日本語コーパスの記述について

名古屋外国語大学　谷澤　まどか

目次

- CSJとI-JASについて
  - 概要
  - 表記と単位
  - 使用タグ
  - まとめ
- 転写作業より
  - 転写内容
  - フィラー
  - 外国語
  - 誤用と意味不明語
- 結論

- 参考文献

日本語話し言葉コーパス（CSJ: Corpus of Spontaneous Japanese）

- 国立国語研究所・情報通信研究機構・東京工業大学による共同開発
- 2004年公開
- 成人日本語母語話者の独話を主対象とするコーパス
- 規模662時間（国立国語研究所，2006）
- 録音内容
  - 「学会講演」：学会での研究発表のライブ録音
  - 「模擬講演」：人材派遣された話者による主に個人的内容に関するスピーチの録音
  - インタビュー対話
  - 朗読音声

多言語母語の日本語学習者横断コーパス
（I-JAS:International Corpus of Japanese As a Second Language）

- 2016年　第一次データ公開
- 12言語の異なる母語の学習者（約1000人分）
  - 日本語能力の客観テスト実施（SPOT：Simple Performance-Oriented Test、TTBJ : Tsukuba Test-Battery of Japanese）
- 日本語母語話者のデータ
- 教室環境、自然環境
- 書き起こしデータ、検索システム、音声データ、（作文データ）
- タスク
  - 発話
    - ストーリーテリング（2 タスク）
    - 対話（30 分程度）
    - ロールプレイ（2 タスク）
    - 絵描写
  - 作文
    - ストーリーライティング（2 タスク）
    - 作文調査（メール文 3 タスク、エッセイ 1 タスク）

表記と単位：CSJ

発話番号　発話開始時間-発話終了時間：原則0.2 秒以上のポーズ（言語音の途切れに相当）に挟まれた音声範囲

```
0013  00025.722-00028.067
話者が                  &   ワシャガ
ちゃんと                &   チャント
コントロールして        &   コントロールシテ
出してるんだけども      &   ダシテルンダケドモ
0014  00028.332-00030.855
…
```

… 一行が文節に相当。
「&」を挟んで左に基本形を，右に発音形を表記

（小磯他,2006：26）

＊発話の重複：両話者の発話において重複している部分を［　］で囲む

表記と単位：I-JAS

- 一般的な漢字仮名交じり文
- 話者交代で改行
- 一文の終わりを示す「。」は使用しない
- ポーズの位置に「、」を打つ
- ポーズや伸ばす音の秒数はカウントしない
- 上昇イントネーションがあると判断した箇所に「？」を付す
- あいづち及び発話の重複は＜　＞で囲み相手の発話の中に挿入

使用タグ：CSJ

タグ	タグの概要	使用例
(F)	フィラー，感情表出系感動詞，（応答表現 *3）	(F あの)，(F うわ)，(F うーん)
(D)	言い直し・言い淀み等による語断片	(D こ) これ，(D チ) チーズ
(D2)	助詞・助動詞・接辞・数字の言い直し	そこ (D2 が) に，(D2 不) 不自然
(?)	聞き取りや語の判断に自信がない場合	(? タオングー)，(? 堆積，体積)
(O)	外国語・古語・方言など	(O ザッツファイン)
(M)	音や言葉に関するメタ的な引用	助詞の (M は) は (M わ) と発音
(R)	話者の名前・差別語・誹謗中傷など	国語研の (R ××) です
(X)	非朗読対象発話（朗読における言い間違い等）	(X 実際は) 実際には，
(A)	アルファベット・算用数字・記号の併記	(A シーディーアール;ＣＤ－Ｒ)
(K)	何らかの原因で漢字表記できなくなった場合	(K たち (F えー) ばな;橘)
(W)	転訛や発音の怠けなど，一時的な発音エラー	(W ギーツ;ギジュツ)
(B)	語の読みに関する知識レベルの言い間違い	(B シブタイ;ジュータイ)
(笑)	笑いながら発話している箇所	(笑 ナニガ)
(泣)	泣きながら発話している箇所	(泣 ドンナニ)
(咳)	咳をしながら発話している箇所	シャ (咳 リン) ノ
(L)	ささやき声や独り言などの小さな声	(L アレコレナンダッケ)

