Towards Sharing of Conversation Corpora: Automatic Transformation between Different Transcription Conventions

国語研プロジェクトレビュー
NINJAL Project Review　Vol.4 No.2 pp.89–99（October 2013）

会話コーパスの共有化に向けて：
転記方式の自動変換

Towards Sharing of Conversation Corpora:
Automatic Transformation between Different Transcription Conventions

伝 康晴（DEN Yasuharu）

1. はじめに

近年，書き言葉コーパスの構築は飛躍的な発展を見せている。国立国語研究所では，1 億語を超える規模の『現代日本語書き言葉均衡コーパス』を開発し，さらに 100 億語を超える規模の超大規模 Web コーパスの開発が進行中である。これに対して，話し言葉コーパスは，学会講演や模擬講演などの独話を中心とする『日本語話し言葉コーパス』を除いて，大規模なものは存在しない。とくに我々の日常の言語行動の中心である会話に関しては，個々の研究プロジェクトごとに数時間～数十時間程度の比較的小規模なデータを独自に収集・利用している状態を脱していない。これは，書き言葉では収集対象となる文書が成文化資料として（場合によっては電子化資料として）前もって存在するのに対して，話し言葉では収集対象の音声の収録・転記をみずから行なわなければならず，その負担が非常に大きいことが一因である。

これに対する一つの解決策として，既存の会話コーパス[1]の共有化というアプローチに着目する。小規模データを所有する研究機関は多くあり，それらは音声収録・転記の段階を一通り終え，開発初期の負担をクリアしている。これらのコーパスを共有すれば，研究に利用できる会話データの量は従来よりも飛躍的に増加する。しかし，これらのコーパスでは，研究機関ごとに転記方式が不統一であり，韻律情報や連鎖構造など会話研究に必要な基本情報は必ずしも完備していない。本プロジェクトでは，これらの基本情報を共通化し，相互利用可能な形で会話コーパスを共有する方法を考案することを目標としている。

本稿では，会話コーパス共有化の出発点として，転記方式の共通化に焦点を当てる。まず，本プロジェクトの参加メンバーが保有する会話コーパスで採用されている転記方式の違いを調査した結果を述べる。その後，主要な転記方式である『日本語話し言葉コーパス』方式と会話分析方式を取り上げ，前者から後者に自動変換する試みについて述べる。

[1] 会話コーパスとは，二人以上の参与者が主体的に発話し合う形式の談話を収集したコーパスのことを指す。複数の話者の発話の時間関係が収録音声や転記テキストから読み取れるものを想定しており，会話場面を収録していても，特定の話者の発話だけを抜き出したものや各話者の音声が同期されていないものは含まれない。

2.　参加メンバーが保有する会話コーパス

本プロジェクトの研究組織を構成する共同研究者・研究協力者たちの専門領域は，会話分析・談話分析・日本語教育学・日本語学・認知心理学・音声言語情報処理など多岐にわたる。いずれも会話データをみずから収集し，研究に利用している研究者たちであり，本プロジェクトにデータを提供し，共有化に伴う問題点を解決し，共有化の利点を検証するという趣旨に賛同していただいている。

会話の諸現象の普遍性と多様性をとらえるためには，参与者数・関係性・様式・内容などがさまざまに異なるコーパス群を集積することが必要である。本プロジェクトで想定している多様性の一端を表 1 に示す。表中で，「参与者数」は一つの会話に参与する者の人数，「関係性」は参与者間の関係，「様式」は対面／非対面の区別，「内容」は会話の内容を示す。

表 1　会話コーパスの多様性

参与者数	関係性	様式	内容
2 人	初対面	対面	雑談
3 人	友人・家族・知人	非対面／電話	インタビュー
4 人以上	上司と部下／教師と生徒など		ビジネス会話
	店員と顧客／医者と患者など		議論・討論
	母語話者と非母語話者		課題指向

本プロジェクトの参加メンバーが所有するコーパスは参与者数・関係性・様式・内容に関して表 1 の多様性を比較的網羅している。これらの会話コーパスの概要を表 2 に示す[2]。

