京都大学言語学研究 (*Kyoto University Linguistic Research*) **41** (2022), 69–91

OTTサービスを利用したパラレルコーパスの構築方法

徐　敏徹

【要旨】 本稿は，over-the-top media service（OTTサービス）の字幕を利用して日韓・韓日パラレルコーパスを構築する方法，そして，その際にどのような点に注意する必要があるのかについて紹介することを目的とする。OTTサービスを利用してパラレルコーパスを構築するためには，良質，かつ，十分な量の字幕を提供しているOTTサービスを選択しなければならない。本稿では，OTTサービスとしてNetflixを選択し，Language Reactorを活用して日本語・韓国語の（翻訳）字幕を同時に収集した。なお，OTTサービスを利用して収集した字幕は，重複・修正・重訳などが問題となりうるので，言語研究に用いる際には注意を要する。

【キーワード】 Netflix，字幕，著作権，準口語，言語資源

1　はじめに

近年，NetflixやDisney+のような動画配信サービスが普及している。このようなサービスを「over-the-top media service（以下，OTTサービス）」とも言う。「OTT」という用語について，柴田 (2016: 2-3) は次のように述べている。

(1) OTTに決まった定義はないが，アメリカのメディアが伝える内容を総合すると“既存のケーブルテレビや衛星放送を介さず，インターネット経由で番組やコンテンツを配信するサービス”といったものになる。

たとえば，先述したNetflixを利用して動画を見るために，テレビのアンテナケーブルのようなものなどは必要ない。パソコンやスマートフォン，タブレット端末などがインターネットにさえつながっていれば，他に特別な装置がなくても，すぐに動画を閲覧することができる。このようなOTTサービスの高い利便性は，OTTサービスが急成長した要因の一つであると考えられる。

本稿では，上述したOTTサービスを利用してドラマや映画の字幕を収集し，「パラレルコーパス（parallel corpus）」を構築する方法について説明する。パラレルコーパスは，たとえば，日本語（あるいは韓国語など）のテキストを韓国語（あるいは日本語など）に訳し，その二つのテキストを語・文などの単位で並列（align）させ

たものである[1]。書籍や新聞などの書き言葉を利用したパラレルコーパスの構築に比べると，話し言葉を利用したパラレルコーパスの構築は，それほど簡単ではない。しかし，OTTサービスで提供されている字幕を言語資源として利用すれば，比較的簡単に，話し言葉のパラレルコーパスを構築することができると考えられる。

本稿の流れは次の通りである。まず，次の2節では，どのような場合に字幕を言語資源として有効活用することができるのかについて述べ，実際に字幕を利用して言語研究を行った例を示す。つづいて，3節では字幕の言語研究への使用と著作権に関して述べる。そして，4節では，OTTサービスの字幕を利用して日韓・韓日パラレルコーパスを構築するために，どのような手順と方法が必要なのかについて説明する。つづく5節で，OTTサービスを利用して収集した字幕にどのような問題点があるのかを説明し，最後の6節で本稿をまとめる。

2　言語研究のための字幕データの利用

話し言葉コーパスの構築は，音声データの収集や音声の文字化などの作業が必要なので，相当な費用と時間がかかる。そのため，研究者個人からすると，話し言葉コーパスの構築は負担になりかねない。このような状況では，ドラマや映画などの字幕を利用した話し言葉コーパスの構築が，一つの打開策となりうる。なぜなら，研究者が直接音声を録音したり，音声データの文字起こしをしたりする必要がないからである。また，ドラマや映画の字幕データに加え，映像資料の使用もできるのであれば，字幕に付いているタイムスタンプ（字幕が画面に現れる時間と消える時間に関する情報）を利用し，どこで字幕に相当する音声が流れるのかを確認することもできる。

むろん，ドラマや映画における話し言葉は，我々が日常生活で使う自然な話し言葉とは異なる。そのため，このような種類の話し言葉に対して「準口語」という用語を用いる研究者もいる。ソ ＝ サンギュ (1999: 253) は準口語を「戯曲・シナリオ・ニュースなど，あらかじめ作成されたスクリプトに基づいて発せられた発話を集めたもの」（筆者訳）としている。このように，ドラマや映画における話し言葉は，自然な話し言葉とは性質が異なるので，言語研究における分析対象として準口語を利用した場合は，分析結果の一般化に慎重である必要がある。

[1] パラレルコーパスの構築には「Source Text（ST）」と，STを翻訳した「Target Text（TT）」も必要となるので，パラレルコーパスのレジスターは限定される場合が多く，規模も一般的なコーパスに比べると，比較的小さい(ユ ＝ ヒョンギョン・ファン ＝ ウンハ 2010)。これらの点は，パラレルコーパスのデメリットとして挙げられる場合もある(Aijmer 2008)。そのため，いわゆる「代表性（representativeness）」のあるパラレルコーパスを構築することは研究者個人のレベルでは困難であり，言語資源として利用するものが，書籍や新聞などの書き言葉ではなく，話し言葉となると，さらにその難易度は高くなる。

聴覚障害者や難聴者がテレビ番組・ビデオなどの映像を楽しめるようにするために開発されたクローズド・キャプション（closed caption，以下CC）という技術の登場に伴い，字幕データを利用してコーパスを構築することができるようになった。磐崎 (2002: 101-102) は字幕の一種であるCCデータを利用すると日英字幕のコーパス，つまり，パラレルコーパスの構築ができると述べている。当時の日本では，CCデータを表示させるために，外付けのCCディコーダ，あるいは，CC機能内臓のビデオデッキやテレビなどを利用する必要があった(磐崎 2002: 98)。さらに，日本語のCCデータは入手できなかったので，手作業で日本語字幕を入力しなければならなかった(磐崎 2002: 101)。

字幕データの入手が容易になり，それを利用した言語間の対照研究も行われるようになった。たとえば，日韓の映画字幕を用いて対照研究を行った曺 (2004) やキム ＝ ボエ (2013)，欧米のテレビドラマの日本語版DVDと韓国語版DVDの字幕を利用して日韓対照研究をした尹 (2016)，Netflix上で配信されているドラマの英語・フランス語字幕を研究に利用したMoreau (2021) などが挙げられる。しかし，これらの研究は，パラレルコーパスの構築に関する研究ではないので，それに関する詳細な記述はない。