（小磯他,2006：80）

使用タグ：CSJ

<FV>	ボーカルフライ等で母音が同定できない場合	だから<FV> & ダカラ<FV>
<VN>	「うん/うーん/ふーん」の音の特定が困難な場合	(F うん) & (F <VN>)
<H>	非語彙的な母音の引き延ばし	ソレデ<H>，スゴ<H>イ
<Q>	非語彙的な子音の引き延ばし	カイ<Q>セキ，ス<Q>ゴイ
<笑>	言語音と独立に生じる話者の笑い	ガクセー<笑>ノ
<咳>	言語音と独立に生じる話者の咳	ソレデ<咳>
<息>	言語音と独立に生じる話者の息	ツマリ<息>
<P>	短単位の内部に生じる 0.2 秒以上のポーズ	オ<P:00453.373-00454.013>モイ

（小磯他,2006：80）

使用タグ：I-JAS

処理	具体的な内容	タグ表記（α は発話通りの文字列、β は解析用の語）	タグの由来
① 解析用の品詞を指定	フィラーを感動詞に指定	[α=F]	フィラー
	外国語を名詞に指定	[α=N]	Noun
	連体詞に指定	[α=R]	連体詞
② 解析用の語を指定	語中の長音、ポーズ	[α=T=β]	訂正
	語や活用や発音の誤り	[α=G=β]	誤用
	PC 入力時の変換ミス	[α=K=β]	漢字・仮名
③ 解析から除外	意味不明語、語の断片	[α=X]	バツ
④ 曖昧性への対応	発音不明瞭（α1 か α2）	[α1/α2=H]	発音
	複数の読みがある漢字語 α	[α（読み）=Y]	読み

（迫田,2016：176）

＊
α：発話通り
β：形態素解析用

CSJとI-JASの転写方法の特徴

CSJ		I-JAS	
メリット	デメリット	メリット	デメリット
・時間情報が正確 ・発話形により音声情報を補足 ・豊富なタグを用いることで解析精度を上げている ・表記方法が事細かに定めてある	・転写に手間がかかる ・*発話すべてに対し発話形を付す必要があるのか？	・必要最低限のタグ ・誤解析になる部分が正しく解析されるようなタグ ・転写者の負担が少ない	・ポーズの記述も含め発話時間の情報が少ない ・転写者によりタグの付与ユレが生じる可能性 ・表記方法が転写者にゆだねられている部分がある

転写方式に関する主な検討項目

- 発話の単位
- 書記法
- 発話内・発話間の時間関係
- 韻律情報・非言語情報
- **非流暢現象**

実際の転写作業より

- 日本語学習者2人による会話　（計2つの録音）
  - フランス語母語話者2人（AとB）の会話　（約18分）
  - トルコ語母語話者2人（CとD）の会話　（約20分）

- Trasncriberに転写

フィラー

- 言い淀み時などに出現する場繋ぎ的な表現
- 短い表現が多い　（ex.「あっ」,「え」）
- 文節や単語の途中に発話される　（ex. D:「あんまり、**うー**役立たないと…」）

→タグなしだと転記テキストの可読性の低下、自動解析の精度の低下につながる

I-JAS

処理	具体的な内容	タグ表記（α は発話通りの文字列、β は解析用の語）	タグの由来
① 解析用の品詞を指定	フィラーを感動詞に指定	[α=F]	フィラー
	外国語を名詞に指定	[α=N]	Noun
	連体詞に指定	[α=R]	連体詞

（迫田,2016：176）

タグF

	例
フィラー	うー んー え／えー えっと／えと／えっとー／えーっと／えーと／えっとねー ま／まあ
感情表出系感動詞	あ／あっ あら あー えっ
応答表現	うん／うーん そう／そー／そうね／そうだね まあね

問題点１：「で」

1. C : えっとー、始めには、｛息｝えっとー、バスで二時間ぐらいかかりました X **で**、最近あんまり寝れないので、｛笑｝始めにちょっと疲れちゃったんですけれども、いい感じでした
2. D :花火は夜だけど、なんか朝行って、　＜八時｛笑｝＞　場所とってくれてなんか、めっちゃありがたい、**で**、うー、なんか場所を結構とっているからもし欲しかったら友達とか誘ってとか言われて…
3. D：…めっちゃ楽しかった

   C：そう

   D：**でー**、＜｛笑｝＞　他の所で、まだ見てないけどなんか初めての花　火だったけど、競馬場だったからあんまり何もなくて、このばーっと見えるから、＜そう＞　それすごくきれいだった　**でー**、その大会終わってから、んーなんか友達と遊びに行って…

＊フィラー？接続詞？

・「それで」に置換可能→接続詞

・「で」を抜いて前後の文のつながりが不自然でないとき→フィラー

問題点１：「で」

1. C : えっとー、始めには、｛息｝えっとー、バスで二時間ぐらいかかりました X **で**（**→フィラー**）、最近あんまり寝れないので、｛笑｝始めにちょっと疲れちゃったんですけれども、いい感じでした
2. D :花火は夜だけど、なんか朝行って、　＜八時｛笑｝＞　場所とってくれてなんかめっちゃありがたい、**で**、（**→接続詞**）うー、なんか場所を結構とっているからもし欲しかったら友達とか誘ってとか言われて…
3. D：…めっちゃ楽しかった

