Sinica Corpus user manual

Introduction to Sinica Corpus

"Academia Sinica Balanced Corpus of Modern Chinese", simplified as Sinica Corpus, is the first Balanced Modern Chinese Corpus with part-of-speech tagging. The preliminary version of Sinica Corpus was developed on a small-scale and opened to the academic community in 1994 with the major purpose of obtaining feedback. Later in 1997 we present the corpus (Sinica Corpus 3.0) with 5 million words and a user-friendly search interface. The new version targeted at 10 million words is completed in 2006. Texts are collected from different areas and classified according to five criteria: genre, style, mode, topic, and source. Therefore, this corpus is a representative sample of modern Chinese language. Sinica Corpus is developed and maintained by Institute of Information Science and CKIP group in Academia Sinica. The CKIP group is managed by Prof. Keh-jiann Chen (Institute of Information Science) and Prof. Chu-Ren Huang (Institute of Linguistics).

In addition to data-collection and data cleaning in the construction of a Chinese Balanced Corpus, we are also concerned with: 1) balancing and classifying collected data, 2) Chinese word segmentation, and 3) the design of pos-tag sets (Chen 1994).

1. Data extraction and classification for a Balanced Corpus
Topical distribution of the Sinica corpus:

Topics	Philosophy	Science	Society	Art	Life	Literature	Total
Percentage	8%	8%	38%	5%	28%	13%	100%

2. Issues of Chinese word segmentation:
The word segmentation standard for Chinese information processing issued by the Central Standards Bureau was adopted as the guideline for segmenting words in the Sinica corpus.

3. The Part-of Speech tagging system and its Interpretation:
In accordance with the Tagset of 178 syntactic categories from the CKIP lexicon(CKIP 1993), a reduced tagset of 46 different tags (43 tags plus 3 features) is applied by Sinica Corpus.

4. Part-of-speech analysis: Technical Report no.93-05. This technical report includes detail PoS analysis and the corresponding argument structures.

Copyright Notice

Using Sinica Corpus interface online

1. Help on the main page

In order to retrieve sentences, matching with the parameters given by the users, from the corpus, Main Page allows our users to set up searching restriction(s). There are two approaches to process your searching: A) Single-restriction searching, an approach that only has one set-up condition. B) Multiple-restriction searching, an approach that is possible to set up several parameters at one time, which includes "AND Searching Restriction"; and "OR Searching Restriction" by applying "Joint Restrictions". The four parameters that you will find in this page are: a) Keyword; b) Reduplication; c) Part of Speech (POS); and d) Feature.

Single-restriction searching, an approach that only has one set-up condition

1.1 **Keyword**: Enter the **keyword** you wish to search for and click on **"GO"**
Your keyword can be composed by the followings：

Chinese character(s)

？ ： standing for any single syllable

＊ ：standing for null or indefinite number of syllables

For example:

- Entering "電話", the system will search for sentences containing "電話".
- Entering "電＊", all sentences that contain words beginning with "電" (words composed by single syllable, disyllable, poly-syllable are all included, e.g., 電, 電話, 電視機) will be displayed.
- Entering "電？", the searching will look for sentences carrying disyllabic words starting with "電" (e.g., 電話, 電腦, 電子).
- Entering "＊電", the searching will focus on sentences taking words ending with "電" (single syllable, disyllable, poly-syllable words are all implied, e.g., 電, 觸電, 無線電).
- Entering "電？？" the system will look for sentences with tri-syllabic words which start with "電" (e.g., 電擊棒, 電療法, 電纜車).
- Entering "＊電＊", sentences carrying any words containing "電" (including single syllable, disyllable, poly-syllable, e.g., 電, 電扇, 靜電, 電壓計, 無線電) will be collected.
- Entering "？電？", tri-syllable words that have "電" occurring in the middle will be regarded as the searching keywords (e.g., 心電圖, 手電筒).
- Entering "？電＊", all sentences that include a non-single syllabic word in which "電" occurs in the second position will be presented (both disyllable and poly-syllable words fall in this category, e.g., 充電, 核電廠).
- Entering "？？？？", the system will search for sentences containing any four-syllable word.

