2018.03.19
オンライン検索システム「中納言」講習会

『日本語話し言葉コーパス』（CSJ），『名大会話コーパス』概説

国立国語研究所 言語変化研究領域　山崎誠

★概要

- 1.『日本語話し言葉コーパス』(CSJ)とは
- 2. CSJの構成と内容
- 3.『名大会話コーパス』とは
- 5.『名大会話コーパス』の構成と内容
- 7. データについて

本資料には，コーパスの説明として必要な情報を盛り込みましたが，時間の関係で口頭による説明は要点のみになります。

★1.『日本語話し言葉コーパス』(CSJ)とは

- 国立国語研究所・情報通信研究機構(旧通信総合研究所)・東京工業大学 が共同開発
- 1999年構築開始，2004年6月公開
- 約661時間，約753万語(短単位)を収録
- 質・量ともに世界最高水準の話し言葉データベース
- 略称 CSJ (Corpus of Spontaneous Japanese)　**Corpus of Spoken Japaneseではない**

1.1. CSJの特徴

- 音声の公開（以下のURLでサンプル音声が聞ける）
- http://pj.ninjal.ac.jp/corpus_center/csj/sample.html

- ２種類の解析単位（短単位と長単位）
- 豊富なアノテーション

★1.2. CSJの公開方法

- アカデミック利用　54,000円（税込）。ただし，学生（個人）の場合は，27,000円（税込）。
- 一般利用　270,000円（税込）
- 商業利用（統計情報利用）　540,000円（税込）
- 『中納言』版CSJ　無料。ただし機能が制限される。

２．CSJの構成・内容

2.1. 音声のタイプと時間

音声のタイプ	タイプ	話者数（異なり）	講演数	時間
学会講演	独話	819	987	274.4
模擬講演	独話	**594	1,715	329.9
その他の講演	独話	***16	19	24.1
学会講演インタビュー	対話	*(10)	10	2.1
模擬講演インタビュー	対話	*(16)	16	3.4
課題指向対話	対話	*(16)	16	3.1
自由対話	対話	*(16)	16	3.6
再朗読	朗読	*(16)	16	5.5
朗読	朗読	*(248)	507	15.5
計		1,417	3,302	661.6

*( )内は全員が学会講演話者もしくは模擬講演話者としてもカウント。
**10名は学会講演話者としてもカウント。
***2名は学会講演話者としてもカウント

★2.2. 形態論的単位数

音声のタイプ	短単位数	長単位数	%長単位
学会講演	3,279,364	2,654,823	81.0
模擬講演	3,605,729	3,115,302	86.4
その他の講演	282,728	239,989	84.9
朗読と再朗読	207,478	172,216	83.0
対話	149,826	131,544	87.8
全体	7,525,125	6,313,874	83.9

- 語数は第1刷に基づくもの。

★2.3. 学会講演

- 理工学，人文，社会の3 領域におよぶ種々の学会における研究発表のライブ録音。
- 講演時間は10 分から25 分程度が大半。
- 学会講演の多くをしめる理工学系の学会では，男性の大学院生であることが多い。
- 発話スタイルは，あらたまり度が高い。

★2.4. 摸擬講演

- できるだけ年齢と性別のバランスをとった一般話者による，日常的話題についての講演。
- 話者の大部分は人材派遣会社からの派遣。
- あらかじめ指定された3つのテーマ（例えば「人生で一番嬉しかったこと」「人生で一番悲しかったこと」「私の住んでいる街」等）に基づいて，具体的な講演内容を決めてタイトルをつけて話す。
- 1 講演の長さは10～15 分程度。
- 聞き手は3，4 名。
- 発話スタイルは，学会講演よりもくだけている。

2.5. その他の講演

- 研究機関が一般聴衆を対象に企画した連続講演会の講演音声。対象は歴史や民俗学など
- 国語研究所が一般聴衆むけに開催した講演会の講演音声，および国語研究所員を聴衆とした識者による講演（1 講演のみ）
- 専門学校における日本語教師養成関係の講義音声

