『現日研・職場談話コーパス』中納言版の語彙表・語数表の作成と公開

大村舞(国語研究所)

『現日研・職場談話コーパス』公開記念シンポジウム

現日研・職場談話コーパス

- 『現日研・職場談話コーパス』は，現代日本語研究会が作成した職場での自然談話を文字起こししたテキストを元に作成したコーパス

前の発表参照

『現日研・職場談話コーパス』中納言版

(詳しくは前の発表参照)

- 中納言によって「特定の」単語（短単位SUW)の検索・事例を調査・確認することができる。

中納言のみでは難しいところ

- コーパス中にどのような単語があるのか？その頻度は？

- ある対話、話者ごとでの全体単語数が知りたい

中納言では難しいところ

- コーパス中にどのような単語があるのか？その頻度は？

語彙表

- ある対話、話者ごとでの全体単語数が知りたい

語数表

語彙表・語数表

- 中納言にて作成したコーパス内の単語の統計値をカウントした表

語彙表

コーパス中のある単語についての頻度表

語数表

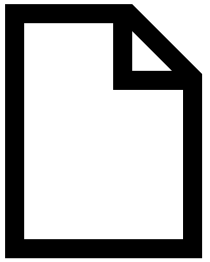
コーパス内のある会話IDと話者ごとについて全体単語数を表した表

入手先

http://pj.ninjal.ac.jp/conversation/shokuba.html

にて公開（予定）

エクセルファイルのほか、
.tsvというテキストファイルで公開
このファイルもメモ帳やエクセルなどで開くことが可能

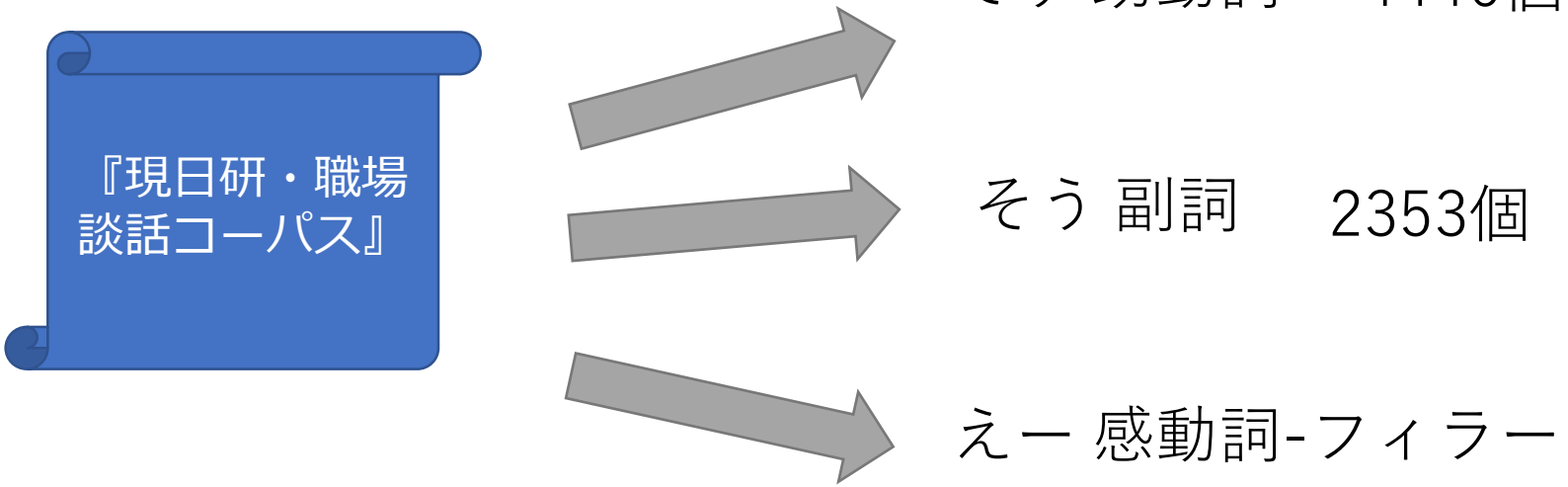


集計方法

- 『現日研・職場談話コーパス』をMecab+UniDic(短単位)で解析した単語によって集計
  - 語彙素、語彙素読み、品詞、語彙素細分類、語種の 5 つの組で見出し語とする
  - 品詞に「空白」「補助記号」「記号」の文字列を含むものは除いている。