また，対照研究ではないが，日本のテレビプログラムを録画し，CCデータを抽出してコーパスを構築したMochizuki and Shibano (2015) も，字幕データを言語資源として利用した例である。しかし，Mochizuki and Shibano (2015: 99-100) が利用した方法はテレビそのものをはじめ，テレビプログラムを録画するためのハードディスクや録画ソフトウェア「EpgDataCap_Bon」，予約録画のための「EpgTimer」，字幕を抽出するための「Caption2Ass_PCR」などが必要なので，それほど簡単な作業ではないと考えられる。

なお，上述した字幕データによるパラレルコーパスの構築は，英語・中国語・日本語・韓国語などの言語に限られた話ではない。ミン ＝ ギョンモ (2020: 196) は，韓国語と，ベトナム語・ロシア語・インドネシア語・クメール語・タイ語・ネパール語・モンゴル語・ミャンマー語・シンハラ語などの言語間にパラレルコーパスを構築するために利用可能な資料の一つとして，ドラマ・映画の字幕を自動抽出する方法を挙げている。2022年現在は，字幕が含まれている映像資料に接することが，比較的容易にできるようになっているので，先述した諸言語のパラレルコーパスの構築も，およそ20年前に比べ，容易になっていると考えられる。

1節で述べたように，OTTサービスはインターネットを経由して利用するので，ドラマや映画が格納されているDVDやBlu-ray Discなどの記録媒体を入手する必要がなく，パソコン以外の特殊な装置も必要としない。しかし，このようなOTTサービスにおける複数言語の字幕を利用し，パラレルコーパスを構築する方法について詳細に記

述している文献は，筆者が調べた限りでは見当たらなかった。以下では，まず次の3節で字幕と著作権に関して述べ，その後，OTTサービスを利用してパラレルコーパスを構築するための手順や方法，注意点について詳述する。

3　字幕と著作権に関して

字幕を利用してコーパスを構築する際に注意すべき点は著作権である。もし，複製の対象となる著作物に，コピープロテクションのような技術的保護手段が施されていることを知りながら，複製などの行為をする場合は，著作権侵害となりうる。OTTサービスの字幕は，一般的なウェブサイトで目にする文字列データとなっており，技術的保護手段は施されていない。そのため，特定のプログラムや技術などを利用しなくても，字幕を複製し，研究者個人のコンピュータに保存することが可能である。

また，複製した字幕をウェブ上に無断転載したり，共有プログラムなどを利用して他人と共有，あるいは，営利的な目的で利用する場合に問題となる可能性がある。むろん，本研究で字幕を複製する理由は，研究者個人の言語研究のためであり，無断転載や共有，営利的な目的での利用は想定していない。著作権法の第三十条（私的使用のための複製）には，次の (2) が書かれている[2]。

(2)　（上略）個人的に又は家庭内その他これに準ずる限られた範囲内において使用すること（以下「私的使用」という。）を目的とするときは，次に掲げる場合を除き，その使用する者が複製することができる。

研究者が個人の研究を目的として字幕を複製する行為は「私的使用」の範囲に含まれると考えられる。また，著作権法の第三十条の四（著作物に表現された思想又は感情の享受を目的としない利用）には，次の (3) が記されている。

(3)　著作物は，次に掲げる場合その他の当該著作物に表現された思想又は感情を自ら享受し又は他人に享受させることを目的としない場合には，その必要と認められる限度において，いずれの方法によるかを問わず，利用することができる。

本研究では，言語研究のために収集した字幕を小説のように読み，ドラマや映画に表現された思想や感情などを享受する，あるいは，享受させることを目的としていない。そのような目的ではなく，字幕を通して，人間が使用する言語を分析することを想定している。このような目的は，次の (4) に示した著作権法の第三十条の四の二と

[2] 著作権法はすべて「e-Gov法令検索」から検索し，引用したものである（https://elaws.e-gov.go.jp/（最終アクセス：2022年11月11日））。

も関連している。これは，上述した著作権法の第三十条の四における「著作物は，次に掲げる場合（中略）利用することができる」に該当する（下線部筆者）。

(4) 情報解析（多数の著作物その他の大量の情報から，当該情報を構成する言語，音，影像その他の要素に係る情報を抽出し，比較，分類その他の解析を行うことをいう。第四十七条の五第一項第二号において同じ。）の用に供する場合

本研究で収集した字幕は，言語情報の解析のために使用するので，上記 (4) に該当すると考えられる。

末吉 (2012: 434) は「（上略）『情報解析』と『著作権』のキーワードを掛け合わせて判例データベースを検索してみると，該当判例はゼロである」と述べている。2022年9月19日，「裁判所」の「裁判例検索」[3]を利用して先述したキーワードをAND検索した結果，「知的財産裁判例」が4件あった。これらの判例は，BitTorrent（ビットトレント）というP2Pファイル共有ソフトウェアを利用して漫画をアップロード・ダウンロードしたことで著作権侵害をした判例である。4件の判例を見ると原告は，不特定多数の間で著作物を共有することは，情報解析や引用などに該当しないため，違法性を阻却する事由は存在しない，と主張している。したがって，これらの判例は情報解析という行為を著作権侵害であるとする判決ではないと言える。

上野 (2021: 749) は，日本の著作権法における「情報解析のための権利制限」規定によると，「著作権等のあるコンテンツを自由に利用できる」と述べ，「たとえ営利目的・商業目的であっても，たとえ違法に入手した著作物等であっても，情報解析に必要な限度といえれば，あらゆる利用行為が許容され得る」とも述べている。このような日本の規定は，他の国における同様の規定に比べると，非常に強力だと言える。

