Manyoshu and Hachidaishu in the Corpus of Historical Japanese ver. 2019.3

小木曽 智信 / Toshinobu OGISO
国立国語研究所 / NINJAL

2019/09/20　EAJRS in Sofia, Bulgaria

Topics

- NINJAL Corpora & CHJ
- Recent Updates to CHJ
- CHJ Manyoshu & ONCOJ etc.
- CHJ Hachidaishu
- (Preliminary Analysis)

Center for Corpus Development, NINJAL

コーパス
開発センター
Center for Corpus Development, NINJAL

https://pj.ninjal.ac.jp/corpus_center/

NINJAL Corpora

NINJAL Corpora

- Spoken 話し言葉 → CSJ
- Written 書き言葉 → BCCWJ
- **History 歴史 → CHJ**
- Web (100 billion words) → NWJC
- Dialects 方言 → COJADS
- Everyday Conversation日常会話 → CEJC

Corpus as a research infrastructure

BCCWJ (Balanced Corpus of Contemporary Written Japanese)

- Registered users : 20,000
- Queries : 500,000/year
- Papers : 70/year

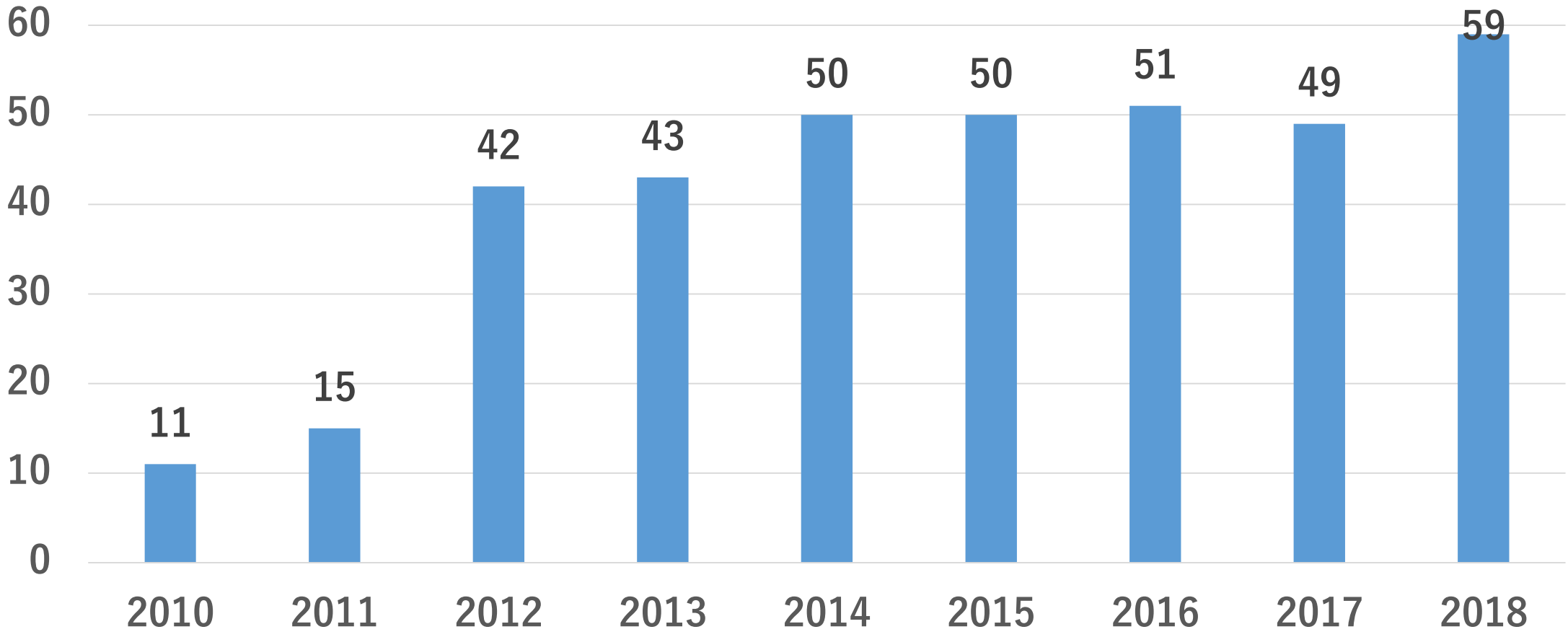
Corpus as a research infrastructure

CHJ (Corpus of Historical Japanese)

- Registered users : 10,000
- Queries : 260,000/year
- Papers : 50/year

  (including proceedings)

Number of papers using CHJ



NINJAL Diachronic Corpus Project

The Construction of Diachronic Corpora and New Developments in Research on the History of Japanese (2016-2022)

「通時コーパスの構築と日本語史研究の新展開」プロジェクト

1

1000年をこえる日本語の歴史を収録する通時コーパスを構築

- 上代（奈良時代以前）から明治・大正時代まで

2

通時コーパスを活用した新しい日本語史研究を展開

- マクロな視点
- 新しい手法の導入

3

日本語史研究者以外の人々にも

- 歴史的研究への参入障壁をコーパスで排除、海外の研究者や関連分野の研究者にオープンに
- ポータルサイトを構築し言葉に関心を持つ全ての人々に情報提供

➢ **総合的日本語研究の骨格となる日本語史の研究基盤の構築と活用**

Corpus of Historical Japanese

- **Diachronic corpus** from 8th to 20th century
  - **Full text** corpus of historical Japanese materials
  - Lemmatized head words, part-of-speech and morphological information are **annotated**
  - **Linked** to related resources on the web

