

 *Corpus Linguistics and Japanese Language*
コーパス言語学と日本語

Takehiko Maruyama
丸山 岳彦

National Institute for Japanese Language and Linguistics /
University of Oxford



18 November 2014
SEMINÁŘ JAPONSKÝCH STUDIÍ
Masarykova Univerzita







 NINJAL 


- National Institute of Japanese Language and Linguistics (“NINJAL”) 国立国語研究所
 - Established in 1948
 - Scientific surveys of Japanese language
 - Creation of Japanese corpora




 Contents 
(10:50 – 11:50)

- Introduction : Japan and Japanese Language
- Japanese Corpus: History
 - What is a “Corpus” ?
 - History of Japanese Corpus
- Japanese Corpus: Present situation
 - Spoken Corpus : *CSJ*  日本語話し言葉コーパス
Corpus of Spontaneous Japanese
 - Written Corpus : *BCCWJ*  少納言 中納言

 Introduction:
Japan and
Japanese Language 



 Where is Japan ? 



 Japan / *Nihon* 日本 



Dialects in Japan

- Dialect surveys by NINJAL since 1966

Fukushima pref. 1949
 Okinawa pref. 1978
 Tottori pref. 1984
 Iwate pref. 1980
 Hachijo island 1949

Dialects in Japan

- Linguistic Atlas of Japan (NINJAL, 1966)

Japanese Writing System

- Three types of characters
 - Kanji 教科書 玉子
 - Hiragana ほん たまご
 - Katakana テキスト タマゴ
- Other types of characters
 - Punctuation, mark 、。？！「」（／＆
 - Alphabet NINJAL
 - Arabic numeral 1,234
 - Roman numeral I IV XIII

Japanese Corpus : History

- What is a “Corpus” ?
- History of Japanese Corpus

What is a “Corpus” ?

- A “corpus” is...
 - an collection of language in “real world”.
 - “a collection of texts assumed to be **representative** of a given language, dialect, or other subset of a language, to be used for linguistic analysis”. (Francis 1982)
- Various corpora
 - Text (written) corpus, Speech (spoken) corpus, Historical corpus, Learner corpus, Dialect corpus...
- “Corpus linguistics” is...
 - a **methodology** of linguistic study using corpora

Various corpora in the world


- Corpus collection/creation started in 1960s

1959		<i>The Survey of English Usage</i> (1 million words)
1964		<i>Brown Corpus</i> (1 million wds)
1991		<i>Bank of English (BOE)</i> (500 million wds)
1994		<i>British National Corpus (BNC)</i> (100 million wds)
2000		<i>Czech National Corpus (CNC)</i> (100 million wds)
2004		<i>Corpus of Spontaneous Japanese (CSJ)</i> (7.5 million wds)
2011		<i>Balanced Corpus of Contemporary Written Japanese (BCCWJ)</i> (100 million wds)


Where is the origin of Japanese corpus?

History of Japanese Corpus 1

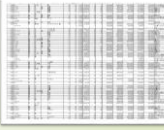
- Surveys of daily vocabulary at NINJAL
 - 1953 *Research on vocabulary in women's magazines*
 - 1957-1958 *Research in vocabulary in cultural reviews*
 - 1962-1964 *Vocabulary and Chinese characters in ninety magazines of today (I, II, III)* **0.5 million words**