しかし，上述した規定が著作物の利用に対する免罪符となることを期待してはいけない。たとえば，「Netflix利用規約」[4]の「4. Netflixサービス 」には「（上略）以下のことを行わないことに同意します」と書かれており，それに続いて次の (5) が示されている。

(5) a. Netflixサービスに含まれる，または同サービスから，もしくは同サービスを通じて取得されるコンテンツおよび情報について，アーカイブ，複製，頒布，改ざん，展示，上演，出版，使用許諾，二次的著作物の創出，販売申し出，または使用すること。

b. Netflixサービスへのアクセスのためにロボット，スパイダー，スクレイパー，その他の自動化手段を使用すること。

[3] https://www.courts.go.jp/app/hanrei_jp/search1（最終アクセス：2022年11月11日）
[4] https://help.netflix.com/legal/termsofuse（最終アクセス：2022年12月11日）

c. 何らかのデータマイニング，データ収集もしくは抽出方法を利用すること。

上記 (5) の行為をした場合は，利用者のNetflixアカウントが制限・停止される可能性がある。「Netflix利用規約」においてもう一つ注目すべきところは，次の (6) に示した「6. 雑則」である。

(6) 準拠法: 本利用規約は日本法に準拠し，同法にしたがって解釈されます。

以上の内容を踏まえると，「Netflix利用規約」は，先述した日本の著作権法（「私的使用」や「情報解析」）に基づいて解釈することができると考えられる。しかし，その「解釈」となるものは，「利用者がどの程度の複製を行ったのか」「どのような自動化手段を使用したのか」などによって異なってくる可能性もあるので，他人の著作物を言語資源として利用する際には，十分な注意を払う必要がある。

4 OTTサービスの字幕を利用したパラレルコーパスの構築手順

ここでは，次の4.1で，OTTサービスを選択する際にどのような点を考慮する必要があるのかについて述べる。その後，4.2ではOTTサービス上にある作品の中から，字幕収集の対象となる作品を選定する基準について説明する。そして，4.3と4.4では，それぞれ，字幕をコンピュータに保存する作業と字幕の前処理に関して記述する。

4.1 OTTサービスの選択

字幕を収集するためには，まず，どのようなOTTサービスを利用するのかを決めておく必要がある。OTTサービスによって，提供されている作品の数・種類が異なるだけでなく，作品によっては字幕を提供していない場合もあるからである。もし，韓国語の字幕を収集するのであれば，言うまでもなく，韓国の作品が多く配信されているOTTサービスを選ばなければならない。日本国内で利用可能なOTTサービスとしては「Netflix（ネットフリックス）」「YouTube（ユーチューブ）」「Disney+（ディズニープラス）」「Amazon Prime Video（アマゾン プライム ビデオ）」などが挙げられる。

本稿では，日本と韓国のドラマや映画の字幕を収集するためにNetflixを選択した。Netflixは，NETFLIXオリジナル作品・ドラマ・映画・ドキュメンタリーなどが視聴できる配信登録制のストリーミングサービスである[5]。複数のOTTサービスがある中でNetflixを選択した理由は，次の通りである。一つは，他のOTTサービスに比

[5] https://help.netflix.com/ja/node/412（最終アクセス：2022年11月11日）

べ，Netflixに日本と韓国のドラマや映画が多かったからである。そのため，日本語と韓国語の字幕を入手することも，他のOTTサービスに比べると容易である。

Netflixを選択した二つ目の理由は，他のOTTサービスに比べると，Netflixの字幕は質がよいと考えられるからである。オ ＝ ギョンハン・ノ ＝ ヨンヒ (2021) は映像翻訳の実態を把握するために，NetflixとYouTubeで配信されている同じ映画を5本選定し，韓国語の翻訳字幕を分析した。その結果，YouTubeの翻訳字幕では台詞の多くを省略しているのに対し，Netflixではなるべく原文を保持する方法，つまり，直訳に近い翻訳をしていたことがわかったと述べている。日本語と韓国語の間には言語間の類似性が見られるので，このようなNetflixの直訳に近い翻訳方法は長所として捉えることができる。一方，言語間の類似性がほとんど見られない場合，直訳に近い翻訳方法は，逆に短所になる可能性がある。

また，Netflixの映画の字幕をプロ翻訳者が作成した字幕として捉えた例も見られている。渡邉・山田 (2020) は，プロ翻訳者が作成した字幕とアマチュアの字幕翻訳者が作成した字幕を比較する際，プロ翻訳者の字幕としてNetflixの字幕を選択している。その理由として，Netflixの映画の字幕は「プロ翻訳者として一定の水準を満たしている者」が翻訳したものだと考えられるから，と述べている(渡邉・山田 2020: 5) 。

一方，Netflix以外の他のOTTサービスの字幕には，次のような問題点があった。まず，韓国におけるDisney+の場合，韓国語の翻訳字幕の質が問題視されている。韓国放送通信電波振興院 (2021) は，Disney+が韓国市場において解決すべき課題の一つとして，韓国語の翻訳字幕の質を挙げている。また，2021年11月に韓国でサービスを開始したDisney+の利用者数が，60万人から同年同月37万人まで急減した主な原因としても，字幕の質が挙げられている[6]。このような点を考慮し，Disney+はOTTサービスの候補から除外した。なお，Amazon Prime Videoは，Netflixに比べると日本・韓国の作品が少なく，日本語と韓国語の字幕を同時に提供している作品も見当たらなかった。

4.2 作品の選定

OTTサービスを選択したら，次は，そこから作品を選定する。すでに述べたように，今回は日韓および韓日パラレルコーパスを構築する。したがって，一つの作品において，日本語と韓国語の字幕がどちらも提供されている必要がある。たとえば，日本の作品の場合，日本語の台詞の字幕と，それを韓国語に訳した韓国語字幕が必要である。どの作品に2種類の字幕が提供されているのか，そのリストを網羅的に表示さ

[6]「디즈니플러스 일일 이용자 수 급감…엉망 번역 탓?（Disney+一日利用者数急減…でたらめな翻訳が原因か？）」（MBN뉴스（2021年12月1日付）https://mbn.co.kr/news/economy/4650646（最終アクセス：2022年11月11日））

