

Mini Demonstration of Japanese Corpora

Toshinobu OGISO

Department of Corpus Studies, NINJAL
togiso@ninjal.ac.jp



Center for Corpus Development, NINJAL

Corpus:

"a collection of written or spoken material stored on a computer and used to find out how language is used"

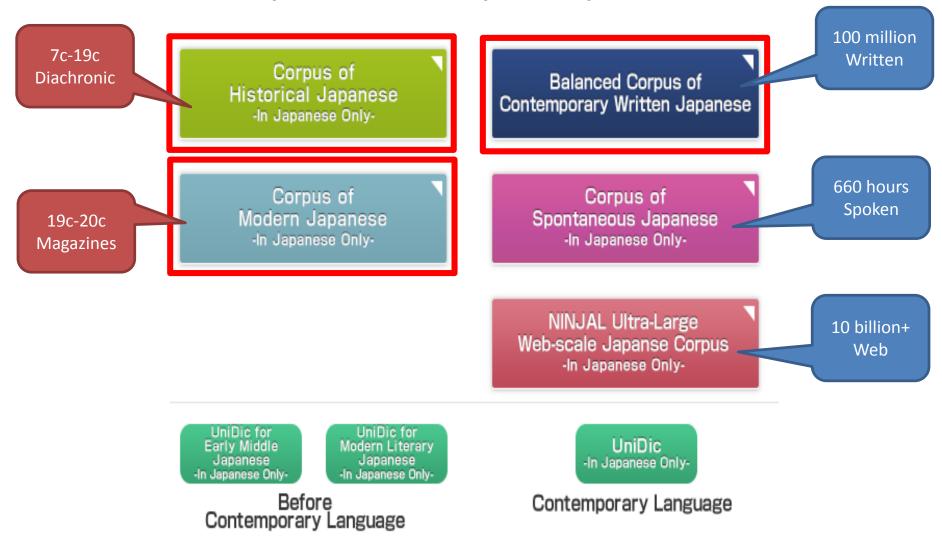
Center for Corpus Development
 http://www.ninjal.ac.jp/corpus center/en/







Corpora developed by NINJAL





1. Balanced Corpus of Contemporary Written Japanese (BCCWJ)

Japan's first balanced corpus of about 100 million words

PUBLICATION SUBCORPUS

Books, magazines, and newspapers published between 2001-2005 35 million words

LIBRARY SUBCORPUS

Books catalogued at more than 13 public libraries in the Tokyo area and published after 1986
30 million words

SPECIAL PURPOSE SUBCORPUS

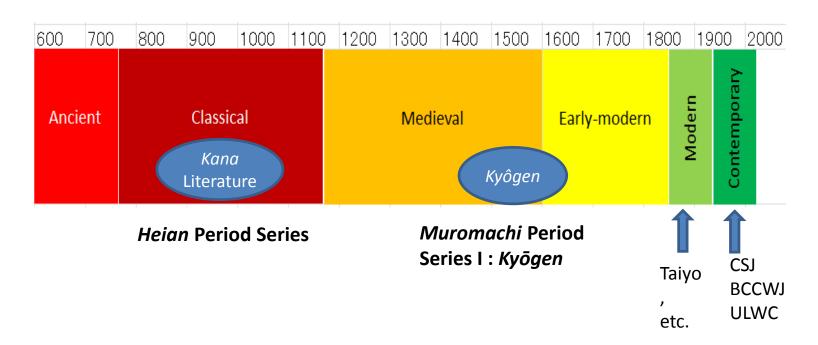
Whitepaper texts, internet texts, diet minutes, best-selling books, etc. 1975-2005

35 million words



2. Corpus of Historical Japanese (CHJ)

- ・『日本語歴史コーパス』
- Diachronic corpus from the 8th century to the 20th century
- Under development, but partially released



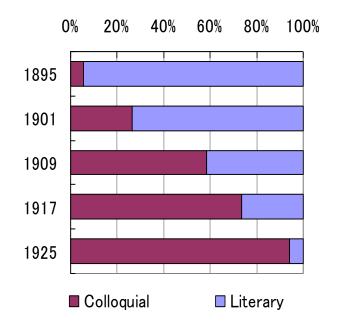


3. *Taiyo* Corpus

- Full-text corpus of the *Taiyo* magazine, published between 1895-1925
- When the grammar of written Japanese shifted from the literary to colloquial system. (言文一致 movement)
- About 9 million words









User Interfaces for NINJAL Corpora

I. Shonagon 少納言

-Online free access without registration

II. Chunagon 中納言

—Online free access with registration

III.Himawari ひまわり

Offline stand-alone Java application



I. Shonagon

- Only for BCCWJ
- Only string query is possible
- http://www.kotonoha.gr.jp/shonagon/







Shonagon for BCCWJ

- Simple and easy to use
 - what you type is what you get



Type your target strings

Click and search

Select registers to search if needed



Shonagon for BCCWJ



Keyword in context



Biographical Info of the author

Bibliographical Info of the source

Shonagon for BCCWJ

- Up to 500 samples are shown. No download.
 - Search Result:3859 results found. Showing 500 results.