Clicking on **"CLEAR"** anytime, you will clear all settings.

1.2 **Reduplication**：Move your cursor to the box next to **"Reduplication"** and click on the arrow. Then, select the type of reduplication you are looking for from the four options or you may simply type in the box, and press **"GO"**.

There are four types of reduplication as illustrated on the interface：

Reduplication AAB－e.g.,試試看、走走路
Reduplication ABB－e.g.,試看看、亮閃閃
Reduplication AABB－e.g.,高高興興、平平安安
Reduplication ABAB－e.g.,高興高興、研究研究

For Example：

- Entering "AAB", sentences that contain reduplicated words in AAB pattern in our

corpus will be displayed.

- Entering "AABB", the system will search for sentences carrying reduplicated words in AABB pattern.

Clicking on **"CLEAR"** anytime, you will clear all settings.

1.3 **Part of Speech (POS)**： Move you cursor to the box right to **"POS"**, and key in the part of speech you wish to search. Or, you may click on the arrow beside the column of **"List of POS"**. A list of 46 parts of speech will come out among which you may make your choice. Then, activate **"GO"**.
The parts of speech are distinguished as 46 types, which are all symbolized in English letters as you may find out in the **"List of POS"**. To understand the POS abbreviations further, please refer to Symbols of Parts of Speech, or refer to the Technical Report no. 92-05, An Introduction to Sinica Corpus* (中央研究院平衡語料庫的內容與說明), and Technical Report no. 93-05, An Analysis of POS in Chinese Lexicon* (中文詞類分析), by CKIP, Academia Sinica.

In the box of **"POS"**, you may enter a sequence of English letters representing a certain type of part of speech or one single English letter that represents a collective type.
For Example：

- Entering "VA", you will retrieve all sentences that contain any words that are tagged in VA (active intransitive verbs).
- Entering "V", sentences in our corpus that contain words tagged with a V-type POS (including any part of speech symbolized with a beginning V: VA, VB, VC, VD, VE, VF, VG, VH, VI, VJ, VK, VL, i.e., all verbs).
- Entering "Ne", the searching will be limited in the restriction of POS, "Ne" (including any part of speech symbolized with a beginning Ne: Nep, Neqa, Neqb, Nes, Neu, i.e., all determiners).

Clicking on **"CLEAR"** anytime, you will clear all settings.

1.4 **Feature**： Move your cursor and click on the arrow besides the box of **"List of Features"**. A list of nine types of feature will appear. Choose your desired feature and press **"GO"**.
There are nine types of feature which are all labeled in English letter(s) as shown in the interface. If you need a further introduction to features, please refer to section F or Technical Report no. 95-02, An Introduction to Sinica Corpus* (中央研究院平衡語

料庫的內容與說明).
You may also type a sequence of English letters in the box which represents a feature.
For example,
Entering "vrv", you may acquire sentences that contain words having a feature of "vrv" (the verb component in a verb-complement construction).

Clicking on **"CLEAR"** anytime, you will clear all settings.

Multiple-restriction searching：an approach that at least two parameters are set up at one time

AND Searching Restriction： The resulting sentences are satisfied with all restriction.
You may set the values under three categories, "Keyword/Reduplication", "POS", and "Feature" in one time, only words matched with all three parameters will be the result. You may set the values under three categories, **"Keyword/Reduplication", "POS", and "Feature"** in one time, only words matched with all three parameters will be the result.
For example,
Entering "？？" in the "Keyword" column; "VC" in the "POS" column, and "vrr" in the "Feature" column, the outcome will be any sentence that contains disyllabic word tagged with VC (active transitive verb) and labeled with a vrr feature.

Clicking on **"CLEAR"** anytime, you will clear all settings.

