NINJAL
National Institute for Japanese Language and Linguistics

職場編資料と名大会話コーパスとの比較

山崎　誠（国立国語研究所）

2018.09.03
『現日研・職場談話コーパス』公開記念シンポジウム

目次

- 1.名大会話コーパス
- 2.職場編資料と名大会話コーパスの語彙的比較
- 3.まとめ

1.名大会話コーパス

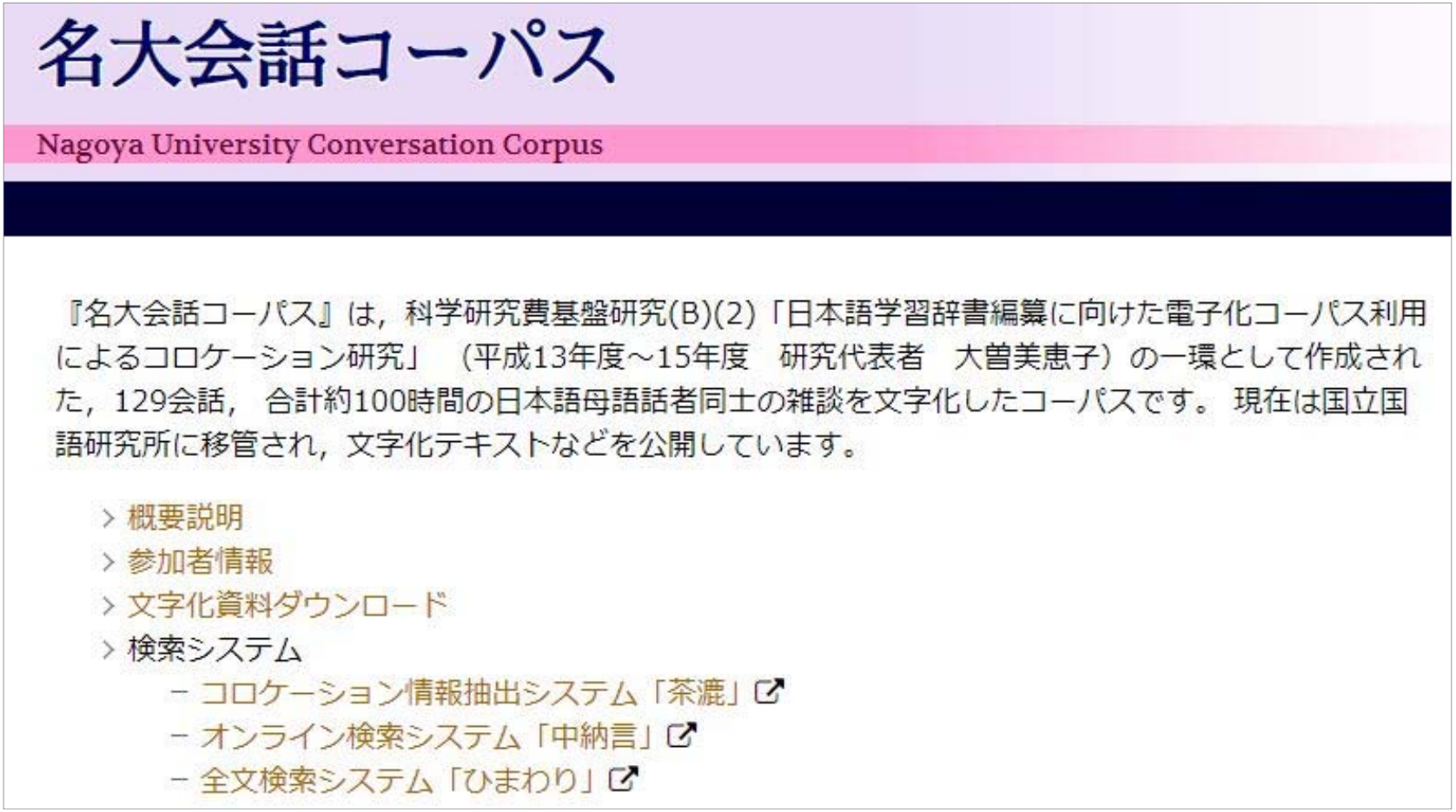
- 日本語母語話者の100時間分の129件の会話（雑談）を収録して，文字化したコーパス。