語彙表

- コーパス中のある単語についての頻度表



語彙表のサンプル

rank	lForm	lemma	pos	subLemma	wType	frequency	pmw	会議_rank	会議_frequency	会議_pmw	朝_rank	朝_frequency	朝_pmw	休憩_rank	休憩_frequency	休憩_pmw	会議_電話_rank	会議_電話_frequency	会議_電話_pmw	朝_電話_rank	朝_電話_frequency	朝_電話_pmw	休憩_電話_rank	休憩_電話_frequency	休憩_電話_pmw
1	ダ	だ	助動詞		和	7753	42162.45	1	2624	38806.81	1	2140	41285.64	1	2989	46389.27	5	35	29711.38	4	182	30665.54	1	34	34239.68
2	ノ	の	助詞-準体助詞		和	5118	27832.76	2	1978	29253	2	1399	26990.01	2	1741	27020.32	3	44	37351.44	5	172	28980.62	4	27	27190.33
3	テ	て	助詞-接続助詞		和	4658	25331.19	3	1882	27833.24	4	1288	24848.56	5	1488	23093.76	14	21	17826.83	8	117	19713.56	9	22	22155.09
4	ネ	ね	助詞-終助詞		和	4543	24705.79	7	1481	21902.78	3	1363	26295.48	3	1699	26368.48	27	12	10186.76	7	151	25442.29	9	22	22155.09
5	デス	です	助動詞		和	4446	24178.29	4	1810	26768.42	5	1278	24655.63	7	1358	21076.16	2	53	44991.51	1	296	49873.63	3	28	28197.38
6	ノ	の	助詞-格助詞		和	4170	22677.34	6	1692	25023.29	7	1094	21105.84	6	1384	21479.68	6	34	28862.48	6	169	28475.15	5	26	26183.28
7	タ	た	助動詞		和	4107	22334.73	11	1256	18575.21	6	1275	24597.75	4	1576	24459.52	16	17	14431.24	13	100	16849.2	11	21	21148.04
8	ハ	は	助詞-係助詞		和	3950	21480.93	5	1711	25304.29	8	997	19234.48	8	1242	19275.84	23	14	11884.55	17	80	13479.36	12	18	18126.89

- 25列で構成、１行1単語を表している

語彙表の見方　その１

最初の列は全体情報

- rank　コーパス全体での順位
- lForm　語彙素読み
- lemma　語彙素
- pos　品詞
- subLemma　語彙素細分類
- wType　語種
- frequency　コーパス全体での頻度
- pmw　コーパス全体での100万語当たりの頻度