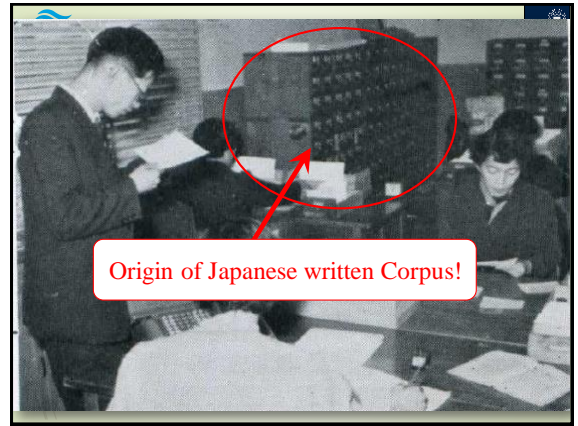
Real text



Sampling




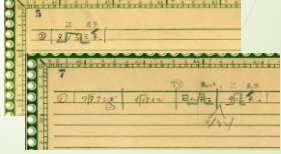
Vocabulary list

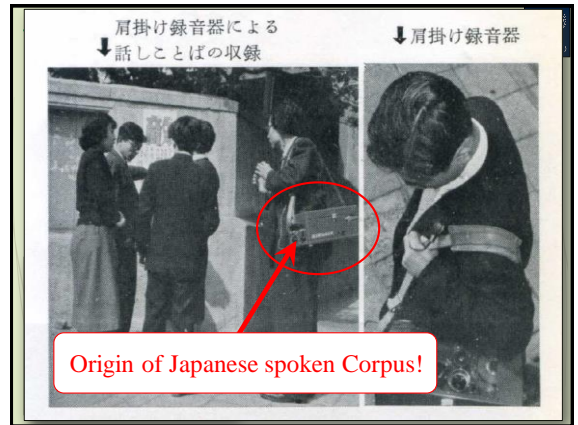


History of Japanese Corpus 2

- Surveys of colloquial speech at NINJAL
 - 1955 *Research in the colloquial Japanese*
 - 30 hours** of colloquial speech were recorded, **83,620 words**
 - 1960, 1963 *A research for making sentence patterns in colloquial Japanese (1: dialog) (2: monolog)*




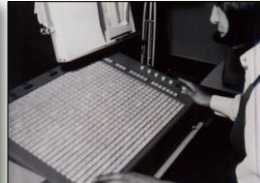


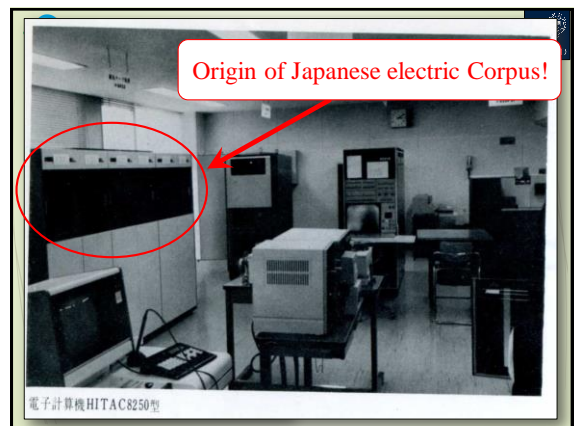


History of Japanese Corpus 3

- Vocabulary surveys using computers
 - 1970-1973 *Studies on the vocabulary of modern newspapers 1-4*
 - 2 million words** from three major newspapers in 1966







Japanese Corpora in 2000s

- NINJAL started creating large sized corpora
 - *Corpus of Spontaneous Japanese (CSJ)* - 2004
 - 651 hours, 752 million words of spontaneous speech
 - *Balanced Corpus of Contemporary Written Japanese (BCCWJ)* - 2011
 - 100 million words of various written text (well balanced)
 - *Corpus of Historical Japanese (CHJ)* - 2013~
 - 14 literary works with 0.79 million words in Heian period
 - *Ultra Large-sized Corpus (ULC)* - under construction
 - 10 billion words of Japanese text extracted from web

Japanese Corpus :

Present situation

- Spoken Corpus : *CSJ*
- Written Corpus : *BCCWJ*

Keywords

Knowledge and Behavior

Znalosti a chování
知識と行動

CSJ :

Corpus of Spontaneous Japanese

『日本語話し言葉コーパス』

日本語話し言葉コーパス
Corpus of Spontaneous Japanese

Question 1

Which one is a correct spell of "communication" in Japanese?

1. コミニケーション
2. コミュニケーション
3. コミニュケーション
4. コミュニケーション


Variable forms in speech

Question 2

How do you read this word?


じてんしゃ
自転車
ji ten sya


Yes!



Question 2-2

How do Japanese people pronounce the word “自転車” in real life?

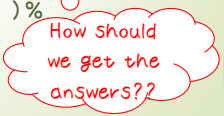
じでんしゃ
自 転 車
 ji den sya  !?



Question 3


Guess the percentages of each pronunciation in real Japanese.

- コミュニケーション ()%
- コミューケーション ()%
- コミニュケーション ()%
- コミュニケーション ()%
- じてんしゃ ()%
- じでんしゃ ()%

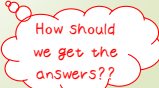
 How should we get the answers??

Question 4


When you hesitate while speaking, you might use **FPS** (filled pauses).


hm... er... uh... 