- 科学研究費基盤研究(B)(2)「日本語学習辞書編纂に向けた電子化コーパス利用によるコロケーション研究」（平成13年度～15年度　研究代表者　大曽美恵子）により作成された。

- 音声は公開していない。

1.名大会話コーパス：HP

- https://mmsrv.ninjal.ac.jp/nucc/
  - データの説明があり，文字化資料がダウンロードできる



1.1.名大会話コーパスの検索方法

- (1)『中納言』

- (2)全文検索システム『ひまわり』/ダウンロード/『名大会話コーパス』パッケージ
  （短単位解析済み全データが含まれる。オフラインで検索可能）

- (3)「茶漉(ちゃこし)」(米国パデュー大学の深田淳氏が作成したWEB検索ツール)
  http://telldev.cla.purdue.edu/chakoshi/public.html

- テキストのダウンロードも可能。

1.2. データの内容

- ＠データ04（35分）
- ＠収集年月日：2001年10月23日
- ＠場所：車中（知立駅より西尾市まで。運転者M018）
- ＠参加者M018：男性20代前半、愛知県半田市出身、同市在住
- ＠参加者F128：女性20代前半、愛知県西尾市出身、同市在住
- ＠参加者の関係：大学時代の部活の友人
- F128：いや、別にいいよ。
- ローソンでいいやろ。
- ちょっと倒していい、これ。
- どうよ、調子は。
- M018：何かね（うん）ソービジー。
- F128：＜笑い＞そうなん、何で？
- M018：何かせん、大学の先生の（うんうん）お仕事を頼まれたりするんで、（うんうんうん）それで何かワープロ打ったりね、（うんうん）何か今度留学生が小学校で何か交流会やるもんで、その留学生集めてこいとか何か、結構ね、（うんうん）わっけわからん仕事押しつけられる。
- F128：え、留学生の世話をしろって？

1.3. 話者の属性：性別と年代

年代	女性	男性	総計
10代	13	2	15
20代	70	18	88
30代	26	1	27
40代	16	8	24
50代	18	4	22
60代	11	4	15
70代～	6		6
不詳	1		1
合計	161	37	198

- 性別では女性，年代では20代～40代が多い。

1.4.話者の属性:出身地

出身地	人数
北海道	11
東北	8
関東	49
中部	86
近畿	21
中国・四国	11
九州・沖縄	11
海外	1
合計	198

- 中部地方出身者が多い。

1.5.話者の属性：出身地と居住地

地域	出身人数	居住人数
北海道	11	18
東北	8	1
関東	49	49
中部	86	120
近畿	21	7
中国・四国	11	1
九州・沖縄	11	0
海外	1	2
合計	198	198

- 出身も居住も中部地方が多い。

1.6.話者の属性:関係性

関係	件数
同級生	51
友人	31
家族	15
先輩	15
同僚	11
初対面	6
知人	5
恋人	4
親族	2
先生	1
合計	141

- 親しい関係が多い。

1.7.話者の属性：場所

場所	件数
飲食店	46
家	30
大学	29
大学の研究室	13
車内	8
職場	2
大学の食堂	2
学校	1
電車内	1
合計	132

- 日常的な場所が多い。

1.8.話者の属性：参加者の人数

参加者の人数	件数
2人	96
3人	28
4人	5
合計	129

- 2～3人の少人数の会話が多い

1.9.話者の属性：収録時間

収録時間	件数
～30分	13
31～60分	99
61～90分	16
91～分	1
合計	129

- 30分～60分の会話が多い。

2.職場編資料と名大会話コーパスの語彙的比較

2.1.基本的な情報

	職場編	名大
延べ語数	183,884	1,129,271
異なり語数	7,189	18,186
Guiraud値	16.76	17.11
会話数	415	129
話者数（延べ）	1324	198
1会話あたりの語数	443.1	8754.0
1話者あたりの語数	138.9	5703.4