表 2　メンバー所有の会話コーパスの概要

名称	参与者数	関係性	様式	内容
千葉大 3 人会話	3 人	友人	対面	雑談
日本語話し言葉コーパス	2 人	初対面（1 人はインタビューア）	対面	インタビュー
FGI	4 人	初対面（1 人はインタビューア）	対面	インタビュー
言語接触場面 3 人会話	3 人	知人（1 人は非母語話者）	対面	雑談
新聞販売店会話	2 人	店員と顧客	電話	ビジネスコール
早稲田大自由対話	2 人	友人（親密度が異なる複数時期）	対面／非対面	雑談
JPN	2〜3 人	友人／家族	対面	雑談
作業療法会話	主に 2 人	療法士とクライエント	対面	作業療法
宇都宮大音声対話	2 人	友人	非対面	課題指向
三重大地図課題対話	2 人	知人／初対面	対面	課題指向
タングラムパズル対話	2 人	知人／初対面	対面	課題指向
ロゴ積み木対話	2 人	知人／初対面	対面	課題指向
北大 2 人会話	2 人	先輩と後輩（1 人は非母語話者）	対面	雑談

[2]　本プロジェクトは会話コーパス共有のための方法論を構築することに主眼があり，表 2 のコーパスをすべて整備・公開することは想定していない。そのため，各コーパスの全体量などは調査していない。

3. 転記方式の調査

会話コーパスの転記方式が研究機関ごとにどの程度異なるか把握するために，メンバー所有の会話コーパスで採用されている転記方式の違いを調査した（伝・土屋・小磯 2012）。

3.1 方法

表 2 に挙げた 13 個のコーパスについて，転記テキストの断片（数分程度）を収集し，以下の調査項目について比較した。

- レイアウト：転記テキストのレイアウトは発話ごとに改行したものか，それ以外か。
- 転記基準：会話分析方式など，広く流通した転記基準を採用しているか。
- 時間情報：発話や休止の時間情報（開始・終了時間や長さ）が与えられているか。
- 文字表記：転記テキストの文字表記は漢字かな混じりか，かなやローマ字のみか。
- 非言語音の転記：笑いや呼気・吸気は転記されているか。
- 非流暢性の注釈：フィラー・語断片や音の延伸は記されているか。
- 音調の注釈：上昇や継続などの音調は記されているか。
- 重複位置の注釈：複数話者による重複発話の開始位置は記されているか。

3.2 転記テキストの例

いくつかのコーパスの転記テキストの例を付録（97–98 ページ）に挙げる。

『新聞販売店会話』は，標準的な会話分析の転記方式（Jefferson 2004）（以下，CA 方式）によって書き起こされている。CA 方式では，各話者の発話が交互に並べられ，テキスト中に上昇・下降・継続音調，音の延伸，語中断などのさまざまな注釈が挿入される。また，発話間や発話内の休止の長さが 0.1 秒単位で示される（0.1 秒以下の短い休止は“(.)”で示される）。CA 方式を採用しているコーパスが他にも 2 つあったが，どの程度詳しい注釈を与えているかはまちまちであった。

『日本語話し言葉コーパス』は，独自に開発された転記方式（小磯・西川・間淵 2006）（以下，CSJ 方式）によって書き起こされている。CSJ 方式では，可読性の高い漢字かな混じり表記による転記（基本形）と，発音に忠実なカナ表記による転記（発音形）が併用されており，発音形転記には音の延伸や言い誤りなどが記されている。また，フィラーや語断片は基本形と発音形の双方に示されている。上昇・下降などの音調の情報は転記テキスト中には注釈として直接記されておらず，韻律情報として別に与えられている。基本形と発音形の転記を併記する方式は他にも 2 つのコーパスで採用されていた。一方，基本形と発音形を統合して（音の延伸や言い誤りを基本形に記して），CA 方式のように各話者の発話を交互に並べた，CSJ 方式の変種を採用しているコーパス（『千葉大 3 人会話』）もあった。

これら以外はいずれも個別の転記方式を採用しており，その中には過去に提案された方式を踏襲したものや独自の方式を採用したものがある。付録の『言語接触場面 3 人会話』はとりわけ独創的な転記方式であり，各話者の発話を交互に並べるのではなく，それぞれ独立し

