『千葉大学 3 人会話コーパス』使用説明書
Release 1

伝 康晴　　榎本 美香

2014 年 6 月 10 日

Chiba Three-party Conversation Corpus Manual
Yasuharu Den and Mika Enomoto

Release 1 10 June 2014

目次

はじめに

『千葉大学 3 人会話コーパス』は、千葉大学で収録された、大学生・院生・ポスドクを含む同性 3 人からなる友人同士 12 組の雑談を収めたものである。本リリースでは、各組の収録データの内、1 会話ずつ（計約 2 時間）の音声データ・転記テキスト・形態論情報を公開する。会話の内容や進行には極力制限を加えず、高い自発性を確保する一方で、話者ごとに個別のヘッドセットマイクを用いて高品質なデータを収録した。「自然生起データ（naturally-occurring data）」ではないため、沈黙を回避したり、無理やり話題を振ったりするといった不自然な言動もときには見られるが、その点に留意すれば、日常会話の高品質なデータとして幅広い用途に利用できる。

本データを用いて行なった研究を公表する場合は、以下の文献を引用すること。

- Den, Y. & Enomoto, M. (2007). A scientific approach to conversational informatics: Description, analysis, and modeling of human conversation. In Nishida, T. (Ed.), *Conversational informatics: An engineering approach*, pp. 307–330. Hoboken, NJ: John Wiley & Sons.

本コーパスの収録・構築には、日本学術振興会科学研究費補助金学術創成研究費「人間同士の自然なコミュニケーションを支援する知能メディア技術」（2001〜2005 年度、代表：西田豊明）、日本学術振興会科学研究費補助金基盤研究（B）「対話における発話単位とその機能の認定に関する研究」（2008〜2010 年度、代表：伝康晴）、日本学術振興会科学研究費補助金基盤研究（B）「発話単位アノテーションに基づく対話の認知・伝達融合モデルの構築」（2011〜2013 年度、代表：伝康晴）から助成を受けた。転記テキスト・形態論情報の作成には、小磯花絵（国立国語研究所）、丸山岳彦（国立国語研究所）、吉田奈央の各氏の協力を得た。また、音声データ中の固有名の伏字化処理には、石本祐一氏（国立国語研究所）の協力を得た。記して感謝します。

質問やバグ報告などは下記までお願いします。

千葉大学文学部
伝 康晴
den@L.chiba-u.ac.jp

1　コーパスの概要

『千葉大学 3 人会話コーパス』は、千葉大学で収録された、大学生・院生・ポスドクを含む同性 3 人からなる友人同士 12 組の雑談を収めたものである。それ以前の対話コーパスが、人工的な状況設定（与えられた課題を解くなど）の元でのやり取りを防音室などの非日常的な空間の中で収録したものや、逆に、日常の会話場面をボイスレコーダなどで収録した低品質なものが中心であったのに対して、本コーパスでは、なるべく日常場面に近い自然なインタラクションを高品質なデータとして収録することを狙いとした。

本リリースでは、12 組の収録データの内、1 会話ずつ（各約 10 分、計約 2 時間）を対象として、

- 音声データ
- 転記テキスト
- 形態論情報

を公開する（映像データは含まれない）。

■公開データの内容

00README.txt	README ファイル
Doc\	マニュアル類を収めたフォルダ
manual.pdf	本マニュアル
Wav3\	音声データ（3 チャンネル、Microsoft WAVE 形式）を収めたフォルダ（3 節）
chiba0132.wav	各会話の 3 チャンネル音声ファイル
⋮	
chiba1232.wav	各会話の 3 チャンネル音声ファイル
Wav2\	音声データ（2 チャンネル、Microsoft WAVE 形式）を収めたフォルダ（3 節）
chiba0132.wav	各会話の 2 チャンネル音声ファイル
⋮	
chiba1232.wav	各会話の 2 チャンネル音声ファイル
Wav1\	音声データ（1 チャンネル、Microsoft WAVE 形式）を収めたフォルダ（3 節）
chiba0132-A.wav	各会話の各話者の 1 チャンネル音声ファイル
⋮	
chiba1232-C.wav	各会話の各話者の 1 チャンネル音声ファイル
Trans\	転記テキスト（独自形式）を収めたフォルダ（4 節）
chiba0132.txt	各会話の転記テキストファイル
⋮	
chiba1232.txt	各会話の転記テキストファイル
Morph\	形態論情報（CSV 形式）を収めたフォルダ（5 節）
chiba0132.csv	各会話の形態論情報ファイル
⋮	
chiba1232.csv	各会話の形態論情報ファイル

ELAN\	転記テキストと形態論情報の統合ファイル（ELAN 形式）を収めたフォルダ（6 節）
chiba0132.eaf	各会話の ELAN ファイル
⋮	
chiba1232.eaf	各会話の ELAN ファイル
Info\	各種メタ情報（CSV 形式）を収めたフォルダ（6 節）
speakers.csv	話者情報を記述したファイル
anonyms.csv	固有名の仮名（かめい）化に関する情報を記述したファイル