II. Chunagon

- for BCCWJ, CHJ, CSJ (coming soon), Taiyo (planned)
- Enables query of morphological information
- http://chunagon.ninjal.ac.jp







Chunagon Searching by Part-of-Speech

- Using morphological information, users can retrieve all examples of the target part-of-speech.
- 品詞(part-of-speech) = "形容詞" (adjective) in CHJ-Heian

短単位検索	•	*
▼ 前方共起条件の追加	*	*
‡─ (10	▼ 語 ▼)	■キーを未指定
品詞 ▼の大分	類 🔻 が 形容詞	■ 短単位の条件の追加
▲ 後方共起条件の追加	*	٠
	*	*



Chunagon Searching by Lemma

- Using morphological information, users can retrieve all examples of the target word, including all conjugation forms.
 - 「読む」→ 「読ま」「読み」「読む」「読め」





Chunagon Searching by Lemma

- Using morphological information, users can retrieve all examples of the target word.
- 語彙素 (lemma) ="玉葱" (onion) in BCCWJ

	タマネギ	446
	たまねぎ	345
Variety of orthography	玉ネギ	94
or enegrapiny	玉ねぎ	1026
	玉葱	176
	Total	2087



Chunagon Searching by Complex Morphological Info

- Search by multiple words (max. 10 words)
 - -ex. verbs in front of the auxiliary verb "ぬ"
 - -ex. adjectives of adnominal form in front of the noun "女"

▼ 前方共起条件の追加	
≑ - (▼ 10 ▼ 語 ▼)	
品詞 ▼ の 大分類 ▼ が 形容詞 ▼ ■ AND 活用形 ▼ の 大分類 ▼ が 連体形 ▼ ■ ■ 短単位の条件の追加	
後方共起1 (キーから ▼ 2 ▼ 語以内 ▼)	
語彙素 ▼ が 女 ■ <u>□ 短単位の条件の追加</u>	
(▲ 後方共起条件の追加)	検索



Chunagon Search Results

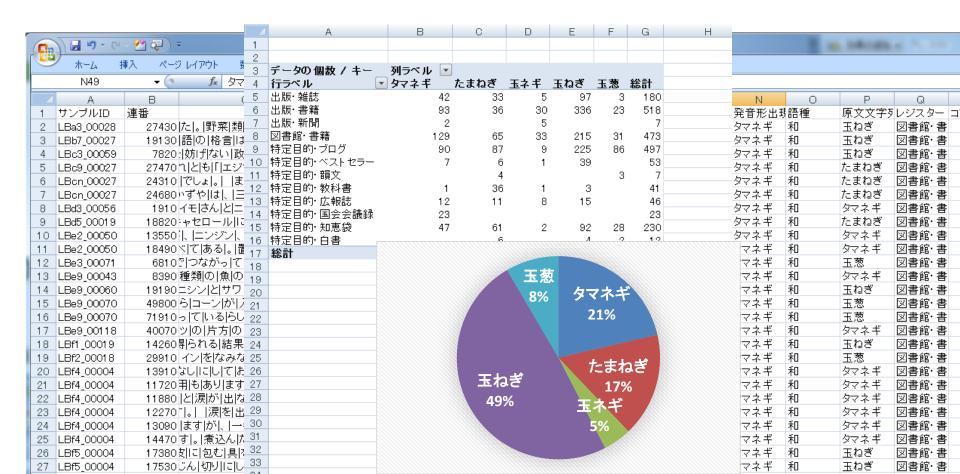
 Keyword in context, morphological info, bibliographical info, Biographical info