OR Searching Restriction： Two searching targets are focused in one search.
"Joint Restrictions"：It is possible searching for several objects in one step. After you set up the conditions of your first target, click on the box in front of**"Joint Restrictions"** (where you will see a check mark after clicking) and then **"GO"**. The Main Page will re-appear in which you are able to set up the second searching parameters. You may repeat such steps to join your restrictions within ten times.
For example：
Entering "高興" in "Keyword", click on "Joint Restriction", then activate "GO";
Entering "快樂" in "Keyword", click on "Joint Restriction", then activate "GO";
Entering "開心" in "Keyword", click on "Joint Restriction", then activate "GO";
Entering "歡喜" in "Keyword", click on "Joint Restriction", then activate "GO";
Any sentence that has "高興" or "快樂" or "開心" or "歡喜" in it will be listed.

Clicking on **"CLEAR"** anytime, you will clear all settings.

2. Help on the advanced search

This page offers some functions that give your search result a further process. Result data undergoes several processes may produce layers of collection of sentences. There are five functions provided in this page: a) "Data Screening": this is an advanced searching function while searching results acquired from the first step is too ponderous. b) "POS Statistics": If you wish to make additional observations on syntactic characteristics of your collection of sentences, this offer serves your purpose. It calculates the sum of the occurrence of each POS. c) "Calculating Collocation": this function enables you to calculate the percentage of the co-occurrence of the keyword and the words or POS's in the context. d) "Ordering": Your searching result is ordered according to the ordering in our database. By using this "Ordering" function, you may re-order the collective data to make your desired observation and comparison. e) "Page Switching": This helps you to go back to Main Page and re-start your searching, or enter the result page of every layer of collection of sentences.
The default setting of function when entering this page is "Data Screening".

2.1 **"Data Screening"**：Taking your searching result as the basic processing domain, this function helps you to modify your result by setting additional conditions or your required domain.

Setting Conditions：The operation is the same as in the settings in Main Page, which includes **keywords, reduplication, POS, and feature**. Having any questions about operation, please refer back to "Help on Main Page"

取或不取：依照設定的條件「挑出」語料或依照設定的條件「去除」語料。預設條件是依照設定的條件「挑出」語料。

Negative Restriction：By undergoing this process, a collection of data can be picked out and displayed or picked out off the display depended on your settings. Your settings, under an ordinary situation, would pick out data that fulfills these settings and shows on the result page. However, if you click on the box in front of **"Negative Restrictions"** (i.e., a check mark appears in the box), the picked out data would be eliminated in the new layer of results.

Setting Domain：The screening step can either be processed centered upon the

keyword itself, or centered upon the adjacent environment. The screening domain is adjustable. It simply depends on the settings in two Window Range columns: **"starts from:"** and **"ends at:"**. The default setting is centered upon the keyword. Thus, if you wish to widen your screening domain, you need to enter numbers in the boxes on the right side of the two columns mentioned above, e.g., 0 means the keyword itself; -1means one word left to the keyword; and 1 means one word right to the keyword. However, you may not set a range that is greater than 10 words.

If you wish to apply two screening restrictions at one time, settings in the columns, **"Restriction 1"** and **"Restriction 2"** may serve your purpose.
For example:
Range setting from 0 to 0, that is, the keyword itself: Enter your settings in **"Restriction 1"** which will constrain the keyword.
Range setting from -x to 0: **"Restriction 1"** constrains the word(s) on the left side of the keyword, whereas **"Restriction 2"** constrains the keyword itself.
Range setting from 0 to x: **"Restriction 1"** controls the keyword, whereas**"Restriction 2"** controls the context on the right side of the keyword.
Range setting from -x to y: **"Restriction 1"** gives the conditions on the left-side context, whereas **"Restriction 2"** gives the conditions on the right-side context.
When you finish setting up the conditions and domains, activate **"GO"**.
Clicking on **"CLEAR"** anytime, you will clear all settings.