2　収録

収録は、学内のラウンジ施設に収録機器を持ち込んでリラックスした雰囲気の中で行なわれた。会話の内容や進行には極力制限を加えず、高い自発性を確保する一方で、話者ごとに個別のヘッドセットマイクやビデオカメラで音声・映像を記録し、高品質なデータを収録した。収録の詳細について、以下で説明する。

■参加者　会話収録には、大学生・大学院生・ポスドクを含む同性 3 人の日本人からなるグループ 12 組（男性グループ 6 組・女性グループ 6 組、18 才〜33 才（収録当時））が参加した。グループの成員はいずれも友人同士であった。各参加者の属性については、話者情報ファイル（6.2 節）に記述されている。

■収録場所　騒音の少ないラウンジ施設（千葉大学人文社会科学系総合研究棟内）に収録機器を持ち込み、テーブルやカーペット・吸音カーテンで簡易スタジオを設営して、収録を行なった（図 1）。ラウンジ内は床や壁からの反響が見られたため、カーペットと吸音カーテンを設置することで軽減した。

■収録機器　ラウンジに設置された 4 台のハンディカム（Sony DCR-VX2000）によって、各話者の正面映像と全話者を捉えた全景映像を記録した（図 1：カメラ A〜D）。音声は、単一指向性ヘッドセットマイクロホン（Sennheiser MKE104）を通じて、各チャンネル独立にデジタルマルチトラックレコーダ（TASCAM MX-2424）に記録した（図 1：マイク A〜C）。対面であるため他の話者の音声が多少回り込むが、サウンドスペクトログラムや基本周波数曲線の抽出などに支障のない程度である。

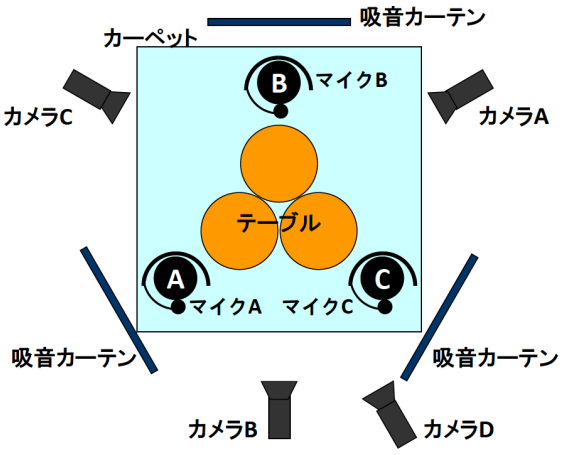

図 1　収録環境

図 2　収録風景

■収録手続き　参加者は 1.5m ほどの距離を隔てて、ほぼ正三角形をなすように対座し、対面で会話を行なった（図 2)。リラックスした雰囲気を作るため、テーブルの上にはジュースとお菓子が置かれた。各グループは、10 分程度ずつに区切って 3 セッションの会話を行なった。収録に要した時間は休憩時間を含めて 40 分程度であった。各会話の開始前に、参加者のいずれかがサイコロを振って会話のトピックを決定した。サイコロの目は「情けない話」「ビックリした話」「びびった話」「恋の話」「腹の立つ話」「あたり目」であり、「あたり目」が出た場合は、別途用意した「臭い話」「大事件」などのリストから選択した。参加者は、そのトピックに固定されることなく、自由に話題を展開してよい旨を伝えられており、実際のトピックは一定でない。約 10 分経過後に、収録者の合図で強制的に会話が打ち切られた。

■コーパス作成　収録後、各映像ファイルと音声ファイルを手動で同期させ、会話開始から 9 分 26 秒後で打ち切って、長さを統一した。各会話には以下の形式の ID を与えた。

chibaXX3Z

XX は参加者グループの番号（01〜12)、Z はセッション番号（1〜3）である。なお、XX と Z の間の数字は各会話の話者数を示し、本コーパスではつねに 3 である。

本リリースでは、各グループ 3 会話の収録データの内、第 2 セッションのものを公開する（したがって Z はつねに 2 である)。参加者の収録環境への慣れと疲労度・集中力や話題の豊富さという点から、第 2 セッションがもっともアクティブなインタラクションとなっていることが多い。

なお、個人情報保護の観点から、映像ファイルについては公開しない。

3　音声データ

3.1　3 チャンネル音声ファイル

Wav3 フォルダには、チャンネルごとに各話者の音声を収めた、3 チャンネルの WAV 形式ファイルが収められている。データ形式は以下の形式の非圧縮・リニア PCM である。