た列に並行して記述している。これによって，異なる話者の発話相互の前後関係が一目でわかるようになっている。

3.3　調査結果

3.1 に挙げた項目に対する調査結果の概要を表 3 に示す。

表 3　転記方式の比較

コーパス	転記基準	時間情報	文字表記
千葉大 3 人会話	CSJ 方式（変種）	発話開始・終了時間	漢字かな混じり
日本語話し言葉コーパス	CSJ 方式	発話開始・終了時間	基本形・発音形併記
FGI	独自方式	発話開始・終了時間	漢字かな混じり
言語接触場面 3 人会話	独自方式	なし	漢字かな混じり
新聞販売店会話	CA 方式	発話内・発話間休止	漢字かな混じり
早稲田大自由対話	CSJ 方式（簡略版）	発話開始・終了時間	基本形・発音形併記
JPN	Du Bois 方式	発話間休止	ローマ字
作業療法会話	独自方式	なし	漢字かな混じり
宇都宮大音声対話	CSJ 方式（簡略版）	発話開始・終了時間／発話内休止	基本形・発音形併記
三重大地図課題対話	千葉大地図課題方式	発話開始・終了時間／発話内休止	ひらがな
タングラムパズル対話	独自方式	発話開始・終了時間	ひらがな
ロゴ積み木対話	CA 方式	発話内・発話間休止	漢字かな混じり
北大 2 人会話	CA 方式	発話内・発話間休止	漢字かな混じり

コーパス	非言語音	非流暢性	音調	重複位置
千葉大 3 人会話	笑	フィラー・語断片・音の延伸	（別ファイル）	なし
日本語話し言葉コーパス	笑・咳・息	フィラー・語断片・音の延伸	（別ファイル）	なし
FGI	笑	なし	上昇	あいづちのみ
言語接触場面 3 人会話	笑・咳	音の延伸	上昇	あり
新聞販売店会話	笑・咳・息	語中断・音の延伸	上昇・下降・継続	あり
早稲田大自由対話	笑	なし	なし	なし
JPN	笑	語中断・音の延伸	上昇・下降・継続	あり
作業療法会話	笑	音の延伸	上昇	なし
宇都宮大音声対話	笑・息	フィラー・語断片	なし	なし
三重大地図課題対話	笑	なし	上昇	あり
タングラムパズル対話	笑	音の延伸	なし	あり
ロゴ積み木対話	笑	語中断・音の延伸	上昇・下降・継続	あり
北大 2 人会話	笑・息	語中断・音の延伸	上昇	あり

CSJ 方式（簡略版・変種を含む）を採用しているコーパスが 4 つあり，そこでは発話ごとに開始・終了時間が与えられている。それ以外にも発話開始・終了時間を記しているコーパスが 3 つあった。これらのコーパスでは発話相互間の時間関係（たとえば話者交替に要する時間）を精密に算出できる。一方，CA 方式を採用しているコーパスも 3 つあり，そこでは発話開始・終了時間の代わりに発話間や発話内の休止の長さを記している。CA 方式に類似

した Du Bois 流の転記方式（Du Bois et al. 1993）でも発話間休止の長さを記している。これらのコーパスでは重複発話の開始位置も記されており，発話相互間の時間関係をある程度精密に知ることができる。なお，CSJ 方式のコーパスの中には，語や音韻のレベルで開始・終了時間を記述したデータを持つものがあり，それらの情報から発話内休止の長さや重複位置を求めることができる。

非言語音については，笑いはすべてのコーパスで転記されていたが，咳や息（呼気・吸気）まで転記しているものは少なかった。非流暢性の注釈は，CSJ 方式のコーパスの多くでフィラーと語断片・言い誤りを記していた。一方，CA 方式では，断片化した語の中断位置を記すという形を取っていた（たとえば CSJ 方式：「（D から）体に」に対して CA 方式：「から–体に」）。これら両方式を用いているコーパスを含め，多くのコーパスで音の延伸（「でー」のような引き伸ばし）が記されていた。

最後に，CSJ 方式以外では，発話末や句末の上昇音調が記されており，CA 方式ではさらに下降や継続の音調まで記しているものもあった。CSJ 方式で転記テキスト中に音調を記していないのは，韻律情報を別途与えることを前提としているからで，『日本語話し言葉コーパス』や『千葉大 3 人会話』では実際に韻律情報データが与えられている。ただし，CSJ 方式の韻律情報と CA 方式の音調注釈は必ずしも一対一に対応しない。