What type of FP do you use most frequently in your daily Czech?
 How about in Japanese?

 How should we get the answers??

How should we get the answers?

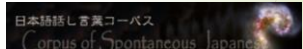
- Think it in your head ! (intuition) 
 - ▶ Your answer may be wrong !!
 - ▶ Who guarantee your answer ?
- Ask the speech corpus ! (survey)
 - ▶ Everyone can get the same answer !
 - ▶ Of course you need a reliable corpus.




**We have knowledge about (at least) a language.
 But we don't know how we behave with it.**

CSJ

- Corpus of Spontaneous Japanese (2004)
 - ▶ Japanese spontaneous speech (mainly monolog)
 - ▶ 651 hours, 7.52 million words
 - ▶ 3,302 lectures by 1,418 different speakers
 - ▶ Rich annotations
 - ▶ 18 DVDs
- Aims
 - ▶ Automatic Speech Recognition (ASR) system
 - ▶ Linguistic study of spontaneous speech





The screenshot shows a software interface for speech analysis. At the top, there's a waveform and spectrogram. Below that, a list of time-coded segments is visible, such as 0120 00305.342-00308.315 L. The interface also includes a small video window on the left showing a person speaking.

Two Ways of Transcription

Basic Form & Pronounced Form

FP

Fragment

Repair

Variable pronunciation

Elongation

0020 00041.527-00043.511 L:	& (W タシ;ワタシ)
私 (F あの)	& (F アフ)
食べるの	& タベルノ
遅く	& スゴク
好きで	& スキデ (KH)
0035 00071.128-00074.357 L:	& ショクヨー (D コン(L トゥ))
食用 (D こんどう)	& キョージュニ
教授に	& ソーダンシタラ (KH)
相談したら	& ショクヨーコン
食用昆虫は	& チョー
0045 00092.141-00092.843 L:	& ソレ (D2 ニ) (KH)
それ (D2 ニ)	& ノ
0047 00093.570-00094.625 L:	& ケンイノ
の	& センセーデ
権威の	
先生で	

Answer to "Communication"

- How do Japanese pronounce "communication" ?
- Corpus : CSJ, 651 hours, 7.52 million words
- Frequency of the word "communication" : **601 times**

コミュニケーション	296
コミュニケーション	136
コミュニケーション	123
コミュニケーション	36
misc	10
Total	601

Answer to "自転車"

- How do Japanese pronounce 自転車 ?
- Corpus : CSJ, 651 hours, 7.52 million words
- Frequency of the word 自転車 : **483 times**

ジテンシャ	349
ジテンシャ	116
misc	18
Total	483

Answer to Filled Pauses (JP)

- What FP do Japanese use most frequently?
- Corpus : CSJ, 651 hours, 7.52 million words
- Frequency of Filled Pauses : **430,472 times**

(F えー) (F e:)	116,772	27.1%
(F え) (F e)	45,665	10.6%
(F ま) (F ma)	44,549	10.4%
(F あのー) (F ano:)	40,695	9.5%
(F あの) (F ano)	33,330	7.7%

(top 5)

Answer to the Filled Pauses (JP)

えー e: (Male)

あのー ano: (Female)

	Male	Female
(F e:)	95,359 30.2%	21,413 18.7%
(F e)	36,078 11.4%	19,393 16.9%
(F ma)	34,643 11.0%	15,954 13.9%
(F ma:)	24,369 7.7%	9,906 8.6%
(F ano:)	21,302 6.8%	9,587 8.4%

Answer to the Filled Pauses (CZ)

- What FP do Czech use most frequently?

Ask CNC by yourself !!

... and tell me the result ☺

Annotations to speech signals

- Two-way Transcription
- Segment Labels
- Intonation Labels
- Morphological Analysis
- Clause Boundary Labels
- Dependency Structure
- Discourse Structure
- Impression Rating
- Speaker Info

→ Phonetics / Phonology
→ Morphology / Lexicon
→ Syntax
→ Discourse analysis
→ Metadata / bibliography

Morphological Analysis

All transcriptions were segmented into words (manually/automatically) with rich information