サンプルレート	16kHz
サンプルサイズ	16bit
データ符号化方式	符号付き整数
バイトオーダー	リトルエンディアン
チャンネル数	3

多くの音声再生ソフトは 3 チャンネル音声ファイルに対応しておらず、そのため、2 チャンネル音声にミックスダウンしたものを別途提供する（3.2 節）。また、各話者の音声を独立に取り出すには、3 チャンネル音声ファイル中の対応するチャンネルを抽出しなければならない。このようにして抽出した各話者の音声データも別途提供する（3.3 節）。

3.2 2 チャンネル音声ファイル

`Wav2` には、3 チャンネル音声ファイルを 2 チャンネル（ステレオ）音声にミックスダウンした WAV 形式ファイルが収められている。ミックスダウンの際には、収録環境（図 1）に合わせ、話者 A, B, C の音声がそれぞれ左・中央・右に定位するようにした。

2 チャンネル音声ファイルは音声再生ソフトで普通に再生できる。ただし、2 チャンネル音声ファイルから各話者の音声を分離して取り出すことはできない。

3.3 1 チャンネル音声ファイル

`Wav1` には、3 チャンネル音声ファイルから各話者の音声を 1 チャンネル音声として抽出した WAV 形式ファイルが収められている。各ファイルは、会話 ID と話者ラベルによって、`chiba0132-A.wav` のように命名されている。

1 チャンネル音声ファイルは音声再生ソフトで普通に再生できる。また、Praat [*1] などによる音声分析に利用できる。

3.4 固有名の匿名化について

本コーパスに収められている会話は友人同士の雑談であるため、参加者やその友人の姓・名・愛称などがしばしば言及されている。個人情報保護の観点から、音声ファイル中でこれらの固有名に対応する箇所は、ビープ音に置き換えることで匿名化した。なお、これらの箇所は転記テキストにおいては、仮名（かめい）に置き換えられている（4.5 節参照）。

4 転記テキスト

`Trans` フォルダには転記テキストが収められている。

[*1] `http://www.fon.hum.uva.nl/praat/`

4.1 フォーマット

転記テキストは、1 行に 1 発話が記されており、各行には以下の情報が記されている。

```
0.8267  2.2868  C:  でね:恋の話なんですけど
開始時間 終了時間 話者      発話内容
```

開始時間　その発話の開始時間（秒）（会話開始時 = 0 秒）
開始時間　その発話の終了時間（秒）（会話開始時 = 0 秒）
話者　その発話の発話者
発話内容　漢字仮名混じりテキストと転記記号による発話内容の書き起こし

4.2 発話区切り

発話は「長い発話単位」と呼ばれる単位によって区切られている。「長い発話単位」は話し手と聞き手が行為や情報を交換する際の基本単位に相当し、統語的・談話的・相互行為的な一まとまりに対応する。詳細は Japanese Discourse Research Initiative による『発話単位ラベリングマニュアル』[*2] を参照のこと。

4.3 転記記号

発話内容の転記に際して表 1 の転記記号（タグ）を用いた。これらは、会話分析で広く用いられている転記記号 [*3] や『日本語話し言葉コーパス』で用いられている転記タグ [*4] を簡略化・改良したものである。ただし、一部のタグ（表 1 中で【暫定的】とあるもの）については、体系的には付与されておらず、本リリースにおいては暫定的である。

■発話単位内休止（(数字) タグ）　発話単位内での 0.1 秒以上の休止は “(*Dur*)” の形で記す。*Dur* は休止の長さ（秒）である。

```
48.1638 50.6475 C: まむかし立ったのは(0.836)あれだね
```

■発話単位内休止（(.) タグ）　発話単位内での 0.1 秒未満の休止は “(.)” の形で記す。

```
212.6837 213.9761 C: でも(.)電化製品もやってるよね
```

[*2] http://www.jdri.org/open-data/ より入手可能
[*3] たとえば、http://www.hituzi.co.jp/books/buntohatuwa_tenkikigou.pdf
[*4] http://www.ninjal.ac.jp/corpus_center/csj/doc/k-report/02.pdf

表 1　転記記号

(0.334)	0.1 秒以上の発話単位内休止（数字は秒）
(.)	0.1 秒未満の発話単位内休止
:	非語彙的な音の引き伸ばし【暫定的】
%	非語彙的な音の詰まり
-	語の中断
?	上昇調【暫定的】
(F_あの)	フィラー
(I_うん)	応答系・感情表出系感動詞
(T_チョ-|ちょっと)	言いさし（意図された語が同定可能）
(D_スハ)	言いさし（意図された語が不明）
(W_ジュオー|授業)	言い誤りや非標準的な発音
(K_り % つ|律)	漢字表記できなくなった文字
(R_木村)	固有名を仮名（かめい）に置き換えたもの
(歌_ハニー)	歌いながらの発話
<声>	聞き取れないか、言語音と見なせない音声
<笑>	発話を伴わない笑い
<息>	呼気・吸気【暫定的】