2.2 **POS Statistics**：This command calculates the frequency of the part of speech of the keyword and/or of the adjacent word(s) in the context. The calculating domain is limited in the search just made..
To activate this command, you need to mark the circle in front of **"POS Statistics"**
Setting Domain: The frequency of either the keyword itself or the word(s) adjacent to the keyword can be counted. The decision is made upon domain setting which you may make adjustment in the Window Range columns **"starts from:"** and **"ends at:"**. By entering numbers in the boxes next to these two columns, you can change your counting domain. (0 means the keyword itself; -1means one word left to the keyword; and 1 means one word right to the keyword. However, you may not set a range that is greater than 10 words.)
After entering the numbers, click on **"GO"**.
Clicking on **"CLEAR"** anytime, you will clear all settings
Calculating Collocation： If you wish to observe the percentage of the collocation of the keyword and it's adjacent character(s) (i.e., word form or punctuation) from the

searching result* just made, this is the command you are looking for. Collocation is presented in Mutual Information Value (MI value). MI value shows the probability of the co-occurrence of two units. While it gives a large amount value, collocation is high and vice versa.

※NOTE: The calculation of collocation can only be valid within unscreened data. That is, you may not activate the screening step before you proceed to calculate collocation. Thus, if you have any constraints to make, you have to complete the step in Main Page.

Click on the circle in front of "Collocation"

Setting Unit： There are three options of calculating Units you may select from: "word & POS", "word", and "POS".

"word & POS", is to calculate regarding on both the adjacent character(s)and its/their part(s) of speech. In other words, once the word form is labeled with more than one part of speech, is the MI value calculated separately.

"word", means the calculation is made regarding to different adjacent character(s) and without looking their part of speech.

"POS",on the other hand, indicates that the MI value is calculated regarded to the different parts of speech. Different word forms or punctuations are not considered as distinctive units under this command.

Setting Domain： Since collocation indicates the relation between the keyword and its adjacent element(s), it is invalid to set the keyword as your calculating domain. That is, while you set your domain in Window Range **"starts from:"** and **"ends at:"**columns, you may not set "0" in both columns. However, your settings must include the keyword: the range you set has to cross zero. The default value is the keyword itself. This means, you need to change the original settings which can be succeeded by entering numbers into the boxes next to Window Range "starts from:" and "ends at:". (0 means the keyword itself; -1means one word left to the keyword; and 1 means one word right to the keyword. However, you may not set a domain that is greater than 10 words.)

Setting Ordering：The result of calculation can be ordered regarding either to the amount of MI value or to the frequency of word forms (Word Freq.).

Minimum Frequency (Min Freq.)： You may request to have your MI value calculated only if the adjacent element(s) reach(es) your desired frequency. The default frequency is 2.

Formula of calculating MI value：

$$I(x,y) = \log P(x,y)/P(x)P(y) = \log \frac{f(x,y)/N}{f(x)/N \cdot f(y)/N}.$$

Ⅰ：mutual information
Ｐ：probability
Ｎ：size of the corpus
freq（x）：the frequency of the keyword occurring within our corpus
freq（y）：the frequency of the calculating unit occurring within our corpus.
freq（x，y）：the frequency of co-occurrence of the keyword and the calculating unit occurring within this certain calculating domain.
How to read MI value?
If MI value is greater than "0", it means the keyword and this certain calculating unit tends to co-occur. The larger the MI value is, the larger the possibility of the collocation would be.
If MI value is smaller than "0", it means the keyword and this certain calculating unit tends NOT to co-occur. The smaller the MI value is, the larger the possibility of the repellency would be.
Sorting：Under this command, searching results can be sorted respecting to the desired target and criterion.
To request our system activating this command, you simply need to click on the circle in front of **"Sorting"**
Setting Target： You may select your target from three options: Keyword, L-to-Keyword, and R-to-Keyword.
Keyword：sorting respects to keyword.
L-to-Keyword：sorting respects to the first word that is in the left side context.
R-to-Keyword：sorting respects to the first word that is on the right side context.
You may select all three options and rank them. To apply your desired sorting hierarchy, you need to enter them into the boxes that are right to **"FIRST to", "SECOND to" and "FINALLAY to"** respectively.
- Setting Criterion：There are only three types of ordering criterion: Syllable-initial, Syllable-final, and POS.
Syllable-initial：sorting respects to the syllable-initial element of the target.
Syllable-final： sorting respects to the syllable-final element of the target.
POS：sorting respects to the part of speech of the target.
Eliminate Repetition：After the step of sorting, you may request a concise result that has all the repeated data deleted.
For example：
If you wish to sort respecting to "Keyword", "Syllable-initial/final", in the result, sentences that have the same keyword will be arranged together. If you have selected "Eliminate Repetition", only one sentence that contains the same keyword will be listed.

