The New Chinese Webcorpus *Hanku*— Origin, Parameters, Usage

Ľuboš Gajdoš
Radovan Garabík
Jana Benická

Nový čínsky webový korpus Hanku—*pôvod, parametre, využitie*

Resumé V príspevku stručne predstavíme dostupné (cez webové rozhranie) korpusy čínskeho jazyka, ich klady ale i nedostatky, ktoré nás viedli k vybudovaniu nového čínskeho (webového) korpusu *Hanku*. Predostrieme tiež detailnejšie zdôvodnenie zvoleného postupu a parametre korpusu. Naznačíme tiež využitie korpusu *Hanku* v jazykovednom výskume a didaktike jazyka.

Abstract In this paper, we describe the initial impetus for the building of the Chinese corpus *Hanku,* briefly summarize available corpora of Chinese language, their strengths and weak sides. We also provide information regarding the solutions chosen and parameters. We show the usage of the *Hanku* corpus in linguistic research and language teaching.

Keywords Chinese, Language · Chinese Corpus Linguistics, Building and Using Corpora

Introduction

Information technologies perforce play an important role in our everyday lives and are widely available. Their application's domain is still on the rise and they

are penetrating into many new areas. One of such is a language and linguistic research. A corpus is one of the implementation of information technology in linguistics research and language teaching. Corpus linguistics uses information technology and systematically works with corpora.

1 *The Primary Motivation*

Let us briefly summarize Chinese corpora available via the web interface, their parameters, advantages and shortcomings, as they were initial impetus for our decision to build a new Chinese web corpus.

We will focus only on (monolingual) Chinese corpora (in simplified Chinese characters) available via the web interface for free, with or without registration.[1]

1.1 *The CCL Corpus—Center for Chinese Linguistics, Peking University*

The corpus is available at: <http://ccl.pku.edu.cn:8080/ccl_corpus/index.jsp>.

Parameters	Status	Notes
Type	synchronous, diachronic, parallel	published texts
Language of interface	Chinese	including explanatory notes
Size (February 2016)	581.794.456 (Modern) 201.668.719 (Classical)	size referred in Chinese characters, not tokens (see shortcomings)
Tokenization	×	1 token = 1 Chinese character
POS annotation[2]	×	
Bibliographic annotation	✓	searchable
Style and genre annotation	✓	searchable
Phonetic annotation	×	

1 The analysis of these corpora is based on our experience and it should be considered only as a brief introduction. From the available corpora, we have excluded the Internet corpus of the

2 POS—Part of Speech.

Statistic tools	✓	only limited usage
Save results directly from the interface	✓	in text format
KWIC[3]	✓	
Collocations search	×	
Advanced search options	✓	to use Boolean operators—conjunction, disjunction, negation
Registration	×	
Other		A stable and fast server response

Strengths

- KWIC in synchronic and diachronic texts
- KWIC in a defined style, genre or in author's works
- co-occurrence of words
- suitable for language teaching and learning
- clean and simple UI[4]

Limitations

- searchable only for characters, not words
- rather limited usage of statistic tools—frequency of occurrence
- only partly suitable for linguistics research

3 KWIC—Key Word in Context.

4 UI—User Interface.

1.2 *The CNC—Institute for Applied Linguistics, Ministry of Education (PRC)*
The corpus is available at: <http://www.cncorpus.org/>.

Parameters	Status	Notes
Type	synchronous, diachronic	published texts
Language of interface	Chinese	including explanatory notes
Size (February 2016)	200.000.000 (MC)[5] 100.000.000 (CC)	size referred in tokens
Tokenization	✓	also available as a standalone tool
POS annotation	✓	also available as a standalone tool[6]
Bibliographic annotation	✓	not searchable
Style and genre annotation	✓	not searchable
Phonetic annotation	✓	also available as a standalone tool
Syntactic annotation	×	
Statistic tools	✓	only limited usage
Save results directly from the interface	✓	in text format
KWIC	✓	
Collocations search	×	
Advanced search options	✓	Boolean operators—conjunction, disjunction, negation; searchable for words according to POS tags
Registration	✓	
Other		a slow server response from Europe

5 MC—Modern Chinese, CC—Classical Chinese.

6 The POS tagset available at <http://www.cncorpus.org/resources/GBT20532-2006StandardofPOSTagofContemporaryChinese.pdf>.

