『名大会話コーパス』中納言版・ひまわり版公開データの作成

著者	柏野 和佳子, 西川 賢哉, 小磯 花絵
雑誌名	言語資源活用ワークショップ発表論文集
巻	1
ページ	324-335
発行年	2017
URL	http://doi.org/10.15084/00001488

『名大会話コーパス』中納言版・ひまわり版公開データの作成

柏野 和佳子（国立国語研究所音声言語研究領域）*
西川 賢哉（国立国語研究所コーパス開発センター）
小磯 花絵（国立国語研究所音声言語研究領域）

Supplemental Arrangement for Public Data Available in the Chunagon and Himawari Versions of "Nagoya University Conversation Corpus"

Wakako Kashino (National Institute for Japanese Language and Linguistics)
Ken'ya Nishikawa (National Institute for Japanese Language and Linguistics)
Hanae Koiso (National Institute for Japanese Language and Linguistics)

要旨

『名大会話コーパス』は，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」（研究代表者：大曾美恵子，平成 13 年度～15 年度）の一環として作成された，120 会話，合計約 100 時間の日本語母語話者同士の雑談を文字化したコーパスである。国立国語研究所に移管後，文字化テキストを公開し，続けて『中納言』版，『ひまわり』版を作成し，公開している。

本稿では，『名大会話コーパス』の概要と特徴を述べる。また，『中納言』版，『ひまわり』版公開データの作成に際して行った，形態素解析結果の人手修正の内容について報告する。

１．はじめに

国立国語研究所の『日本語話し言葉コーパス』（CSJ）は，独話を主対象とするコーパスである。また，『現代日本語書き言葉均衡コーパス』（BCCWJ）及び，『国語研日本語ウェブコーパス』（NWJC）は，いずれも書き言葉のコーパスである。日常会話場面を対象とした大規模な『日本語日常会話コーパス』の構築は，国立国語研究所の機関拠点型基幹研究プロジェクト「大規模日常会話コーパスに基づく話し言葉の多角的研究」（プロジェクトリーダー：小磯花絵）により，着手されたところである(小磯ほか 2017)。そのような状況の中，現時点で広く利用可能である自然会話のコーパスが『名大会話コーパス』である。

『名大会話コーパス』は，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」（研究代表者：大曾美恵子，平成 13 年度～15 年度）の一環として作成された，120 会話，合計約 100 時間の日本語母語話者同士の雑談を文字化したコーパスである。国立国語研究所に移管後，文字化テキストを公開している。さらに，「大規模日常会話コーパスに基づく話し言葉の多角的研究」のプロジェクトにおいて，文字化テキストを対象に，形態素解析用辞書『UniDic』 と形態素解析器『MeCab』を用いて形態論情報（短単位）を自動付与し，メタ情報として発話者の属性（性別・年代・出生地など）と会話の情報（収録日・収録場所など）を整理した上で，オンライン検索システム『中納言』，及び，全文検索システム『ひまわり』にて 2016 年 12 月より一般公開

* waka ＠ninjal.ac.jp

している。

本稿では，『名大会話コーパス』の概要と特徴を述べる。また，『中納言版』，『ひまわり版』公開データの作成に際して行った，形態素解析結果の人手修正の内容について報告する。

２．『名大会話コーパス』概要と特徴

２．１　『名大会話コーパス』の概要

はじめに，文字化テキストの例(data001 の冒頭)を示す。

> ＠データ１（約３５分）
> ＠収集年月日：２００１年１０月１６日
> ＠場所：ファミリーレストラン
> ＠参加者 F107：女性３０代後半、愛知県幡豆郡出身、愛知県幡豆郡在住
> ＠参加者 F023：女性４０代後半、岐阜県出身、愛知県幡豆郡在住
> ＠参加者 M023：男性２０代前半、愛知県西尾市出身、西尾市在住
> ＠参加者 F128：女性２０代前半、愛知県西尾市出身、西尾市在住
> ＠参加者の関係：英会話教室の友人
> F107：＊＊＊の町というのはちいちゃくって、城壁がこう町全体をぐるっと回ってて、それが城壁の上を歩いても１時間ぐらいですよね。
> F023：１時間かからないぐらいだね。
> ４、５０分で。
> F107：そうそう。
> ほいでさあ、ずっと歩いていたんだけど、そうすと上から、なんか町の中が見れるじゃん。
> あるよね。
> ほいでさあ、なんか途中でワンちゃんに会ったんだね。
> （ふーん）散歩をしてるワンちゃんに会ったんだ。
> F023：城壁の上をやっぱ観光客なんだけどワンちゃん連れてきてる人たち結構多くて。