Recent Updates to CHJ

Sub-corpora of CHJ

奈良時代	☑万葉集　☐宣命	
平安時代	☑仮名文学	☑**和歌**
鎌倉時代	☑説話・随筆　☑日記・紀行　☐軍記	
室町時代	☑狂言　☑**キリシタン資料**	
江戸時代	☑洒落本　☑**人情本**　☐近松	
明治・大正	☑雑誌 ☑**教科書** ☐文学作品 ☐新聞 ☑**明治初期口語資料**	

Update of CHJ / 2018

March 2018

- Muromachi Series II: *Kirisitan Shiryo* (Amakusa Edition of Feiqe, Esopo) thanks t o BL

October 2018

- Meiji/Taishō Series II: Textbooks （国定読本）

Update of CHJ / March 2019 (1)

Meiji/Taishō Series

- Magazines
  +Toyo Gakugei Zasshi 1881, 1882
- Early Meiji Spoken Language Materials (明治初期口語資料)
  +10 works

Update of CHJ / March 2019 (2)

- Edo Series
  +洒落本 Sharebon (30 books)
  +人情本 Ninjobon (7 works=93 books)
- Wakashu Series
  +**八代集 Hachidaishu**

Web Interface Chunagon 中納言

- Online concordancer for NINJAL corpora
- Enables query of morphological information

https://chunagon.ninjal.ac.jp

Registration

- Free of charge, Registration needed
- Online registration using email and SMS (text)

https://chunagon.ninjal.ac.jp/useraccount/register

How to use Chunagon?

Resource providers workshop

TODAY 15:00-18:00

NINJAL 国立国語研究所

CHJ Manyoshu 万葉集

CHJ奈良時代編Ⅰ万葉集

- 「万葉集」全20巻　4,516首
  - 約98,000語
  - 小学館「新編日本古典文学全集」
  - 万葉仮名の原文本文と対応付け

万葉仮名

籠毛与　美籠母乳　布久
思毛与　美夫君志持　此
岳尓　菜採須児　家告閑
名告紗根　虚見津　山跡
乃国者　押奈戸手　吾許
曾居　師吉名倍手　吾己
曾座　我許背歯　告目
家呼毛名雄母

籠もよ　み籠持ち　ふくし
もよ　みぶくし持ち　この
岡に　菜摘ます児　家告ら
せ　名告らさね　そらみつ
大和の国は　おしなべて
我こそ居れ　しきなべて
我こそいませ　我こそば
告らめ　家をも　名をも

https://japanknowledge.com/hvst/display/?lid=80110V00060023

CHJ万葉集の特長

検索結果に万葉仮名の原文を併記

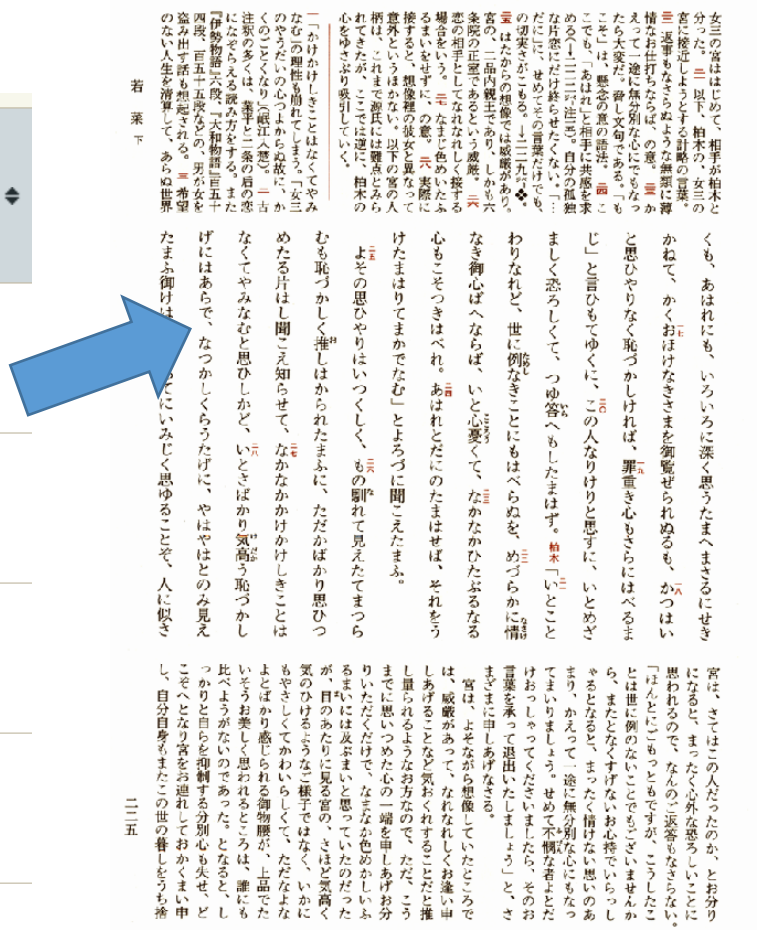
10-万葉 0759_00008	37020	葛花|なでしこ|が|花|をみなへし|また|藤袴|朝顔|が|花 #	秋	|の|日|の|穂田|を|雁がね|暗けく|に|夜|の	
		花瞿麦之花姫部志又藤袴朝貝之花 #	秋	日乃穂田乎鴈之鳴闇尓夜之穂杼呂	
10-万葉 0759_00004	22630	しきたへ|の|衣手|交へ|て|自妻|と|頼め|る|今夜|	秋	|の|夜|の|百夜|の|長|さ|あり|こせ|ぬ	
		而敷細乃衣手易而自妻跡憑有今夜	秋	夜之百夜乃長有与宿鴨	
10-万葉 0759_00019	54900	万代|に|記し|継が|む|そ|やすみしし|我|が|大君|	秋	|の|花|し|が|色々|に|見し|たまひ|明らめ|たまひ	
		仁万世尓記続牟曾八隅知之吾大皇	秋	花之我色ゑ尓見賜明米多麻比酒見	
10-万葉 0759_00010	41020	が|衣|摺ら|む|に|にほひ|こそ|島|の|榛原|	秋	|立た|ず|とも # 風|に|散る|花橘|を|袖|に	
		男子之衣摺槻年年保比与嶋之榛原	秋	不立友 # 風散花橘叫袖受而為君御跡	