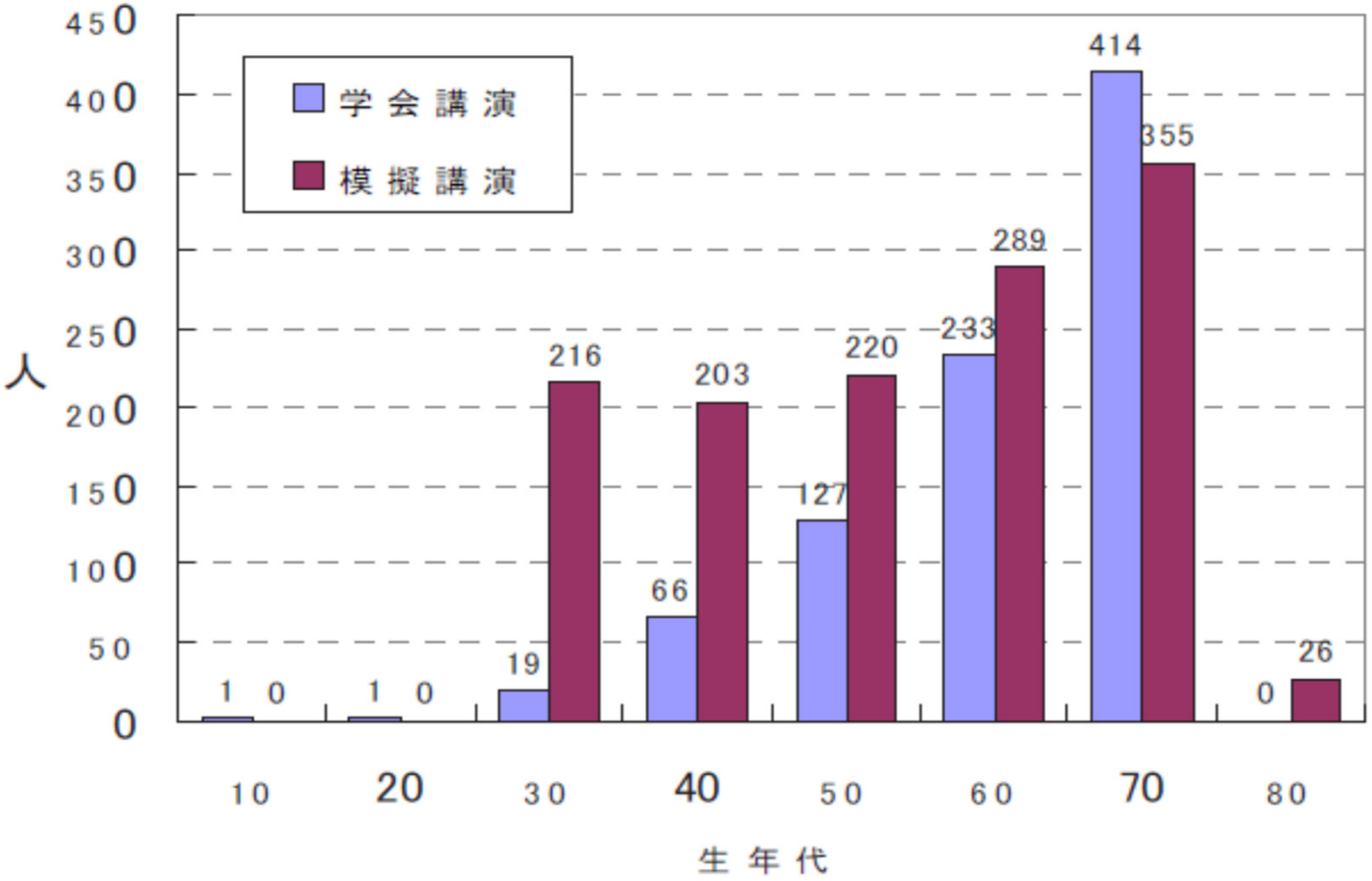
2.6. 朗読

- 自発的な音声の特徴を明らかにする比較データ
- 「朗読音声」と「再朗読音声」の2種類
- 朗読音声は模擬講演話者の一部が，書き言葉のテキストを朗読したもの
- 再朗読音声は，学会講演ないし模擬講演として収録された音声の転記テキストを同一の話者が朗読した音声。フィラーや言い直しも朗読の対象。
- 再朗読の話者は，学会講演話者から選ばれた10 名と模擬講演話者から選ばれた6 名の合計16 名。