上記のような会話データが全部で 129 会話(data001～data129)ある。それぞれに以下の情報が付与されている。以下，これら情報の内訳を概観する。

データ情報：収録時間，収録年月日，収録場所
参加者情報：性別，年代，出身地，居住地
参加者の関係：参加者間の関係

２．１．１　データ情報：収録時間，収録年月日，収録場所

まず，表 1 に収録時間，表 2 に収録年，表 3 に収録場所の内訳を示す。収録時間は，31～60 分のものが最も多く，平均は 47 分である。収録年月日は 2001 年 10 月 16 日～2003 年 2 月 17 日の間である。2001 年のデータが最も多い。収録場所は，テキストには例えば，「レストラン」「うどん屋」「喫茶店」「○○の実家」「○○宅」といったように示されているが，

それらを大まかに分類しなおしてみると，表 3 の通り，飲食店，自宅，大学で多く収録されている。なお，場所については，二重分類 3 件を含んでいる。

表 1　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

表 2　収録年

年	件数
2001年	78
2002年	48
2003年	3
合計	129

表 3　収録場所

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

２．１．２　参加者情報：性別，年代，出身地，居住地

次に，表 4 に年代別の性別，表 5 に出身地，表 6 に居住地の内訳を示す。性別は女性が多い。また，年代は 20 代が最も多い。出身地と居住地は，テキストには都道府県に加え市まで示してあるものもある。また，途中の引っ越し歴の記載があるものもある。出身地，居住地ともに中部が多いが，それ以外もある。

表 4　年代別の性別

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

表 5　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

表 6　居住地

居住地	人数
北海道	18
東北	1
関東	49
中部	120
近畿	7
中国・四国	1
九州・沖縄	0
海外	2
合計	198

２．１．３　参加者の関係：参加者間の関係

最後に，表 7 に参加者の関係，表 8 に会話の参加者の人数の内訳を示す。参加者の関係は，テキストには例えば，「英会話教室の友人」「アルバイトの友人」「中学の同級生、F106 の母親」「F154 と F130 は友人。M004 は初対面の人。」といったように示されているが，それらを大まかに分類しなおしてみると，表 7 の通り，同級生，友人，家族，先輩，同僚の関係が多く収録されている。なお，関係については，重複分類 12 件を含んでいる。表 8

より，本データのほとんどが 2 名の対話であることがわかる。

表 7　参加者の関係

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

表 8　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

２．２　『名大会話コーパス』の特徴

２．２．１　上位語

書き言葉の代表として『現代日本語書き言葉均衡コーパス』（以下，『BCCWJ』）の語彙表を用いて，『名大会話コーパス』の話し言葉としての特徴を概観する。まず，上位語の比較を表 9 に示す。

