last but not least, the Litchi mainly reflects the language use of speakers born in recent decades, as shown in Figure 11. Therefore, this corpus is also appropriate for studies focusing on some specific features of the most recent language use.

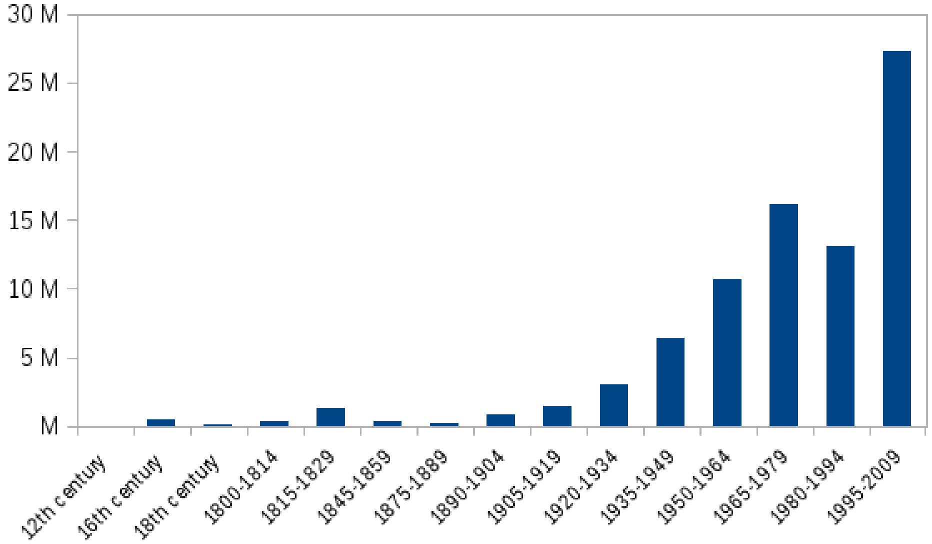


Figure 11: Distribution of author birth dates, by the number of words (词) in the corpus

5 Conclusion

The Litchi corpus is the third corpus of a family of Chinese language corpora used at the Comenius University. It adds a corpus of a different language variety and register to the existing corpora of Chinese (texts of laws, web corpus), while keeping compatible structure and annotations. The corpus manager offers the possibility of quantitative/ qualitative analysis of various Chinese language registers - comparison of the three corpora, but it can also be used for comparison between. Chinese language usage in different situations or contexts (e.g. between translation and original texts; analysis of different expressions used by authors based on their gender, historical period etc.). The corpus is accessible through a web interface upon registration and aims to be a valuable resource for both teachers and students of Chinese as a foreign language, but also for linguistic research.

To conclude, the Litchi is a unique corpus in many respects. It provides a rich bibliographic, phonological, morphological and syntactic annotation and thus offers wide range of possibilities for linguistics research, e.g. lexicography/lexicology, morphology, syntax and to some extend also sociolinguistics.

Acknowledgments

This work was partially supported by the Slovenian Research Agency under research program Asian languages and cultures (P6-0243).

References

Council of Europe. (2011). *Common European framework of reference for languages: Learning, teaching, assessment*. Cambridge, U.K: Cambridge University Press. https://rm.coe.int/1680459f97

Gajdoš, Ľ. (2019). Retrieving Linguistic Information from a Corpus on the Example of Negation in Chinese. *Acta Linguistica Asiatica*, *9*(2), 103-115. https://doi.org/10.4312/ala.9.2.103-115

Gajdoš, Ľ., Garabík, R., & Benická, J. (2016). The New Chinese Webcorpus Hanku – Origin, Parameters, Usage. *Studia Orientalia Slovaca, 15*(1), 53–65.

Kilgarriff, A. et al. (2014). The Sketch Engine: Ten Years on. *Lexicography*, 1.1, 7-36.

Kilgarriff, A. (2009). Simple maths for keywords. In M. Mahlberg, V. González-Díaz & C. Smith (Eds.), *Proceedings of Corpus Linguistics Conference CL2009*. University of Liverpool, UK.

Lunde, K. (2009). CJKV Information Processing: Chinese, Japanese, Korean & Vietnamese Computing, 2nd edition. O'Reilly Media.

Rychlý, P. (2007). Manatee/Bonito-A Modular Corpus Manager. In P. Sojka & A. Horák (Eds.), *RASLAN 2007* (pp. 65-70). Brno: Masaryk University.

Zhang, Y., & Clark, S. (2011). Syntactic Processing Using the Generalized Perceptron and Beam Search. *Computational Linguistics*, *37*(1), 105-151.