2.7. 対話

- 「学会講演インタビュー」「模擬講演インタビュー」「課題指向対話」「自由対話」の4 種類。
- インタビューは，学会講演ないし模擬講演の内容に関する対話形式。インタビュアーは20代と30 代の女性各1 名。
- 課題指向対話は，特定のタスクを与えて行った2名の間の対話。
- 自由対話では，話題の制約なしに行った10分程度の対話。

2.8. 話者の属性(生年:延べ人数)

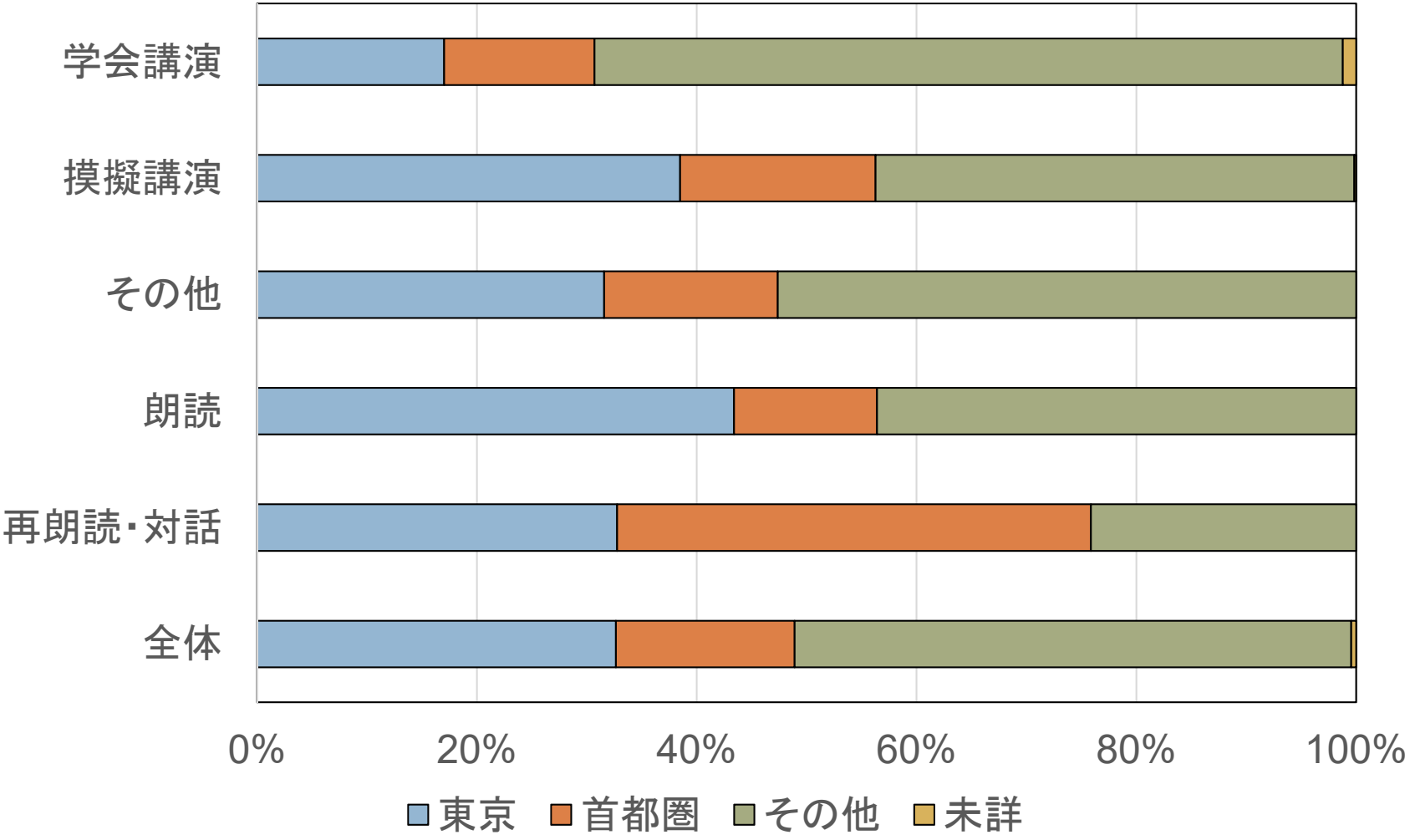