   C：そう

   D：**でー**、（**→フィラー**）＜｛笑｝＞　他の所で、まだ見てないけどなんか初めての花　火だったけど、競馬場だったからあんまり何もなくて、このばーっと見えるから、＜そう＞　それすごくきれいだった　**でー**、（**→接続詞**）その大会終わってから、んーなんか友達と遊びに行って…

問題点２：「なんか」

- 「なんか」が場繋ぎ的な表現として多用されている

4. D : じゃあここは、**なんか**、思い出っていうか、初めての花火について話したいと思います…
   ＊例文４では「何か」の発音の怠けともとれる
5. A : うん、けど、**なんか**、思ってるのは、高校のときにも自分の意見は、まあ、もう考えて、なんだろう、表すよ
6. D : このホストファミリーの父さんは、**なんか**朝の、八時に行って、そこで**なんか**場所をとってくれた、で**なんか**…
7. A : それか、ちょっと、高校生、**なんか**、女子高生から、聞いたのは（うん）、バイトする理由はお金をもらいたくて、
   ＊例文５,６,７では「なんか」自体には意味はなく、場繋ぎ的な表現
   →フィラーとしてタグ付けするべきか？

問題点３：「なんだろう」

・「なんだろう」が場繋ぎ的な表現として多用されている

8. C : これは X **なんだろう** X 山梨県の X 小さな町で {笑} ＜うーん＞　富士山の隣に、超可愛い {笑} ちっちゃい町です
9. C : えっとー、ネズミとウサギの物語の神社がありましたので、この所でちょっと遊んで、また X **なんだろう** X 景色を遊んで、その後また帰りました、…
10. A : 日本人はいつもすごい、**なんだろう**、え、**なんだろう**、あのー、日本人は、大学に行くため（うん）、すっごい時間かかる

→フィラーとしてタグ付けするべきか？

問題点４：母語干渉

D（トルコ語母語話者）の発話
「ずー」？「うー」？
トルコ語の影響か？

https://www.youtube.com/watch?v=xlkfCGZbZ3E

問題点４：母語干渉

B（フランス語母語話者）の発話
“euh”というフランス語で言い淀んだ時に用いる音を日本語の発話でも用いている

どのように表記するか？
・(o) ウー （CSJでの表記）
・（仏 euh ）
・euh：ウー
・（母euh）/（MT： euh ）　MT : Mother Tongue

問題６：誤用と意味不明語

- C：駅について、是非**食べり**（*食べたり？）もしようと思って、この町の、なんだろう、いか天という天ぷらの X 天ぷら屋に行った
- C：このお姫さが、様が、ちょっと、悲しいからまた結婚したくないから、**自分を焼けた**、X **自分を焼けた**、（*自分を焼いた？）そのままでまた結婚することができなかったということ、
- B：あ、そうだ、いつも**すじして**、（*数字して？）そのすじ勉強して、えー、**理系して**（*理系の勉強をして？）

＊誤用に加え、話者が何を意図していったのかが推測困難

・誤用タグGにとどめる

・意味不明語タグXとして解析から除外

・誤用タグG＋意味のつながり上エラーがあることを示すことができると、学習者特有の語の使用（誤用及び母語干渉含む）が見えるのでは？

結論

- 学習者の発話には発音や活用の誤り、予測不可能な誤用、意味不明の語、多様な外国語、母語話者よりも多様なフィラーがみられる
- 母語による影響

- 多言語母語の日本語学習者による日本語の発話に見られる特徴を知る必要がある
- 日本語学習者の各母語に関するある程度の知識が必要

参考文献

- 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.　pp.23-132,国立国語研究所.
- 小磯花絵『講座日本語コーパス　3.話し言葉コーパスー設計と構築ー』，朝倉書店,2015
- 迫田久美子,小西円,佐々木,藍子,須賀和香子,細井陽子（2016）.

「多言語母語の日本語学習者横断コーパス」,国立国語研究所プロジェクトレビューVol.6,No. 3,pp.93-110,国立国語研究所

- 迫田久美子（研究代表者）,（2016）「海外連携による日本語学習者コーパスの構築 ―研究と構築の有機的な繋がりに基づいて― I-JAS 構築に関する最終報告書」,『**平成 24～27年度科学研究費助成事業（基盤研究 A）研究成果報告書**』.pp.1-410 ,国立国語研究所