行	A	B	C	D	E	F	G	H	I	J
番号	発音ID	所属ID	発音ID	時間情報	出辞	発音形	品名	活用型	活用形	品詞情報
2	474496	SO2M001	005	0000215-00002207	(Fスー)	(Fエー)	感動詞			
3	474497	SO2M001	005	0000215-00002207	チヤ	チヤ	名詞			
4	474498	SO2M001	005	0000215-00002207	は	ワ	助詞			係助詞
5	474499	SO2M001	005	0000140-00002012	(Fオ)	(Fア)	感動詞			
6	474500	SO2M001	005	0000140-00002012	タトル	タトル	名詞			
7	474501	SO2M001	005	0000140-00002012	は	ワ	助詞			係助詞
8	474502	SO2M001	005	0000213-00004612	絶望的	ゼツボウテキ	名詞			
9	474503	SO2M001	005	0000213-00004612	な	ナ	助動詞			連体形
10	474504	SO2M001	005	0000213-00004612	沖縄行き航空	オキナワイキョウカイ	名詞			
11	474505	SO2M001	004	00005317-00005591	です	デス	助動詞			
12	474506	SO2M001	005	00007211-00011038	(Fスーとでずほ)	(Fエートデスホ)	感動詞			終止形
13	474507	SO2M001	005	00007211-00011038	僕	ボク	代名詞			
14	474508	SO2M001	005	00007211-00011038	は	ワ	助詞			係助詞
15	474509	SO2M001	005	00007211-00011038	(Fスー)	(Fエー)	感動詞			
16	474510	SO2M001	005	00007211-00011038	旗	ハタ	名詞			
17	474511	SO2M001	005	00007211-00011038	を	オ	助詞			係助詞
18	474512	SO2M001	005	00007211-00011038	する	スル	動詞	ワ行変格		連体形
19	474513	SO2M001	005	00007211-00011038	の	ノ	助詞			係助詞
20	474514	SO2M001	005	00007211-00011038	が(の)	ガ(ノ)	助詞			係助詞
21	474515	SO2M001	005	00007211-00011038	好き	スキ	名詞			
22	474516	SO2M001	005	00007211-00011038	です	デス(ロ)	助動詞			終止形

XML Encoding

Various annotations were encoded into XML file.

```

<Open xmlns="http://www.csj.or.jp/voice/xml/1.0" soundings="01-01">
  <Talk TalkID="25F1600" SpeakerID="514" SpeakerBirthPlace="神奈川県" SpeakerBirthGeneration="70s74" SpeakerSex="女">
    <Text>
      <Comment CommentStrings="舞臺ID:25F1600"/>
      <Comment CommentStrings=""/>
      <Comment CommentStrings=""/>
      <IPU [PUID="060"] [PUStartTime="0000.313" [PUEndTime="0001.103" Channel="L"]>
        <LUM [LUID="2" LUMPOS="名詞" [LUMLabelTime="0001.001" LUMLabelForm="チヤ" LUMLemma="チヤ" LUMSense="チヤ"]>
          <LUMLabelWord Time="0.615750" PerceInedocPos="0" [F_Wh]/>
          <LUMLabelWord Time="0.615750" PerceInedocPos="0" [F_Wh]/>
          <LUMLabelBreak Time="0.615750" F_/>
        </LUM>
        <Phoneme [PhonemeID="14" Phoneme="チヤ" PhonemeClass="vowel" PhonemeStartTime="0.35341" PhonemeEndTime="0.48944" EndInSyllable="1"/>
        </Phoneme>
        <Mora [MoraID="22" Mora="チヤ"]>
          <Phoneme [PhonemeID="14" Phoneme="チヤ" PhonemeClass="special" PhonemeStartTime="0.48944" PhonemeEndTime="0.615750" EndInSyllable="1"/>
          <MoraLabelWord Time="0.48944" FWH="07.878" ToneClass="F111" F1/F2/F3/LabelTime>
            <MoraLabelWord Time="0.615750" PerceInedocPos="0" [F_Wh]/>
            <MoraLabelWord Time="0.615750" F_/>
          </MoraLabelWord>
        </MoraLabelWord>
        </Mora>
      </LUM>
    </Text>
  </Talk>
</Open>
  