せることはできないので，日本と韓国の作品を一つずつ再生し，提供されている字幕を確認しなければならない。

作品の選定をする際には，OTTサービスの「サービス開始日」に注意する必要がある。なぜなら，OTTサービスが開始された時期と字幕の質の間に相関関係があると考えられるからである。韓国人映画翻訳家のファン ＝ ソクフィは，Netflixが韓国でサービスを開始する前に，多くの韓国の翻訳業者が1,500～2,000本程度の映像翻訳に努めたと述べている[7]。その時期には人手不足によって，翻訳経歴の足りない翻訳家まで呼び寄せることになったと言う。OTTサービスでは，提供可能な作品と字幕の数が多ければ多いほど，利用者が増える可能性も高くなる。Netflixではこの点を意識し，サービスを開始する前に少しでも多くの韓国語字幕が提供できるように，短い期間の間に多くの翻訳作業を進めたのではないかと考えられる。この点を考慮すると，Netflixがサービスを開始して，しばらく時間が経ってからストリーミングが始まった作品の字幕の質の方が，より優れている可能性がある。サービスが安定的に提供されるようになってからは，翻訳に必要な翻訳業者も少なくなり，翻訳経歴が足りない翻訳業者が翻訳に努める可能性も，その分低くなると考えられるからである。以上の点を踏まえ，日本の作品は2015年10月以降，韓国の作品は2016年2月以降にストリーミングが始まった作品を対象として字幕を収集した[8]。

なお，作品を選ぶ際，研究目的によっては，作品の時代背景を考慮する必要がある。たとえば，『るろうに剣心』は日本の明治時代，『킹덤（キングダム）』は韓国の朝鮮時代が作品の時代背景となっている[9]。『るろうに剣心』を例にすると，この作品には「拙者」や「でござる」，「かたじけない」など，現代日本語ではほとんど使われていないと考えられる表現が頻出する。したがって，研究対象が現代の日本語や韓国語であれば，このような作品の字幕は収集対象外とするのが望ましいであろう。

上述した点は研究の方針によって，柔軟に対応した方がいいかもしれない。なぜなら，現代が時代背景となっている作品の中でも，異なる時代背景が登場する場合もあるからである。たとえば，『野武士のグルメ』という作品は現代の日本が主な舞台となっている。しかし，主人公が戦国時代を想像する場面では，現代日本語ではないような台詞も一部登場する。現代日本語を研究する場合，このような作品の字幕を収集

[7] https://www.facebook.com/story.php?story_fbid=2675766735812795&id=129904367065724（最終アクセス：2022年11月11日）

[8] Netflixが日本でサービスを開始したのは2015年9月（https://youtu.be/jswbeCROpzc?t=28（最終アクセス：2022年11月11日）），韓国でサービスを始めたのは2016年1月からである（https://about.netflix.com/ko/news/sknetworkfee（最終アクセス：2022年11月11日））。

[9] 本稿では，日本の作品の場合，タイトルを日本語で表記する。そして，韓国の作品の場合，タイトルを韓国語で表記し，日本におけるタイトルを添える。

してもいいのだろうか。一つの拠り所として考えられるのは，現代日本語が作品全体に占める割合である。『野武士のグルメ』で戦国時代を意図した日本語が登場するのは，作品の全体の長さからすると，ほんの一部に過ぎない。言い換えると，作品における台詞はほとんどが現代日本語である。もし，現代日本語の分析をすることを目的として，このような作品の字幕も収集したのであれば，用例を分析する際には細心の注意を払う必要があろう。本研究では，このような作品の字幕も収集対象としている。一方，『るろうに剣心』のように，作品全体の時代背景が現代ではない場合は，作品を選定する段階で除外している。

4.3 Language Reactorを利用した複数言語の字幕収集

作品の選定が終わったら，次の手順は字幕を収集することである。字幕を収集する際には，日本語字幕とそれを訳した韓国語翻訳字幕（またはその逆）を同時に表示させ，いっしょに収集する方法を利用した。Netflixでは本来，一つの字幕しか表示されない。しかし，ウェブブラウザの拡張機能である「Language Reactor」を利用すると，二つの字幕を同時に表示させることができる[10]。Language Reactorは，語学学習を目的として開発されたもので，Netflixだけでなく，YouTubeで動画を閲覧する際にも利用可能である。次の図1は，Language Reactorを利用してNetflix上で日本語字幕と韓国語翻訳字幕を同時に表示させた例である。

図1　Language Reactorを利用して日本語字幕と韓国語翻訳字幕を同時に表示させた例（『深夜食堂: Tokyo Stories（シーズン1，第4話）』一部改変）

このような機能は，Language Reactor以外に「NflxMultiSubs」[11]や「Netflix dual

[10] この機能は現在，「Google Chrome」「Brave」「Microsoft Edge」など，Chromiumをベースとしたウェブブラウザで使用可能である（ https://www.languagereactor.com/（最終アクセス：2022年11月11日））。

[11] https://github.com/gmertes/NflxMultiSubs（最終アクセス：2022年11月11日）

subtitle for learning languages」[12]にもある。しかし，前者は以下で説明する字幕のエクスポートができず，後者は有料となっている。なお，Amazon Prime Videoの場合も「Subtitles for Language Learning (Prime Video)」[13]を利用すると，二つの字幕を同時に表示させることができる。

Language Reactorには，字幕をエクスポートする機能がある。図1のように二つの字幕を表示させた状態でエクスポート機能を利用すると，タイムスタンプ付きの日韓の字幕をテキストファイルなどの形式で保存することができる。この方法を利用して収集した字幕は，台詞ごとに並列されるようになる。むろん，すべての翻訳字幕が原文と完全に一対一で並列されているわけではない。翻訳の手法によっては，原文と同じ内容を表す翻訳字幕が，原文とはまったく異なる位置に現れる場合もある。ゆえに，構築したコーパスを利用して翻訳字幕を分析するときには，該当字幕がある行だけでなく，その上下の字幕にも目を向ける必要がある。