国立国語研究所報告No.124『日本語話し言葉コーパスの構築法』p.7

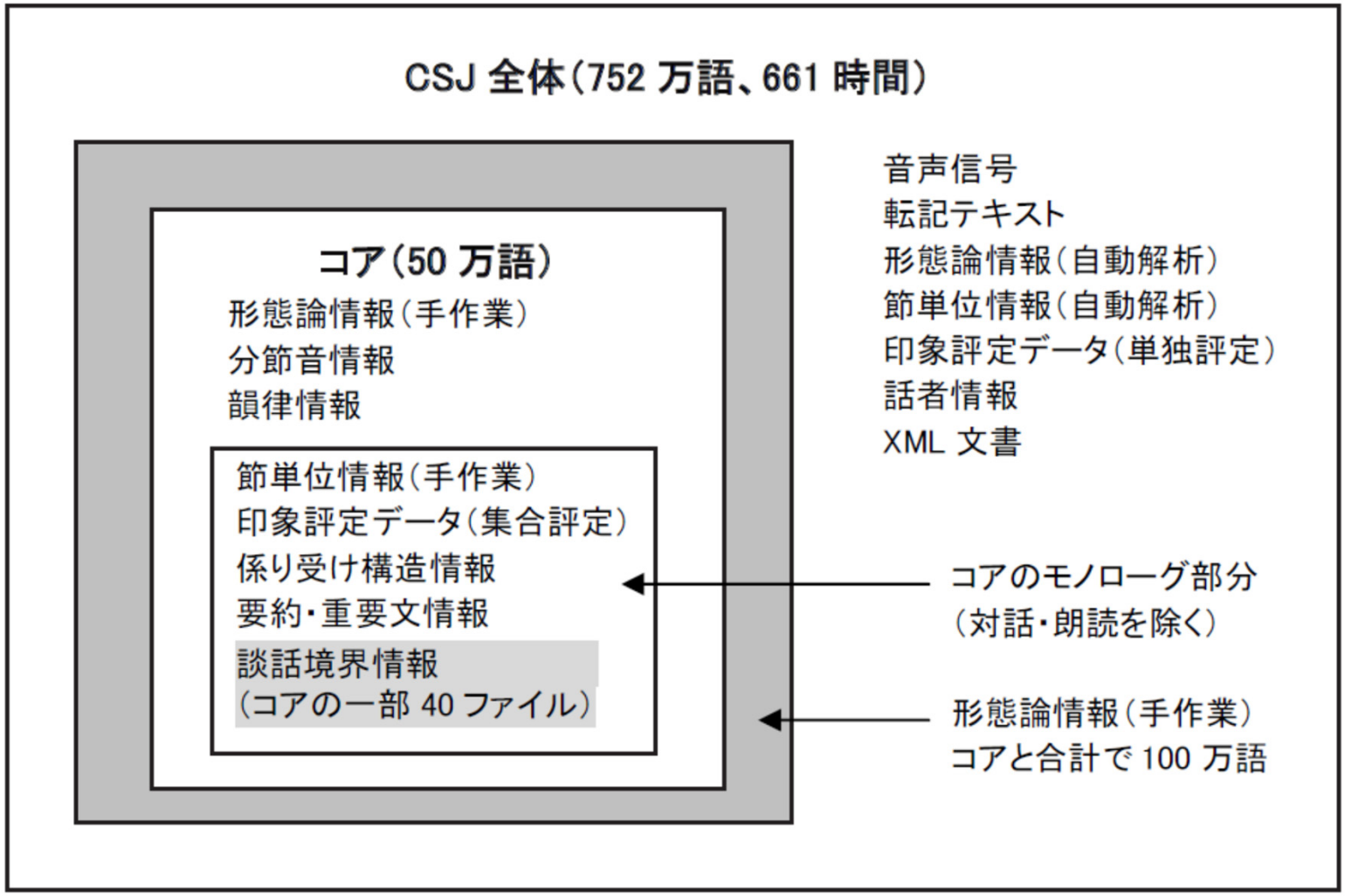
2.9. 話者の属性：性別（延べ）

性別	学会講演	摸擬講演	その他	朗読	再朗読	対話	全体
女	173	910	9	252	8	29	1381
男	814	805	10	255	8	29	1921
計	987	1715	19	507	16	58	3302

