ATSC規格のテレビ字幕に基づく話し言葉コーパスの構築

YI Yeong-il
東京大学・日本学術振興会特別研究員
yi-yeong-il@g.ecc.u-tokyo.ac.jp

1 はじめに

コーパスを用いた言語研究によって母語話者の内省では発見できなかった事実が数多く発見されてきたが，話し言葉コーパスの活用に関しては書き言葉ほど研究環境が整備されているとは言い難い．その理由の一つにコーパス規模の格差がある．話し言葉は書き言葉よりアノテーションのコストが格段に高く，結果として言語資源が制限されるのはある種の必然と言える．

以上の課題を解決する試みとして，本研究ではテレビ字幕を活用する．実際に韓国で放送されたATSC規格の地上波デジタル放送を受信し，その中に含まれる字幕情報を抽出してコーパス構築を試みた．本稿ではその実装と課題について記述する．

2 先行研究

テレビ字幕のコーパス構築については日本の地上波デジタル放送を活用した事例がすでに存在する [1][2]．ただし世界的には研究報告が少ない状況で，管見の限り韓国に関しても先行事例は確認できない．

韓国の地上波デジタル放送はATSC規格[1]を採用し，字幕はCEA-708に準拠する．日本とは仕様が異なるので先述の先行研究と一定の共通を持ちつつも字幕抽出にいたる過程に自ずと差異が生じること，ATSC規格の放送ならば国・地域・言語を問わず手法を援用できることに本報告の意義があろう．

3 デジタル放送の字幕

3.1 文字データの種類

デジタル放送の文字データは大きく2種類存在する．一つは文字が映像と一体になっているテロップ（telop, open captionとも），今一つはピクチャユーザデータ（picture user data）領域にテキスト形式で埋め込まれる字幕（closed caption）であるが，本研究は後者を対象とする[2]．

字幕は聴覚障碍者への音声情報伝達を第一義とする．この目的から，テロップが映像演出的にも活用されるに対し，字幕は音声を（可能な範囲で）忠実に再現する傾向にある．両者の違いについて実際に以下の用例(1)を見てみたい[3]．1行目が音声の書き起こし，2行目が字幕，3行目がテロップである．字幕では音声がそのまま書き起こされているが，テロップでは単語の省略・挿入・変更が見受けられる[4]．

(1) 이게 그 튀겨지는 맛이 너무 맛있었어요
이게 그 튀겨지는 맛이 너무 맛있었어요
되게 구운 맛이 너무 맛있었어요

3.2 字幕コーパスの言語学的特徴

まず，ニュース・実況・インタビュー・トーク・ドラマなどの各場面から多様な言語変種（variety）を入手できることが大きな特徴である．出演者の人口構成に偏りはあるものの話者の年齢層は幅広く，独話から対話まで，豊富なレジスター（register，言語使用域）が現れることは想像に難くない．話し言葉コーパスが課題として抱えてきた均衡性（balance）や代表性（representativeness）が大きく改善する可能性がある．

多様性に加えてコーパスの規模（scale）の確保が容易なことも重要である．これは数多くの番組が毎日放送されること，そして字幕抽出過程の大半は自動化可能なので人手の書き起こしに比べて低コストであることに起因する．標本コーパスとモニターコーパスの単純比較はできないが，日本の「日本語話し言葉コーパス[5]」

[1] https://www.atsc.org/standard/a53-atsc-digital-television-standard/

[2] 一般的に字幕という用語は文字データの総称として使われることが多いが，本稿ではテロップと字幕を区別して用いることとする．

[3] 2020年1月4日21時00分SBS放送『정글의 법칙（ジャングルの法則）』より引用．

[4] さらにテロップには擬態語のような音声以外の情報も現れる．

[5] https://pj.ninjal.ac.jp/corpus_center/csj/

の 700 万語，韓国の「21 世紀世宗計画 現代口語コーパス[6]」の 80 万語，イギリスの「Spoken BNC2014[7]」の 1000 万語に対し，[1] の字幕コーパスは 1 億語に達する [3]．本コーパスも 1 日あたり 30 万語のデータが日々追加されており，将来的には 1 億語規模を目指す．

以上の利点を持ちつつも，字幕コーパスにも難点は存在する．[4] は話し言葉コーパスの多くは書き起こしの段階で韻律情報や非言語情報が欠落することに留意すべきだと指摘するが，本コーパスもこの立場からの批判を免れるものではない．実際にフィラー（filler）や言い間違いなどの非流暢現象が必ずしも書き起こされていないことが今回のデータ観察から明らかになった．

