Guide to using the KOTONOHA 少納言 (Syoonagon) corpus

- KOTONOHA (少納言/Syoonagon) corpus — demo version of the Balanced Corpus of Contemporary Written Japanese (BCCWJ) [https://shonagon.ninjal.ac.jp/]
- English-language overview [https://clrd.ninjal.ac.jp/en/tool.html]
- Description of corpus project [https://link.springer.com/article/10.1007/s10579-013-9261-0]
  Maekawa, Kikuo et al. (2014). Balanced corpus of contemporary written Japanese. *Language Resources and Evaluation* 48(2): 345-371. (Open access journal article.)

The KOTONOHA corpus (Balanced Corpus of Contemporary Written Japanese; BCCWJ) was created by The National Institute for Japanese Language and Linguistics (NINJAL; 国立国語研究所). The Syoonagon interface is publicly available and allows full-text string-match searches. (Access to the morphologically parsed version of the corpus requires a license for the Tyuunagon interface.)

---

1. Accessing the corpus; Terms of Use

(1) **Access the Syoonagon corpus** here: [http://www.kotonoha.gr.jp/shonagon/]

- The main page has introductory and background information about the corpus; for an English-language introduction and overview, see Maekawa et al. (2014)
- Scroll down to the red button at the bottom of the main page and click through to the Terms of Use:

利用条件を読んで少納言を使う

Read the Terms of Use and begin using Syoonagon

(2) The **Terms of Use** are relatively straightforward; I'll give the key points here, but anyone who wants a fuller translation is welcome to contact me for more information.

1. (著作権の帰属) — Copyright: NINJAL holds copyright to the corpus, and the sources of the materials sampled in the corpus hold copyright to those materials.
2. (許諾の範囲等) — Limitations: The use of Syoonagon is limited to research and educational uses; duplication is prohibited; infringement of the rights of the copyright holders is prohibited.
3. (研究成果の公表) — Publication: Research results and knowledge obtained by using the corpus in accordance with point (2) may be published; please acknowledge use of the Balanced Corpus of Contemporary Written Japanese (BCCWJ).
4. (免責) — Disclaimers: NINJAL is not responsible for damages resulting from the use of the corpus. NINJAL may change access to Syoonagon at any time without prior notice.
5. (利用条件の更新について) — These Terms of Use are subject to change.

- If you agree with the Terms of Use, click OK to enter the corpus.

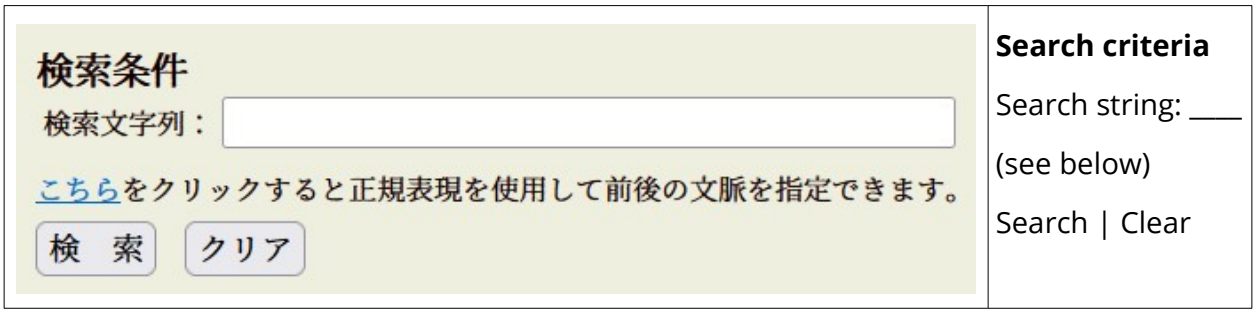
2. Searching the corpus

(3) Overview — please read (4)–(6) for details.

(a) Type or paste your search term in the text box. This must be an exact-string match (no options or alternatives are available — do a separate search for each distinct string you want to find).

(b) You can optionally specify the preceding or following context for your search term. This function allows you to specify alternatives at the single-character level! Use the format ^[ABC], with or without spaces, to mean "A or B or C". (This may be most useful for finding conjugated forms of verbs.)

(4) At the top of the search page is a text box. Type or paste your search term here. Note that the text you type will be matched exactly: 子供 <kodomo> 'child' and こども <kodomo> 'child' are two different searches.



(5) The third line in the view above says:

こちらをクリックすると正規表現を使用して前後の文脈を指定できます。

Click here to specify preceding and/or following contexts, using regular expressions.

- If you click, you see the following version of the search interface:

- Here, lines 3 and 4 say:

  前文脈：　Preceding context

  後文脈：　Following context

(6) The link to the right side in the above view says:
前後文脈の指定について About specifying the preceding and following context

- If you click, you go to a screen with the following information. (Translations paraphrased; please see the actual web site for the search examples referred to in the chart below.)