今回は，上述した方法を利用して日本と韓国のドラマ30作品，映画10作品の字幕を収集した[14]。ドラマの場合，一つの作品の話数や，一話あたりの平均的な長さが異なるので注意を要する。たとえば，『賭ケグルイ（シーズン2）』は全5話，『グッドモーニング・コール（シーズン1）』は全17話からなるドラマで，両者にはかなりの開きがある。また，一つの話の長さの側面からすると，『野武士のグルメ』の第9話は14分52秒に最後の字幕が表示される。一方，『Jimmy～アホみたいなホンマの話～』の最終話では，最後の字幕が表示されるのは1時間1分39秒である。つまり，単にドラマ30作品といっても，ドラマによって話数が異なり，映像の長さにも相違がある，ということである。したがって，二つ以上のコーパスを比較するような研究では，コーパスの容量・語数・字数などを明記し，読者に誤解が生じないようにすべきである。

4.4　字幕の前処理

NetflixとLanguage Reactorを利用して収集した生データ（raw data）は，次の図2のような形式となっている。

この生データを用いて，日本語や韓国語の調査・分析をすることは，むろん可能である。しかし，生データを対象として語や構文の検索をすると，漏れが生じる可能性が高くなる。たとえば，「ただでもらう」における「ただ」を検索する場合や「軽

[12] https://niko-pay.appspot.com/（最終アクセス：2022年11月11日）
[13] https://www.subtitlesfll.com/（最終アクセス：2022年11月11日）
[14] 収集した作品のリストは，次のウェブページで確認することができる。https://github.com/kr-jp/JpKrAndKrJpParallelCorpora（最終アクセス：2022年11月11日）

15:44	（受付）おかえりなさいませ	어서 오세요
15:51	ただいま	나 왔어
15:55	なんか… 軽く食べる？	뭐 좀 먹고 싶어?

図2　NetflixとLanguage Reactorを利用して日韓の字幕を抽出した例（『金魚妻（第1話）』）

い」の活用形まで含めて検索をする場合を考えてみよう。テキストエディタなどを利用して「ただ」という文字列を検索すると，上の図2にある感動詞「ただいま」の「ただ」も検索されてしまう。また，「軽い」という形容詞が使われている例を網羅的に抽出するためには，「軽い」「軽く」「軽か（った）」などで複数回検索をしなければならない。図2に示したような生データは，字幕をあらかじめ形態素解析器を利用して形態素ごとに分け，品詞付与を行っておくと便利である。なお，生データには台詞以外，環境音や登場人物の動作などを描写する字幕ガイドが含まれている。もし，研究をするうえで，そのような情報が必要ないのであれば，あらかじめ削除しておくことも一つの方法である。

今回は上述した一連の作業を，プログラミング言語「Python」を利用して自動的に行った。日本語と韓国語の形態素解析には「PORORO（Platform Of neuRal mOdels for natuRal language prOcessing）」(Heo et al. 2021) を利用した。POROROは形態素解析器として「MeCab」を使用している。MeCabは，日本語（IPA辞書）だけではなく，韓国語の形態素解析もできる[15]。このような字幕の前処理を行い，上の図2に示した生データは，次ページの図3のような形式となった。

このようなデータ形式は，JSON（JavaScript Object Notation）形式と呼ばれている。JSON形式はほとんどのプログラミング言語において使用可能であるため，字幕の前処理をこのような形式にしておくと，（ウェブ）アプリケーションを作ってデータを読み込んだり，利用する際に便利である。

なお，作品の字幕を図3のように形態素解析した結果，日本語は約76万形態素，韓国語は約293万形態素が得られた。このような差が生じる理由の一つは，韓国のドラマが日本のドラマに比べ，平均的に話数が多いからである。日本の作品を対象として集めた字幕（生データ）のサイズは約10MBであるが，韓国の作品の場合，約30MBとなっている。

[15] 「mecab-ko-dic」は，21世紀世宗計画の成果物を使用している（https://bitbucket.org/eunjeon/mecab-ko-dic/src/master/README.md（最終アクセス：2022年11月11日））。

```
},
{
        "Time": "15:44",
        "Subtitle": "（受付）おかえりなさいませ",
        "Translation": "어서 오세요",
        "POS": "[おかえりなさい/感動詞_ませ/助動詞]"
},
{
        "Time": "15:51",
        "Subtitle": "ただいま",
        "Translation": "나 왔어",
        "POS": "[ただいま/感動詞]"
},
{
        "Time": "15:55",
        "Subtitle": "なんか… 軽く食べる？",
        "Translation": "뭐 좀 먹고 싶어?",
        "POS": "[なんか/フィラー_軽く/形容詞_食べる/動詞]"
},
```


図3　字幕の前処理を行った例（『金魚妻（第1話）』）

5　Netflixの字幕をコーパスの構築に用いる際の問題点

ここでは，Netflixの字幕を利用してコーパスを構築・使用する際に，どのような問題点が生じうるのかについて述べる。筆者がNetflixから字幕を収集し，コーパスを構築する過程で気づいた問題点や注意点として，次の (7) が挙げられる。

(7)　a. 字幕に重複がある。
　　b. 字幕が修正される可能性がある。
　　c. 字幕が重訳されている可能性がある。
　　d. 字幕の翻訳者に関する情報がほとんどない。
　　e. その他の注意点（方言・外国語の台詞）。

ここに示した問題点の一部は，Netflix以外のOTTサービスの字幕においても観察されうる。DVDやBlu-ray Discなどで販売されている作品の字幕に，上記 (7e) 以外の問

題点が見られる可能性は低いと考えられるので，(7a-d) は，OTTサービスにおける翻訳字幕特有の問題なのかもしれない。以下では，(7a) から順に，例を挙げながら説明する。