しかしながら話し言葉コーパスの利用目的は多岐にわたる．音声研究に限らず，例えば大量の用例が持つ言語教育への活用可能性は既に [1] が指摘するところであり，テキストのみの国会議事録も早くから研究に取り入れられてきた事実がある [5]．サンプルサイズの都合で計量的分析の導入が難しかった分野に大規模な字幕コーパスが与える影響は大きいと言えよう．

4　デジタル放送の受信と解析

韓国の地上波デジタル放送は日本と同じく UHF (Ultra High Frequency) アンテナで受信する．マルチメディアフォーマットは MPEG2-TS (Transport Stream) 形式であり，188 バイトの固定長で分割されたデータを各パケットのヘッダー情報をもとに復号することで視聴が可能となる．本研究では録画ハードウェアを自作して韓国ソウルに設置し，Korean Broadcasting System 局から 2 チャンネル，Seoul Broadcasting System 局と Munhwa Broadcasting Corporation 局からそれぞれ 1 チャンネル，計 4 チャンネル（略称はそれぞれ KBS1，KBS2，SBS，MBC）のデータを取り込む．

多くの録画管理ツールには EPG (Electronic Programming Guide，電子番組表) をもとに番組単位で予約を管理する機能が備わっている．番組延長などで以後の番組編成に変更があったときに番組表の更新に従って開始時間と終了時間を自動で調整する機能も EPG に基づく．ただし番組単位での予約録画は，録画ソフトの起動時と終了時に生じるオーバーヘッドの影響により，同一チューナーによる同一チャンネルの連続番組の録画においてパケットの喪失が避けられない．

本研究では 1 日の大半の番組を常に連続して録画するため，番組境界でのパケット喪失が多くなる．そこで予約録画の形式は採らず，チューナーから常時パケットを受信し続け[8]，バイナリをリアルタイム解析し，ATSC 規格が A/65[9]で定義する STT (System Time Table) や EIT (Event Information Table) などの必須テーブルの情報をもとにパケット出力先を動的に変更するプログラムを作成した．これによって番組変更時にもシームレスな録画切り替えが可能となる[10]．

5　字幕データの構造化

5.1　SRT 形式での抽出

字幕の抽出には CCExtractor[11] というソフトウェアを用いる．CCExtractor では様々な出力形式を選択できるが，本システムでは SRT (SubRipText) 形式に変換する．以下はマルチメディアデータから実際に抽出した字幕の一例である[12]．

22_20200104_6.p2.svc01.srt

```
1  31
2  00:01:40,935 --> 00:01:51,544
3    이라크에서는 미국과 이란 간 분쟁이
4    이라크 상황을 악화시킬 수 있다는 우려가
5    나오고 있습니다.
6
7  32
8  00:01:52,279 --> 00:01:54,948
9    <font color="ffff00">-(특파원) 다만 이란의 지원을
         받는 시리아</font>
10
11 33
12 00:01:54,949 --> 00:01:56,483
13   <font color="ffff00">-(특파원) 다만 이란의 지원을
         받는 시리아</font>
14   <font color="ffff00">정부군과 내전을 치르고 있는
         이들립</font>
```

SRT 形式では空行（l.6，l.10）で各字幕が区切られ，各 1 行目（l.1，l.7，l.11）が通し番号，各 2 行目（l.2，l.8，l.12）が表示時間，それ以降（ll.3–5，l.9，ll.13–14）がテキストの内容となっている．

CCExtractor は字幕のロールアップ (Roll up) 機能のエミュレーションのために，複数の時間帯にまたがって同一字幕を重複して記述する（l.9 と l.13 が同

[6] https://ithub.korean.go.kr/user/guide/corpus/guide1.do

[7] http://corpora.lancs.ac.uk/bnc2014/

[8] パケットの受信には MPEG-TS ツールキットである TSDuck を利用した．

[9] https://www.atsc.org/standard/a652013-program-and-system-information-protocol-for-terrestrial-broadcast-and-cable/