表 9　『名大会話コーパス』と『BCCWJ』の上位語の比較

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

表 9 において赤四角で囲んだ赤字の箇所で示した語が上位語であることが，すなわち『名大会話コーパス』の話し言葉としての特徴を表すと思われるものである。

２．２．２　品詞の分布

続いて，同じく『BCCWJ』の語彙表を用いて，『名大会話コーパス』と品詞の分布を比較する。

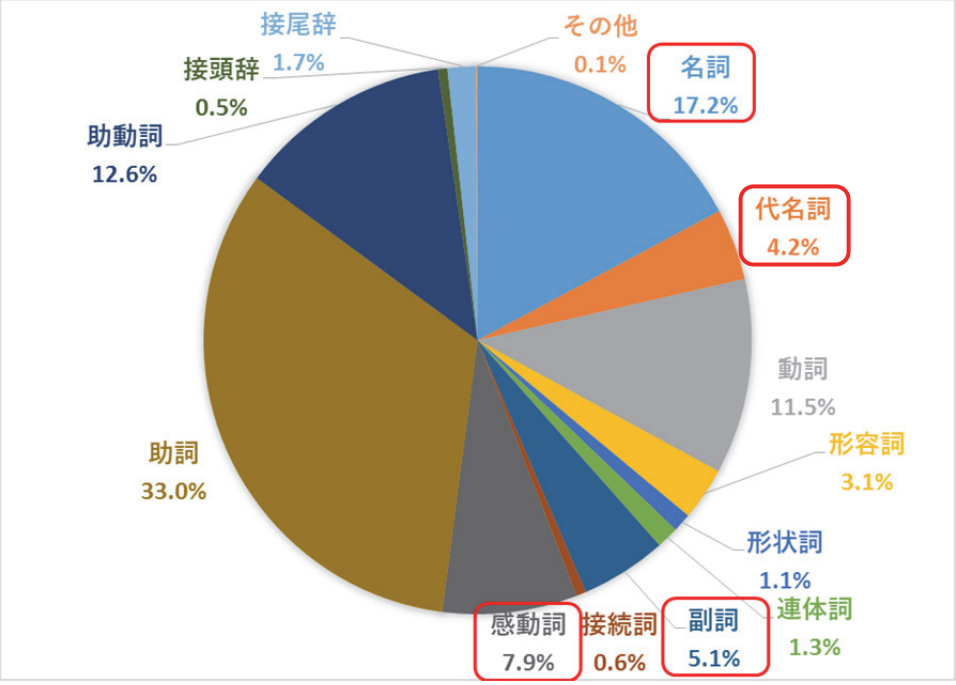

図 1　『名大会話コーパス』の品詞の分布

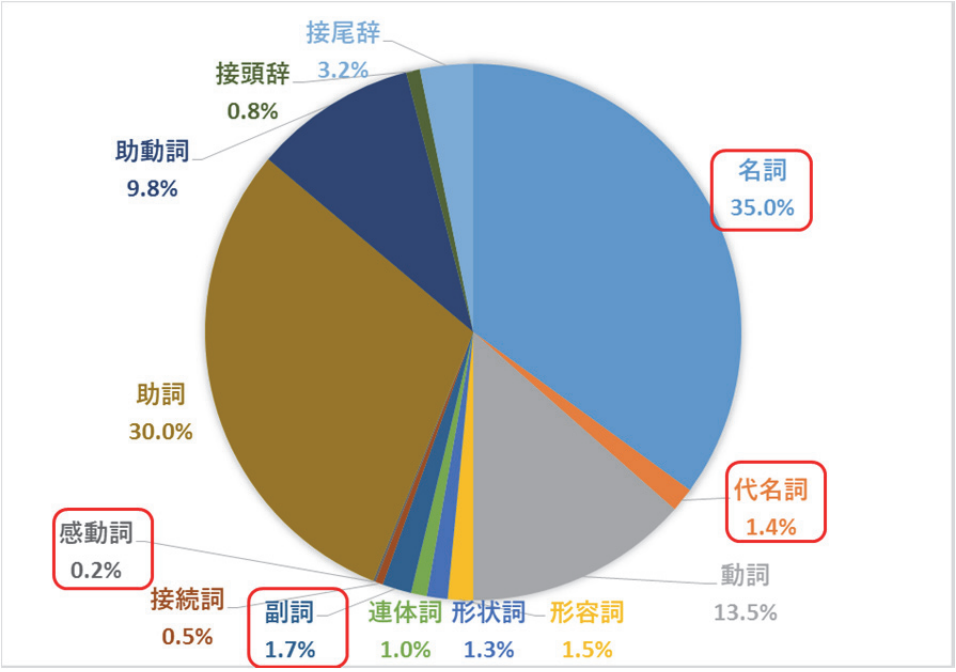

図 2　『BCCWJ』の品詞の分布

図 1 と図 2 を比べると，赤の四角形で強調している通り，『名大会話コーパス』は名詞が少なく，感動詞，副詞，代名詞が多い。ここに話し言葉としての特徴が表れていると考えらえる。