CHJ万葉集の特長

- 万葉仮名の原文（漢字）での検索が可能
- 用例を参照するpermalink https://chunagon.ninjal.ac.jp/chj/permalink?unit=short&position=10-万葉0759_00005,13660

Japan Knowledgeへのリンク

サンプルID	開始位置	連番	コア	前文脈	キー	後文脈	語彙素読み	語彙素	品詞	原文文字列	本文種別	歌番号	作品名	巻名等	底本	ページ番号	底本リンク
10-万葉 0759_00001	4190	2610	1	落ち|ず|家|なる|妹|を|かけ|て|偲ひ|つ #	秋	|の|野|の|み|草|刈り|葺き|宿れ|り|し	アキ	秋	名詞-普通名詞-副詞可能	金	歌	7	万葉集	巻第一	新編全集<6>	28	JK
				自見寐夜不落家在妹乎懸而小竹櫃 #	金	野乃美草苅葺屋杼礼里之兎道乃宮												
10-万葉 0759_00001	19040	12050	1	の|奉る|御調|と|春へ|に|は|花|かざし|持ち|	秋	|立て|ば|黄葉|かざせ|り|行き|沿ふ|川|の|神	アキ	秋	名詞-普通名詞-副詞可能	秋	歌	38	万葉集	巻第一	新編全集<6>	47	JK
				山山神乃奉御調等春部者花插頭持	秋	立者黄葉頭刺理逝副川之神母大御												
10-万葉 0759_00001	37190	23340	1	いつ|か|越え|な|む|妹|が|あたり|見|む #	秋	|さら|ば|今|も|見る|ごと|妻恋ひ|に|鹿|鳴か	アキ	秋	名詞-普通名詞-副詞可能	秋	歌	84	万葉集	巻第一	新編全集<6>	69	JK
				浪立田山何時鹿越奈武妹之当見武 #	秋	去者今毛見如妻恋尓鹿将鳴山曾高												
10-万葉 0759_00002	920	560	1	なびく|我|が|黒髪|に|霜|の|置く|まで|に #	秋	|の|田|の|穂|の|上|に|霧ら|ふ|朝霞	アキ	秋	名詞-普通名詞-副詞可能	秋	歌	88	万葉集	巻第二	新編全集<6>	80	JK
				者将待打靡吾黒髪尓霜乃置万代日 #	秋	田之穂上尓霧相朝霞何時辺乃方二												
10-万葉 0759_00002	8310	5440	1	愛しき|か|も|君|が|み言|を|持ち|て|通はく #	秋	|の|田|の|穂向き|の|寄れ|る|片寄り|に|君	アキ	秋	名詞-普通名詞-副詞可能	秋	歌	114	万葉集	巻第二	新編全集<6>	93	JK
				思吉香聞君之御言乎持而加欲波久 #	秋	田之穂向乃所縁異所縁君尓因奈名												



× 令和

- コーパスのアノテーション対象は歌のみ
- 「令和」の出てくる左注は対象に含まれず
- https://chunagon.ninjal.ac.jp/chj/permalink?unit=short&position=10-万葉0759_00005,13660