4. 転記方式の自動変換～音調に注目して～

3 節の調査から，本プロジェクトの参加メンバーが所有するコーパスは，CSJ 方式か CA 方式の転記方式を採用しているものが多かった。両者は見かけ上は大きく異なっているが，発話相互間の時間関係や非流暢性など，ほぼ同等な情報を与えていることも多い。その一方で，CSJ 方式で転記テキストとは別に記述している韻律情報と，CA 方式で転記テキスト中に記述している音調注釈とは，必ずしも一対一に対応しない。そこでまず，相互行為の分析で重要な役割を果たす発話末や句末の音調に注目し，CSJ 方式の韻律情報から CA 方式の音調注釈に自動変換する試みを行なった（土屋・伝・小磯 2012，2013）。

4.1 データ

CSJ 方式の変種を採用している『千葉大 3 人会話』の中から 2 会話（各 10 分）を選択し，会話分析を専門とする 3 名の研究者（X 氏・Y 氏・Z 氏）に独立に CA 方式の転記テキストを作成してもらった。X 氏と Y 氏はそれぞれ 1 会話ずつ転記し，Z 氏は双方の会話を転記した。ここでは，CA 方式の種々の注釈記号のうち，発話末や句末の音調を示す下降・上昇・継続・平坦の 4 種類の音調注釈に注目する。

4.2 CSJ 方式の韻律情報と CA 方式の音調注釈の対応

『千葉大 3 人会話』には転記テキストに加え，（簡略版）X-JToBI（五十嵐・菊池・前川 2006）に基づく韻律情報が与えられている。そこで，ピッチレンジのリセットや複合境界音調を伴うアクセント句末を対象に，CSJ 方式の韻律情報（句末境界音調）と CA 方式の音調

注釈との対応を調べた。X 氏と Y 氏の転記に対する結果を図 1 に示す。L%，H%，HL%，LH% は各アクセント句末における CSJ 方式の句末境界音調であり，棒グラフは各句末境界音調の何割が CA 方式の下降・上昇・継続・平坦音調（および音調注釈なし）に対応するかを示している。

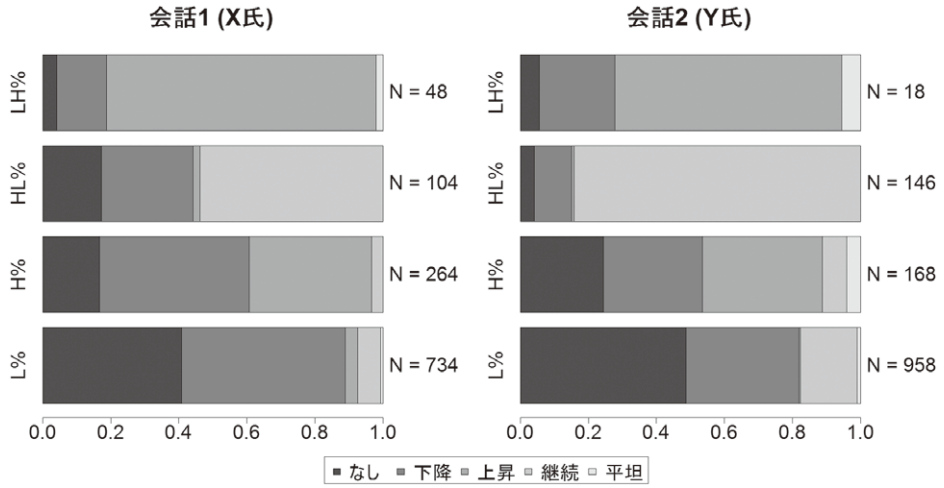


図 1　CSJ 方式の句末境界音調と CA 方式の音調注釈との対応

CSJ 方式の LH%（上昇前の低ピッチ区間を伴う上昇調）の大半が CA 方式の上昇音調に対応したり，HL%（上昇下降調）の多く（会話 2 では大多数）が継続音調に対応したりしているものの，全般的に CSJ 方式の句末境界音調と CA 方式の音調注釈との対応は一対一とは言えない。とくに，H%（通常の上昇調）は CA 方式の下降と上昇の両方の音調に対応している。