■非語彙的な音の引き伸ばし（:タグ）【暫定的】　標準的な発音に含まれない、語中・語末音の引き伸ばしは ‘:’ を付けることで記す。

```
10.9446 14.6697 C: で:(1.547)なんか:(0.151)国語担当してて:
                    ↑           ↑                    ↑語末音の引き伸ばし
```

```
315.8914 316.7150 A: 怖:い
                       ↑語中音の引き伸ばし
```

```
46.8164 47.6117 C: 何やっ:てた
                         ↑促音の引き伸ばし（長い無音部分）
```

通常は漢字表記する語の途中にこのタグを用いる必要がある場合は、後述する K タグを併用する。

```
373.7102 374.7834 C: 五(K_ま:ん|万)
                       ↑通常「万」と表記される語の途中で音の引き延ばし
```

なお、会話分析の転記記号とは異なり、引き伸ばされた音の相対的な長さは示されない。つまり、どれだけ長く引き伸ばされても ‘:’ の数は一つである。

本タグの使用は本リリースでは暫定的であり、すべての引き伸ばし箇所が必ずしも網羅されていない。

■非語彙的な音の詰まり（% タグ）　標準的な発音に含まれない、語中での音の詰まりは **'%'** を付けることで記す。

```
529.6075 530.4498 B: き%ついね
                        ↑語中音の詰まり
```

無声破裂音（カ行・タ行・パ行）で始まる語の語頭の閉鎖区間が通常より長いものもこのタグで示す。

```
278.5398 279.4977 A: なん%かね:
                          ↑通常より長い語頭の閉鎖区間
```

通常は漢字表記する語の途中にこのタグを用いる必要がある場合は、後述する K タグを併用する。

```
306.7040 308.2081 C: なんか十(K_ま%ん|万)とか言ってました
                                ↑通常「万」と表記される語の途中で音の詰まり
```

■語の中断（-タグ）　言いかけた語を途中で止めているものは、中断箇所に '-' を付けることで記す。多くの場合、後述する T タグ（言いさし）と併用する。

```
408.5781 409.8154 C: (T_ヨ-|四)四階なんて書いてあるの
                            ↑「四」と言いかけて「ヨ」で中断
```

動詞・形容詞の未然形・連用形や接頭辞など、後続要素との形態的結びつきが強い箇所での中断もこのタグを用いる。この場合、T タグは用いず、-タグを単独で用いる。

```
229.6021 231.4242 A: (D_チヨ)それあげ-あげ-あげたっていうけど
                                    ↑    ↑動詞連用形での中断
```

```
151.0790 152.1934 C: 未-(0.156)未成年が:
                        ↑接頭辞での中断
```

■上昇調（?タグ）【暫定的】　質問などの上昇調は '?' を付けることで記す。

```
103.3147 103.8362 A: まじで?
```

本タグの使用は本リリースでは暫定的であり、すべての上昇調が必ずしも網羅されていない。

■フィラー（F タグ）　フィラーは “(F_*Word*)” の形で記す。

```
282.6478 283.7975 A: (F_あの)あるじゃないですか
```

```
520.1492 522.1126 B: (F_うんと:)(F_あの)(0.276)歌やってるから
                     ↑          ↑連続したフィラーは別々にくくる
```

```
280.1062 281.8245 A: (F_あのね)(0.138)送った直前まで寝てた
                            ↑後続する「ね」「さ」「ですね」類も含める
```

以下の形式のものがフィラーに該当する。

- 「あの」「あんの」「あーの」「その」「そーの」
- 「あと」「あっと」「あーと」「あーっと」「あーんと」「えと」「えっと」「えーと」「えーっと」「と」
- 「うんと」「うーんと」「ん」「んと」「んっと」
- 「あ」「あー」「い」「いー」「う」「うー」「え」「えー」「お」「おー」

後述する応答系・感情表出系感動詞や言いさしとの区別については、4.4 節を参照。

■応答系・感情表出系感動詞（I タグ）　「うん」「はい」などの応答系感動詞や「あっ」「えっ」「ふーん」「へー」などの感情表出系感動詞は “(I_*Word*)” の形で記す。

```
438.5458 439.7271 B: (I_へえ:)
```

```
114.1687 115.2067 C: (I_うん)(I_うん)(I_うん)(I_うん)
                     ↑      ↑      ↑      ↑連続した感動詞は別々にくくる
```

```
56.2324 57.8817 B: あれが:(0.255)(I_うん)(.)くさかった
                                ↑自己発話内でも生じうる
```

以下の形式のものが応答系・感情表出系感動詞に該当する。

- 「あ」「あっ」「ああ」「あー」「あーん」「あい」「あら」
- 「うお」「うむ」「うわ」「うん」「うーん」
- 「え」「えっ」「ええ」「えー」
- 「お」「おっ」「おお」「おー」
- 「はあ」「はー」「はい」「ふう」「ふー」「ふむ」「ふん」「ふーん」「へえ」「へー」「ほい」「ほう」「ほー」
- 「まあ」「まー」「むむ」「もう」
- 「わあ」「わー」「わーい」「わーん」