If you wish to sort respecting to "Keyword", "POS", in the result, keywords that has the same part of speech will be gathered. A request of "Eliminate Repetition" will generate the result that sentences carrying same keyword which is also tagged with same part of speech reduced to one only.
If you wish to sort respecting FIRST to "Keyword", and "Syllable-initial/final", and wish to sort respecting SECOND to "R-to-Keyword", and "Syllable-initial/final", in the result, sentences which have the same keyword and the same right-side word would be gathered together. If activating "Eliminate Repetition", each group that has the same characteristic will be represented by one example only.
If you wish to sort respecting FIRST to "Keyword" and "POS", and wish to sort respecting SECOND to "R-to-Keyword" and "POS", in the result, sentences taking the same keyword with the same part of speech and the words right to the keyword that have the same part of speech would be grouped together. Requesting "Eliminate Repetition" would display only one in each group.
If you wish to sort respecting FIRST to "Keyword" and "Syllable-initial/final", and to sort respecting SECOND to "R-to-Keyword" and "POS", in the result, data that contain the same keyword and its right-side word having the same POS would be listed in the same group. "Eliminate Repetition" would make sentences in one group reduced to one example only.
After you set up the parameters of sorting target, criterion, and "Eliminate Repetition", click on **"GO"**.
Clicking on **"CLEAR"** anytime, you will clear all settings.

Page Switching

The collective data that results from the Main Page search is the first layer data. If the data is reduced after advanced processing, it becomes the second layer data. The second layer data can be processed again and produce the third layer data and so on. You can produce ten layers of data the most.
On the top of your result page, it has a command of **"Main Page"**. You can enter Main Page simply by clicking on that command. Entering Main Page means you wish to make a new search. Thus, your previous settings would be invalid and disappeared.
Also on the top of the result page, you would find **"Upper Layer Data", "Present Layer Data", or "Lower Layer Data"**. Click on "Upper Layer Data", you would enter the page that contains the searching result belonging to a higher layer. Click on "Present Layer Data", you would enter the Data Page of this current result. Click on "Lower Layer Data", you would retrieve the collective data from the lower layer.

3. Help on Result Page

There are two types of result page, including Data Page and Statistics Page. Data Page provides three functions: a) options to whether showing POS or not, b) switching pages, and c) description of layers. Statistics page offers the last two functions.

3.1 Options to whether showing POS and features

Only Data Page has such function. If you click on **"Display POS & Feature"**, every word will have its part of speech and feature labels shown next to it. If you wish to cancel the labels, likewise, click on the same command, which has its name now changed to **"Hide POS & Feature"**, the labels would disappear.

3.2 Page switching

There are two commands: **"Main Page" and "Advanced Processing"** listed on the top of your result page. If you wish to start a new search, click on "Main Page" and the current result on your result page would be all deleted. If you wish to give an advanced search or process based on your current result, you would like to select "Advanced Processing" instead
Besides, you may find commands of **"Upper Layer Data", "Present Layer Data", or "Lower Layer Data"** on the top of your screen as well. You may enter the Data Page in a higher layer by clicking on "Upper Layer Data"; enter the current Data Page by clicking on "Present Layer Data"; and enter the Data Page containing a lower layer result by clicking on "Lower Layer Page". There are ten layers you may retrieve at most. As to the layering method, please refer to the next section, "Description of Layers".