CHJ Manyoshu & ONCOJ

ONCOJ

Oxford-NINJAL Corpus of Old Japanese

Oxford-NINJAL上代日本語コーパス

ONCOJ

Prof. Bjarke Frellesvig (Univ. of Oxford) のプロジェクトが構築した，統語情報付きの上代日本語コーパス（OCOJ）

↓

オックスフォードと国語研との協力で整備拡充して国語研から公開中

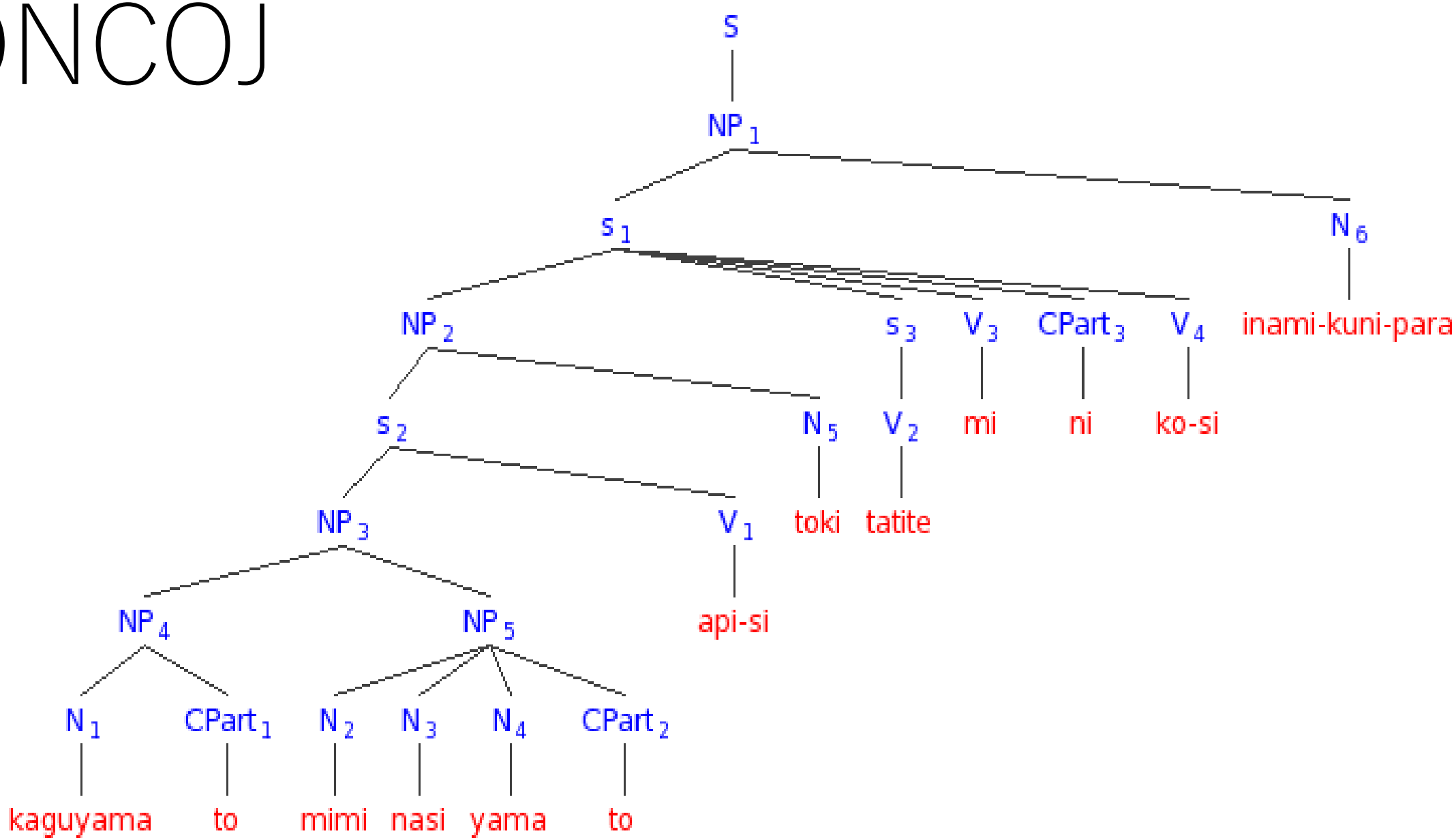
https://oncoj.ninjal.ac.jp

ONCOJ

- **MYS.1.14**
- 高山与
- 耳梨山与
- 相之時
- 立見尓来之
- 伊奈美國波良

- kaguyama to
- mimi nasi yama to
- apisi toki
- tatite , mi ni kosi
- inamikunipara .

ONCOJ


S
NP 1
s 1
N 6
NP 2
s 3
V 3
CPart 3
V 4
inami-kuni-para
s 2
N 5
V 2
mi
ni
ko-si
NP 3
V 1
toki
tatite
NP 4
NP 5
api-si
N 1
CPart 1
N 2
N 3
N 4
CPart 2
kaguyama
to
mimi
nasi
yama
to

ONCOJとCHJ

- ONCOJ
  - 上代日本語の共時的なコーパス
  - 詳細な文法情報付き
- CHJ
  - 上代から近現代までの通時的なコーパス
  - 単語情報＋外部リンク

CHJ↔ONCOJ

歌番号による相互リンク

☑ ONCOJ → CHJ

☐ CHJ → ONCOJ　/ coming soon

万葉集の写本

- 万葉集の原本は失われ、平安時代中期以降の写本が残る（最古の写本、桂本は巻第四のみ）
- 全巻揃ったものは鎌倉時代後期（西本願寺本）

万葉集校本DBへのリンク

万葉集校本データベース
（万葉集校本データベース作成委員会）
- https://www.manyou.gr.jp/SMAN_1/

- CHJから歌番号でリンク
  coming soon

万葉集校本DBへのリンク

歌首表示画面

●各古写本表示

任意の歌句表示□で囲まれている各部分を選択すると、本文異同を表示。

●本文表示

各注釈本に採用されている本文一覧が表示されます。

●訓み表示

各注釈本に採用されている訓み一覧が表示されます。

●注釈リスト

(※現在データ不備のためご使用いただけません)

注釈本リストが表示されます。

この中から見たい注釈データを選択できます。

●口訳表示

各注釈本に採用されている口訳一覧が表示されます。

寛永版本

万葉集校本DBへのリンク

- 注釈本の本文、古写本の画像表示

C03	万葉集拾穂抄	金野乃　美草刈葺　屋杼礼里之　兎道乃宮子能　借五百磯所念
C04	万葉代匠記（初稿本）（精撰本）	金野乃 美草刈葺 屋杼礼里之 兎道乃宮子能 借五百●所念
C05	万葉集僻案抄・童蒙抄・劄記	金野乃　美草苅葺　屋杼禮里之　兎道乃宮子能　借五百磯所念
C06	万葉考	金野乃　美草刈葺　屋杼禮里之　兎道乃宮子能　借五百磯所念

寛永版本	元暦校本	広瀬本	類聚古集	紀州
				

CHJ Hachidaishu 八代集

コーパス化の対象

- 底本
  国文学研究資料館所蔵　正保版本
  『二十一代集』テキストデータ
  http://base1.nijl.ac.jp/~selectionfulltext/21textpagelist.html
- 二十一代集より八代集のみ
  古今、後撰、拾遺、後拾遺、金葉、詞花、千載、新古今