前述のフィラーや後述する言いさしとの区別については、4.4 節を参照。

■言いさし（T タグ）　語を言いさしたものの内、意図された語が同定可能なものは、“(T_*Phon*|*Word*)” の形で記す。*Word* は意図された語、*Phon* は実際の音列である。語の中断を伴うため、前述の-タグを併用する。

```
408.5781 409.8154 C: (T_ヨ-|四)四階なんて書いてあるの
                     ↑「四」と言いかけて「ヨ」で中断
```

```
259.6934 260.5750 C: (T_コド-|子供)(0.146)小さい時
                     ↑「子供」と言いかけて「コド」で中断
```

```
68.1518 70.0826 A: (T_ボ-|某)(T_ボ-|某)某ＳＳゼリア
                   ↑          ↑連続した言いさしは別々にくくる
```

本タグが使用されるケースの多くは、上例のように直後に繰り返しや言い換えを伴う場合である。

■言いさし（D タグ）　意図された語が不明な音列は “(D_*Phon*)” の形で記す。

```
109.6374 110.0502 C: (D_ドフエ)
```

```
191.2408 192.5371 C: (D_ナ)家電はやんない感じ
```

■非標準的な発音（W タグ）　言い誤りや発音の怠けなど、意図された語の標準的な発音から外れている場合は、“(W_*Phon*|*Word*)” の形で記す。*Word* は意図された語、*Phon* は実際の音列である。

```
191.6988 193.3493 B: あと(.)(W_マゼ|松)屋食券制じゃん
                           ↑「松屋」を「マゼヤ」と発音
```

```
106.7944 107.1688 B: (W_ワアン|わかん)ない
                     ↑「わかんない」を「ワアンナイ」と発音
```

■漢字表記不能箇所（K タグ）　語中への ‘:’ や ‘%’ の挿入などによって、通常は漢字表記する語が漢字表記できなくなった箇所は、“(R_*Kana*|*Kanji*)” の形で記す。*Kanji* は本来の漢字表記、*Kana* は仮名表記であり、転記記号は *Kana* に挿入する。

```
373.7102 374.7834 C: 五(K_ま:ん|万)
                       ↑通常「万」と表記される語の途中で音の引き延ばし
```

■仮名（かめい）（R タグ）　個人情報保護のため仮名（かめい）に置き換えられている箇所は“(R_*Anonym*)”の形で記す。

```
337.0270 338.1160 B: (R_小野)さんと(R_荒木)さん
                      ↑              ↑収録データ中では別の名前
```

仮名化の詳細については 4.5 節を参照。

■歌いながらの発話（歌タグ）　歌いながら発せられた部分は“(歌_*Text*)”の形で記す。

```
388.7642 390.7111 B: (歌_イッツソーリーイヤー)
```

■非言語音声（<声>タグ）　まったく聞き取れないか、言語音と見なせないものは、“<声>”で記す。

```
286.7260 286.7647 A: <声>
```

```
508.8814 511.7307 A: 市長の頭に<声>大仏の置物とか置いてきたんちゃうん
                                ↑発話途中や発話冒頭・末尾に生じることもある
```

■笑い（<笑>タグ）　発話を伴わない笑いは“<笑>”で記す。

```
4.8301 5.6575 B: <笑>
```

```
39.9419 45.9860 C: なんか(0.802)電車ん中で(0.438)(D_コ)<笑>(.)一番角に座ってたんだけど:
                                                    ↑発話中でも笑い声だけの部分は<笑>とする
```

会話分析の転記記号とは異なり、笑いの相対的な長さは示されない。つまり、どれだけ長く笑っていてもたんに“<笑>”と書かれる。

■呼気・吸気（<息>タグ）【暫定的】　笑い以外の呼気・吸気は“<息>”で記す。

```
108.3010 108.4310 C: <息>
```

会話分析の転記記号とは異なり、呼気と吸気は区別されず、相対的な長さも示されない。つまり、どれだけ長くはいたり吸ったりしていてもたんに“<息>”と書かれる。

本タグの使用は本リリースでは暫定的であり、部分的にしか付与されていない。

以下は、本リリースでは未対応であるが、今後、導入予定の転記記号である。

■重複開始位置（[タグ）【未対応】 複数の話者間で発話が重複している場合は、重複開始位置を ‘[’ で記す。

```
151.8181 152.8377 B: 何分ぐらい発表[したの
152.4927 153.6165 A:              [発表って何が
```

```
218.6195 220.0683 A: なんかね怪しいんだよね:
220.2610 221.2766 A: [いつのだろうみたい[な
220.2751 220.8008 C: [怪しい
                     ↑同時開始
221.1112 221.3750 B:                  [<笑>
```

```
84.5632 85.7242 A: 裾直[し  [ね
85.1340 85.8670 C:     [裾直[し
85.5297 85.8532 B:          [うん
                            ↑３人の話者による重複
```