Strengths

- KWIC in synchronic and diachronic texts
- precise tokenisation
- sorting results according to frequency
- phonetic annotation
- suitable for Language teaching and learning
- simple and clean UI

Limitations

- rather limited possibility for advanced search
- rather limited usage of statistic tools—frequency of occurrence
- only partly suitable for linguistics research

1.3 *The BCC—Beijing Language and Culture University*

The corpus is available at <http://bcc.blcu.edu.cn>.

Parameters	Status	Notes
Type	synchronous, diachronic	published texts, blogs
Language of interface	Chinese[7]	including explanatory notes
Size (June 2016)	15 billion thereof 2 billion of Classical Chinese	size referred in tokens
Tokenization	✓	
POS annotation[8]	✓	
Bibliographic annotation	✓	searchable
Style and genre annotation	✓	searchable
Phonetic annotation	×	
Syntactic annotation	✓	as tree structure
Statistic tools	✓	frequency of absolute occurrence
Save results directly from the interface	✓	in text format
KWIC	✓	

7 There is an English (French) interface, but reserved for the corpus of English (French) langauge.

8 Tagset available at <bcc.blcu.edu.cn/help#intro>.

Collocations search	×	
Advanced search options	✓	Boolean operators—conjunction, disjunction, negation; possibility to use regular expressions[9] etc.
Registration	?	optional
Other		sometimes a slow server response from Europe

Strengths

- KWIC in synchronic and diachronic texts
- KWIC in synchronic texts according e.g. subcorpus of journalistic texts
- KWIC in synchronic texts with a time frame
- precise tokenisation
- syntactic annotation
- sorting results according to frequency
- suitable for language teaching and learning
- simple way of using regular expressions
- suitable for linguistics research

Limitations

- rather limited possibility for advanced search
- rather limited usage of statistic tools—frequency of occurrence in absolute occurrence and not in IPM[10]
- collocation search missing.

Taking into consideration strengths and weaknesses of the above corpora, none of these can fullfil both functions—linguistics research *vs.* language teaching since both areas have different requirements for a corpus. Nevertheless, we hope that our approach might gradually converge towards a sustainable solution.

9 Regular expressions are queries that use a well-established, fairly standard and extremely powerful search syntax that was originally developed within the field of computer science. Cf. Tony McEnery and Andrew Hardie, *Corpus Linguistics* (Cambridge, England: Cambridge University Press, 2012), 255.

10 IPM—Instances Per Million, a number of occurrences normalized by the size of the corpus.

2 *The* Hanku—The Chinese Internet Corpus, *Confucius Institute at Comenius University in Bratislava*

Initial requirements for the new corpus from the perspective of linguistics research might be listed below (by relevance):

- an accurate and reliable tokenizator
- an accurate and reliable POS annotation with a complex tagset
- powerful statistical data analysis software
- advanced capabilities of search, possibility to use regular expressions
- collocation search
- style and genre annotation
- syntactic annotation
- phonetic annotation

As the requirements for the language teaching are distinctly different—a simple UI, a quick and easy access to KWIC and a collocations tool etc.—the solution is in many way challenging and an acceptable compromise between the two approaches.

2.1 *Realisation*

The *Hanku* is available via the website of the Confucius Institute at Comenius University in Bratislava at <konfuciovinstitut.sk>. The process of the building has begun in spring 2016 and the authors decided to use open-source tools.

Parameters	Status	Notes
Type	synchronous	published texts, texts from the Internet
Language of interface	Slovak, English, Chinese	explanatory notes in English
Size (June 2016)	2 billion plus	size referred in tokens
Tokenization[11]	✓	