表 「和歌集編」作品別語数（短単位）

和歌集名	成立年	語数
古今和歌集	905	29,943
後撰和歌集	955	38,881
拾遺和歌集	1005	34,164
後拾和歌集	1087	40,250
金葉和歌集	1128	20,543
詞花和歌集	1151	12,376
千載和歌集	1188	37,095
新古今和歌集	1205	49,160
合計		262,412

本文校訂

- 濁点（゛）の付与
- 踊り字（ゝ／＼）の処理
- 送り仮名（久恋→久しき恋）修正
- 誤り（幡磨→播磨）訂正

NIJL原文画像へのリンク

- NIJL正保版本『二十一代集』影印画像

- CODHのIIIF「日本古典籍ビューア」経由でリンク
  http://codh.rois.ac.jp/pmjt/book/200007092/

NIJL原文画像へのリンク

脈	キー	後文脈	語彙素読み	語彙素	品詞	原文文字列	本文種別	歌番号	作品名	巻名等	底本	ページ番号	底本リンク	参考リンク
へ|の|よよ|の| ず|春|の|花|の| |	秋	|の|月|の|夜|ごと|に|さふらふ|人々|を|めし	アキ	秋	名詞-普通名詞-副詞可能	秋			古今和歌集	仮名序	古典選集本文DB　正保版本「二十一代集」	7	Nijl	Maruzen
ゝのみかと春 のあした	秋	の月の夜ことにさふらふ人々をめ												
の|あした|に| ちる|を|見|	秋	|の|夕ぐれ|に|に| の|は|の|落る|を|き き	アキ	秋	名詞-普通名詞-副詞可能	秋			古今和歌集	仮名序	古典選集本文DB　正保版本「二十一代集」	8	Nijl	Maruzen
のあしたに花の 見	秋	の夕くれにこのはの落るをきゝ、												
を|あはせ|たり| 小|なる|べし|、	秋	|の|ゆふべ|竜田| 川|に|ながるる|もみち|を|ば|みかど| の	アキ	秋	名詞-普通名詞-副詞可能	秋			古今和歌集	仮名序	古典選集本文DB　正保版本「二十一代集」	9	Nijl	Maruzen
ゝはせたりとい へし、#	秋	のゆふへ竜田川 になかるゝもみち												
はじめ|をはり| に|なら|ず|、# ば|	秋	|の|月|を|見る|に| あかつき|の|雲|に| あへ	アキ	秋	名詞-普通名詞-副詞可能	秋			古今和歌集	仮名序	古典選集本文DB　正保版本「二十一代集」	10	Nijl	Maruzen
しめをはりた らす、#い	秋	の月を見るにあかつきの雲にあへ												

日本古典文学大系本文へのリンク

「新日本古典文学大系」岩波書店
Maruzen eBook Library
(https://elib.maruzen.co.jp/)
のページにリンク

サンプルID	開始位置	連番	コア	前文脈	キー	後文脈	語彙素読み	語彙素	品詞	原文文字列	本文種別	歌番号	作品名	巻名等	底本	ページ番号	底本リンク	参考リンク
20W古今0905_00001	19440	11250	1	いにしへ|の|よよ|の|みかど|春|の|花|の|あした|	秋	|の|月|の|夜|ごと|に|さぶらふ|人々|を|めし	アキ	秋	名詞-普通名詞-副詞可能	秋			古今和歌集	仮名序	古典選集本文DB　正保版本「二十一代集」	7	Nijl	Maruzen
				へのよゝのみかと春の花のあした	秋	の月の夜ことにさふらふ人々をめ												
20W古今0905_00001	21980	12720	1	又|春|の|あした|に|花|の|ちる|を|見|	秋	|の|夕ぐれ|に|こ|の|は|の|落る|を|きき	アキ	秋	名詞-普通名詞-副詞	秋			古今和歌集	仮名序	古典選集本文DB　正保版本	8	Nijl	Maruzen
				又春のあしたに花の	秋	のタくれにこのは												