```

Concordancer "Himawari"

no	前次語	出現形	語彙表	品名	活用型	代表表記	代表形
1	感傷を覚えます	チヤ	おつれさし	感動詞		チヤ	チヤ
2	った	チヤ	おつれさし	感動詞		チヤ	チヤ
3	ハルガリ	チヤ	おつれさし	感動詞		チヤ	チヤ
4	と	チヤ	おつれさし	感動詞		チヤ	チヤ
5	ドライブ	チヤ	おつれさし	感動詞		チヤ	チヤ
6	行き	チヤ	おつれさし	感動詞		チヤ	チヤ
7	な	チヤ	おつれさし	感動詞		チヤ	チヤ
8	か	チヤ	おつれさし	感動詞		チヤ	チヤ
9	そう	チヤ	おつれさし	感動詞		チヤ	チヤ
10	い	チヤ	おつれさし	感動詞		チヤ	チヤ
11	い	チヤ	おつれさし	感動詞		チヤ	チヤ
12	の	チヤ	おつれさし	感動詞		チヤ	チヤ
13	F	チヤ	おつれさし	感動詞		チヤ	チヤ

What CSJ offers

- Variations in spontaneous speech
 - Pronunciation, Accent, Intonation, Grammar...
- Disfluency in spontaneous speech
 - FP, Word Fragments, Elongation, Self-repair...

Resource to analyze behaviors in spontaneous JP
Linguistic knowledge never tells us our behavior !

Future work: to create a large dialog corpus

BCCWJ : Balanced Corpus of Contemporary Written Japanese

『現代日本語書き言葉均衡コーパス』

少納言 中納言 KOTONOHANA
コーパス検索アプリケーション
KOTONOHANA『現代日本語書き言葉均衡コーパス』

BCCWJ

- Contents: balanced corpus for general purpose
- Corpus Size: 100 million words
- Period: 1976 - 2005 (-2009)
- Media: Books, Magazines, Newspapers, Whitepapers, Textbooks, Web Documents, Law, Verse, Diet minutes...
- Method: Stratified random sampling
- Aim: Vocabulary survey, Grammatical study, Lexicography, Natural language processing...

Structure of BCCWJ

Publication sub-corpus Books, Magazines, Newspapers 35 million words 2001-2005	Library sub-corpus Books stored in many public libraries 30 million words 1986-2005
Special-purpose sub-corpus Whitepapers, Textbooks, Public Relation, Best-Seller book, Web documents, Verse, Law, Diet minutes 40 million words 1976-2005	

Publication Sub-corpus

- Population:
 - All the books, magazines, and newspapers published in the years 2001 to 2005.
 - defined by the number of characters.

Population (# of chars) → Sample (35M words)

Actual state of Publication

Definition of Population

- Investigated number of chars in 2001 - 2005

	Titles	Pages	Chars
Books	317,117	74,911,520	48,539,925,351
Magazines	55,779	10,414,955	10,515,681,636
Newspapers	49,625	1,198,189	6,416,070,114

Powered by :
 National Diet Library
 Japan Magazine Publishers Association
 Japan Newspaper Publishers Association

Stratification and Each Ratio

Media	Strata	# of chars	Ratio	Media	Strata	# of chars	Ratio
Book	0. General works	1,636,414,548	2.50%	Magazine	General	7,421,447,806	11.34%
	1. Philosophy	2,597,610,813	3.97%		Education	877,875,592	1.34%
	2. General history	4,301,204,340	6.57%		Politics	456,459,405	0.70%
	3. Social sciences	12,408,321,943	18.95%		Industry	110,640,958	0.17%
	4. Natural sciences	5,069,594,034	7.74%		Technology	1,468,293,360	2.24%
	5. Technology	4,615,929,967	7.05%		Medical	180,964,513	0.28%
	6. Industry	2,196,387,437	3.35%		National	2,417,622,461	3.69%
	7. The arts	3,258,432,447	4.98%		Block	1,296,592,154	1.98%
	8. Language	888,800,128	1.36%		Local	2,701,855,499	4.13%
	9. Literature	9,341,275,486	14.27%		Total	65,471,677,100	100%
n. Unclassified	2,225,954,208	3.40%					

Distribution of # chars = Compositional Ratio

Extracting sample

Sample starts here

Figures, old Japanese are omitted

A character randomly chosen in a page

Compilation of BCCWJ

- Sampling (as shown above)
- Copyright solution
 - We identified almost 30,000 copyright holders.
 - 70-80% of them approved to the request.
- Text digitalization and XML tagging
 - Logical structure of text
- Annotation of Part of Speech information
 - 98% accuracy with an electronic dictionary UniDic
 - 99.9% with annotator's modification for 1 million wd