２．２．３　特徴語

Fujimura et al.(2012)では，『名大会話コーパス』にみられる話し言葉の特徴を，書き言葉の代表として『現代日本語書き言葉均衡コーパス』（モニター公開データ 2008/2009）と比較して述べている。LLR（対数尤度比）により，特徴語 10 語として「うん，だ，ね，の，か，そう，って，た，何，言う」を挙げている。表 9 で示した上位語にないものが「そう」と「言う」である。この「言う」は，次のような例（下線，太字部分）で頻出するものである。

午前中はずーっと部屋に、部屋って<u>いう</u>か玄関に入ってたんだけどー　(data019)

なお，この「っていうか」は「てか」という短縮形でも頻出する（以降にその用例を示す)。いずれも話し言葉の特徴を示す語であると言える。

山崎(2016)では，『名大会話コーパス』の特徴を『日本語話し言葉コーパス』（CSJ）の学会講演と模擬講演，『BCCWJ』の小説会話文と比較し，示している。LLR により『名大会話コーパス』の特徴語として示しているのは，例えば，10 代では「超」「うざい」などの若者語，60 代では「いらっしゃる」「おっしゃる」といった敬語や，「何しろ」「随分」などの特定の副詞である。

本稿では，いわゆる「話し言葉」であるため，『BCCWJ』では用例が得にくいが，『名大会話コーパス』で頻出することが期待される語を特徴語の例として，次の a)〜h)の 8 語を取り上げる。以下に示す方法にて，『中納言』で検索をした。その結果得られた検索結果数と用例とを示す。

a)　微妙　156 件（語彙素「微妙」で検索）※ほか，「びみょー」1 件あり
b)　やば　168 件（文字列「やば」で検索）※「やば」「やばい」同時に検索
c)　まじ　197 件（語彙素「まじ」で検索）※「まじか」は 1 例のみ
d)　無理　273 件（語彙素「無理」で検索）
e)　てか、　60 件（文字列「てか､」検索）※ほか，「､」以外にも用例あり
f)　すごい＋形容詞　344 件（書字形出現形「すごい」＋形容詞で検索）
g)　うける　37 件（語彙素「受ける」の終止形で検索）
h)　みたいな 473 件（文字列「みたいな[､。？]」で検索）※「！」はなかった。

なお，以上に示す検索結果数は，当該語の「話し言葉」ならではの用法例の正確な件数ではない。検索もれ，あるいは，別語，別用法の例が少々混じっている。より正確に検索するためには，語彙素検索と文字列検索とを併用し，さらに検索結果を絞り込むことが望ましい。