4.3　CA 方式の音調注釈の転記者間でのゆれ

次に，CA 方式の音調注釈の転記者間でのゆれを検討するため，会話 1 における X 氏と Z 氏，会話 2 における Y 氏と Z 氏の音調注釈の対応を調べた。結果を表 4 に示す。

表 4　音調注釈の転記者間でのゆれ

会話 1（一致率 = 76.0%，κ = .66）

	Z 氏				
X 氏	なし	下降	上昇	継続	平坦
なし	**130**	12	4	2	1
下降	9	**140**	15	0	0
上昇	2	9	**58**	0	1
継続	33	20	2	**26**	0
平坦	0	1	1	0	**1**

会話 2（一致率 = 69.9%，κ = .58）

	Z 氏				
Y 氏	なし	下降	上昇	継続	平坦
なし	**184**	8	0	0	0
下降	30	**126**	1	3	0
上昇	5	11	**29**	1	0
継続	92	14	1	**89**	5
平坦	0	13	0	0	**0**

全般的に，X 氏と Z 氏のほうが一致が高く（κ = .66），Y 氏と Z 氏では一致がより低かった（κ = .58）。X 氏／Y 氏が継続音調を付与している箇所で，Z 氏がそうしていない例が多く見られた。総じて，Z 氏は音調注釈を付けない場合が多く，とくに Y 氏との違いが顕著である。Y 氏は Du Bois 流の転記法（Du Bois et al. 1993）を学んだ経験があり，最初にイントネーションユニットを同定し，その末尾ごとに音調注釈を付けるという作業方略を採用していた。このため，Z 氏よりも音調注釈を付ける箇所が多くなっているものと思われる。

4.4　CSJ 方式の韻律情報から CA 方式の音調注釈への自動変換

4.2 と 4.3 の結果から，CSJ 方式の韻律情報と CA 方式の音調注釈は一対一に対応しないこと，CA 方式の音調注釈には転記者間でゆれがあることがわかった。では，特定の転記者が付与した音調注釈をどの程度自動的に再現できるであろうか。ここでは，CSJ 方式の句末境界音調に加え，以下の特徴量[3]を用いて，CA 方式の音調注釈への自動変換を試みた。

- アクセント句末尾／次末の語の品詞（lastPOS, penultPOS）
- アクセント句全体／末尾単語の最大・最小 F0（f0MaxAP/Word, f0MinAP/Word）
- アクセント句全体／末尾単語の最大パワー（pwrMaxAP/Word）
- アクセント句全体／末尾単語の平均モーラ長（amdAP/Word）
- アクセント句内の最右 F0 抽出点の位置と値（lastF0Loc, lastF0Val）
- アクセント句の発話冒頭・末尾から測った位置（loc, revLoc）

自動変換には機械学習の一種であるランダムフォレスト法（Breiman 2001）を用いた。ランダムフォレスト法は，多数の決定木の多数決によって判別を行なう手法で，個々の決定木は元のデータからランダムに復元抽出したブートストラップサンプルとランダムに選択された数個の特徴量を用いて学習される。ブートストラップサンプルの一部を評価用に取り置きする（OOB データ）ことで，未知データに対する誤判別率や特徴量の重要度を推定することができる。

2 つの会話データに対して Z 氏が付与した音調注釈を正解事例として用い学習を行なった。誤判別率の OOB 推定値はそれぞれ 24.6%，21.1% であった。特徴量の重要度を図示したものを図 2 に示す。ドットが右にあるほど重要度が高く，図では重要な特徴量から順に並べてある。いずれの会話においても，もっとも重要な特徴量は発話末尾から測った位置（revLoc）であり，発話末に位置するアクセント句には音調注釈が付きやすいことを反映している。句末境界音調（tone）の重要度は二番目であるが，他にも重要度の高い特徴量があり，句末境界音調だけでは音調注釈を決定できないことを示している。両会話に共通して重要度が高いのは，句末単語の品詞（lastPOS）や最右 F0 抽出点の位置と値（lastF0Loc，lastF0Val）などである。この傾向は会話 1 に対する X 氏の音調注釈から得られた結果と共通し