なお、本リリースには形態論情報（5 節）が含まれており、単語ごとに開始・終了時間が与えられている。その情報を用いて重複開始位置を近似的に求めることができる。

■切れ目のない接続（=タグ）【未対応】 同一話者内あるいは異なる話者間で切れ目なく発話が続いている場合は、先行発話末と後続発話頭に ‘=’ を記す。

```
8.6010 8.7302 A: (I_え)=           ←同一話者内での
8.7302 9.9707 A: =じゃプラハと他は?  ←切れ目のない接続
```

```
121.6627 124.8539 B: 仕事のことんなると:や[たら:(0.553)うるさくて:=  ←異なる話者間での
123.0570 123.4237 A:                    [(I_うん)
124.8135 125.6038 A: =(I_あー:)=                                    ←切れ目のない接続
125.5432 125.9139 B: =(I_うん)                                      ←連続して生起
```

■終端音調（. タグ）【未対応】 終端下降（final lowering）などの発話の終端を示す音調は ‘.’ で記す。

```
16.5313 17.4036 B: またそのふりかよ.
```

■継続音調（, タグ）【未対応】 上昇下降調などの発話の継続を示す音調は ‘,’ で記す。

```
51.4504 57.2316 C: (F_あの)目玉焼きを作ってもらうときに:,[(0.568)いつも俺は半熟に,(0.262)
                   してって言ってんだけど:,
53.9184 54.1964 B:                                    [(I_うん)
57.6782 59.9572 C: いつもなぜか親がこう完熟で出してくるのね
```

■笑いながらの発話（笑タグ）【未対応】　笑いながら発せられた部分は“**(笑_*Text*)**”の形で記す。

```
78.2965 79.8760 C: 完熟って(笑_言い)方が(笑_おかし)いけどさ
```

4.4　フィラー・感動詞・言いさしの区別

以下の形式のものは、フィラーか応答系・感情表出系感動詞かで迷うことがある。また、これらの音列を言いさしと判断したほうがいい場合もある。

- 「あ」「あっ」「ああ」「あー」「え」「えっ」「ええ」「えー」「お」「おっ」「おお」「おー」「ん」「んっ」「んー」

この判断は以下の手順で行なう（図 3・表 2）。判断に際しては、必ず音声を再生し、音の長短や音調を確認する。

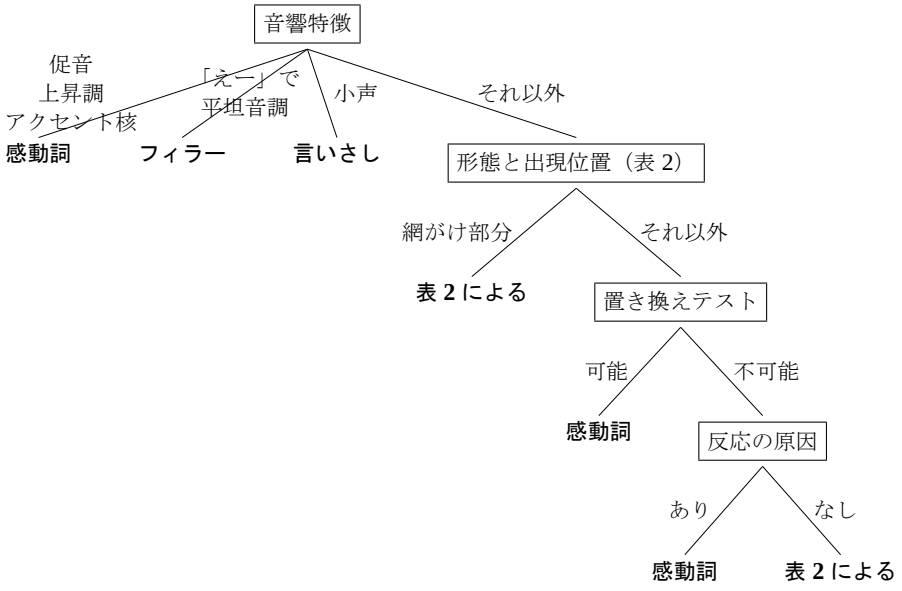

図 3　フィラー・感動詞・言いさしの区別

表 2　形態と出現位置による判断

形態 \ 出現位置	自己発話後続	先行発話あり	先行発話なし
ん	言いさし	言いさし	言いさし
んー	フィラー	フィラー	フィラー
あ・え	フィラー	言いさし	言いさし
お	言いさし	言いさし	言いさし
あー・えー	フィラー	フィラー	感動詞
おー	フィラー	フィラー	感動詞