- 語数は，短単位で示す。
- Guiraud値は，異なり語数÷（延べ語数の平方根）

2.2.上位語の比較

職場編			
rank	語彙素	品詞	PMW
1	だ	助動詞	42162.5
2	の	準体助詞	27832.8
3	て	接続助詞	25331.2
4	ね	終助詞	24705.8
5	です	助動詞	24178.3
6	の	格助詞	22677.3
7	た	助動詞	22334.7
8	は	係助詞	21480.9
9	に	格助詞	18413.8
10	と	格助詞	18201.7
11	が	格助詞	17032.5
12	で	格助詞	16423.4
13	も	係助詞	15575.0
14	言う	動詞	15210.7
15	よ	終助詞	13399.8
16	為る	動詞	13068.0
17	そう	副詞	12796.1
18	てる	助動詞	12241.4
19	うん	感動詞	12067.4
20	か	副助詞	12056.5

名大			
rank	語彙素	品詞	PMW
1	だ	助動詞	47523.6
2	うん	感動詞	36315.5
3	た	助動詞	26669.4
4	て	接続助詞	26286.9
5	ね	終助詞	26197.4
6	の	準体助詞	23372.6
7	か	副助詞	20955.1
8	と	格助詞	18535.0
9	の	格助詞	18329.5
10	も	係助詞	17740.6
11	で	格助詞	17414.8
12	が	格助詞	17330.6
13	に	格助詞	17263.3
14	は	係助詞	16519.5
15	そう	副詞	16179.5
16	言う	動詞	15868.6
17	って	副助詞	15857.1
18	何	代名詞	15618.9
19	てる	助動詞	13157.2
20	よ	終助詞	13068.6

対数尤度比

- 対数尤度比（LLR）の計算式は以下のとおり（金愛欄ほか（2008：201）から引用）。

- $LLR=2(a\ln a+b\ln b+c\ln c+d\ln d-(a+b)\ln(a+b)-(a+c)\ln(a+c)-(b+d)\ln(b+d)-(c+d)\ln(c+d)+(a+b+c+d)\ln(a+b+c+d))$

- a：Aでの単語 W の頻度 b：Bでの単語 W の頻度
- c：Aの延べ語数−a d：Bの延べ語数−b ※ln(x)=x の自然対数
- ただし、単語 W のAでの使用率がBでの使用率より低い場合、×(−1)の補正を行う。

2.3.職場編資料で対数尤度比(LLR)の高い語

rank	語彙素	品詞	LLR
1	です	助動詞	2301.7
2	ます	助動詞	2022.7
3	ゼロ	名詞	1326.5
4	はい	感動詞	1190.4
5	えー	感動詞	923.9
6	あの	感動詞	852.7
7	えーと	感動詞	657.5
8	此れ	代名詞	422.1
9	御早う	感動詞	394.8
10	が	接続助詞	387.7

rank	語彙素	品詞	LLR
11	頂く	動詞	314.3
12	ええ	感動詞	313.6
13	御座る	動詞	309.9
14	一(いち)	名詞	259.9
15	点検	名詞	259.5
16	は	係助詞	216.2
17	分(ぶん)	名詞	157.3
18	の	格助詞	153.0
19	形(かたち)	名詞	151.0
20	今日(きょう)	名詞	149.9