しかしながら今回の検索方法でも，下記に示す通り，当該語の「話し言葉」ならではの用例が得られることが確かめられる。

ドラえもんとかあ。ドラえもんは超うめー。ハットリ君とかあ。ハットリ君はそれは	微妙	。めっちゃ微妙。もうね、もうね、みんなね、みんながねすごい頭ん中	data005
おる？この子は、Ｅ短の子だよ。あっ、そうなんだー、	微妙	、微妙。うんそういうのばっかり。はい、ん、これはだめだな。うん	data077
何かね、ストレス感じると食べちゃうのね、すごい。うんうんだから絶対	やばい	。これ４年生の二の舞になると思って、ちょっと制限しようと思ってるん	data003
だけど、なかなか時間がないんだよね。ねーあたしもだよ。	やばー	い。あっ、ＴＯＥＩＣさ、こないだあったけど、一番初めに受けたやつさ、	data072
受かっちゃったもん、最初。５級だったしね、一番最初受けたの。	まじ	？６年のときに５級。そっか、そんなん、だって、小学校	data011
のなんなの、これ。知らないよ。大体電池がないじゃん。うそ、	マジ	？なんか電池があと２つみたいになってる。本当？うーう、携帯でも	data046
んだ、あそこで？そうそうそうそうそうそうこえーなー。私、あのカーペット系がもっと	無理	。あ、あれ、超酔う。あれ、なんかさ、なんかもう、抜けそう	data072
無理だよ。だからなんで。うーん。無理。だから、なんで。とにかく	無理	。それは、そういうことしたら、そういうことをやったっていうことを	data046
ゆくんだね。そうだねえ、ほんと。どんどんみんな大人になっちゃう、	てか、	社会人になっちゃう。うん。わせさいんときの先輩なんかもみんなもう決まって	data066
抽象画じゃないのよ。じゃないの。うんどこで見たの。	てか、	あの、日本に来たときの。昨日か、昨日かおといもテレビであった	data056
ないと、あれだね。１人でお鍋セットとかもらっても、うれしく	なくない	？そっかー。じゃ、友だち呼んで、お鍋パーティすればいい。うーん、そう	data055
そういうことは全然ないけど、それでもさ、そんなに聞きたい話じゃ	なくない	？うんうん、うんうんうん。全然。うん。でさ、何か、その、	data094
のー、なんか、中からしか選べないからとか言ってたよね。うん	すごい	(高い)んだろうねー。だってなんかゴージャスだからねー。うーんじゅうたん１つとって	data120
鼻血出そうだ。本気で出そうだ。んー、かっこいい。＊＊＊。	すごい	(かわいい)、この絵。すげーな才能のある人っていうのはすごい。うん。＊	data103
に言ったんだって。うんうんうんうんうんすごい懇願したんだって。	うける	ー。ほんで、私、ハッて思いついたんだ。うんあ、F114 ちゃんって	data066

でーとか、あ、君は日本文学専攻か、ふーん、とか言って。	受ける	ー。うーん話しかけやすい雰囲気なんじゃん。なのかね。うん困っちゃうね。	data065
いろいろ言ってたのー。うん日本語教師とかそっち系に進むには	みたいな、	でも、結構ね。うーんでも日本じゃ無理だってさ。うん、だから	data011
てー。うんいいよねー結構いろんな人にー、こういうの、どう？	みたいな。	うんだからさ、写真で生きていけるって、F141 の場合。うん。だから	data123
じゃない？うん雰囲気。今日、さむ。な、何となく寒そう、	みたいな？	うん。雪が降ったときとか、雪のお水がまだ解けきらないとき	data085

３．形態素解析結果の人手修正の内容

３．１　人手修正した範囲

『名大会話コーパス』を，オンライン検索システム『中納言』，及び，全文検索システム『ひまわり』にて公開するに際し，形態素解析用辞書『UniDic』 と形態素解析器『MeCab』を用いて形態論情報（短単位）を自動付与し，その結果を人手修正した。129 会話全ての一部分（各会話 1,500 形態素以上）を目視で修正する範囲に定め，まずはその範囲内のものを修正した。加えて，全範囲に対する一括修正も行った。『名大会話コーパス』の形態素数と人手修正した形態素数を表 10 に示す。記号・補助記号・空白を除いた形態素数に対し，人手修正で目視した作業範囲の形態素数は 26.8%にあたる。また，人手修正した形態素数は 2.5%にあたる。

表 10　『名大会話コーパス』の形態素数と人手修正した形態素数（短単位）

全形態素数	1,419,729
記号・補助記号・空白を除いた形態素数	1,131,891
人手修正で目視した作業範囲の形態素数	303,282
人手修正した形態素数	27,931

３．２　人手修正の内容

３．２．１　口語表現の誤解析の修正

以下に，口語表現に関して誤解析を修正した具体例を示す。例は左から，「対象」「テキスト（誤）」「修正後（正）」で示す。形態素の区切りは「｜」で示す。語彙素，品詞，活用形等は着目する部分のみ簡易表示する。

①「なん」

なん	そう｜なん：「何」｜だ	そう｜な：「だ」助動詞｜ん：準体助詞｜だ
なん	一緒｜なん：「など」｜じゃん。	一緒｜な：「だ」助動詞｜ん：準体助詞｜じゃん。
なん(なる)	男｜の｜人｜も｜便秘｜に｜なん：「何」｜だ｜ね。	男｜の｜人｜も｜便秘｜に｜な：「成る」連体形-省略｜ん：準体助詞｜だ｜ね。