【列の表示】設定を隠す

□ コーバス† □ サブコーバ	青報 ス名 ☑ サンブルID	■連種	番																
	報 ☑ キー ☑ 後文脈 『 見形 語種 原ご			□語彙	素細分	类 頁 ▽	語形	☑ 品;	a ✓	活用型	<u>!</u> •;	舌用形	■書	字形	□仮	名形出:	現形		
□ 本文情報☑ 本文種別	☑話者 □本文属性	ŧ	□ 作品情報☑ ジャンル☑ 作品	品名 🖫	☑成立年	年 ☑	巻名等	F	巻順		』作者 作者		年	性別					
□ 底本情報	ベージ番号 🔲 校注者	. <u> </u>	出版社																
32751 件の結果	が見つかりました。その	らち 5 (00 件を表示しています													テーブ	ルの幅	を固定	短🔻
				語								_							ペ
サンブルID ¢	前文脈 ◆	‡ _ ‡	後文脈 幸	乗 素 ≑ 読 み	語 乗 ‡ 素	語 ◆ 形	品章	活 用 ≑ 型	用幸	本 文 種別	話 幸	ジャンル	作 品 ≑ 名	成 立 ≑ 年	巻 名 \$ 等	作 者 [‡]	生 年 [‡]	本	ー ジ ⇒ 番 号
22_源氏物語 03_033_藤裏葉	と 思ひなり たまひ で 、 「 こ の をり に 添へ た てまつり たまへ 。 まだ	_ ◆ うしろ めた き	に、 さぶらふ 人 とて も、 若々しき のみに	素 み ウシ ロメタ	乗 幸 素 後ろ	ウシ ロメタ	形容 詞-一	文語 形容	用幸	文種別	紫の	ャキンル	作品名 源物語	成 立 年	藤裏	& 寸		新編	番

Chunagon Search Results

- Up to 10,000 samples can be downloaded.
- Users can analyze downloaded data using pivot tables in Excel.



Chunagon (for CHJ) Linking to the Original Page

 Search results of CHJ will be linked to the original pages of Shimpen-Zenshū in the Japan Knowledge webservice.

3等 \$	巻♠順	作 者	生 年	性 別	底 💠	ページ番号	校注者 ◆	出 版 ◆ 社
	2	紫式 部	978	女	新編 全集 <20>	105	阿部秋生/秋山虔/ 今井源衛/鈴木日出 男	小学館
	35	紫式 部	978	女	新編 全集 <23>	225	阿部秋生/秋山虔/ 今井源衛/鈴木日出 男	小学 館
		紀 貫 之	865	男	新編 全集 <13>	19	菊地靖彦	小学 館
	14	紫式 部	978	女	新編 全集 <21>	309	阿部秋生/秋山虔/ 今井源衛/鈴木日出 男	小学 館
	9	炸式 部	978	女	新編 全集 <21>	17	阿部秋生/秋山虔/ 今井源衛/鈴木日出 男	小学 館
	7	紫式 部	978	女	新編 全集 <20>	319	阿部秋生/秋山虔/ 今井源衛/鈴木日出 男	小学館
								$\overline{}$

III. Himawari

- A full text retrieval system for *Taiyo*, *Meiroku-Zasshi*, and *Sharebon*, etc.
 - + Aozora-bunko and other free materials
- Developed with JAVA, and works on Windows, MacOS and Linux







*Himawari*Data Packages

- Users can download the program of Himawari and data packages for Himawari from the web
 - Corpus of modern women's magazines, Meiroku-Zasshi, Kokumin-no-Tomo (from NINJAL)
 - + Aozora Bunko, Wikipedia, Minutes of the National Diet, etc.