歌人情報の整備とリンク

- 『勅撰集 付新葉集作者索引』和泉書院索引叢書(1986) により歌人名を標準化
- NDL Authoritiesの当該歌人にリンク

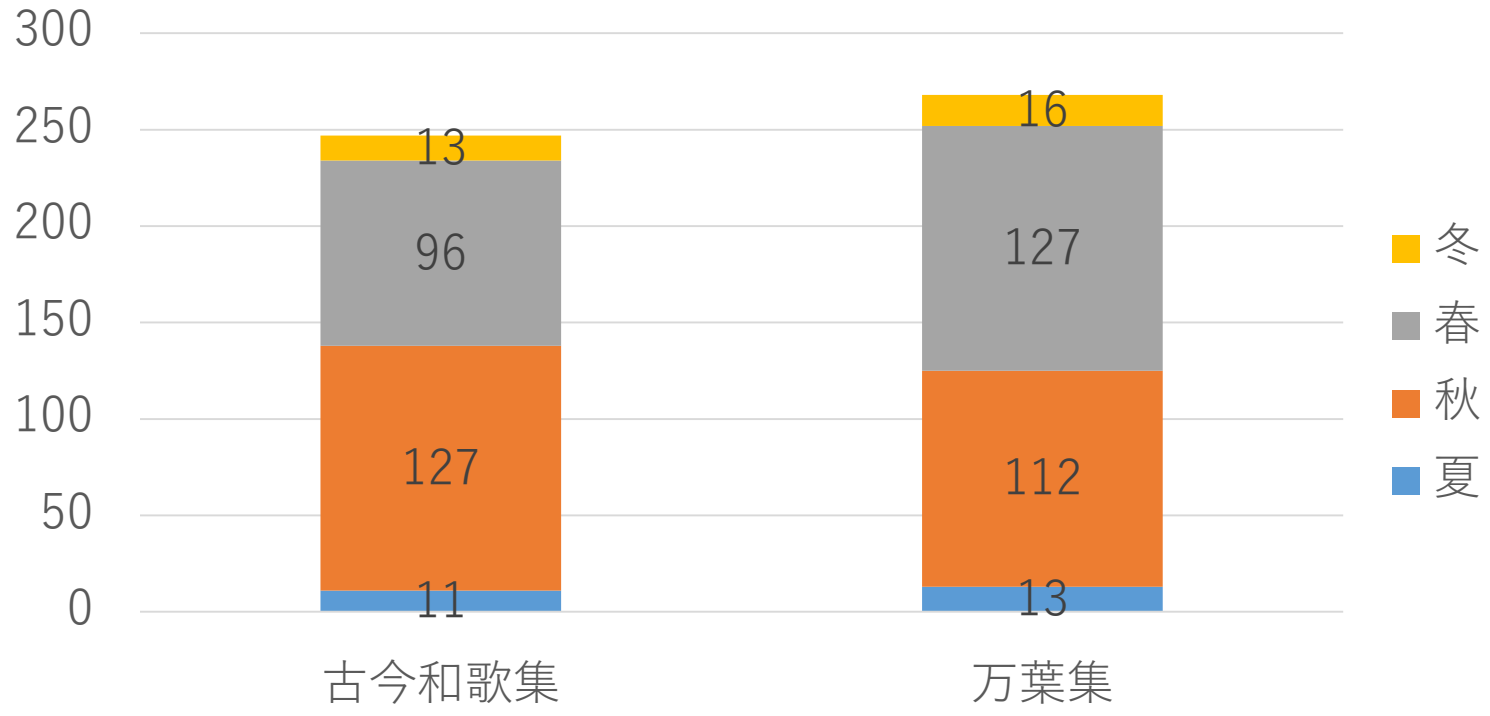
既存の資料と比較した CHJ八代集の長所

国歌大観、岩波「八代集」CD-ROM etc.

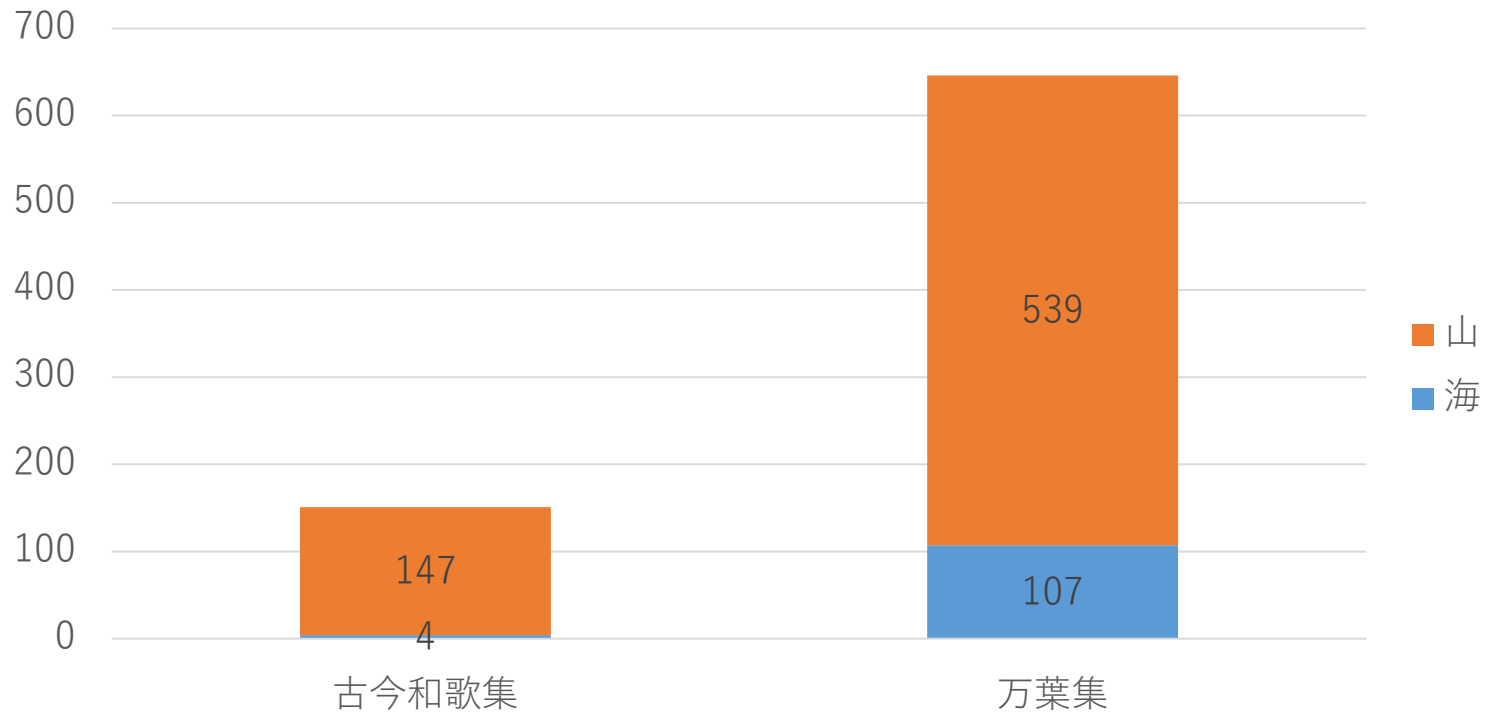
- 同一基準による単語情報付き
  - 句や文字列でなく見出し語で検索可能
  - 散文を含む他の作品と通時的に比較可能

