データ内容について

はじめに

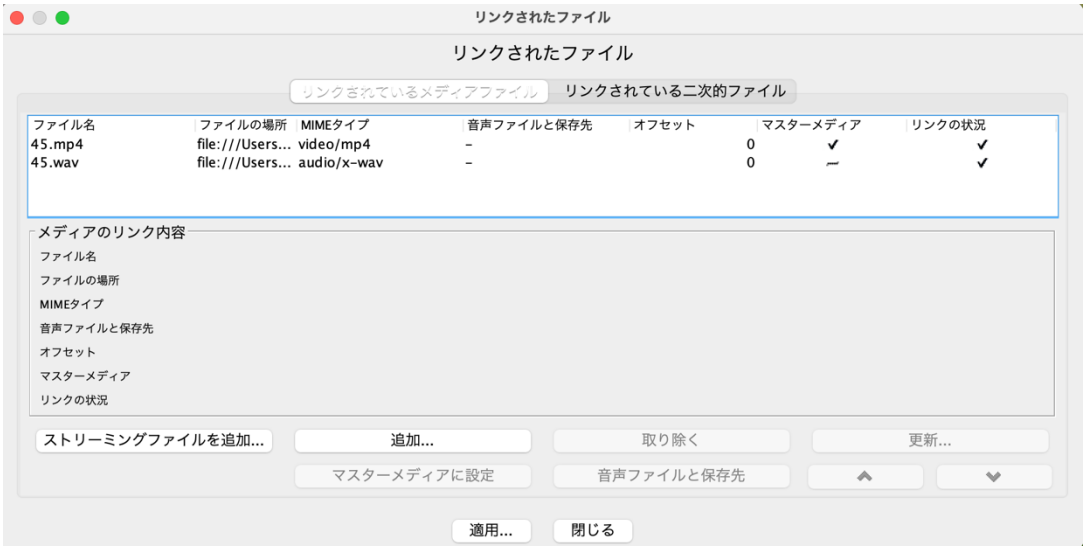
各フォルダに格納されている３種類のファイル（.eaf, .wav, .mp4）を、フリーソフト ELAN を使って開きます。ELAN を事前にインストールしておいてください。

注意：

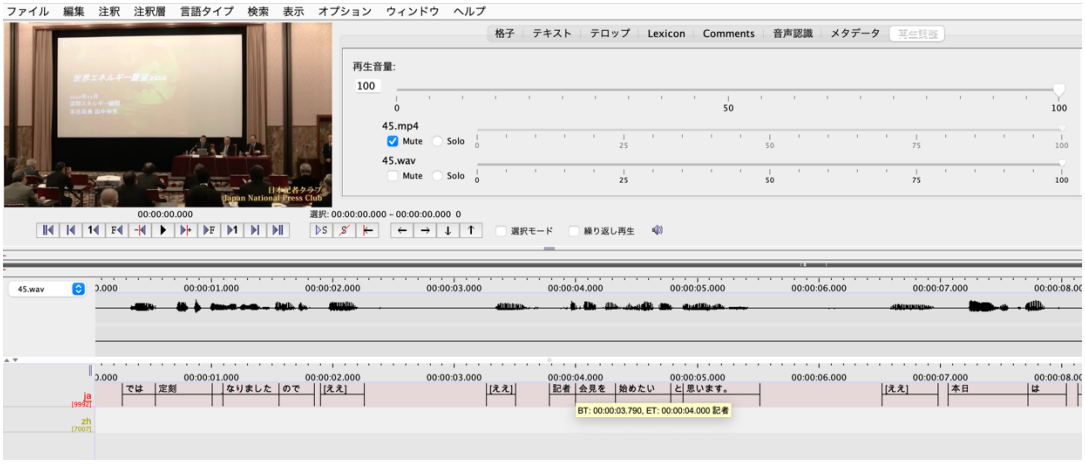
- フォルダ名は、「ZS001」や「ZC001」のような通し番号になっています。「Z」は中国語を示す zh の略、「S」は同時通訳、「C」は逐次通訳を意味します。
- フォルダ内に格納されているファイル名は、フォルダ名とは異なります。例えば、「ZS001」フォルダ内の３つのファイル名は、「C12.eaf」「C12.wav」「C12.mp4」です。
- フォルダ名は「収録番号」、ファイル名は「作業番号」（データ作成上の部品番号のようなもの）になります。これらの対応にいては、別添の「収録データ一覧.xlsx」を御覧ください。

設定方法

1. まず、ELAN から.eaf ファイルを開きます。
2. 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、そのまま使用できます。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。
3. 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。
4. 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、.eaf 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（上の場合は、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 100 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、中国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 人の異なる作業者により 3 巡確認作業を行っていますが、基本的には自動生成テク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、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してご使用ください。
- 音声データ（.wav）はステレオ録音されており、中国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。書き起こしのテクストも、ELAN 上で、中国語（zh）と日本語（ja）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「zh」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」

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

の１つの区切りが入っている言葉です。

中国語の場合は単語が単位になります。例えば、「各位新闻界的朋友,」という表現は「各位」「新闻界」「的」「朋友」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、「中国の」「エネルギー需要は」となり、形態素では「中国」「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、中国語にはほとんどみられませんので記載しておりませんが、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、中国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る
  - どの漢字が使用されるのか不明な場合。例：「あずま」「吾妻」「東」
- ファイルごとの注意事項は以下の通りです。
  - ZC001 は逐次通訳で、冒頭の司会者が話す日本語を通訳者がマイクを使わずにウィスパリング通訳しているため中国語音声が拾えておりません。冒頭、日本語の音声及び書きおこしだけが 12 分余り続きますのでご留意ください。
  - ZS001 は同時通訳ですが、冒頭の一文「では定刻になりましたので記者会見を始めたいと思います」の中国語訳が収録時に録音できておらず、講演のタイトルからの訳出となっています。
  - ZS001 から ZS004 は、プロの通訳者に依頼して日中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中国語への通訳がなされたわけではありません。
  - ZS005 は、冒頭の開始部分が途切れたスタートとなっております。中国語から推測するに、「時間が過ぎましたので始めさせていただきます。今日は中国」という日本語があるはずですが、元々の変換前ファイルの不備によるものですので、ご容赦ください。
- 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。書き起こしテクストは、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 2 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