5.1 字幕の重複

本稿で示した手順でコーパスを構築する場合，次のような理由によって字幕に重複が生じる可能性がある。まず一つは，ドラマの「あらすじ」「次回予告」や，作品中に登場する登場人物の「回想シーン」などによって，字幕に重複が生じる例である。たとえば，『コタローは1人暮らし』というドラマの第2話を見ると，すぐに第2話が始まるわけではない。最初は，第1話のあらすじが流れ，それが終わってから第2話が始まるのである。そのため，第1話の2分21秒に現れる「隣の203号室に越してきたさとうと申す」という字幕は，第2話の3秒にも，繰り返し現れる。

また，台詞以外のナレーションや，手紙・看板などに書かれている文字が翻訳され，台詞と同時に表示されるときにも，字幕の重複が生じる。たとえば，次の図4には複数箇所に字幕の重複が生じている。この重複は，映像に映っている登場人物の台詞と，映っていない登場人物のナレーションが同時に流れることによって生じた重複である。

11:07　秋はまた 恋の季節でもある 秋はまた 恋の季節でもある （小暮）どうですか？ できました？　가을은 사랑의 계절이기도 하다
11:11　（小暮）どうですか？ できました？ 何となく肌寒くなって 人恋しくなるもんだ 何となく肌寒くなって 人恋しくなるもんだ　날씨가 추워지면
11:14　何となく肌寒くなって 人恋しくなるもんだ 何となく肌寒くなって 人恋しくなるもんだ わあ ありがとうございます　사람이 그리워진다

図4　重複字幕の例（『深夜食堂: Tokyo Stories（シーズン 1，第10話）』）

図4の字幕が現れる場面を見ると，屋台にラーメンを食べにきた交番勤務の警官と屋台の店主が会話をしている映像・音声が流れ，さらに，映像には映っていない人物のナレーションも同時に流れる。図4において「どうですか？ できました？」「わあ ありがとうございます」は警官の台詞であり，それ以外の台詞はナレーションとして流れるものである[16]。このような場面では，次の図5に示したように，横書きの字幕に加えて縦書きの字幕も同時に表示される。

[16] この場面には，屋台の店主の台詞として「お待ちどお」があり，それがNetflix上では字幕として表示されている（図5を参照）。しかし，収集した字幕にはそれが含まれていなかった。なぜこのような現象が起こるのか，その原因はわからない。

図5　横書きの字幕と縦書きの字幕が同時に表示されている例（『深夜食堂: Tokyo Stories（シーズン 1，第10話）』一部改変）

図4における登場人物の台詞やナレーションは，本来，それぞれのタイムスタンプが異なるはずである。しかし，Language Reactorを利用して字幕をエクスポートする際，タイムスタンプのミリセカンド単位は切り捨てられてしまうので，図4のような字幕の重複が発生するのである。

タイムスタンプが異なることによって字幕に重複が発生する例は，他にもある。次の図6は，Language Reactorを利用して「Netflix字幕言語」を韓国語，「翻訳言語」を日本語に設定して収集した字幕の一部である。ここでは，日本語の翻訳字幕である「先生 カメラを意識しないで」が重複して現れている。

7:27　（하영）그리고 쌤　先生 カメラを意識しないで
7:28　카메라 의식 좀 그만해요! 아，사진 처음 찍어 봐요?　先生 カメラを意識しないで 撮影は初めて？
7:32　미안해，나 지금 최선을 다하고 있어　ごめん 頑張ってるんだが

図6　日本語の翻訳字幕の重複例（『도도솔솔라라솔（ドドソソララソ（第13話））』）

この重複は，韓国語字幕と日本語字幕におけるタイムスタンプが一致しないことによって生じるものである。図6と同じ部分を，「Netflix字幕言語」を日本語，「翻訳言語」を韓国語に設定して字幕を表示させ，エクスポートすると，次ページの図7のようになる。

図6と図7を比べてみると，図6の韓国語字幕のタイムスタンプ（7分28秒）と，図7の日本語字幕のタイムスタンプ（7分30秒）にずれがあることがわかる。このようなタイムスタンプのずれは，原文（ここでは韓国語）を翻訳した結果，翻訳字幕においては字幕の量が調整されることによって生じると考えられる。

7:27 先生 カメラを意識しないで （하영）그리고 쌤 카메라 의식 좀 그만해요! 아，사진 처음 찍어 봐요?
7:30 撮影は初めて？ 카메라 의식 좀 그만해요! 아，사진 처음 찍어 봐요?
7:32 ごめん 頑張ってるんだが 미안해，나 지금 최선을 다하고 있어

図7 韓国語字幕の重複例（『도도솔솔라라솔（ドドソソララソ（第13話））』）

以上のような字幕の重複は，マクロやプログラムなどを作成し，重複している文字列を機械的に削除することで対処することができる。しかし，一定の範囲に同じ文字列が重複して現れているとしても，それが図4や図6，図7に示したような重複なのか，それとも，「はい はい はい わかりました わかりました」のような反復（repetition）という言語現象なのかを，機械に区別させることは困難であると考えられる。そのため，上述した重複を完全に取り除くためには，人が字幕を確認し，どのような種類の重複なのかを判断してから削除するしかない。なお，冒頭で述べた「あらすじ」や「次回予告」などによる重複の問題は，字幕ファイル全体を対象として，最初と最後の数行を削除することで，ある程度は対処することができると考えられる。

5.2 修正される可能性がある字幕

OTTサービスの字幕は修正される可能性がある。この点に関しては，筆者が直接発見した例は現在のところないが，既存の字幕が修正された例が報告されている。次の図8に示した字幕は，Netflixで配信されている『사냥의 시간（狩りの時間）』という韓国映画の日本語・ドイツ語の翻訳字幕である。

53:42 それで今はどこに？ Wo seid ihr?
53:44 東海(トンヘ)にいます Wir sind am Ostmeer.