2.4.職場編資料で対数尤度比(LLR)の低い語

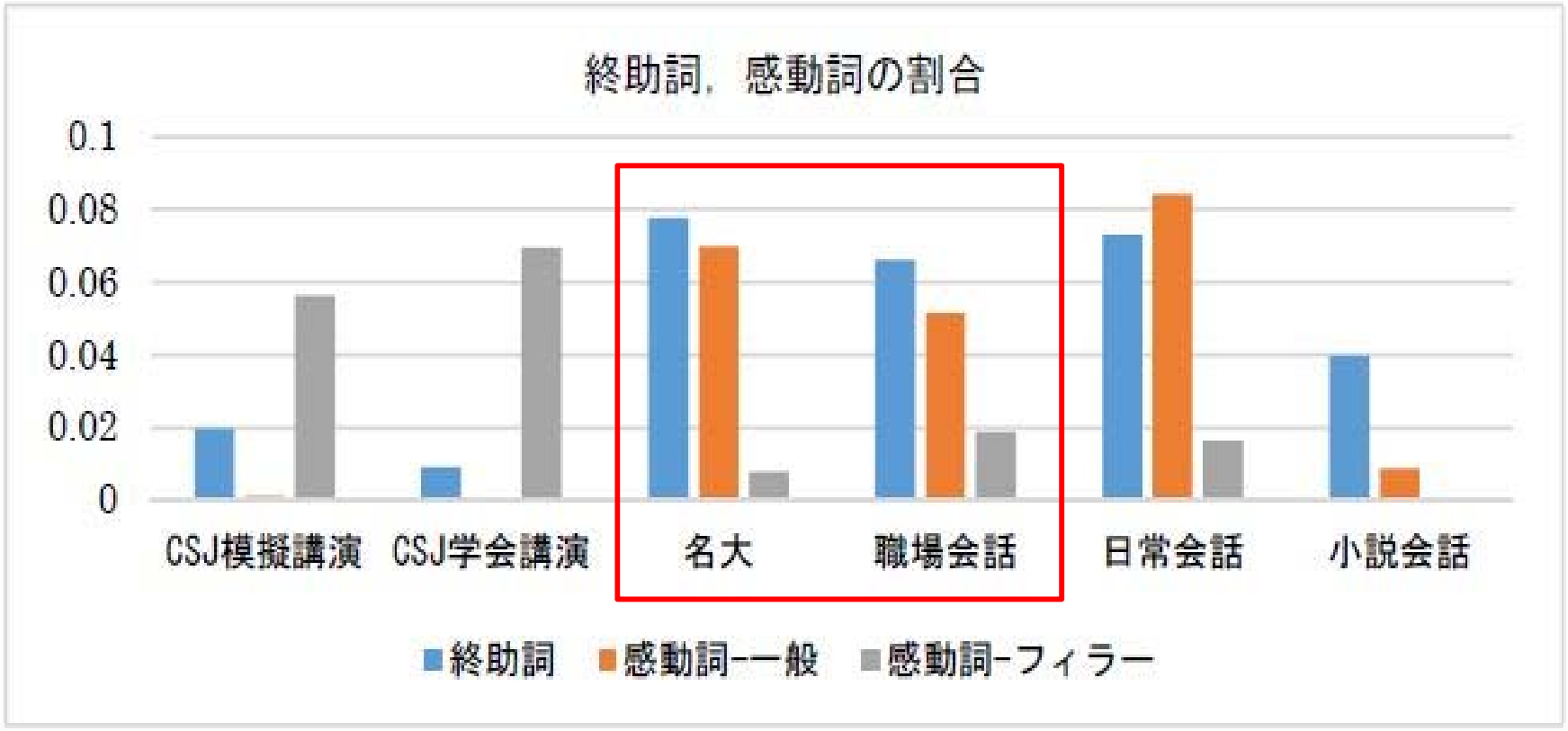
rank	語彙素	品詞	LLR
1	うん	感動詞	-3693.2
2	何(なに)	代名詞	-758.4
3	か	副助詞	-735.1
4	さ	終助詞	-542.1
5	私(わたし)	代名詞	-364.0
6	や	助動詞	-316.9
7	ふん	感動詞	-314.7
8	凄い	形容詞	-281.8
9	へえ	感動詞	-275.9
10	然う然う	感動詞	-228.7

rank	語彙素	品詞	LLR
11	って	副助詞	-206.8
12	じゃん	終助詞	-185.2
13	本当	名詞	-149.7
14	そう	副詞	-123.7
15	の	終助詞	-123.6
16	た	助動詞	-122.0
17	人(ひと)	名詞	-109.0
18	だ	助動詞	-104.8
19	そんな	連体詞	-101.5
20	もう	副詞	-97.3

2.4.品詞の割合の比較

品詞	職場編	名大
名詞	20.03	17.13
代名詞	3.55	4.16
動詞	11.61	11.51
形容詞	2.50	3.13
形状詞	0.82	1.11
副詞	4.54	5.10
連体詞	1.10	1.35
接続詞	0.80	0.66
感動詞	6.95	7.78
接頭辞	0.53	0.54
接尾辞	1.92	1.64
助詞	31.95	33.13
助動詞	13.70	12.77
計	100.00	100.00