Compilation of BCCWJ

```

<?xml version="1.0" encoding="UTF-8"?>
<sample sampleID="LBe2_00005" version="1.0" type="fixedLength">
<article articleID="LBe2_00005_F001">
<paragraph>
<sentence>やがて、後<sampling type="start" />燕は漢人の<ruby rubyText="ひょう">馮</ruby><ruby rubyText="ばつ">跋</ruby>に乗っ取られてしまいます。</sentence>
<sentence>西暦四〇九年のことですが、この翌年前記の南燕が東晋の<ruby rubyText="りゅう">劉</ruby><ruby rubyText="ゆう">裕</ruby>によって、ほろぼされてしまいました。</sentence>
<paragraph>
<paragraph>
<sentence>四〇九年には、いろいろなことがおこっています。</sentence>
<sentence>さしもの拓跋珪も、この年、思わぬことで、あろうことか息子の一人、<ruby rubyText="たく">拓</ruby><ruby rubyText="はつ">跋</ruby><ruby rubyText="しゅう">紹</ruby>によって殺されました。</sentence>
<paragraph>
    
```

Release of BCCWJ

- In 2011, completed BCCWJ is released
 - 少納言 *Shonagon*
<http://www.kotonoha.gr.jp/shonagon/>
 Character-based Concordance on the web
 Free, max 500 examples (randomly chosen)
 - 中納言 *Chunagon*
<https://chunagon.ninjal.ac.jp/>
 Word-based Concordance on the web
 Registration is needed, all the examples downloadable
 - DVD
 All the morphologically analyzed text, bibliographic data
 Academic Use, 52,500 YEN

Collocations in BCCWJ

- NINJAL-LWP for BCCWJ NINJAL-LWP for BCCWJ (NLB)
 - <http://nlb.ninjal.ac.jp/>
 - Shows collocation (common word combinations)

Question 5

How do Japanese write  in daily life?

tamago

katakana

which is most frequent?

kanji


What BCCWJ offers

- The first balanced corpus of written Japanese
 - Actual situation of published / spread written text
 - Various types of written text
- Easy access to 100 million words corpus
 - Everybody can use a large-sized corpus
 - Objective tests for linguistic analyses

Infrastructure for Japanese corpus linguistics

Conclusion before Lunch

- Japanese corpora
 - NINJAL stated creating a series of large corpora rapidly since 2000.
 - Infrastructures for Japanese corpus linguistics
- Knowledge and Behavior
 - There are many linguistic questions we can not answer with our linguistic knowledge.
 - Linguists need reliable corpora to investigate the linguistic behavior in actual life.

Use corpora ! 

Workshop after Lunch

- BCCWJ demonstrations
 - 少納言 *Shonagon* 
 - 中納言 *Chunagon* 
 - NINJAL-LWP for BCCWJ 
- CSJ demonstration
 - ひまわり *Himawari* 
- Other resources
 - 青空文庫 *Aozora Bunko* on ひまわり *Himawari* 

Corpus Linguistics and Japanese Language (2) Workshop

Takehiko Maruyama
National Institute for Japanese Language and Linguistics / University of Oxford
18 November 2014
SEMINÁŘ JAPONSKÝCH STUDIÍ
Masarykova Univerzita

Contents (14:10 – 15:45)


- BCCWJ demonstrations
 - 少納言 *Shonagon* 
 - 中納言 *Chunagon* 
 - NINJAL-LWP for BCCWJ 
- CSJ demonstration
 - ひまわり *Himawari* 
- Other resources
 - 青空文庫 *Aozora Bunko* on ひまわり *Himawari* 

BCCWJ : demonstrations

『現代日本語書き言葉均衡コーパス』


 少納言  中納言 コーパス検索アプリケーション KOTONOHA「現代日本語書き言葉均衡コーパス」

what is this ?




すいか
スイカ
西瓜

How do they write it?



すいか
スイカ
西瓜

How do they write it?




