Utilization of Praat in the Development of the Corpus of Everyday Japanese Conversation

『日本語日常会話コーパス』構築における Praat の利用

西川 賢哉 (国立国語研究所コーパス開発センター) *

Utilization of Praat in the Development of the Corpus of Everyday Japanese Conversation

Ken'ya NISHIKAWA (National Institute for Japanese Language and Linguistics)

要旨

国立国語研究所で構築を進めている『日本語日常会話コーパス』(CEJC) のアノテーション作業（書き起こし・短単位情報付与作業）を支援するために，無償の音声分析ソフトウェア Praat を利用したツールをいくつか開発した：(i)［Praat 起動］必要な情報（ファイル名・時刻情報等）が記された Emacs バッファ，あるいは形態論情報修正ツール「大納言」の検索結果画面から Praat を起動し，転記情報とともに当該箇所を表示するツール，(ii)［転記保存］Praat TextGridEditor 上で変更した転記を，CEJC 転記ファイル（タブ区切り形式）に上書き保存するツール，(iii)［メモ］TextGridEdior 上で選択された区間にある転記情報を，その他必要な情報（ファイル名・時刻情報等）とともにクリップボートにコピーするツール，(iv)［別音声聴取］当該会話に参加している別の話者の音声ファイルを追加で開くツール，など。これらのツールを用いることで，音声聴取をはじめとする，話し言葉コーパス構築に不可欠な作業が簡単な操作で行なえるようになり，作業の効率化および精度の向上が期待できる。

1. はじめに

コーパス開発センターでは，音声コーパス構築における作業者の負担軽減や作業の効率化を目指し，作業支援手法の開発を進めている。本稿では，無償の音声分析ソフトウェア Praat (Boersma & Weenink 2018) を利用したアノテーション（書き起こし・短単位情報付与作業）支援ツールを紹介する(1)。

2. 『日本語日常会話コーパス』(Corpus of Everyday Japanese Conversation; CEJC)

本稿で紹介するツールは，現在のところ，『日本語日常会話コーパス』(Corpus of Everyday Japanese Conversation; 以下 CEJC) 構築作業で使用されている(2)。ツールの紹介に先立ち，

* nishikawa[AT]ninjal.ac.jp

(1) 本稿では Praat の機能についてはほとんど触れない。Praat を基礎からわかりやすく解説したものとして北原・田嶋・田中 (2017)，話し言葉コーパスの構築・分析の観点から Praat の機能を簡潔に紹介したものとして西川 (2015) を参照されたい。なお，以下に紹介するツールにおいては，Praat を外部から操作するプログラム `sendpraat` を使用している。同プログラムは，Praat 公式サイト内で配布されているが，目立たない場所に置かれているため (`http://www.fon.hum.uva.nl/praat/sendpraat.html`)，非常に有益なものにもかかわらず，広く知られているわけではないと思われる。`sendpraat` については，Praat の Help（あるいは，`http://www.fon.hum.uva.nl/praat/manual/Scripting_8_2__The_sendpraat_program.html`）を参照。

(2) 内部で CEJC に特有の処理も行なっているが，できるだけ（最小限の修正を施すだけで）他のコーパスに対しても使用できるよう配慮しつつツールを作成した。

CEJC について必要な範囲で簡単に触れておく。

2.1 収録

日常生活において自然に生じる活動に埋め込まれた多様な会話を収録するために，調査協力者にビデオカメラや IC レコーダーなどの収録機材を 2-3 か月間ほど貸し出し，日常生活における多様な場面での会話を自ら収録してもらう。プロジェクトメンバーは収録場面に立ち会わない。IC レコーダーは会話者全員が装着する。個々の発話に加え，会話全体を録音するために，別の IC レコーダを中央に配置する。したがって，一つの会話に対し，複数の音声ファイルが存在することになる。同時に動画も収録している。収録についての詳細は，田中他 (2018) を参照。