11 The tokenizator and POS tagger »ZPar« are available at <people.sutd.edu.sg/~yue_zhang/doc/doc/joint_seg_tag.html>.

POS annotation		✓	Penn Chinese Treebank tagset[12]
Bibliographic annotation		✓	searchable, (only partly)
Style and genre annotation		✓	searchable, (only partly)
Phonetic annotation		✓	
Syntactic annotation		✓	only syntactic functions are searchable, dependency trees and phrase structure are available separately
Statistic tools		✓	frequency in IPM, average reduced frequency
Save results directly from the interface		✓	in text or XML format
KWIC		✓	
Collocations search		✓	many collocation measures
Advanced search options		✓	Boolean operators—conjunction, disjunction, negation; possibility to use regular expressions at the character, word, pinyin, and metadata level; full CQL[13] etc.
Sorting by		✓	Left, right, node, references etc.
Registration			optional
Other			

12 Tagset available at <cs.brandeis.edu/~clp/ctb/posguide.3rd.ch.pdf>.

13 CQL—Computer Query Language.

2.2 *Structure of the Corpus*

The logical (as presented to the end user) structure of the corpus is based on documents. Each document represents more or less one continuous standalone piece of written language. Typically, one document corresponds to one webpage (»s« referenced by a URL), or a newspaper article, a book etc. The set of documents form the corpus directly, there is no higher hierarchy level included. Lower hierarchy levels include text structures (paragraphs, sentences) and tokens with their positional attributes.

The paragraphs in the corpus correspond to common text paragraphs, as present in the original text documents. In case of HTML documents, paragraphs correspond directly to the <p> element. Sentences are delimited automatically, given the hierarchical structure, a sentence cannot span more paragraphs.

The basic block of the corpus is a token—one single position in the text. Traditionally in corpus linguistics, one token represents one word in the source text, with additional information, such as lemma, part of speech or syntactical function. For Chinese, there are two possibilities—the more natural is to tokenize text into characters (*Hanzi* 漢字), which closely follows the mental model of written Chinese and is natural and intuitive for native speakers. Such a tokenization is also technically rather easy, though the presence of numerals or Latin characters complicates the matters

However, for learners of the language, the transfer from a character (morpheme) to a lexeme is opaque and requires substantial effort. Therefore, it is better to analyse the language in term of words, themselves composed of syllables (=characters), even if the division of text into words is often fuzzy and subject to individual interpretation.

In the *Hanku* corpus, each token is annotated for part of speech (POS), its composition into characters and the *Hanyu pinyin* transcription.

2.3 *Positional Attributes*

There are following positional attributes associated with each token: word, lemma, tag, *pinyin,* »Npinyin« (cf. page below).

»Word« is the direct representation of the source text—either a sequence of (most often one or two) Chinese characters, a numeral in Arabic digits, a punctuation character or a word in Latin letters.

»Tag« is the part of speech tag, following the tagset used in the Penn Chinese Treebank.

»Pinyin« is the transcription into *Hanyu pinyin*, using diacritical marks for the tones. In case of a multicharacter word, the syllables are joined together (without spaces). Npinyin is the transcription into *Hanyu pinyin*, using numbers to denote tones, the neutral tone (*qingsheng* 輕聲) using the number 5.

»Lemma«[14] is a special attribute designed to facilitate easy and hassle free access to the corpus. By using the multivalue facility of the positional attributes, it is possible to encode several values at the same position. These values are the word itself, separate characters comprising the word (if the word consists of more than one character), POS tag, *Hanyu pinyin* transcription with tonal diacritics, *Hanyu pinyin* transcription with numeric tones, *Hanyu pinyin* transcription without tones, and the same variants of *Hanyu pinyin* transcription of separate characters. This enables to query either by a word, single character, *Hanyu pinyin* transcription of either the whole word or a single character, with or without the tones.

Example of Token Annotation[15]

Word	POS	Pinyin	Npinyin
名	M	míng	ming2
男子	NN	nánzǐ	nan2zi3
毫	AD	háo	hao2
不	AD	bù	bu4
避讳	VV	bìhuì	bi4hui4
谈论	VV	tánlùn	tan2lun4
当天	NT	dāngtiān	dang1tian1

14 Here, »lemma« is just a name of the second positional attribute, not the basic form of the word, as used in other (inflectional) languages.