1. 音響特徴を調べる。
   (a) 末尾に促音や上昇調が認められるか、アクセント核に類似した下降が認められる
      → **感動詞**
   (b)「えー」で、低い平坦音調である
      → **フィラー**
   (c) 声が小さい
      → **言いさし**
   (d) それ以外
      → ステップ 2 へ
2. 形態と出現位置 *5 に基づき表 2 を参照する。
   (a) 網がけ部分に該当
      → **表 2 により決定**
   (b) それ以外
      → ステップ 3 へ
3. 代替要素で置き換えてみる。
   (a) 驚き・悲しみ・喜び・感心・疑問・応答などを表わす感動詞で置き換え可能
      → **感動詞**
   (b) 置き換え不可能
      → ステップ 4 へ
4. 反応の原因を推測してみる（相手発話中に反応を誘出した要素があるか、自身で何かに気づいた様子があるかなど)。
   (a) 原因が推測可能
      → **感動詞**
   (b) 原因が推測不可能
      → **表 2 により決定**

4.5 固有名の仮名（かめい）化について

本コーパスに収められている会話は友人同士の雑談であるため、参加者やその友人の姓・名・愛称などがしばしば言及されている。個人情報保護の観点から、転記テキスト中でこれらの固有名に対応する箇所は、仮名（かめい）に置き換えた。仮名化は以下の方針で行なった。

■仮名化の対象　仮名に置き換える対象としたのは、個人の特定性が高い、参加者やその友人などの姓・名・愛称、および、出身高校の名前である。以下のものは、仮名化の対象とせず、実名のままとした。

*5 出現位置の分類
   (a) 自己発話が後続する → **自己発話後続**
   (b) 自己発話が後続しない
      i. 直前に先行発話がある → **先行発話あり**
      ii. 直前に先行発話がない（文脈上孤立している）→ **先行発話なし**

- 著名な人物の名前：「原辰徳」「小堺一機」など
- 参加者に縁のある地名：「名古屋」「仙台」など
- 参加者のバイト先など：「ユニクロ」「吉野家」など

■仮名の選択基準　仮名を選択する際には、元の実名と「モーラ数」「アクセント型」が一致することを条件とした。モーラ数を一致させてあるため、元の実名で語中の詰まり（% タグ）がある場合は、仮名中の対応する位置に '%' を挿入してあるし、言いさしている場合（T タグ）は、同じモーラ数のところで中断してある。

5　形態論情報

Morph フォルダには形態論情報が収められている。

5.1　フォーマット

形態論情報は、1 行 1 単語からなる CSV 形式で記されており、各行には以下の情報が記されている。

開始時間（startTime）　その単語の開始時間（秒）（会話開始時 = 0 秒）
終了時間（endTime）　その単語の終了時間（秒）（会話開始時 = 0 秒）
話者（who）　その単語の発話者
発話単位区切り（luuLabel）　発話単位の区切りを示す IOB2 ラベル [*6]
発話内容（text）　発話内容の書き起こしを単語ごとに分割したもの
書字形（orth）　発話内容から転記記号を除去したもの
発音形（pron）　その単語の発音をカタカナで記したもの
語彙素読み（lForm）　語彙素（下記）の読みをカタカナで記したもの
語彙素（lemma）　書字形に対して、表記の揺れや異形態を正規化したもの（漢字仮名混じり）
品詞（pos）　その単語の品詞
活用型（cType）　その単語が活用語である場合、活用の型
活用形（cForm）　その単語が活用語である場合、活用の形

なお、言いさしや非言語音声・笑い・呼気・吸気に対しては、「書字形」以降の情報は付与されていない。また、仮名（かめい）化された固有名（R タグ）については、仮名に対して形態論情報を付与した。

5.2　形態論情報の認定

単語区切りや語彙素読み・語彙素・品詞・活用型・活用形は、国立国語研究所が定める「短単位」に基づいて認定した。詳細は『『現代日本語書き言葉均衡コーパス』形態論情報規定集』[*7] を参照のこと。

[*6] 'B' は発話単位の開始位置を示し、'I' は開始された発話単位の内部（末尾を含む）にあることを示す。また、'O' は「笑い」など通常の発話単位に含まれないことを示す。この情報を使って転記テキストを再構成することができる。
[*7] http://www.ninjal.ac.jp/corpus_center/bccwj/doc/report/JC-D-10-05-02.pdf