すいか
スイカ
西瓜

Question 6


- すいか / スイカ / 西瓜
- Which is the most frequent in **Newspapers** ?
- Ask **少納言 Shonagon!**

<http://www.kotonoha.gr.jp/shonagon/>



Question 7

- Give an example of writing variation like すいか, and ask **少納言 Shonagon!**
- For example...
 - バイオリン・ヴァイオリン
 - ダイヤモンド・ダイアモンド
 - 買い物・買物
 - 打ち合わせ・打合わせ・打合せ
 - にんじん・ニンジン・人参
 - ひふ科・ヒフ科・皮ふ科・皮フ科・皮膚科



Question 5

How do Japanese write  in daily life?

tamago

katakana

which is most frequent?



kanji

Question 5


- たまご タマゴ 玉子 卵
- Which is the most frequent in BCCWJ?
- Is it a good way to ask **少納言 Shonagon!**
- Example of search result “卵”

「バター、黒糖、卵黄をよくすり混ぜる。」
(Butter, brown sugar, **yolk**, mix them well.)

らん おう (ran o:) たまご
卵黄 (yolk) → It's not the case of 卵!

Question 5


- Ask 中納言 *Chunagon*, in which Part-of-Speech information can be used.
- <https://chunagon.ninjal.ac.jp/login>
- Registration is needed to log in !



Question 5



- Settings for the corpus search

『語彙素』が『卵』 ← Lemma
AND 『語彙素読み』が『タマゴ』 ← Reading



Question 8

- Give an example of writing variation like たまご, and ask 中納言 *Chunagon* !
- For example...
 - 買い物・買物
 - ねこ・ネコ・猫
 - いぬ・イヌ・犬






Collocations in BCCWJ

- NINJAL-LWP for BCCWJ NINJAL-LWP for BCCWJ (NLB)
- <http://nlb.ninjal.ac.jp/>
- Shows collocation (common word combinations)



Question 9

- Ask NLB about Japanese collocations !
- 「X を飲む」(to drink X)
- What is the most frequent word for X in BCCWJ?

Question 10

- Give an example of collocation like 「X を飲む」, and ask NLB !
- For example...
 - 「X を食べる」 eat X
 - 「X を聞く」 listen to X
 - 「X を読む」 read X
 - 「X を書く」 write X
 - 「X を話す」 speak X

CSJ :
Corpus of Spontaneous Japanese

『日本語話し言葉コーパス』

Distribution of CSJ

- CSJ (with 18 DVDs) is distributed at the Center for Corpus Development, NINJAL.

http://www.ninjal.ac.jp/corpus_center/csj/

Himawari

- Himawari is a character-based concordance system for Japanese linguistics

<http://goo.gl/nBcPO>

Answer to “Communication”

- How do Japanese pronounce “communication” ?

 - Corpus : CSJ, 651 hours, 7.52 million words
 - Frequency of the word “communication” : **601 times**

コミュニケーション	296
コミュニケーション	136
コミュニケーション	123
コミュニケーション	36
misc	10
Total	601

Answer to Filled Pauses (JP)

- What FP do Japanese use most frequently?

 - Corpus : CSJ, 651 hours, 7.52 million words
 - Frequency of Filled Pauses : **430,472 times**

(F えー)	(F e:)	116,772	27.1%
(F え)	(F e)	45,665	10.6%
(F ま)	(F ma)	44,549	10.4%
(F あのー)	(F ano:)	40,695	9.5%
(F あの)	(F ano)	33,330	7.7%


(top 5)

Aozora Bunko

『青空文庫』

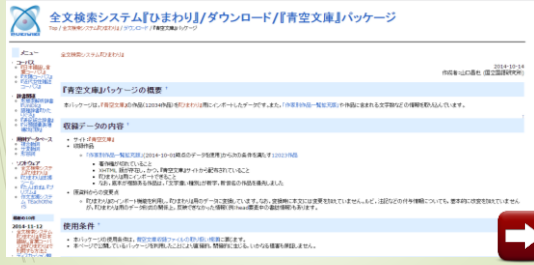
Aozora Bunko

■ *Aozora Bunko* (青空文庫) is a Japanese digital library. This online collection has several thousands of works of Japanese-language fiction and non-fiction. *Aozora Bunko* has digital copies of many out-of-copyright books.



Aozora Bunko on Himawari

■ *Aozora Bunko* Package can be downloaded.
<http://goo.gl/Re73C>



Instead of Conclusion...

ありがとう	7085
有難う	419
ありがと	337
有り難う	102
アリガト	26
アリガトウ	24
ありがとう	3
ア・リ・ガ・ト	2
ありがと	2
アリ、(*^▽^*)ノガトウ	1
総計	8001