2.5.終助詞，感動詞の割合



2.6.感動詞-フィラーの特徴語（職場編資料）

語彙素	品詞	LLR
えー	感動詞-フィラー	923.9
あの	感動詞-フィラー	852.7
えーと	感動詞-フィラー	657.5
んー	感動詞-フィラー	47.6
まー	感動詞-フィラー	30.3
その	感動詞-フィラー	22.9
うんと	感動詞-フィラー	14.7
あーと	感動詞-フィラー	14.3
うー	感動詞-フィラー	0.8
おー	感動詞-フィラー	0.1
あー	感動詞-フィラー	-0.1
と	感動詞-フィラー	-1.5
いー	感動詞-フィラー	-15.6

2.7.代名詞の特徴語（職場編資料）

語彙素	品詞	LLR
此れ	代名詞	422.1
此処（ここ）	代名詞	92.6
此方（こちら）	代名詞	91.7
僕	代名詞	88.4
私（わたくし）	代名詞	34.7
何方（どなた）	代名詞	24.3
我々	代名詞	15.6
てまえ	代名詞	11.0

語彙素	品詞	LLR
何（なに）	代名詞	-758.4
私（わたし）	代名詞	-364.0
貴方（あなた）	代名詞	-22.2
うち	代名詞	-13.7
彼	代名詞	-12.6
彼女	代名詞	-11.8

- 対数尤度比（LLR）が0.1%で有意である値10.83（高見(2003:89)）を目安に抽出。

2.8.形容詞の特徴語（職場編資料）

語彙素	品詞	LLR
宜しい	形容詞	36.4
狡(ずる)い	形容詞	13.6
望ましい	形容詞	11.3
ぼろい	形容詞	11.0

語彙素	品詞	LLR
凄い	形容詞	-281.8
面白い	形容詞	-74.2
可愛い	形容詞	-53.6
怖い	形容詞	-42.8
美味しい	形容詞	-38.3
楽しい	形容詞	-31.1
痛い	形容詞	-23.8
眠い	形容詞	-15.4
嬉しい	形容詞	-15.0
煩(うるさ)い	形容詞	-14.7
若い	形容詞	-14.1
難しい	形容詞	-13.9
遠い	形容詞	-13.2

2.9.副詞の特徴語（職場編資料）

語彙素	品詞	LLR
まあ	副詞	126.9
例えば	副詞	109.6
宜しく	副詞	61.4
直接	副詞	42.4
一寸	副詞	32.8
きちんと	副詞	26.5
ちゃっちゃっ	副詞	25.5
若し	副詞	23.3
たた	副詞	19.7
先ず	副詞	19.1

語彙素	品詞	LLR
そう	副詞	-123.7
もう	副詞	-97.3
こう	副詞	-50.9
全然	副詞	-44.7
多分	副詞	-37.0
余り	副詞	-24.2
急度	副詞	-18.5
良く	副詞	-14.9
結構	副詞	-14.6
ゆっくり	副詞	-11.7

2.10.数詞

- 職場編資料に現れる数詞64語のうち，対数尤度比（LLR)10.83以上のものが22語。体数尤度比（LLR）が-10.83より小さい語はなかった。

3.まとめ

- 同じ会話データでも場面が違うと，様々な部分に違いが現れる。
- 特徴語による比較が有効。

参考文献

- 金愛欄・桐生りか・近藤明日子・田中牧郎(2008)「一般向け専門用語」抽出の試みー医療用語を例にー, 『日本語学会2008年度春季大会予稿集』199-206. http://pj.ninjal.ac.jp/byoin/tyosa/corpus/zuhyo/corpus_siryo2.pdf
- 高見敏子(2003)「高級紙語」と「大衆紙語」のcorpus-drivenな特定法『北海道大学大学院国際広報メディア研究科言語文化部紀要』44.
- 山崎誠(2017)レジスター・位相の違いによる会話文の語彙的多様性, 言語資源活用ワークショップ発表論文集, 278-289
  info:doi/10.15084/00001529

- ご清聴ありがとうございました。