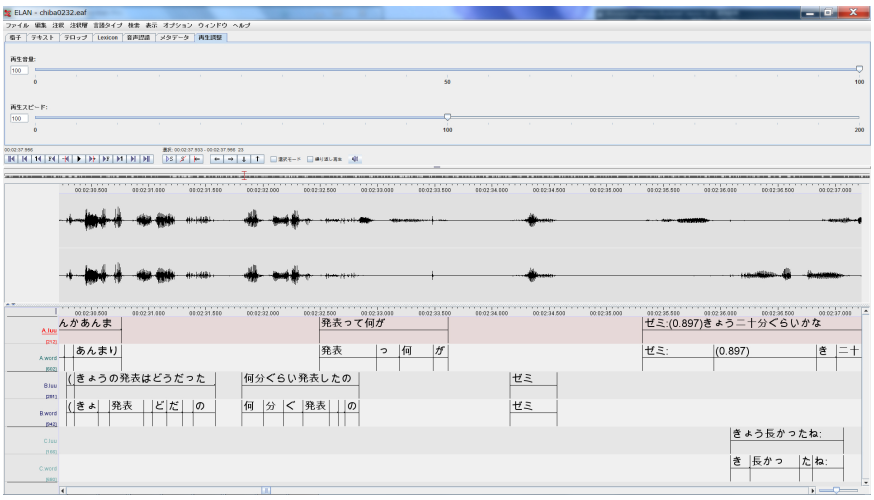

図 4　ELAN ファイルの表示例（注釈層を話者ごとに並べ替え、音声波形の縦ズームを 1000% に設定）

6　その他のファイル

6.1　ELAN ファイル

`ELAN` フォルダには、転記テキストと形態論情報（単語に区切られた発話内容）を統合して ELAN ファイルで表わしたものが収められている。これらはフリーのアノテーションソフト ELAN [*8] で閲覧できる。

ELAN の表示画面の例を図 4 に示す。これらの ELAN ファイルでは、音声ファイルとして、`Wav2` フォルダにある 2 チャンネル音声ファイルを読み込むように設定されている [*9]。そのため、話者 B の発話部分は左右チャンネルにまたがって波形が表示されている。注釈層は各話者についてそれぞれ 2 層からなり、発話単位と単語の 2 つの粒度で区切られている。

6.2　話者情報ファイル

`Info` フォルダ中の `speakers.csv` というファイルには、収録に参加した 36 名の話者の属性が記述されている。話者情報は、1 行 1 話者からなる CSV 形式で記されており、各行には以下の情報が記されている。

会話 ID（dID）　その話者が参加した会話の ID
話者（who）　その話者の話者ラベル（`A`, `B`, `C` のいずれか）
性別（sex）　その話者の性別
年齢（age）　その話者の収録当時の年齢
学年（grade）　その話者の収録当時の学年（`B1`: 学部 1 年, `M1`: 修士 1 年, `D1`: 博士 1 年, `PD`: ポスドクなど）

これらの情報は、各会話の内容や話者間の人間関係を理解する手掛かりとして利用できる。

[*8] `https://tla.mpi.nl/tools/tla-tools/elan/`
[*9] 3 チャンネル音声ファイルを読み込んだ場合はハングアップすることがある。

6.3 仮名（かめい）化情報ファイル

`Info` フォルダ中の `anonyms.csv` というファイルには、会話中で言及されている、仮名化した固有名（R タグが付与された固有名）の情報が記述されている。仮名化情報は、1 行に 1 つの固有名からなる CSV 形式で記されており、各行には以下の情報が記されている。

会話 ID（dID）　その固有名が言及されている会話の ID
書字形（orth）　その固有名の書字形
読み（kana）　その固有名の読み（カタカナ）
説明（description）　その固有名の簡単な説明

「説明」列の記載内容の例を以下に挙げる。

- 話者 A の姓
- 話者 B の愛称
- 話者 C の高校の名前
- 先輩の姓
- 友人の名
- chiba0432 の話者 B と同一人物
- chiba0632 で言及されている教員と同一人物

このように、同一人物への言及に際しては、会話内を通して同じ仮名を用いるだけでなく、別の会話でも同じ仮名を用いている。

7 将来のリリースについて

本コーパスは今後、以下の 2 つの方向で拡張していく予定である。

■他の付加情報の公開　本リリースに含まれる 12 会話（第 2 セッション）には、転記テキスト・形態論情報以外にも以下のような付加情報が付与されている（一部、作業中を含む）。

- X–JToBI による韻律情報
- 文節・節などの統語情報
- 視線・頭部動作などの非言語情報
- あいづち表現とその反応先
- 順番構成単位と順番移行関係

これらの付加情報については、準備ができ次第、随時公開する予定である。

■他のセッションの公開　本リリースでは第 2 セッションの 12 会話のみを公開しているが、これは収録したデータの 3 分の 1 に過ぎない。他の 2 セッション（第 1 セッションと第 3 セッション）については、音声ファイルと粗い転記テキストは存在するが、転記記号は付与されていないし、形態論情報もない。今後、これらのデータについても整備を進め、将来的には全 36 会話（約 6 時間）を公開したい。

付録 A　変更履歴

Release 1　一般公開

- 音声データ・転記テキスト・形態論情報を公開