[3] 特徴量の説明の末尾のカッコ内の lastPOS などの記号は各特徴量の略称を示す。これらは図 2 中で用いられている。

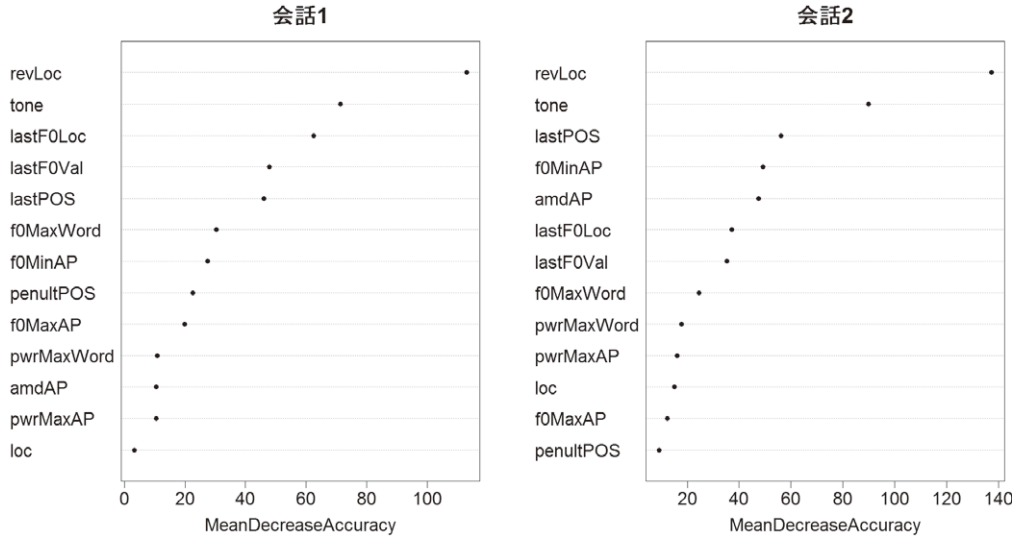


図 2　特徴量の重要度

ていた。これに対して，会話 2 に対する Y 氏の音調注釈から得られた結果では，品詞よりも平均モーラ長（amdAP）や最小 F0（f0MinAP）などの音響特徴のほうが重要度が高く，先に述べた作業方略の違いをうかがわせた（詳しくは土屋・伝・小磯 2013）。

最後に，Z 氏が転記した会話 1 を用いて学習した判別器で会話 2 の音調注釈を再現し，また逆に，会話 2 の学習結果から会話 1 の音調注釈を再現した。結果を表 5 に示す。正解率はそれぞれ 75.1%，72.4% であった。CSJ 方式の句末境界音調ごとにより詳しく見ると，L% では正解率が 80% を超えているものの，H% では 50 数 % とかなり低い（土屋・伝・小磯 2013）。とくに，H% の典型的な機能と思われる上昇音調を正しく予測できないことが多かった。CSJ 方式で採用している X-JToBI の上昇調 H% には疑問上昇調だけでなく，強調上昇調なども含まれており，ここから CA 方式の上昇音調（主に疑問上昇）だけを正しく抽出し変換するのは現状では難しいようである。

表 5　一方の会話で学習し，他方の会話を予測した結果

会話 1 で学習，会話 2 を予測（正解率 = 75.1%）

	正解				
予測	なし	下降	上昇	継続	平坦
なし	**269**	24	5	35	5
下降	34	**129**	13	13	0
上昇	2	16	**12**	1	0
継続	1	2	0	**45**	0
平坦	0	0	0	0	**0**

会話 2 で学習，会話 1 を予測（正解率 = 72.4%）

	正解				
予測	なし	下降	上昇	継続	平坦
なし	**133**	9	2	1	0
下降	30	**161**	57	3	2
上昇	0	2	**20**	0	0
継続	11	10	1	**24**	1
平坦	0	0	0	0	**0**

5. おわりに

本稿では，会話コーパスの共有化に向けて，転記方式の共通化について検討した。まず，本プロジェクトの参加メンバーが保有する会話コーパスの転記方式を調査したところ，CSJ 方式と CA 方式が多く用いられていた。次に，発話末や句末の音調に注目し，CA 方式における音調注釈の転記者間でのゆれや，機械学習を用いた CSJ 方式からの自動変換について検討した。その結果，音調注釈には転記者間でゆれが存在すること，また，自動変換の精度はある程度高いものの，上昇音調などでさらなる精度向上が必要なことがわかった。

会話コーパスの共有化には転記方式の共通化以外にも取り組むべき課題が多くある。とくに，相互行為の分析においては，音調以外にも会話の連鎖構造に言及することが多い。本プロジェクトでは，隣接ペアや遡及的連鎖といった連鎖構造をタグ付けする仕様について検討を進めている。

●参照文献●

Breiman, Leo（2001）Random forests, *Machine Learning* 45: 5−32.