②「そっか」「そやね」「てゆーか」

そっか	あ｜、｜そっ：副詞｜か｜そっ：「其処」代名詞｜か	あ｜、｜そっ：副詞｜か｜そっ：副詞｜か
そっか	そっ：「そう」副詞｜か：副助詞｜そっ：「そう」副詞｜かそっ：「貸す」｜か：終助詞	そっ:「そう」副詞｜か:終助詞｜そっ：「そう」副詞｜か：終助詞｜そっ：「そう」副詞｜か：終助詞
そやね	そや：「粗野」｜ね。	そ：副詞｜や：助動詞｜ね。
てゆーか	て：「で」接続詞｜ゆー：固有名詞一人名｜か：副助詞	て：副助詞｜ゆー：「言う」｜か：終助詞

③「やっとく」「やっとって」「やって」「おって」

やっとく	今日｜、｜やっと：副詞｜かん：「彼」｜と：格助詞	今日｜、｜やっ：「遣る」｜とか：「とく」助動詞｜ん：「ず」｜と：接続助詞
やっとって	３｜級｜の｜問題｜やっと:副詞｜って：副助詞	３｜級｜の｜問題｜やっ：「遣る」｜とっ：「とる」助動詞｜て：接続助詞
やって(だって)	３１｜日｜は｜休み｜やっ：「遣る」｜て。	３１｜日｜は｜休み｜やっ：助動詞｜て。
おって	同じ｜屋敷｜に｜おっ：「追う」｜て｜ね	同じ｜屋敷｜に｜おっ：「居る」｜て｜ね

④「しよー」「よ」「あーあ」「あーん」「えっと」

しよー	これ、｜どー：副詞｜し：「為る」｜よー：終助詞｜。	これ、｜どー：副詞｜しよー：「為る」｜。
よ	朝｜起き｜たら｜ちょっと｜頭痛｜いよ：感動詞	朝｜起き｜たら｜ちょっと｜頭｜痛い｜よ：終助詞
あーあ	あー：感動詞｜あっ：「有る」｜て：接続助詞｜思い｜ながら	あーあ：感動詞｜って：副助詞｜思い｜ながら
あーん	あー：感動詞｜ん：感動詞｜ラジオ｜って｜いう｜か	あーん：感動詞｜ラジオ｜って｜いう｜か
えっと	あたしはー、｜えっ：感動詞｜と：格助詞	あたしはー、｜えっと：「えーと」感動詞

⑤「あら そう」「こら」「あの」「いいやん」

あら そう[あいずち]	あらそう：「争う」｜。	あら：感動詞｜そう：副詞｜。
こら	うん｜こら：感動詞｜やばい	うん｜こら：代名詞｜やばい
あの	あの：感動詞｜人｜も？	あの：連体詞｜人｜も？
いいやん	いいや：感動詞｜ん：感動詞｜、	いい：「良い」｜やん：終助詞｜、

⑥「ねーねー」「ねー」「とか」

ねーねー	ねー：終助詞｜ねー：終助詞｜、	ねー：感動詞｜ねー：感動詞｜、
ねー	すごく｜ねー：「無い」｜。	すごく｜ねー：終助詞｜。
とか	おおーっと：感動詞｜か｜言っ｜て	おおーっ：感動詞｜と：格助詞｜か｜言っ｜て

とか	忙しかっ｜た｜ん｜じゃ｜ない｜か｜な｜っ：補助記号｜と：格助詞｜か	忙しかっ｜た｜ん｜じゃ｜ない｜か｜な｜っと：格助詞｜か

⑦「と」

と[接続助詞]	～し｜ない｜と：格助詞｜。｜	～し｜ない｜と：接続助詞｜。｜
と[格助詞]	歩い｜て｜登る｜と:接続助詞｜おっしゃる｜から	歩い｜て｜登る｜と：格助詞｜おっしゃる｜から