2.10. 話者の属性:出身地(延べ)



2.11. CSJの付加情報の構造



国立国語研究所報告No.124『日本語話し言葉コーパスの構築法』p.9

★2.12. CSJコア

- 形態素解析等の付加情報は基本的に自動で行っているが，人手で解析精度を高め，より多くの付加情報を付与した部分。約50万語。

音声のタイプ	女性話者	男性話者	計
学会講演	24	46	70
摸擬講演	54	53	107
対話	9	9	18
朗読	3	3	6
計	90	111	201

★2.13. CSJ語彙表，語数表

- 中納言データ Ver. 2018.3.1 に基 づく語彙表，語数表を以下のURLで公開しています。

- http://pj.ninjal.ac.jp/corpus_center/csj/chunagon.html#data

★3. 名大会話コーパスとは

- 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。

★3.1. 名大会話コーパスの利用法

- (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。

★4.『名大会話コーパス』の構成と内容

4.1. 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？

★4.2. 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。

★4.3. 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。

4.4. 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。

★4.5. 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。

4.6. 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。

4.7. 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

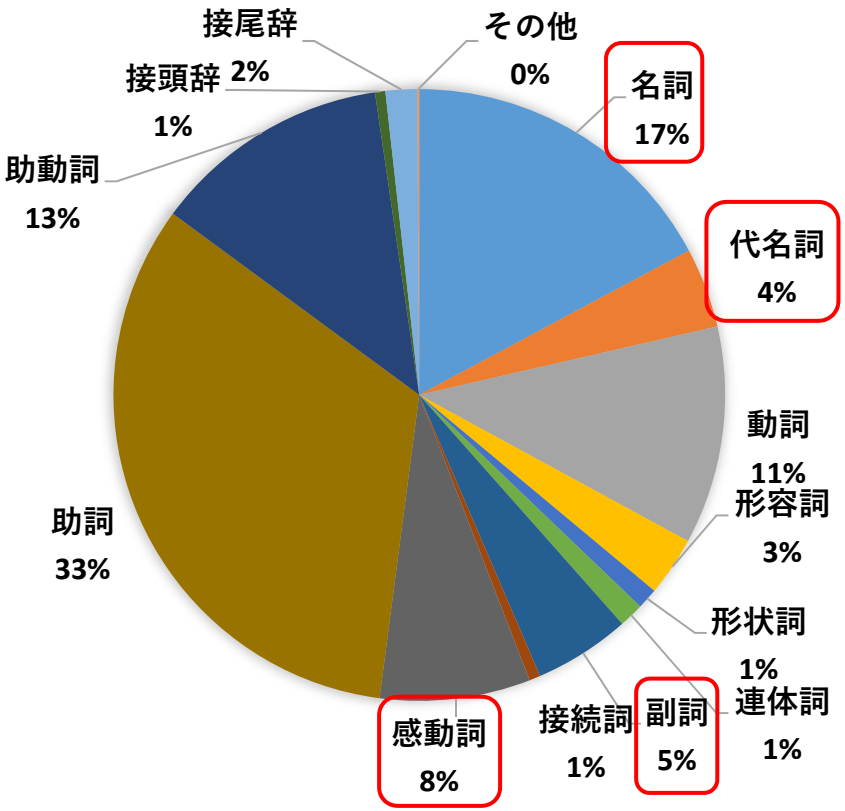
4.8. 話者の属性：収録時間

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

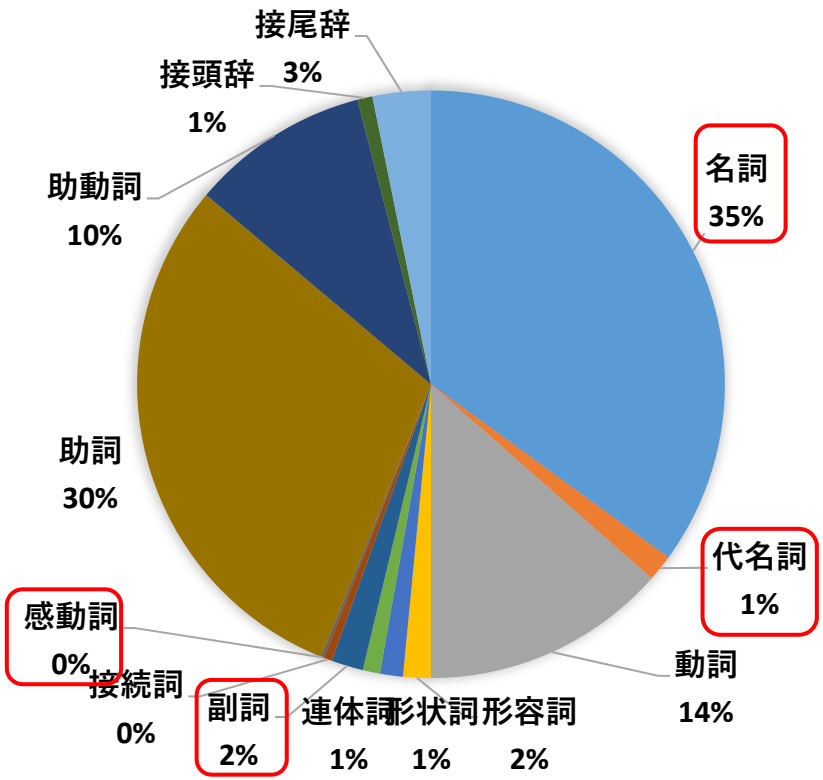
- 30分～60分の会話が多い。

4.9. 品詞の割合（延べ語数）

名大会話コーパス



現代日本語書き言葉均衡コーパス（BCCWJ)



4.10.上位語

順位	名大会話			BCCWJ		
1	ダ	だ	助動詞	ノ	の	助詞-格助詞
2	ウン	うん	感動詞-一般	ニ	に	助詞-格助詞
3	タ	た	助動詞	テ	て	助詞-接続助詞
4	テ	て	助詞-接続助詞	ハ	は	助詞-係助詞
5	ネ	ね	助詞-終助詞	ダ	だ	助動詞
6	ノ	の	助詞-準体助詞	ヲ	を	助詞-格助詞
7	カ	か	助詞-副助詞	タ	た	助動詞
8	ト	と	助詞-格助詞	スル	為る	動詞-非自立可能
9	デ	で	助詞-格助詞	ガ	が	助詞-格助詞
10	ノ	の	助詞-格助詞	ト	と	助詞-格助詞
11	モ	も	助詞-係助詞	デ	で	助詞-格助詞
12	ガ	が	助詞-格助詞	モ	も	助詞-係助詞
13	ニ	に	助詞-格助詞	イル	居る	動詞-非自立可能
14	ハ	は	助詞-係助詞	マス	ます	助動詞
15	ナニ	何	代名詞	ノ	の	助詞-準体助詞

- 「うん」「ね」など会話によく用いられる語が上位に来ている。

５．データについて

5.1. データの精度

- 形態論情報に基づく解析やアノテーションは基本的に自動的に付与されている。
- CSJの形態素解析の精度は約97～98%。（コア部分は約99.9%）。
- コーパスの分析は，一定の誤差があることを前提として行う必要がある。

5.2. データの見方

- 比較しようとするデータのサイズが異なる場合、出現した生の頻度ではなく，per million wordなどの相対頻度で比較する必要がある。
- 分析の過程がブラックボックスにならないようにすることが望ましい。

参考文献

- 国立国語研究所(2006)『日本語話し言葉コーパスの構築法』
- http://pj.ninjal.ac.jp/corpus_center/csj/document.html