Preliminary Analysis

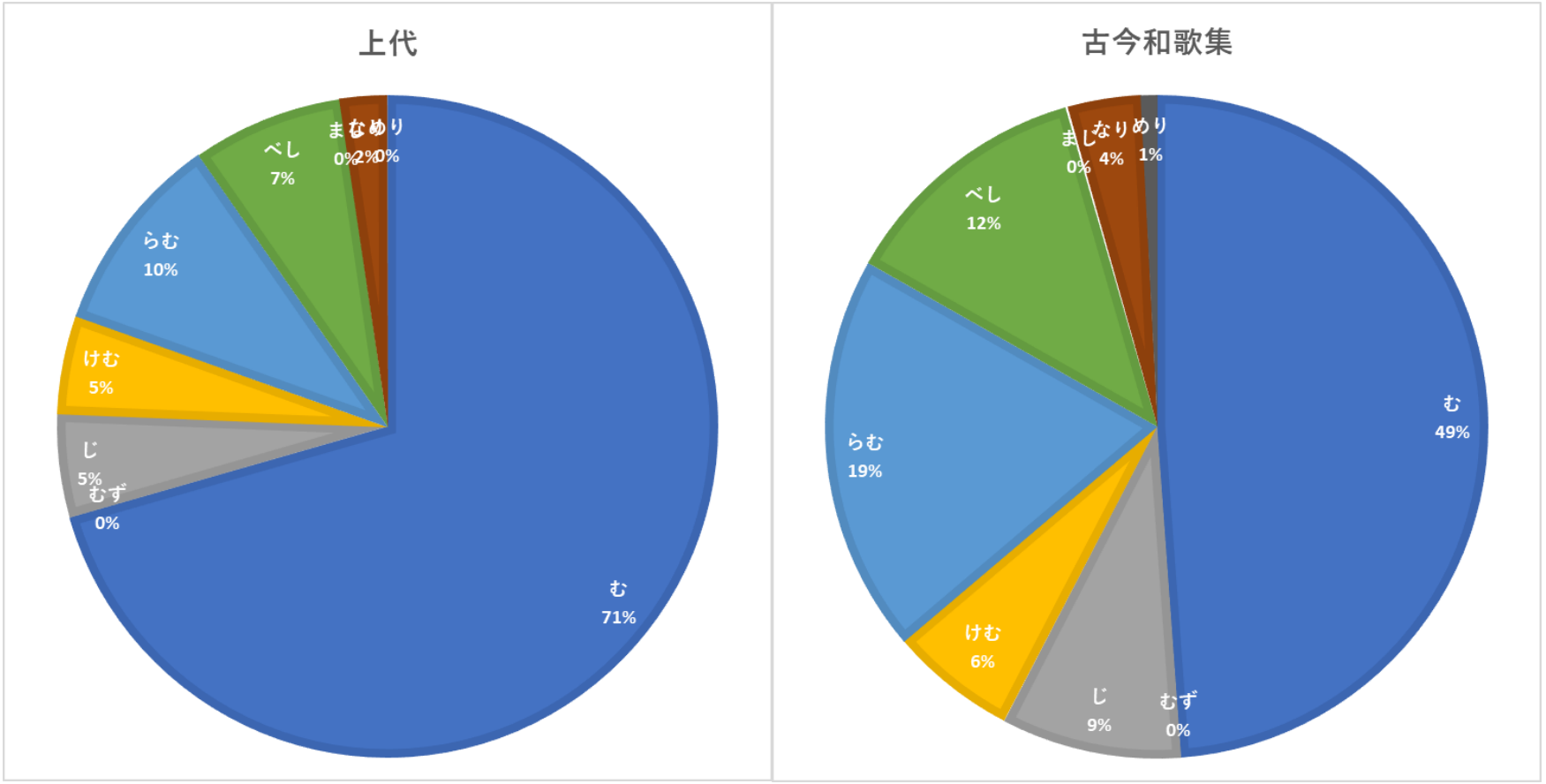
春夏秋冬



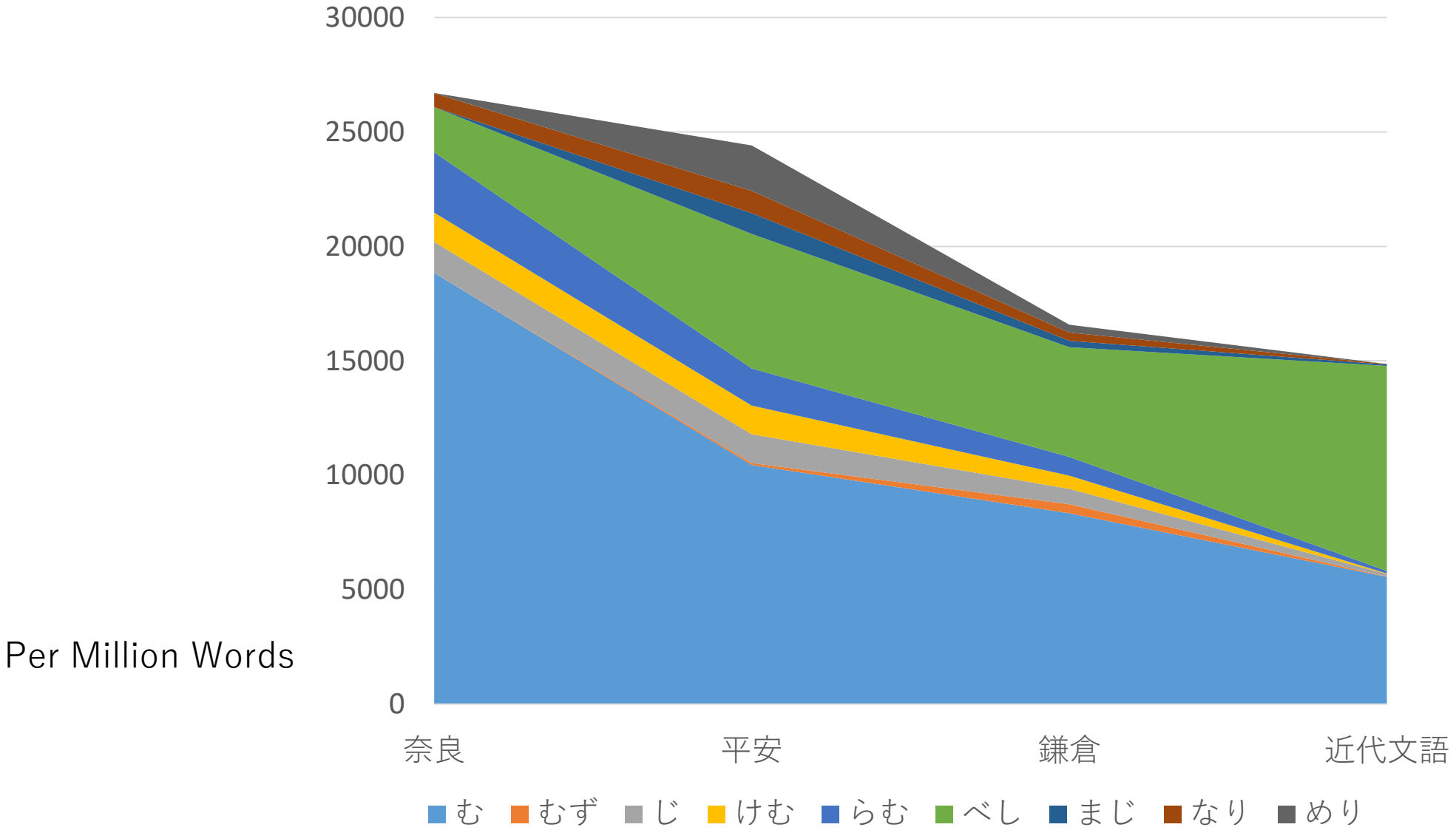
山と海



Modality (Auxiliary verbs)



Modality (Auxiliary verbs)



特徴語の抽出

- 対数尤度比 (Log-Likelihood Ratio, LLR)

$$G/2 = \sum \text{観察値}\,(\log \text{観察値} - \log \text{期待値})$$

$$= a \log \frac{aN}{(a+b)(a+c)} + b \log \frac{bN}{(a+b)(b+d)}$$

$$+ c \log \frac{cN}{(a+c)(c+d)} + d \log \frac{dN}{(b+d)(c+d)}$$

- 他から期待される頻度と比べて、どれだけ多いか少ないか

歌集ごとの特徴語（形容詞）

語	金葉和歌集	古今和歌集	後拾遺和歌集	後撰和歌集	詞花和歌集	拾遺和歌集	新古今和歌集	千載和歌集	万葉集
賢い	-9.94876	-16.7869	-2.11058	-26.3492	-6.33119	-7.64882	-9.95342	-7.71912	131.3103
乏しい	-4.4072	-7.43597	-9.44282	-11.6708	-2.80474	-9.56765	-10.9757	-7.91391	104.0594
術無い	-4.60807	-2.99195	-9.87325	-12.2028	-2.93257	-10.0038	-11.476	-8.27463	97.52756
清い	-0.63127	-4.44909	-18.2387	-12.046	-0.14476	-14.1112	-4.58569	-3.47072	96.8924
良い	-4.5296	0.003582	-1.01714	-18.5622	-6.20255	-3.34575	-9.57577	-7.43956	70.8016