『日本語日常会話コーパス』でこれらの口語表現を全てそのまま表記するとは限らないが，口語表現を積極的に採用する方針であり，これらの修正履歴は今後の解析精度向上のために参考にしていく。

３．２．２　発音の誤認定の修正

発音の誤認定のタイプは主に３つに分けられる。１つ目は「数字」である。以下では一例しか示さないが，１～９すべての数字の読みについて複数ある読みの中から適切なものを自動判定することは難しいため，多く誤認定が生じている。２つ目は「清濁」である。これも多くの誤認定が生じている。このうち，以下の「座布団」のような連濁を自動判定するのには限界がある。しかしながら，その次の「橋の上」のようなものが語頭で「バシ」と濁音になることを解析時に避けることは可能であると考え，その対応を検討中である。３つ目は「音訓」である。熟語が湯桶読みや重箱読みの場合，訓が複数ある場合，熟字訓がある場合などに選択誤りがある。しかし，これらも辞書を整備することで少しでも解析精度を上げることを考えている。

以下に例を示す。着目する箇所の発音をカタカナで表記する。

①数字

一	１　イチ｜個	１　イッ｜個
四	４　ヨン｜人	４　ヨ｜人

②清濁

半濁音	５｜分　プン	５｜分　フン
濁音	座　ザ｜布団　フトン	座　ザ｜布団　ブトン
濁音	橋　バシ｜の｜上｜から	橋　ハシ｜の｜上｜から

③音訓

音訓	洗濯｜物　ブツ	洗濯｜物　モノ
音訓	大　ダイ｜掃除	大　オオ｜掃除
音訓	紅　コー｜ショウガ	紅　ベニ｜ショウガ
訓と訓	小麦｜粉　コナ	小麦｜粉　コ
訓と訓	いつ｜の｜間　アイダ｜にか	いつ｜の｜間　マ｜にか
訓と熟字訓	お｜父　チチ｜様｜と｜お｜母　ハハ｜様	お｜父　トウ｜様｜と｜お｜母　カア｜様

以上のほか，読みが複数あり，音声がないと正誤の判断がつかないものもある。例えば，

「その他（ソノタ／ソノホカ）」「毎年（マイネン／マイトシ）」「白髪（ハクハツ／シラガ）」のようなものである。なお，本データでは，「ソノタ」「マイトシ」は統一されているが，「ハクハツ／シラガ」は統一されていない。

実は，『BCCWJ』のときに整備した発音の後処理(伝ほか 2002)を今回は実施していない。そのため誤認定が多くあった。今後，『日本語日常会話コーパス』では，発音の後処理を積極的に導入する予定でいる。また，『日本語日常会話コーパス』でも漢字仮名交じりで表記するため，発音が一意に同定できないケースも生じるが，後処理として発音を確認する工程を設けることで正確な発音を保証する(川端ほか 2017)。

３．２．３　その他の修正

Unidic に登録のないオノマトペが多く出てきた。例えば，「ぴしゃーっ」「ビヨーン」「ごっしゃごしゃ」「ずきっ」などである。これらはただちに「オノマトペ」としての登録はせず，今回は「新規未知語」として今後の課題としている。

また，長音や撥音がそのまま語の途中に記述されているものが少なくなかった。例えば，「吉祥ー寺」「ひどーい」「つらーい」「なさーい」「すっげー」などである。このようなものは自動解析で語の途中で切られてしまう。そこで，人手修正した範囲内で見つけたものは一短単位に修正し，「新規未知語」としている。しかしながら，「吉祥ー寺」のようなものは同じものが複数例はないが，あとの４語は複数例あるものである。人手修正で見逃してしまったものは，語断片の誤解析のままとなっている。例えば，次の通りである。

ひどーい	ひ：「日」｜どー：「どう」副詞｜い：「いー」フィラー
	ひ：「ひい」感動詞｜どー：「どう」副詞｜い：終助詞
	ひど：「ひどい」形容詞｜ー：補助記号｜い：「いー」フィラー
つらーい	つらー：「つらー」副詞｜い：終助詞
なさーい	な：「だ」助動詞｜さー：「さ」終助詞｜い：終助詞
すっげー	す：「すっ」副詞｜げ：「気」：接尾辞｜ー：補助記号