15 Lemma is omitted for brevity.

荷兰	NR	hélán	he2lan2
与	CC	yǔ	yu3
斯洛伐克队	NR	sīluòfákèduì	si1luo4fa2ke4dui4
比赛	NN	bǐsài	bi3sai4
的	DEC	de	de5
盘口	NN	pánkǒu	pan2kou3

Lemma attribute for the word »lema« 勒马*('rein in a horse')*[16]

勒马	勒马|勒|马|VV|lèmǎ|le4ma3|lè|mǎ|le4|ma3|lema|le|ma

3 *Usage in Linguistics Research*

In this paragraph, we only suggest some possibilities for linguistics research, as it is a very complex issue.

Naturally, the first area of linguistic research seems to quantitative analysis. NoSketch Engine provides various statistical tools, e.g. to evaluate different language hypotheses, explore words usage, syntactic patterns etc.

As already mentioned, style and genre annotation of a Chinese corpus is quite rare. In many ways, this is crucial for the description of discrepancy between written and spoken register of any language with written tradition. As the research has already demonstrated the written register of English is to a certain extent very different from the spoken register and *vice versa*.[17] In Chinese, the situation is particularly tricky, as there is not any adequate corpus

16 Note the use of U+007C VERTICAL LINE as a separation character.

17 See for example Douglas Biber, *University Language: A Corpus-Based Study of Spoken and Written Registers* (Amsterdam: John Benjamins Publishing Company, 2006).

of spoken Chinese language. On the other hand, quantitative (qualitative) research on written registers might be conducted by means of e.g. style and genre annotation and it thus undoubtedly contribute to this language phenomenon.[18]

So far, the *Hanku* corpus is equipped with the following style and genre annotation:

- *baokan* 報刊—journalistic texts from the PRC
- *falv* 法律—legal texts from the PRC; texts of laws and regulations
- *wenxue* 文學—texts of Modern literature
- *zhuanye* 專業—professional texts
- *duihua* 對話—dialogues (direct speech) of modern literature (experimental)

Texts of one particular style (e.g. *falv*) then can be searched alone as a subcorpus and the statistical data are valid only for this register (style). Needless to say, this is very useful not only for linguistics but for language teaching too.

As for the limitations, the texts in the corpus is only partly annotated with style and genre tags. The POS annotation, tokenization and syntactic annotation are results of automatic processing.

4 *Usage in Language Teaching*

Based on above parameters and our experience, we assume that the *Hanku* common usage scenarios in language teaching are as follows:

- basic word usage—KWIC[19]
- collocation[20] preferences of a word
- sentence pattern search

18 See more Ľuboš Gajdoš, »The discrepancy between spoken and written Chinese: methodological notes on linguistics«, *Studia Orientalia Slovaca* 10,1 (2011), 155–159.

19 For instance—a preferencies of a noun *ren* 人 with classificators might be written in CQL as follows: [tag="M"] [word="人"], the result is then sorted by frequencies.

20 The query for the collocation of the word *piaoliang* might be [word="漂亮"], then search for collocation from right side with distance 1.

- register's specific usage of a word
- register's preference of synonyms etc.

Performing the KWIC and collocation's search are basic tasks which an ordinary user of corpora is familiar with. Using a regular expression, e.g. the sentence pattern, might be seen as an advanced level. With the help of style and genre annotation, students of Chinese language can more easily grasp the distinction of written registers. The last aspect of using the corpus is sometimes the only way to obtain information of a word because there are only a few dictionaries of Chinese synonyms on the market and no grammar of written Chinese for foreigners.

Bearing in mind the sophisticated statistical software of the NoSketch Engine on one hand and the criterion of simplicity for language teaching on the other, sometimes less is more and this is especially true for the second area. Based upon our experience with the NoSketch Engine during the classes, even a basic task caused troubles and the NoSketch Engine UI needs time to get used to. Yet the system of the corpus (under the query type »lema«) allows a user to search for Chinese words or characters by writing them in *Hanyu pinyin* with or without the tones.

5 *Summary*

In this article we wanted to briefly describe the genesis and objectives of the new Chinese web corpus *Hanku*. We have also suggested the possible usage scenarios of the corpus in linguistics research as well as language teaching, have shown its strengths. Yet, the *Hanku* is an ongoing project and we seek to remedy all existing shortcomings.

Slovak Academy of Sciences, Ľudovít Štúr Institute of Linguistics, Bratislava
Comenius University in Bratislava, Department of East Asian Studies