[10] 別の対処法として，チューナー数を増やして番組の切替時にチューナー切り替えを行なう方法もあるが，録画ハードウェアの物理的な制約から採用しなかった．

[11] https://www.ccextractor.org

[12] 2020 年 1 月 4 日 9 時 30 分 KBS1 放送『KBS 뉴스（KBS ニュース）』より一部を引用．

一）．これらの重複行は前処理の段階で適切に削除する必要がある．

画面サイズの都合で字幕が 1 行に収まりきらない場合は適度な文字数で複数字幕に分割される（字幕 32 と字幕 33 に分割）．しかし実際の音声がそこで途切れているわけではないので，これらは 1 文（1 発話）として扱うのが適切である．その場合は適宜，開始時間と終了時間の調整も要する．

発話者が交代する場合は文頭に半角ハイフンが挿入され，フォントカラーが変更される．また丸括弧で簡単な属性が記述されることもある（l.9 では「(特派員)」）．人名のような細かい発話者情報のアノテーションは難しいものの，以上の情報は発話単位の認定に活用可能である．

5.2 TEI P5 準拠のマークアップ

テキストデータは適切な構造化を経ることで機械可読性が向上する．本研究では，Text Encoding Initiative が策定し，欧米のデジタル・ヒューマニティーズ分野で広く導入されている TEI P5 ガイドライン[13]に準拠してマークアップする[14]．

22_20200104_6.xml

```
<TEI xmlns="http://www.tei-c.org/ns/1.0">
  <teiHeader>
    <fileDesc>
      <titleStmt>
        <title>22_20200104_6</title>
      </titleStmt>
      <sourceDesc>
        <recordingStmt>
          <recording type="video" dur="PT606S">
            <equipment>
              <p>WinTV-quadHD</p>
            </equipment>
            <broadcast>
              <bibl>
                <title>KBS 뉴스</title>
                <author>KBS</author>
                <date from="2020-01-04T00:29:56" to
                    ="2020-01-04T00:40:02"/>
              </bibl>
            </broadcast>
          </recording>
        </recordingStmt>
      </sourceDesc>
      <extent>
        <measure unit="sentence" quantity="2"/>
        <measure unit="word" quantity="6"/>
        <measure unit="morpheme" quantity="14"/>
      </extent>
    </fileDesc>
    <encodingDesc>
      <transcriptionDesc>
        <desc>
          <bibl>
            <author>YI Yeong-il</author>
            <date when="2020-01-05"/>
          </bibl>
        </desc>
      </transcriptionDesc>
    </encodingDesc>
    <profileDesc>
      <catDesc/>
      <langUsage>
        <language ident="ko-KR">Korean</language>
      </langUsage>
    </profileDesc>
  </teiHeader>
  <text>
    <body>
      <u n="1" who="앵커" start="00:00:09,043" end
          ="00:00:10,844">
        <w n="1" value="여러분,">
          <m n="1" pos="NP">여러분</m>
          <m n="2" pos="SP">,</m>
        </w>
        <w n="2" value="안녕하십니까?">
          <m n="3" pos="NNG">안녕</m>
          <m n="4" pos="XSV">하</m>
          <m n="5" pos="EP">시</m>
          <m n="6" pos="EP">ㅂ니까</m>
          <m n="7" pos="SF">?</m>
        </w>
      </u>
      <u n="2" who="앵커" start="00:00:10,845" end
          ="00:00:15,715">
        <w n="3" value="토요일">
          <m n="8" pos="NNG">토요일</m>
        </w>
        <w n="4" value="아침">
          <m n="9" pos="NNG">아침</m>
        </w>
        <w n="5" value="KBS">
          <m n="10" pos="SL">KBS</m>
        </w>
        <w n="6" value="뉴스입니다.">
          <m n="11" pos="NNG">뉴스</m>
          <m n="12" pos="VCP">이</m>
          <m n="13" pos="EF">ㅂ니다</m>
          <m n="14" pos="SF">.</m>
        </w>
      </u>
    </body>
  </text>
</TEI>
```

ヘッダー（<teiHeader>，ll.2–45）は番組および字幕の属性に関する情報を中心に構成され，字幕の分量についても文数・語数・形態素数でそれぞれ記述する（<extent>，ll.23–27）．テキスト（<text>，ll.46–79）は発話単位（<u>）で区切り，発話内容は形態素単位（<m>）でアノテーションする．なお，本研究では韓国ソウル大学が提供する「꼬꼬마 한글 형태소 분석기（ココマハングル形態素分析器）[15]」を形態素解析に用いている．