伝康晴・土屋智行・小磯花絵（2012）「多様な様式を網羅した会話コーパスの共有化」『第 1 回コーパス日本語学ワークショップ予稿集』227−234.

Du Bois, John W., Stephan Schuetze-Coburn, Susanna Cumming, and Danae Paolino（1993）Outline of discourse transcription. In: Jane A. Edwards and Martin D. Lampert（eds.）*Talking data: Transcription and coding in discourse research*, 45−89. Hillsdale, NJ: Lawrence Erlbaum.

五十嵐陽介・菊池英明・前川喜久雄（2006）「韻律情報」『日本語話し言葉コーパスの構築法』（国立国語研究所報告 124），347−453.

Jefferson, Gail（2004）Glossary of transcript symbols with an introduction. In: Gene Lerner（ed.）*Conversation analysis: Studies from the first generation*, 13−31. Amsterdam/Philadelphia: John Benjamins.

小磯花絵・西川賢哉・間淵洋子（2006）「転記テキスト」『日本語話し言葉コーパスの構築法』（国立国語研究所報告 124），23−132.

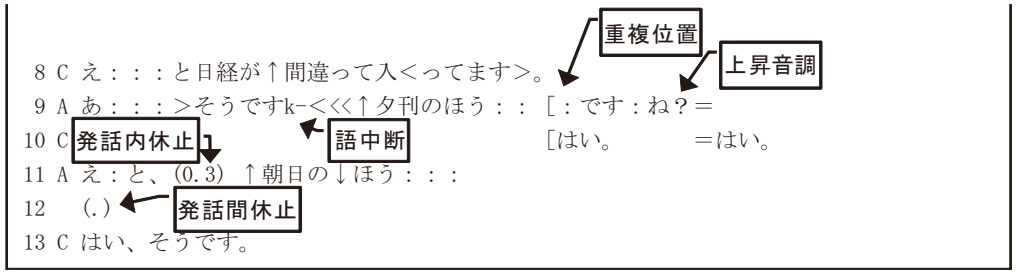
土屋智行・伝康晴・小磯花絵（2012）「会話コーパスの転記方式の相互変換に向けて―イントネーションに着目して―」『第 2 回コーパス日本語学ワークショップ予稿集』117−126.

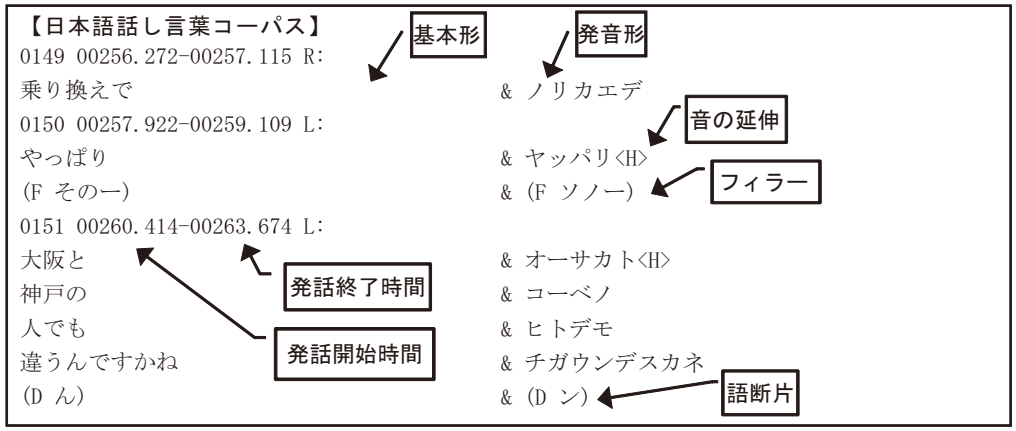
土屋智行・伝康晴・小磯花絵（2013）「会話分析方式への転記変換におけるデータ間・個人間のゆれに関する分析」『第 3 回コーパス日本語学ワークショップ予稿集』417−424.