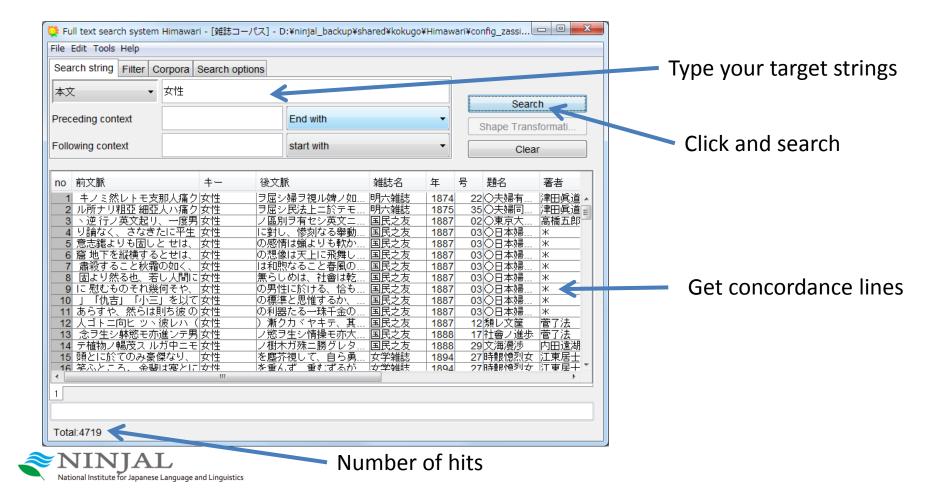


- 『ひまわり』
 - 。 ver.1.5 (2014-12-20) ... 安定版
 - Windows, Linux, Mac OS X 上で動作します。
 - ver.1.3 と比較して、テキストのインボート、形態素解析結果などのアノテーションの機能が強化されています。 詳しくは、リリースノート、履歴をご覧ください。
 - 。 ver.1.3.3 (2014-10-07) ... 旧安定版
 - Windows, Linux, Mac OS X 上で動作します。
 - リリースノート, 履歴
- 言語資料
 - 。『国会会議録』バッケージ(2014-04-01)
 - 。『Wikipedia』パッケージ(2013-11-12)
 - 。『青空文庫』バッケージ(2014-10-14)
 - 。「日本文学テキスト」バッケージ(2007-04-23)
 - 。「人情本」バッケージ(2007-04-23)
 - 。『ふみくら』バッケージ(2014-10-15)
 - 。『日本語話し言葉コーバス』サンブルデータ (2005-01-18)



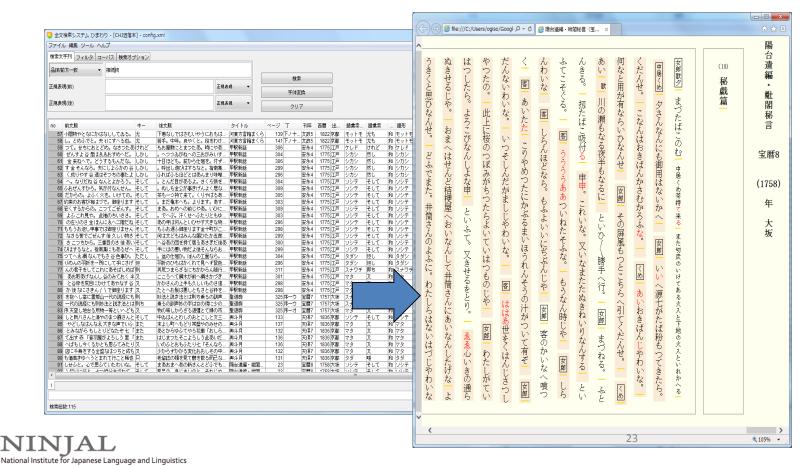
Himawari How to Search

- Has an English user interface (but contents are in Japanese)
- Easy to use for researchers who are unfamiliar with computers



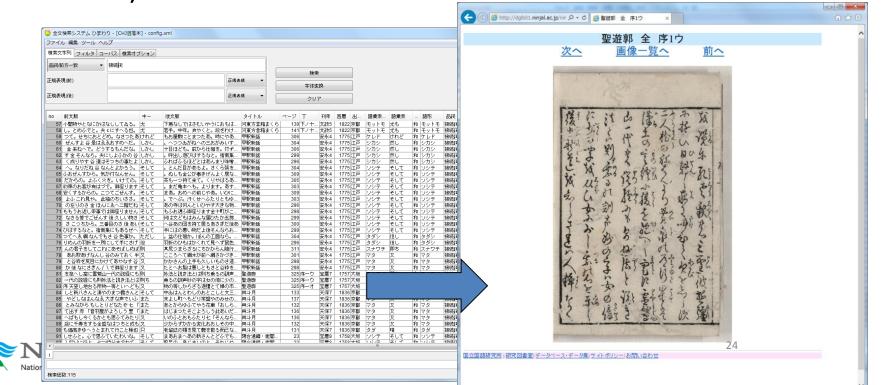
Himawari Browsing the Text in HTML

 By clicking on the concordance line, you can browse the text with rich annotation



Himawari Viewing the Image of the Page

 Each concordance line and HTML document is linked to the original photo image on the web (only Meiroku-Zasshi and Sharebon corpora are currently available).



Registration for Chunagon

- *Chunagon* is the most powerful tool to use NINJAL Corpora.
- Chunagon can be used free of charge, but user registration is required.
- For now, users must contract with NINJAL via postal mail

(this will be improved).

Please contact me!
 e-mail: togiso@ninjal.ac.jp



Thank you for your attention.