2.2 アノテーション

収録した音声に対し，図 1 に示すようなアノテーションを施す。そこに示されている通り，「コア」と呼ばれるサブセットに対しては，より詳細な情報を人手で付与する。

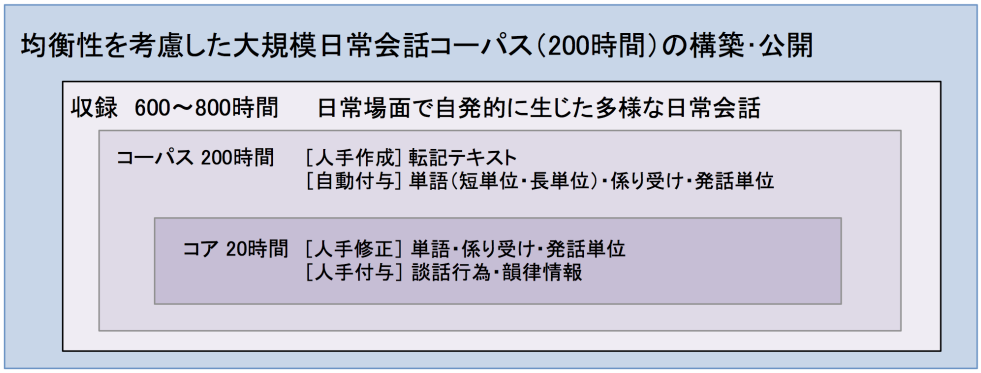

図 1 CEJC のアノテーション（プロジェクトサイトより引用）

以下，転記テキストと単語（短単位）について簡単に述べる。

2.2.1 転記テキスト

映像分析ソフトウェア ELAN[(3)]や Praat を用いて，音声を書き起こす。作業上は，図 2 に示される通り，タブ区切りのテキストファイル (tab-separated values; tsv) で管理されている。1 行は転記単位と呼ばれる単位で区切られており，発話の開始時間と終了時間が割り当てられている。転記テキストには必要に応じて各種タグが付与される。転記テキストについての詳細は，臼田他 (2018) を参照。

2.2.2 単語（短単位）

形態素解析器 MeCab（工藤他 2004)[(4)]と形態素解析用辞書 UniDic[(5)]を用いて，転記テキストを短単位解析したのち，形態論情報管理ツール「大納言」（小木曽・中村 2014）を用いて形態論情報を修正する（図 3)。短単位の規定については，小椋他 (2011) を参照。

(3) http://tla.mpi.nl/tools/tla-tools/elan/
(4) http://taku910.github.io/mecab/
(5) http://unidic.ninjal.ac.jp/

fileID	speakerID	startTime	endTime	pause	text
T004_003	IC01	23.733	24.403	2.716	いいよ いいよ (D ＃＃)。
T004_003	IC02	23.851	24.172	2.09	うん。
T004_003	IC02	26.262	26.947	1	でかいんだよ。
T004_003	IC01	27.119	27.302	1.798	うん。
T004_003	IC02	27.947	28.506	0.002	だから。
T004_003	IC02	28.508	28.99	23.849	あれが。
T004_003	IC01	29.1	29.59	0.097	そうだね。
T004_003	IC01	29.687	30.603	14.262	(W デシ｜出し) にくいんだ。
T004_003	IC03	38.577	39.252	1.109	あー。
T004_003	IC03	40.361	41.516	0.196	雲取も:。
T004_003	IC03	41.712	41.968	0.736	(D イ)
T004_003	IC03	42.704	44.705	0.541	一組だけ外人のご一行みたいの
T004_003	IC01	44.865	45.601	0.899	えー。
T004_003	IC03	45.246	45.935	2.32	帰る時。

図 2　転記テキスト例（タブ区切り）

図 3　形態論情報管理ツール「大納言」

3.　Praat を用いたアノテーション支援ツール

CEJC アノテーション作業を支援するためにこれまでに開発したツールを紹介する。

3.1　Praat 起動 (1): Emacs から

もっとも基本的なツールとして，必要な情報（ファイル名・時刻情報等）が記されたテキストから，Praat を起動し，さらに転記情報とともに当該箇所を表示するツールを作成した。

CEJC 構築作業では，テキストエディタとして Emacs を使用しているため，Emacs Lisp で実装した。この機能は，Emacs 初期化ファイル (.emacs あるいは init.el) で定義してある特定のキー（例えば `C-c C-c C-f`）により実行される。

2.1 節に述べた通り，CEJC では一つの会話に対して音声は複数存在するが，このツールでは起動元のテキストに記されている話者情報を参照し，その話者の IC レコーダーで収録された音声を Praat で開くようにしてある。また，このツールで Praat を起動すると，音声だけでなく，転記も同時に表示される。2.2.1 節に述べた通り，転記ファイルはタブ区切りのテキストファイルで管理されているが，このツールが実行されると，そのタブ区切りファイルから動的に（その場で）TextGrid ファイル（Praat アノテーション形式）が生成され，それが Praat で開かれる。

本ツールは，単なる音声再生機能と比べて，

- 音声だけでなく，波形やスペクトログラムも参照することができる
- Praat TextGridEditor 上で区間を選択し直すことで，特定の部分だけを，繰り返し再生することができる

といった利点がある。

このツールでは，オリジナルの転記テキストからも Praat を起動することができる。ただし，このツールを実行した時点で，TextGrid のほうがマスターデータとなるので，転記ファイルを開いたバッファは自動的に書き込み禁止とするようにしてある。