3.3 Description of Layers

You may read a description of**¡]Xth layer, Y examples, Z layer(s) in total¡^**The interpretation of this description is: By all the advanced processes you have done, you have produced Z layer(s) of data. This current Data Page is the Xth layer which contains Y examples.
The Data page that is produced from the constraints set up in Main Page is the first layer data. Data that is reduced after advanced process becomes the second layer data. The second layer data can be processed again and produce the third layer data and so on. You can produce ten layers of data the most.

4. Appendix

Table 1: Symbols of Parts of Speech

Abbreviation	Corresponded symbols in CKIP	Interpretation
A	A	Non-predicative adjective
Caa	Caa	Conjunctive conjunction， e.g.和、跟
Cab	Cab	Conjunction, e.g.等等
Cba	Cbab	Conjunction, e.g.的話
Cbb	Cbaa, Cbba, Cbbb, Cbca, Cbcb	Correletive Conjunction
D	Dab, Dbaa, Dbab, Dbb, Dbc, Dc, Dd, Dg, Dh, Dj	Adverb
Da	Daa	Quantitative Adverb
DE		的, 之, 得, 地
Dfa	Dfa	Pre-verbal Adverb of degree
Dfb	Dfb	Post-verbal Adverb of degree
Di	Di	Aspectual Adverb
Dk	Dk	Sentential Adverb
FW		Foreign Word
I	I	Interjection
Na	Naa, Nab, Nac, Nad, Naea, Naeb	Common Noun
Nb	Nba, Nbc	Proper Noun
Nc	Nca, Ncb, Ncc, Nce	Place Noun
Ncd	Ncda, Ncdb	Localizer
Nd	Ndaa, Ndab, Ndc, Ndd	Time Noun
Nep	Nep	Demonstrative Determinatives
Neqa	Neqa	Quantitative Determinatives
Neqb	Neqb	Post-quantitative Determinatives

Nes	Nes	Specific Determinatives
Neu	Neu	Numeral Determinatives
Nf	Nfa, Nfb, Nfc, Nfd, Nfe, Nfg, Nfh, Nfi	Measure
Ng	Ng	Postposition
Nh	Nhaa, Nhab, Nhac, Nhb, Nhc	Pronoun
P	P*	Preposition
SHI		是
T	Ta, Tb, Tc, Td	Particle
VA	VA11,12,13,VA3,VA4	Active Intransitive Verb
VAC	VA2	Active Causative Verb
VB	VB11,12,VB2	Active Pseudo-transitive Verb
VC	VC2, VC31,32,33	Active Transitive Verb
VCL	VC1	Active Verb with a Locative Object
VD	VD1, VD2	Ditransitive Verb
VE	VE11, VE12, VE2	Active Verb with a Sentential Object
VF	VF1, VF2	Active Verb with a Verbal Object
VG	VG1, VG2	Classificatory Verb
VH	VH11,12,13,14,15,17,VH21	Stative Intransitive Verb
VHC	VH16, VH22	Stative Causative Verb
VI	VI1,2,3	Stative Pseudo-transitive Verb
VJ	VJ1,2,3	Stative Transitive Verb
VK	VK1,2	Stative Verb with a Sentential Object
VL	VL1,2,3,4	Stative Verb with a Verbal Object
V_2	V_2	有

Table 2: Symbols of Features

Label	Definition	Example
fw	the feature of a foreign word	卡拉 OK (Na [+fw])
nom	the feature for verbal nominalization	他的不講理(VA[+nom])
p1	the first part of a separated compound	初 Nc[+p1]、高中(Nc)
p2	the second part of a separated compound	星期六(Nd)、日(Nd[+p2])
prop	a Na used as a proper noun	蘋果(Na[+prop])電腦
spo	N of a separable V N compound	吃了他的虧 Na[+spo]
spv	V of a separable V N compound	吃 VC[+spv]了他的虧
vrv	V of a sperable VR compound	叫 VC[+vrv]不醒
vrr	R of a separable VR compound	叫不醒 VC[+vrr]