6 おわりに

6.1 今後の課題

第一に，本稿執筆時においてジャンル情報の付与を実装できていない．日本の地上波デジタル放送では

[13] https://tei-c.org/guidelines/p5/

[14] 以下で取り上げる例は紙幅の都合上一部のみの抜粋であり，適宜見やすさのために要素の削除・変更を施している．

[15] http://kkma.snu.ac.kr/documents/index.jsp

EIT にジャンルが記述されているが，韓国の放送では番組開始時間，長さ，番組タイトル，字幕言語などの提供に留まる．この課題については Wikipedia による解決を計画している．韓国語版 Wikipedia には韓国のテレビ番組の記事が多く存在するが，その中のジャンルコードを援用する方法である．EIT に記述された番組タイトルと Wikipedia の記事タイトルのズレについては正規表現などによって吸収する必要があろう．

第二に，今後は音声・映像データの活用を視野に入れたい[16]．例えば既述の通り字幕には詳細な発話者情報が記述されていないが，音声データを用いた発話者のアノテーションが提案されている [6]．ただし音声を用いる場合，字幕の時間的なズレが問題となりうる．生放送番組でリアルタイム字幕放送が行われる場合は発声から字幕表示までに遅延が発生するが，その時間幅は一定ではない．収録番組の場合には遅延が発生しないことも考慮する必要がある．また映像データについては，音声がテロップで書き起こされるに留まり字幕として記述されていない場合に，画像認識で映像からテキストデータを抽出することが一つの解決策となりうるが，その場合テロップに含まれる音声以外の諸情報が障害となる．抽出したテキストの取捨選択，音声の実態に近づける修正のためには，音声との比較も不可欠だろう．

第三に，将来的には大量のデータのハンドリングが大きな課題となる．韓国は放送法 69 条 8 項で字幕放送を義務付けていることから，番組の字幕付与率が高い．本録画システムでは 1 日に 100 番組程度が自動録画されるが，そのほぼすべてに字幕データが存在する．1 番組に対して 1 つの字幕 XML が生成される現在の方法では，仮に今後検索インターフェースなどを導入することになった場合に対応が難しい．TEI P5 によるマークアップと並行し，リレーショナルデータベースなどによるデータベース化も検討していきたい．

6.2 コーパスの公開性について

最後に，コーパスの公開可能性および今後の展望に触れて本稿の結びとする．

本研究が扱うデータは著作権および著作隣接権が深く関わり，韓国著作権法や韓国不正競争防止法に抵触するかが大きな争点となると思われる．本コーパスの構築およびその公開は，非営利的利用であって公益的な側面が強いこと，市場需要の代替性が低いことに鑑みれば，検索インターフェースを介して全体の一部を引用して表示する形式ならば実現可能であると考えている．細かい法律要件の検討などは稿を改めるが，学説および判例を精査し，法律専門家の見解に依拠して権利侵害のない範囲で公開を目指したい．

[16] 本稿執筆時において，音声・映像データは FFmpeg を利用してフィルタ処理（インターレース解除，リサイズ，フレームレート調整など）やトランスコードをしてアーカイブしている．

謝辞

本研究は JSPS 科研費 JP19J14192 の助成を受けたものである．

参考文献

[1] MOCHIZUKI Hajime and SHIBANO Kohji. Building very large corpus containing useful rich materials for language learning from closed caption tv. In *Proceedings of E-Learn: World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education 2014*, pp. 1381–1389, New Orleans, LA, USA, 2014. Association for the Advancement of Computing in Education.

[2] 奥貴裕, 一木麻乃, 尾上和穂, 小林彰夫, 佐藤庄衛. 放送音声と字幕テキストを利用した音声言語コーパスの開発. 研究報告音声言語情報処理 (SLP), Vol. 2014, No. 2, pp. 1–5, 2014.

[3] MOCHIZUKI Hajime. Investigation of words in a japanese closed caption tv corpus. In *2019 STEAM Education PROCEEDINGS*, Honolulu, HI, USA, 2019. Hawaii University International Conferences.

[4] 小磯花絵. 話し言葉コーパス : 設計と構築. 講座日本語コーパス, No. 3. 朝倉書店, 2015.

[5] 松田謙次郎. 国会会議録を使った日本語研究. ひつじ書房, 2008.

[6] 山室慶太, 伊藤克亘. デジタル放送の字幕情報を用いた発話者のアノテーション. 第 73 回全国大会講演論文集, 第 2011 巻, pp. 123–124. 情報処理学会, 2011.