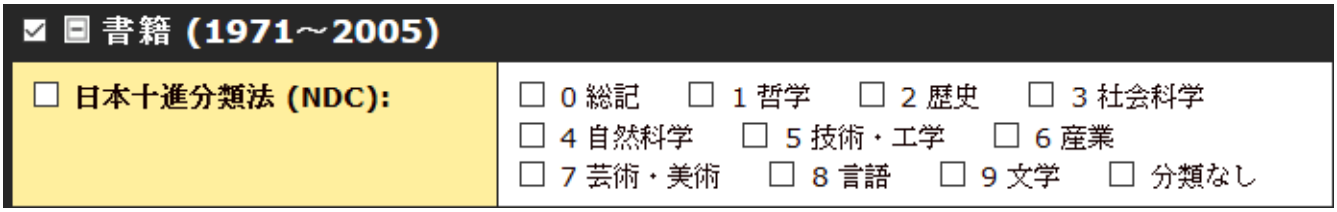
前後文脈の指定について	**Specifying preceding/following context**
前文脈および後文脈では、コーパス内において検索文字列の前後に現れる文字列を検索条件として指定できます。 例えば、以下に示すように、検索文字列に「銀行」、前文脈に「日本」を指定して検索すると、「銀行」を含む文のうち、前に「日本」が現われている文のみを検索結果として表示します。	You can specify text strings that appear before and after the search string as search criteria. For example, as shown below*, you can specify the preceding context「日本」<nihon> 'Japan' for the search string「銀行」<ginkoo> 'bank'. This will return only those sentences containing「銀行」<ginkoo> preceded by「日本」<nihon>. (*see actual web site for search examples)
また、前文脈および後文脈は、正規表現を使用して指定できます。以下に示す検索条件を指定して検索すると、「取」を含む文のうち、直後に「ら」から「っ」のいずれかが現われている文のみを検索結果として表示します。	Preceding and following contexts can be specified using regular expressions. If you search using the search criteria shown below*, this will return sentences containing「取」<to[r]> 'take' in which the following character is one of [らりるれろっ] <ra ri ru re ro *small*-tu>.

3. Restricting your search by genre or time period

(7) The first set of black bars below the search box allows you to restrict your search to particular genres of text. Select/deselect the relevant options. To include all subcategories of a particular genre, check the box in the left corner of the corresponding black bar.

メディア／ジャンル Media/genre

- Buttons say 全てのチェックを外す | 全てにチェックを入れる
  Deselect all | Select all
- Click on the plus (+) symbol to expand the subcategories (as shown below)
- Screenshots of each category, with translations provided except where noted:



Books (1971–2005)	
Nippon Decimal System (NDC):	0 General 1 Philosophy 2 History 3 Sociology 4 Natural sciences 5 Tech/engineering 6 Industry 7 Arts 8 Language 9 Literature Unclassified

☑ ⊟ 雑誌 (2001～2005)	
☐ 総合:	☐ 総記／マスコミ　☐ 一般　☐ 家庭／生活　☐ 児童 ☐ 娯楽／芸能　☐ レジャー／趣味　☐ スポーツ
☐ 教育・学芸:	☐ 教育　☐ 学習／語学　☐ 文学／芸術　☐ 社会科学 ☐ 自然科学　☐ 人文科学
☐ 政治・経済・商業:	☐ 政治／外交　☐ 経済／経営　☐ 金融／財政 ☐ 商業／消費者　☐ 国勢／民力
☐ 産業:	☐ 農林水産　☐ 運輸／通信
☐ 工業:	☐ 機械　☐ 電気機／電子
☐ 厚生・医療:	☐ 厚生　☐ 医学

Magazines/journals (2001–2005)	
Comprehensive:	General/mass comm.　General　Home/lifestyle　Juvenile Entertainment/arts　Leisure/hobby　Sports
Education/Liberal arts:	Education　Language　Literature/arts　Social sciences Natural sciences　Human sciences
Politics/Economics/Business:	Politics/diplomacy　Economics/management　Finance Business/consumers　National resources
Industry:	Agriculture/forestry/fisheries　Transportation/communication
Manufacturing:	Machines　Electronics
Public welfare/Medicine:	Public welfare　Medicine

☑ ⊟ 新聞 (2001～2005)	
☐ 全国紙:	☐ 朝日新聞　☐ 読売新聞　☐ 毎日新聞　☐ 産経新聞
☐ ブロック紙:	☐ 北海道新聞　☐ 中日新聞　☐ 西日本新聞
☐ 地方紙:	☐ 河北新報　☐ 新潟日報　☐ 京都新聞　☐ 神戸新聞 ☐ 中国新聞　☐ 高知新聞　☐ 琉球新報

Newspapers (2001–2005)	
National:	Asahi　Yomiuri　Mainiti　Nikkei
Blogs (regional online news):	Hokkaidoo　Chuuniti　Nisi-Nihon
Regional:	Kahoku　Niigata　Kyooto　Koobe Tyuugoku　Kooti　Ryuukyuu

☑ ⊟ 白書 (1976～2005)

White papers (1976–2005)

'White papers' are official (typically government) research and policy reports. There are many subcategories listed here; please let me know if you need the subcategories translated.