語彙表の見方　その２

以降の列は部分情報

- 会議、朝、休憩の３種類の場面
  - 全体
  - 電話

- _rank　その場面での順位
- _frequency　その場面での頻度
- _pmw　その場面での100万語当たりの頻度

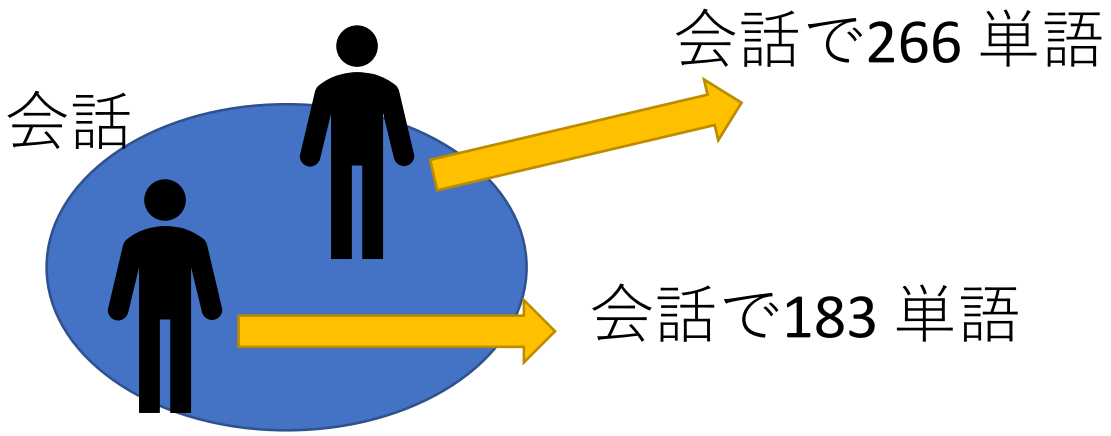
語彙表から分かる統計量

- 単語の種類数 7189個
- 全単語数 183884個
- 最も多い単語「だ（助動詞）」 7753個

他にも統計ソフト（エクセルやR）などを用いて細かい統計計算ができる

語数表

- コーパス内のある会話IDと話者ごとについて全体単語数を表した表
- 会話ID、話者についての情報も一緒に収録



※数単語しか混ざってない話者については
情報自体無いが単語数は掲載

語数表のサンプル

会話 ID	場面 1	場面 2	調査日	場所	会話参加者数	発話者コード	性別	年齢層	職業	職種	役職	出身	最長居住地	語数(全て)	語数(記号等除外・全て)
F01A011	朝	電話	1993年10月	室内	1	F01A	女	28	会社員	イベント企画開発	無	*	*	448	337
F01A021	朝	仕事中の雑談	1993年10月	室内	2	F01A	女	28	会社員	イベント企画開発	無	*	*	266	206
F01A021	朝	仕事中の雑談	1993年10月	室内	2	F01B	女	31	会社員	社長秘書・一般事務	?	*	*	183	142
F01A021	朝	仕事中の雑談	1993年10月	室内	2	Inf(女)								2	2
F01A021	朝	仕事中の雑談	1993年10月	室内	2	他者(女)								1	1
F01A031	朝	電話	1993年10月	室内	1	F01A	女	28	会社員	イベント企画開発	無	*	*	93	73
F01A041	朝	仕事中の雑談	1993年10月	室内	3	F01A	女	28	会社員	イベント企画開発	無	*	*	678	534
F01A041	朝	仕事中の雑談	1993年10月	室内	3	F01C	男	32	会社員	営業	?	*	*	607	428
F01A041	朝	仕事中の雑談	1993年10月	室内	3	F01D	男	45	会社員	イベント企画制作	部長	*	*	305	236
F01A041	朝	仕事中の雑談	1993年10月	室内	3	F01？								1	1
F01A041	朝	仕事中の雑談	1993年10月	室内	3	Inf(女)								9	9
F02Q011	休憩	休憩時雑談	1993年10月	室内	3	F02A	女	27	会社員	事務	無	*	*	1400	1070
F02Q011	休憩	休憩時雑談	1993年10月	室内	3	F02C	女	27	会社員	?	?	*	*	1544	1205
F02Q011	休憩	休憩時雑談	1993年10月	室内	3	F02F	男	43ca	会社員	?	取締役専務	*	*	6	3

語数表のサンプル

会話 ID	場面 1	場面 2	調査日	場所	会話参加者数	発話者コード	性別	年齢層	職業	職種	役職	出身	最長居住地	語数(全て)	語数(記号等除外・全て)
F01A011	朝	電話	1993年10月	室内	1	F01A	女	28	会社員	イベント企画開発	無	*	*	448	337
F01A021	朝	仕事中の雑談	1993年10月	室内	2	F01A	女	28	会社員	イベント企画開発	無	*	*	266	206
F01A021	朝	仕事中の雑談	1993年10月	室内	2	F01B	女	31	会社員	社長秘書・一般事務	?	*	*	183	142
F01A021	朝	仕事中の雑談	1993年10月	室内	2	Inf(女)								2	2
F01A021	朝	仕事中の雑談	1993年10月	室内	2	他者(女)								1	1
F01A031	朝	電話	1993年10月	室内	1	F01A	女	28	会社員	イベント企画開発	無	*	*	93	73
F01A041	朝	仕事中の雑談	1993年10月	室内	3	F01A	女	28	会社員	イベント企画開発	無	*	*	678	534
F01A041	朝	仕事中の雑談	1993年10月	室内	3	F01C	男	32	会社員	営業	?	*	*	607	428
F01A041	朝	仕事中の雑談	1993年10月	室内	3	F01D	男	45	会社員	イベント企画制作	部長	*	*	305	236
F01A041	朝	仕事中の雑談	1993年10月	室内	3	F01？								1	1
F01A041	朝	仕事中の雑談	1993年10月	室内	3	Inf(女)								9	9
F02Q011	休憩	休憩時雑談	1993年10月	室内	3	F02A	女	27	会社員	事務	無	*	*	1400	1070
F02Q011	休憩	休憩時雑談	1993年10月	室内	3	F02C	女	27	会社員	?	?	*	*	1544	1205
F02Q011	休憩	休憩時雑談	1993年10月	室内	3	F02F	男	43ca	会社員	?	取締役専務	*	*	6	3