繁い	-0.69435	0.016517	-19.9639	-9.73172	-3.76372	-8.13308	-2.35568	-0.4346	65.67397
欲しい	-6.4181	0.138568	-13.7519	-4.47346	-4.08443	-1.29771	-3.85446	-11.5251	53.22038
日長い	-2.20072	-3.71305	-4.71506	-5.82745	-1.40055	-4.77739	-5.4804	-3.95168	51.90292
痛い	-4.4072	-0.8731	-9.44282	-3.20438	-0.11791	-4.34593	-2.77134	-7.91391	49.17338
尊い	0.15263	-5.23487	-6.64761	-8.21598	-1.97455	-6.73548	-2.92735	-5.57131	48.58252
鬱しい	-2.00042	-3.37509	-4.2859	-5.29703	-1.27308	-4.34255	-4.98157	-3.592	47.17404
麗しい	-2.40107	-4.05109	-1.18714	-6.35801	-1.52806	-5.21234	-5.97935	-4.31145	46.68617
間無い	-3.80486	-0.46106	-8.15217	-0.80118	-2.42142	-3.33663	-4.26108	-6.83225	42.01455

歌集ごとの特徴語（形容詞）

- **万葉集**：賢し、乏し、すべなし、清し、良し、繁し etc.
- **古今集**：恋し、憂し、正し、強し、侘し、疾し、あやなし　etc.
- **新古今集**：深し、涼し、儚し、憂し、空し、脆し、凄し etc.

CHJ 今後の公開予定

2020年３月

- 奈良時代編 II **宣命**（続日本紀）
- 江戸時代編 III **近松浄瑠璃**（世話物）