●付録●

【新聞販売店会話】

1 A はい朝日新聞です。
2 C あっ、すいません、しんまちどおりの：、　← 音の延伸
3 A ↑はい
4 C おしこうじあがるのかたやまです：。　← 下降音調
5 A え：としんまち＜どおり＞の、え：とおしこうじあがる：のかたやまさまです［ね？
6 C　［はい　← 継続音調
7 A はい





【言語接触場面3 人会話】

行	NS1	NNS	NS2
11			ヘードイツとかってもっと
12			北の方にあるんでしたっけ↑
13			日本より
14	うんけっこう北の方に		
15			けっこう北ですよね
16	けっこう北		
17			あじゃ相当寒いーんです
18	/ん/		よね↑/2月って一番
19		[[んー]]	寒いんですか[ねやっぱ]
20	2月とか寒かった		
21		2月はー寒かった	

上昇調
重複発話
重複位置

《要旨》 話し言葉コーパスでは，音声収録・転記といった開発初期の負担が大きく，とくに会話に関しては大規模なコーパスは皆無である。国語研プロジェクト「多様な様式を網羅した会話コーパスの共有化」では，既存の会話コーパスの共有化というアプローチに着目し，コーパスに記述する基本情報を共通化し，共有するための方法論の構築を目指して

いる。その手初めとして，プロジェクト内の会話コーパスの転記方式の違いを調査し，主要な転記方式である『日本語話し言葉コーパス』方式と会話分析方式の間の自動変換を試みた。変換精度はある程度高いものの，さらなる精度向上が必要な部分もあった。

Abstract: Developing spoken language corpora is difficult because of the tremendous effort required for recording and transcription, and this has hindered the construction of large-scale spoken language corpora. Our project aims at developing a methodology for sharing existing conversation corpora that cover diverse styles and settings. As a first step in this endeavor, we examined the different transcription conventions for corpora that have been developed by various researchers, and then attempted automatic conversion between CSJ-style and CA-style transcriptions. The accuracy of our method was quite high, although there is still room for improvement.

伝 康晴（でん・やすはる）

千葉大学文学部教授。博士（工学）（京都大学）。国際電気通信基礎技術研究所研究員，奈良先端科学技術大学院大学助教授，千葉大学文学部助教授を経て，2008 年 4 月より現職。2011 年 4 月より国立国語研究所言語資源研究系客員教授。
主な著書・論文：『談話と対話』（共著，東京大学出版会，2001），『講座社会言語科学 6：方法』（共編，ひつじ書房，2006），『文と発話』1 ～ 3 巻（共編，ひつじ書房，2005 ～ 2008），A scientific approach to conversational informatics: Description, analysis, and modeling of human conversation（共著，*Conversational informatics: An engineering approach*, John Wiley & Sons, 2007），Prolongation of clause-initial mono-word phrases in Japanese（*Linguistic patterns in spontaneous speech*, Academia Sinica, 2009）.
社会活動：社会言語科学会理事，日本認知科学会常任運営委員，人工知能学会代議員，日本認知心理学会理事.

独創・発展型共同研究プロジェクト「多様な様式を網羅した会話コーパスの共有化」
プロジェクトリーダー　伝 康晴
（千葉大学 文学部 教授／国立国語研究所 言語資源研究系 客員教授）

プロジェクトの概要

本研究の目的は，さまざまな機関・研究者が所有する既存の会話コーパスを対象に，共通の基本情報を付与し，相互利用可能な形で共有することである。とくに，会話の諸現象の普遍性と多様性をとらえるために，参与者数・関係性・様式・内容などがさまざまに異なるコーパス群の集積を目指す。そのために，

1. 共同研究者が所有する会話コーパスの調査と，共有化の上での問題点の洗い出し
2. 共通に付与する基本情報の仕様の策定
3. この仕様に基づく共通基本情報の付与と，共同研究者間での共有
4. 共有されたコーパスの基礎的な分析と，多様な様式のコーパス共有の有効性の確認
5. 本プロジェクト外のデータを対象とした，本手法で共有化できるデータの調査

を行なう。これらの活動を通じ，将来の大規模会話コーパス開発のための足掛かりとする。