一行が話者を表現しており、

オレンジの部分（会話IDが同じ）が同一会話を表現

エクセルでIDでフィルタして合計すれば会話の全体単語数も出せる

語数表の見方

- 「会話ID」「場面 1」「場面 2」「調査日」「場所」「会話参加者数」「発話者コード」「性別」「年齢層」「職業」「職種」「役職」「出身」「最長居住地」
  ：会話と話者についての情報
- 語数(全て)：すべての単語数
- 語数(記号等除外・全て)：
  記号を除いたすべての単語数

語彙表・語数表を使った事例

- 語彙表と語数表を用いることで客観的な統計を取ることができる。
  - いくつか事例を紹介
    - 品詞の分布
    - 特徴語抽出

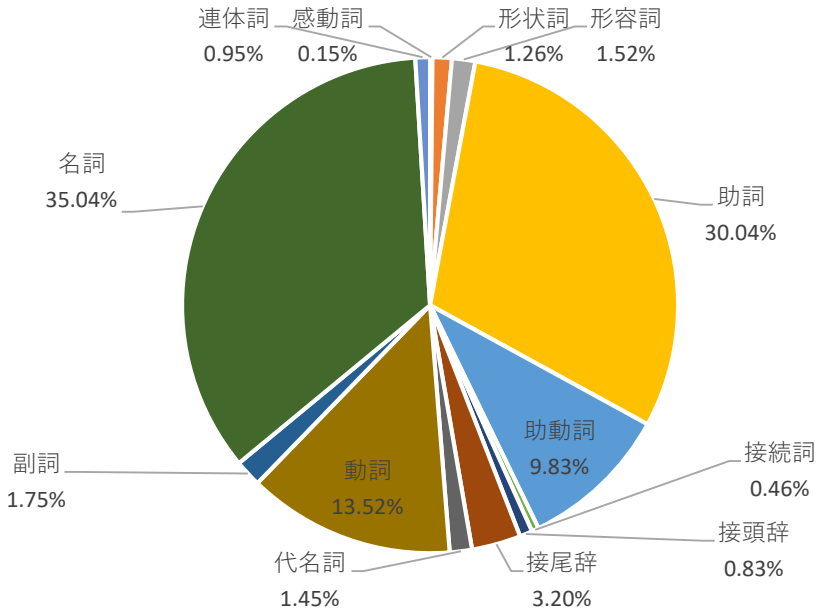
語彙表を使って:品詞の分布

- 品詞の情報があるため、数えることで計算可能
- 語彙表、語数表とともに公開予定

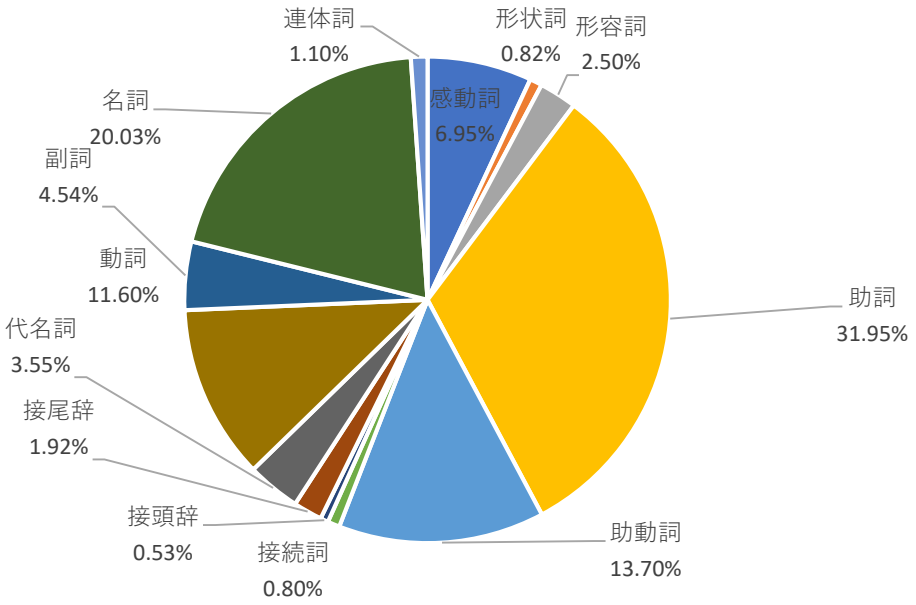
POS	Shokuba_token_suw	会議_token_suw	朝_token_suw	休憩_token_suw	会議_電話_token_suw	朝_電話_token_suw	休憩_電話_token_suw
助詞	58744	21593	16090	21061	275	1681	258
名詞	36838	14126	9902	12810	279	991	239
助動詞	25195	8866	7433	8896	183	988	136
動詞	21340	8155	6134	7051	124	684	119
感動詞	12774	4470	4164	4140	189	693	100
副詞	8355	3056	2264	3035	34	313	43
代名詞	6533	2207	2019	2307	25	162	24
形容詞	4599	1350	1266	1983	5	93	14
接尾辞	3530	1309	991	1230	25	86	20
連体詞	2027	919	504	604	13	63	8
形状詞	1502	538	373	591	6	37	8
接続詞	1478	640	400	438	5	63	13
接頭辞	969	388	294	287	15	81	11
合計	183884	67617	51834	64433	1178	5935	993