図8 日本語・ドイツ語字幕の例（『사냥의 시간（狩りの時間）』）

この字幕に登場する「東海」という語は，ドイツ語字幕において「日本海」という意味の「Japanischen Meer」と訳されていた[17]。しかし，該当箇所は現在，図8に示したごとく，「Ostmeer」に修正されている。Netflixには，映像の字幕に不具合がある場合，それを報告する機能がある。次の (8) は，その機能の「字幕とキャプション」という項目に設けられている選択肢である。

(8) a. 字幕またはキャプションに間違いがある (例: 誤字脱字，句読点，文法，または誤訳)。

[17] https://japanese.joins.com/JArticle/265265（最終アクセス：2022年11月11日）

b. 字幕またはキャプションが音声と合っていない。

c. 字幕やキャプションが正しく表示されない (例: 表示が速すぎる，または音声と同時に進まない)。

d. 字幕やキャプションが希望する言語で利用できない。

この機能による利用者の報告が，実際にどのくらいの字幕の修正に反映されているのかはわからない。しかし，一度公開された字幕が修正される可能性は常にあることを，念頭に置いておく必要はある。

このような字幕の修正に関する問題に対応するための一つの方法として，字幕を収集してから一定の期間を置いて，改めて同じ作品の字幕を収集する方法が考えられる。このように同じ作品の字幕を2回以上収集することで，同じ字幕のどのような部分が修正されているのかを，テキストエディタや（ウェブ）アプリケーションなどを使い，差分（diff）を出力することで見つけることができる。

5.3 重訳の可能性に関する問題

重訳（relay translation）とは，「翻訳された訳文（文字および口頭）を別の言語に（例えば中国語から英語に訳し，その英語をフランス語に）翻訳することである」(モナ・ガブリエラ 2013: 180)。篠原 (2018: 36) は，英語以外の言語の字幕を分析する際には，起点言語（たとえばスペイン語）から翻訳された目標言語（たとえば日本語）の字幕が重訳によるものなのかどうかを検討し，重訳による影響はないのか検証する必要があると述べている。

Netflixで配信されている作品の中には，日本語を韓国語に訳すのではなく，英語を介し，重訳をしたと思われる例がある。次の (9) は，干し芋を食べた女性が，隣に座っている男性に対して言う台詞の日本語字幕と英語の翻訳字幕である。

(9) a. 干し芋って 1 個でおなか膨れる（『やれたかも委員会（第1話）』）

b. The potato jerky really fills you up.（同上）

上記 (9b) の翻訳字幕では，*sweet potato*ではなく，*potato*となっている[18]。干し芋は，ジャガイモやサトイモではなく，サツマイモを原材料として作られるのが一般的だと考えられる。干し芋は韓国でも売られている食品であり，「kokwuma mallayngi（고구마 말랭이）」[19]と呼ばれている。ここで「kokwuma（고구마）」という語は，日本語で「サツマイモ」である。ところで，上記 (9a) に示した字幕を韓国語の

[18] *potato*という語は，この話で2回出現する。最初の登場は，*I found some potato jerky. You want some?*においてである。

[19] 韓国語のローマ字表記には，イェール式を用いる。

翻訳字幕で見ると，次の (10) のように，「kokwuma（サツマイモ）」ではなく，「kamca（감자，ジャガイモ）」となっている。

(10) kamca mallayngi mek-umyen sok-i kkway tuntun-hay
ジャガイモ 切り干し 食べる-れば 腹-が かなり ひもじくない-わ
‘切り干しジャガイモ食べると，結構お腹いっぱいになる。’

もし，この字幕が日本語を韓国語に翻訳したものではなく，日本語の字幕を英語に翻訳し，さらにそれを韓国語に翻訳した重訳だとすると，パラレルコーパスのサンプルとしては不適切だと考えられる。本研究では，この作品の字幕を収集対象から取り除いている。

上記 (10) の韓国語字幕は「양미정（ヤン ＝ ミジョン）」という人が作ったものである。名前からすると，この人は韓国人だと考えられる。しかし，この人がどの言語の翻訳を専門としているのか，あるいは，韓国語を母語としているのかなどを確認する術はない。次の5.4では，このような翻訳者に関する問題について述べる。

5.4 翻訳者に関する問題

Netflixの翻訳字幕の問題点として，翻訳された字幕が誰による翻訳なのか明らかではない場合がある点が挙げられる。5.3の最後の方で言及したように，字幕を誰が作ったのかを示している場合，名前から国籍を推測することができる。しかし，Netflixで提供されている翻訳字幕は，それを作成した人の名前を先述したように示している場合もあれば，公開していない場合もある。翻訳者の名前が公開されていない場合，たとえば，日本語の作品にある韓国語の翻訳字幕を作成したのが「韓国人なのか」「韓国語のできる日本人なのか」「日本語と韓国語のどちらも母語とするバイリンガルなのか」「在日朝鮮人なのか」「延辺朝鮮族なのか」推測をすることすらできない。

むろん，字幕を作成した人の名前が記されているからといって，字幕の質が保証されるわけではない。しかし，名前が公開されていると検索エンジンを利用して名前を検索し，他にどのような作品の翻訳を行ったのか，あるいは，どの言語の翻訳を専門としているのかを調べることができる場合もある。たとえば，5.3の最後の方で示した「양미정」という名前を検索エンジンを利用して検索した結果，英韓・韓英翻訳を専門とする翻訳者であることがわかったとすると，5.3の (10) に示した例も，日本語を韓国語に翻訳したものではなく，英語を韓国語に重訳したと考えることができよう。

翻訳者に関する二つ目の問題点として，一つの作品の字幕を，複数の翻訳者が作成している場合がある，という点が挙げられる。シリーズものの作品の翻訳は，一貫性が重要だと考えられる。たとえば，ある作品の登場人物が，相手の年齢に関係なく，いつも常体を使うとする。その登場人物の台詞を，シーズン1では常体で翻訳し，シー

ズン2では敬体で翻訳すると，一貫性のない翻訳になりかねない。このような一貫性のない翻訳は，翻訳字幕を利用して登場人物の台詞を分析するような研究では妨げとなりうる。Netflixで配信されているドラマの字幕翻訳には，二人以上の翻訳者が関わっている場合もある。たとえば，韓国の作品である『스위트홈（Sweet Home − 俺と世界の絶望 −（シーズン1，第6話））』の日本語字幕は小西朋子氏，『스위트홈（Sweet Home − 俺と世界の絶望 −（シーズン1，第7話））』の日本語字幕は福留友子氏が作成しており，同じドラマの字幕でも複数の翻訳者が作業を行っていることがわかる。このような場合，一貫性のある翻訳のために，翻訳者同士がどのくらいコミュニケーションを行っているのかは不詳である。