なお，『日本語日常会話コーパス』では，強調や言い淀みのために一時的に付加された非語彙的な長音や撥音は，タグによって表現する。よって，上述の問題に『日本語日常会話コーパス』の解析時に対応することは考えていないが，ほかの話し言葉を対象とする解析時には注意が必要である。

４．おわりに

現時点で広く利用可能である自然会話のコーパスである『名大会話コーパス』の概要と特徴を述べた。書き言葉コーパスである『BCCWJ』と比べ，「うん」，終助詞の「ね」「か」，「何」が頻出し，また，感動詞，副詞，代名詞が多く，話し言葉の特徴語の用例が多く得られるコーパスであることを示した。様々な話し言葉の研究利用が期待できる。

また,『中納言』版,『ひまわり』版の公開データ作成のために, 形態素解析用辞書『UniDic』と形態素解析器『MeCab』による形態素解析を行った結果に対する人手修正の内容を報告した。「そうなんだ」「そっか」「てゆーか」などの口語表現の誤解析の修正,「１（イチ）個」「５分（ブン）」「いつの間（アイダ）にか」などの発音の誤認定の修正など，具体例を示し，今後の『日本語日常会話コーパス』などの話し言葉を対象とする形態素解析の精度

をあげるための留意事項を述べた。

謝　辞

本研究は国立国語研究所の機関拠点型基幹研究プロジェクト「大規模日常会話コーパスに基づく話し言葉の多角的研究」（プロジェクトリーダー：小磯花絵）の研究成果を報告したものです。また，オリジナルの『名大会話コーパス』は，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」（研究代表者：大曾美恵子，平成 13 年度～15 年度）による研究成果です。形態素解析結果の人手修正をはじめ，本コーパスの構築，公開にご協力くださった皆さまに感謝します。

文　献

Itsuko Fujimura, Shoju Chiba, Mieko Ohso (2012) Lexical and Grammatical Features of Spoken and Written Japanese in Contrast: Exploring a lexical profiling approach to comparing spoken and Written corpora , *Proceedings of the VIIth GSCP International Conference. Speech and Corpora,* pp.393-398.

川端良子・臼田泰如・西川賢哉・徳永弘子・小磯花絵(2017) 「『日本語日常会話コーパス』の転記基準と作業工程」『言語資源活用ワークショップ 2016 予稿集』(収録予定).

小磯花絵・居關友里子・臼田泰如・柏野和佳子・川端良子・田中弥生・伝康晴・西川賢哉(2017) 「『日本語日常会話コーパス』の構築」『言語処理学会第 23 回年次大会発表論文集』.

伝康晴・宇津呂武仁・山田篤・浅原正幸・松本裕治(2002) 「話し言葉研究に適した電子化辞書の設計」『第 2 回話し言葉の科学と工学ワークショップ講演予稿集』pp. 39-46.

山崎誠(2016) 「レジスターの違いによる話し言葉の変容」『シンポジウム「日常会話コーパス」I』発表資料(http://pj.ninjal.ac.jp/conversation/pdf/sympo2016-3.pdf).

関連 URL

国立国語研究所共同研究プロジェクト「大規模日常会話コーパスに基づく話し言葉の多角的研究」 http://pj.ninjal.ac.jp/conversation/

『UniDic』 https://ja.osdn.net/projects/unidic/

『MeCab』 http://taku910.github.io/mecab/

『日本語自然会話書き起こしコーパス（旧名大会話コーパス）』 https://nknet.ninjal.ac.jp/nuc/templates/nuc.html

全文検索システム『ひまわり』 http://www2.ninjal.ac.jp/lrc/

コーパス検索アプリケーション『中納言』 https://chunagon.ninjal.ac.jp/

『現代日本語書き言葉均衡コーパス』語彙表 http://pj.ninjal.ac.jp/corpus_center/bccwj/freq-list.html