

データ内容について

はじめに

各フォルダに格納されている3種類のファイル（.caf, .wav, .mp4）を、フリーソフト ELAN を使って開きます。ELAN を事前にインストールしておいてください。

注意：

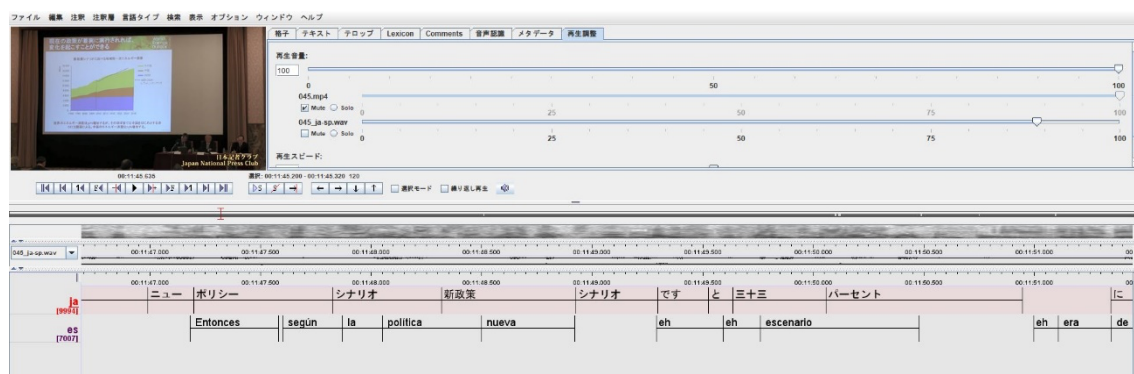
- フォルダ名は、「ES001」や「EC001」のような通し番号になっています。「E」はスペイン語、「S」は同時通訳、「C」は逐次通訳を意味します。
- フォルダ内に格納されているファイル名は、フォルダ名とは異なります。例えば、「ES001」フォルダ内の3つのファイル名は、「45.caf」「45.wav」「45.mp4」です。
- フォルダ名は「収録番号」、ファイル名は「作業番号」（データ作成上の部品番号のようなもの）になります。これらの対応については、別添の「収録データ一覧.xlsx」を御覧ください。

設定方法

- まず、ELAN から.caf ファイルを開きます。
- 音声ファイルと動画ファイルが組み込まれた画面が表示されれば、そのまま使用してください。うまく表示されない場合は、以下の作業を行なってください。
- 音声ファイル（.wav）と動画ファイル（.mp4）を設定します。[編集]-[リンクファイル]を選択します。
- 下のダイアログボックスが表示されたら、[追加]ボタンをクリックして、.caf ファイルと同じフォルダ内にある、.wav と.mp4 を選択します（2つのファイルそれぞれに対して追加を行う）。追加できたら、[適用]ボタンを押します。



5. 正しく選択されると、下のように動画と音声波形、文字情報が表示されるので、確認してください。



mp4 の音声には英語の同時通訳音声も含まれていますので、ELAN 上で再生するときには、[再生調整]上の mp4 の音声の[Mute]をチェックして、wav ファイルの音量を上げてください（上の場合、045.mp4 の下の Mute ボタンをチェックして音を消し、wav の音量を 90 あたりにしています）。

免責事項

- 本データの書き起こし(トランスクリプト)は、スペイン語と日本語ともに同梱した.wav ファイルを音声認識ソフト（speech to text）にかけて自動生成したものです。使用したシステムは IBM Watson Speech to Text、Speechmatix、happyscribe の 3 種類¹です。自動生成した後、誤認識によるエラーを、人手によって修正しています。3 人の異なる作業員により 3 巡確認作業を行っていますが、基本的には自動生成テキストをできるだけ活かす方針としているため、句読点の有無を含む記述の不統一や表記揺れ等があります。また、空白の注釈セルが生成されてしまう場合があります、なるべく消去しているものの、残っている場合があるかもしれません。研究の目的に合わせ、ご自身で修正や調整を行なってご使用ください。
- 音声データ (.wav) はステレオ録音されており、スペイン語と日本語の音声は独立したチャンネルに収録されています。書き起こしのテキストも、ELAN 上で、日本語 (ja) とスペイン語 (es) のように、異なる注釈層に記録しています。注釈層は、画面左下の「ja」と書かれている箇所をマウスで左クリックし、そのまま上下に移動させることで「es」と入れ替えることができます。
- 書き起こしの「単位」に関する基本ルールは以下の通りです。「単位」とは「開始」と「終了」の時間が記録されている言葉のかたまりを表します。ELAN 上では、「注釈」の 1 つの区切りが入っている言葉です。