語彙表を使って:品詞の分布

BCCWJ



職場コーパス



コーパスにおける品詞の分布を出すことができる

語彙表を使って：特徴語抽出

- 特徴づける語を抽出するための客観的な指標として対数尤度比（log-likelihood ratio、LLR 値）がある

コーパスにおける
特徴的な単語を計算できる

語彙表と対応する参照コーパスと
比較することで計算できる。

より詳細は(柏野 2018)参照

特徴語抽出:
参考コーパスについて

現代日本語書き言葉均衡コーパス (BCCWJ)

（書籍PB）のものを使用

書き言葉のコーパス

名大コーパス(NUC)

大学内における雑談のコーパス

参考：対数尤度比の計算

語彙表によりある単語Wについて以下の分割表が作れる

	対象コーパス	参照コーパス	
単語W	a	b	a+b
単語W以外	c	d	c+d
	a+c	b+d	(a+b+c+d)=n

LLR = 2(a*log(a)+b*log(b)+c*log(c)+d*log(d)- (a+b)*log(a+b) - (a+c)*log(a+c) -(b+d)*log(b+d) - (c+d)*log(c+d) + n*log(n))

ただし a/c < b /d だった場合、-1をかける。

特徴語抽出（BCCWJ編）

語彙素読み	語彙素	品詞	LLR
ネ	ね	助詞-終助詞	20122
ウン	うん	感動詞-一般	17666
アノ	あの	感動詞-フィラー	14262
ハイ	はい	感動詞-一般	10070
テル	てる	助動詞	9007.7
ッテ	って	助詞-副助詞	8551.7

砕けたような表現が上位にくる

特徴語抽出（NUC編）

語彙素読み	語彙素	品詞	LLR
デス	です	助動詞	2301.7
マス	ます	助動詞	2022.7
ゼロ	ゼロ	名詞-数詞	1326.5
ハイ	はい	感動詞-一般	1190.4
エー	えー	感動詞-フィラー	923.87
アノ	あの	感動詞-フィラー	852.75

敬語表現が特徴的

おわりに

- 中納言版『現日研・職場談話コーパス』の語彙表、語数表について紹介
- その利用用途などについて紹介

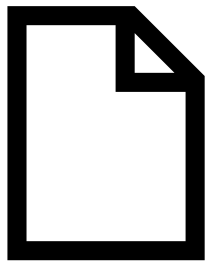
- 語彙表、語数表を用いることで統計的にコーパスの特徴を得ることができるので、研究に活用してもらいたい。

（再掲）入手先

http://pj.ninjal.ac.jp/conversation/shokuba.html

にて公開（予定）

エクセルファイルのほか、
.tsvというテキストファイルで公開
このファイルもメモ帳やエクセルなどで開くことが可能



参考文献

- 柏野和佳子・大村舞・西川賢哉・小磯花絵(2018)「『現日研・職場談話コーパス』中納言データの作成」『言語資源活用ワークショップ2018発表論文集』
- 内山将夫・中條清美・山本英子・井佐原均（2004）「英語教育のための分野特徴単語の選定 尺度の比較」『自然言語処理』11 巻 3 号, pp. 165-197, 自然言語処理学会.
- BCCWJ：http://pj.ninjal.ac.jp/corpus_center/bccwj/
- 名大コーパス：https://mmsrv.ninjal.ac.jp/nucc/