☑ ⊟ 教科書 (2005～2007)

☐ 国語:	☐ 小 ☐ 中 ☐ 高
☐ 数学:	☐ 小 ☐ 中 ☐ 高
☐ 理科:	☐ 小 ☐ 中 ☐ 高
☐ 社会:	☐ 小 ☐ 中 ☐ 高
☐ 外国語:	☐ 中 ☐ 高
☐ 技術家庭:	☐ 小 ☐ 中 ☐ 高
☐ 芸術:	☐ 小 ☐ 中 ☐ 高
☐ 保健体育:	☐ 高
☐ 情報:	☐ 高
☐ 生活:	☐ 小

Textbooks (2005–2007)

Japanese language:	Elementary Middle school High school
Mathematics:	Elementary Middle school High school
Science:	Elementary Middle school High school
Society:	Elementary Middle school High school
Foreign language:	Middle school High school
Home economics:	Elementary Middle school High school
Fine arts:	Elementary Middle school High school
Health/physical education:	High school
Information:	High school
Daily life:	Elementary

☑ ⊟ 広報紙 (2008)

[Local government] newsletters (2005–2007)

The subcategories here are regions in Japan; please let me know if you need translations.

☑ ⊟ Yahoo!知恵袋 (2005)	**Yahoo! answers (2005)** (Please let me know if you need sub-subcategories translated.)
☐ エンターテインメントと趣味:	Entertainment and hobbies
☐ インターネット、PCと家電:	Internet/PCs/Home electronics
☐ ビジネス、経済とお金:	Business, economics, and money
☐ 職業とキャリア:	Employment and careers
☐ ニュース、政治、国際情勢:	News, politics, international
☐ スポーツ、アウトドア、車:	Sports, outdoors, autos
☐ 暮らしと生活ガイド:	Lifestyle guides
☐ 健康、美容とファッション:	Health, beauty, and fashion
☐ 子育てと学校:	Parenting and education
☐ マナー、冠婚葬祭:	Etiquette, ceremonial occasions
☐ 教養と学問、サイエンス:	Culture, education, science
☐ 地域、旅行、お出かけ:	Geography, travel, excursions
☐ Yahoo! JAPAN:	Yahoo! JAPAN
☐ その他:	Other

☑ ⊟ Yahoo!ブログ (2008)	**Yahoo! blogs (2008)** (Please let me know if you need sub-subcategories translated.)
☐ ビジネスと経済:	Business and economics
☐ コンピュータとインターネット:	Computers and internet
☐ 生活と文化:	Lifestyle and culture
☐ エンターテインメント:	Entertainment
☐ 家庭と住まい:	Family and home
☐ 政治:	Politics
☐ 健康と医学:	Health and medicine
☐ 学校と教育:	School and education
☐ 科学:	Science
☐ 出会い:	Dating/marriage
☐ 地域:	Regional
☐ 特集:	Specialized — *single subcategory is* Hobbies and sports
☐ 芸術と人文:	Arts and culture
☐ Yahoo!サービス:	Yahoo! service
☐ 趣味とスポーツ:	Hobbies, sports — *subcategories are* gambling, sports, leisure, hobbies, vehicles

☑ ⊟ 韻文 (1980～2005)	**Poetry (1980–2005)** (These subcategories have no further sub-subcategories.)
☐ 短歌:	Tanka
☐ 俳句:	Haiku
☐ 詩:	Poems

☑ ⊟ 法律 (1976～2005)
Laws (1976–2005)
There are many subcategories listed here; please let me know if you need the subcategories translated.

☑ ⊟ 国会会議録 (1976～2005)	
☐ 衆議院:	☐ 本会議 ☐ 常任委員会 ☐ 特別委員会 ☐ その他
☐ 参議院:	☐ 本会議 ☐ 常任委員会 ☐ 特別委員会 ☐ その他

Records of Diet [national legislature] proceedings (1976–2005)	
Lower house:	Regular session Standing committees Special committees Other
Upper house:	Regular session Standing committees Special committees Other

(8) The final black box at the bottom of the search page allows you to restrict your search by time period.

- You can select entire decades (年代) or individual years
- To search all years, check the box in the black bar (全期間 = 'all time periods')

☑ ⊟ 全期間	
☐ 1970年代:	☐ 1971 ☐ 1972 ☐ 1973 ☐ 1974 ☐ 1975 ☐ 1976 ☐ 1977 ☐ 1978 ☐ 1979
☐ 1980年代:	☐ 1980 ☐ 1981 ☐ 1982 ☐ 1983 ☐ 1984 ☐ 1985 ☐ 1986 ☐ 1987 ☐ 1988 ☐ 1989
☐ 1990年代:	☐ 1990 ☐ 1991 ☐ 1992 ☐ 1993 ☐ 1994 ☐ 1995 ☐ 1996 ☐ 1997 ☐ 1998 ☐ 1999
☐ 2000年代:	☐ 2000 ☐ 2001 ☐ 2002 ☐ 2003 ☐ 2004 ☐ 2005 ☐ 2006 ☐ 2007 ☐ 2008

(self-explanatory!)

Last update: September 28, 2023