3.2 Praat 起動 (2):「大納言」から

上と同様の機能を形態論情報修正ツール「大納言」にも実装した。その結果，「大納言」における短単位検索結果画面からも Praat を起動できるようになった。実行方法は，対象とするレコードの「ファイル名」のセルをダブルクリックするだけである。話し言葉コーパス構築作業においては，短単位解析結果から音を聴取したいというケースは，意外に多い。

3.3 転記保存

Praat で表示される転記に誤りが発見された場合，Praat 上で修正を施しファイルに保存できれば便利だが，単純に Praat の保存機能を使うと，TextGrid 形式（Praat のアノテーション形式）でファイルが保存されてしまう。そこで，変更した転記（Praat では TextGrid オブジェクト）を，CEJC 転記テキストの形式（タブ区切り形式；図 2 参照）で上書き保存するツールを作成した(6)。これにより，作業者はわざわざ転記ファイルに戻る必要がなく，Praat 上で自由に転記を修正できる。

3.4 メモ：クリップボードにコピー

転記で対処不明な箇所があった場合など，メモを取っておき，作業者間でその個所を共有したい，といったケースがある。そのとき，そのメモから 3.1 節に述べたツールを用いて，Praat で当該箇所を表示できれば便利である。そこで，TextGridEdior 上で選択された区間にある転記情報を，その他必要な情報（ファイル名・時刻情報等）とともにクリップボートにコピーす

(6) このツールを導入したことにより，CEJC 構築作業において，TextGrid ファイルを管理する必要がなくなった。

るツールを作成した。このツールを実行後，Praat からテキストエディタ等（例えば Emacs）に移動し，ペーストを実行すれば，ファイル名などとともに当該転記が張り付けられる。

3.5 別音声聴取

CEJC のように，複数の話者が参加している会話の音声をアノテーションしている際，別の話者の（同じ個所の）音声を聴取したくなる場合がある。例えば，Praat で IC01 の音声を聞いている最中に，IC02 の音声を聞きたい，といった具合である。そこで，当該会話に参加している別の話者の音声ファイルを追加で開くツールを作成した。このツールを実行すると，別の TextGridEditor が起動するが，転記は同じものが開かれるので，どちらの TextGridEditor でも転記の修正が可能である。

4. おわりに

CEJC アノテーション支援ツールを紹介した。これらのツールを用いることで，音声聴取をはじめとする，話し言葉コーパス構築に不可欠な作業が，簡単な操作で行なえるようになり，作業の効率化および精度の向上が期待できる。

ここに紹介したツールのほかにも，Praat から，そこで選択されている区間の動画を再生するツールなど，追加のツールを現在作成中である。作業者のフィードバックを得ながら，より便利なツールの開発を進めたい。

謝　辞

本研究は国立国語研究所共同研究プロジェクト「コーパスアノテーションの拡張・統合・自動化に関する基礎研究」および「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」の成果である。形態論情報修正ツール「大納言」への Praat 起動機能実装にあたり、中村壮範氏（国立国語研究所コーパス開発センター）の協力を得た。記して感謝する。

文　献

Boersma, Paul and Weenink, David (2018). Praat: doing phonetics by computer [Computer program]. Version 6.0.40, retrieved 11 May 2018 from http://www.praat.org/

北原真冬・田嶋圭一・田中邦佳 (2017) 『音声学を学ぶ人のための Praat 入門』ひつじ書房.

工藤拓・山本薫・松本裕治（2004）「Conditional Random Fields を用いた日本語形態素解析」『情報処理学会研究報告自然言語処理（NL）』47, pp. 89-96.

小木曽智信・中村壮範 (2014)「『現代日本語書き言葉均衡コーパス』形態論情報アノテーション支援システムの設計・実装・運用」『自然言語処理』21 巻 2 号, pp. 301–332.

西川賢哉 (2015)「音声分析ソフトウェア「Praat」」小磯花絵（編）『話し言葉コーパス：設計と構築』朝倉書店, pp. 152–167.

小椋秀樹・小磯花絵・冨士池優美・宮内佐夜香・小西光・原裕 (2011)「『現代日本語書き言葉均衡コーパス』形態論情報規程集第 4 版（下）」特定領域研究「日本語コーパス」平成 22 年度研究成果報告書（JC-D-10-05-02）（http://pj.ninjal.ac.jp/corpus_center/bccwj/doc/report/JC-D-10-05-02.pdf よりダウンロード可能）

田中弥生・柏野和佳子・角田ゆかり・伝康晴・小磯花絵 (2018)「『日本語日常会話コーパス』の構築：会話収録法に着目して」『国立国語研究所論集』14, pp. 275–292.
臼田泰如・川端良子・西川賢哉・石本祐一・小磯花絵 (2018)「『日本語日常会話コーパス』における転記の基準と作成手法」『国立国語研究所論集』15, pp. 177–193.

関連 URL

「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」プロジェクトサイト
http://pj.ninjal.ac.jp/conversation/