スペイン語の場合は単語が単位になります。例えば、「Muchas gracias a ustedes.」とい

¹ <https://www.ibm.com/watson/jp-ja/developercloud/speech-to-text.html>
<https://www.speechmatics.com/>
<https://www.happyscribe.com/automatic-transcription-software>

う表現は「Muchas」「gracias」「a」「ustedes」のそれぞれの単語を単位として、それぞれの発話時間（開始と終了の時間）が記録されています。

日本語の場合は、わかち書き、もしくはそれより小さい形態素になります。例えば、わかち書きでは、「スペインの」「エネルギー需要は」となり、形態素では「スペイン」「の」「エネルギー」「需要」「は」となります。

※日本語の書き起こしの「単位」は統一されていないため、研究の目的に合わせて調整してください。

- 言い淀み (hesitation) については、スペイン語では、「ah」「eh」のように「h」を含めたスペルで表示されていて、前置詞や接続詞の「a」や「e」と区別できるように記載しています。日本語では「D_エー」のように記していることもあります。しかし、音声認識ツールによっては、言い淀みが削除される場合があるため、すべてが網羅されているわけではありません。また、スペイン語のフィラーや日本語の冗語の一部は、自動音声認識で「不要」と判断され、書き起こされていない場合があります。こうした現象についての研究を予定されている場合は、ご自身でデータの確認と追加・修正をお願いします。
- 表記の不統一。以下については、表記にバラツキがございます。
 - 漢字・ひらがなの使用。例：「いただいて」「頂いて」
 - 数字の表記。例：アラビア数字と漢数字。例：1995 年、千九百九十五年
 - 不要な空白。例：「現役 の」（同一単位内に空白がある）
 - 人名や固有名詞。例：「ふじた」と「藤田」が混在する場合があります
 - どの漢字が使用されるのか不明な場合。例：「あずま」「吾妻」「東」
- 英語での発話は書き起こされていません。
- 英語での発話を日本語へ同時通訳した音声をスペイン語に同時通訳している場合があります。
- 逐次通訳の会見で、日本語からスペイン語への通訳がウィスパリングで行われている（話者の耳元で囁くように小声で訳している）場合は、音声を拾うことが難しく書き起こされていない場合があります。
- 質疑応答で質問者の発話が不明瞭な場合＜聞き取り不能＞と記載している場合があります。
- 明らかな言い間違いは発話どおりに書き起こした後、正しい表記・単語を（ ）内に記載してある場合があります。
- ファイルごとの注意事項は以下の通りです。
 - ES001 から ES004 は、プロの通訳者に依頼して日西の同時通訳音声を新たに録音したものです。日本記者クラブのもともとの会見時にスペイン語への通訳がなされたわけではありません。
 - ES002「天野国際原子力機関 (IAEA) 事務局長」では、00:43:41 あたりから 00:44:35 あたりまでスペイン語通訳がなされていない箇所があります。

- ES005「ホセ・ルイス・ロドリゲス・サパテロ スペイン首相」では日本語発話のスペイン語への同時通訳が音声として収録されていないため、書き起こしありません。
- EC005「ミゲール・ルイスカバーニャス・メキシコ大使」では冒頭の司会の挨拶はスペイン語に逐次通訳がなされなかったため、収録されておらず書き起こしありません。
- その他、正しく書き起こしされていない箇所もあるかもしれませんが、その場合は適宜、修正・調整してください。書き起こしテキストは、ELAN から[ファイル]-[別ファイル形式保存]で任意の形式（スプレッドシートなど）に書き出すこともできます。
- ELAN でファイルを開く際、数十秒から 1 分ほど時間を要する場合があります。根気強くお待ちいただけたら幸いです。

以上

2023 年 4 月