ここでは複数の翻訳者が字幕を作ることの問題点について述べているが，それがむしろ，長所になることもあるので，簡単に説明しておく。翻訳作品を利用した言語研究の場合，翻訳者の文体が問題点として指摘されることがある。ファン ＝ ウンハ (2021) は，対照研究に用いられるコーパスの妥当性を明らかにするために，韓国語と中国語の対照研究に使われたパラレルコーパスを調査した研究である。そこでファン ＝ ウンハ (2021: 268) は，複数話からなるドラマをサンプルとして構築したコーパスであっても，著者と翻訳者はそれぞれ1名に過ぎない点を指摘している。つづいて，この点は著者と翻訳者の言語特徴がテキストに影響を与えてしまう可能性があるとも述べている。もし，この点を問題として考えるのなら，先述したNetflixの複数人による翻訳字幕は，逆に問題点ではなくなると考えられる。なぜなら，翻訳者数が多ければ多いほど，翻訳者個人の言語特徴がテキストに与える影響は少なくなると考えられるからである。

以上のような重訳や翻訳者に関する問題に対処するための方法として，字幕を収集する前に映像の最後の方にあるクレジットを確認し，翻訳者に関する情報がある作品の字幕のみを収集対象とする方法が考えられる。

5.5 その他の注意点

Netflixの字幕を利用して構築したコーパスを言語研究に用いる際に，上述した問題点以外，次の (11) に示した点に注意する必要がある。

(11) a. 方言は標準語・共通語に訳される。
b. 日本の作品中の外国語の台詞が，日本語の翻訳字幕で表示される。

(11a) は，日本語や韓国語の方言が，韓国における標準語や日本の共通語に訳される点である。以下の (12) は，父が息子の部屋のドアを開ける前に，ドア越しに息子に確認をとる場面である。

(12) ほんまにええんか？父さんほんまに開けるぞ。（『ファイナルファンタジーXIV 光のお父さん（第2話）』）

「ほんま」や「ええ」は関西方言であるが，韓国語に翻訳された該当箇所を見ると，次の (13) のように「cengmal（本当）」や「kwaynchanhta（大丈夫）」といった韓国語の標準語となっている。

(13) cengmal kwaynchanh-a? appa cinccalo mwun yenta?
本当 大丈夫-か 父さん 本当に ドア 開ける
‘本当に大丈夫か？父さん本当にドア開けるよ？’

このように，日本語の方言が韓国語の標準語に訳される例は他の作品にも見られる。そして，韓国語の方言，たとえば，済州島方言は日本語に訳されるとき，共通語となる。野原 (2014: 105) は映像字幕における日本語の翻訳字幕の特徴に関して，次の (14) のように述べている。

(14) 字幕翻訳の場合，商業翻訳で実際にスクリーンに出る日本語訳を見てみると，アメリカの地域方言を日本のどこかの地域方言で差し替える手法はそう頻繁には見られない。地域方言はイメージが強いので使用にはリスクも高い。しかし敢えてリスクを冒すことで新しい翻訳の可能性が見つかるかもしれない。

もし，日本語の方言の音声・字幕を，韓国語の方言の音声・字幕に訳したドラマや映画の本数が，言語研究に耐えうる量に達すれば，それを利用した日韓の方言の対照研究を試みることができるであろう。しかし，現時点ではOTTサービスで提供されているドラマや映画の字幕を利用して方言の対照分析をすることは困難である。

もう一つの注意点として，作品中の外国語の台詞に関する問題が挙げられる。ほとんどの台詞が日本語である日本の作品に，一部，外国語の台詞が入っている場合がある。たとえば，『深夜食堂: Tokyo Stories（シーズン1，第4話）』という日本のドラマは，台詞のほとんどが日本語となっている。しかし，次ページの図9に示したように，台詞の一部には韓国人の役者による韓国語の発話も入っている。このようにドラマの中で短時間登場する外国語に対しては，それが何語なのかを「（韓国語）」のように示し，日本語の翻訳字幕を提供している。

このドラマの字幕を収集し，日本語の分析をする場合，図9にある日本語字幕は問題となりうる。計量的な研究をする際には，この日本語字幕も日本語の例として扱ってしまう可能性があるからである。このような誤りを防ぐための方法として，4.4で述べたように，生データの前処理をする段階で図9のような字幕を，あらかじめ削除する方法が考えられる。なお，図9には字幕ガイド（「（若者：韓国語）」）が挿入されているが，他の作品の字幕や，日本語以外の言語においても，このように規則的な形で

外国語であることが示されているのかはわからない。

19:05 （若者：韓国語） いい雰囲気だね〜　우와，분위기 좋은데
19:09 （若者：韓国語） おじさん 年考えてよ　아저씨，나이를 생각해야지
19:12 （若者：韓国語） 昼間から何やってんだよ　대낮부터 뭐 하는 거야?

図9　日本のドラマにおいて登場する韓国語の台詞が日本語の翻訳字幕で表示されている例（『深夜食堂: Tokyo Stories（シーズン1，第4話）』）

6　おわりに

本稿では，OTTサービスの一つであるNetflixの字幕を利用し，日韓・韓日パラレルコーパスを構築するための手順と方法を示した。そして，字幕を言語研究に用いることが，日本の著作権法とどのように関連しているのかについても述べた。また，OTTサービスの字幕を収集してコーパスを構築・使用する際に，どのような点が問題となりうるのかについても，具体例を示しながら説明した。

OTTサービスでは，一つの作品に対して多様な言語の字幕を提供している場合がある。そのような作品を対象として複数言語の字幕を収集すれば，多言語パラレルコーパスを構築することも，不可能ではない。また，OTTサービスの字幕を利用してコーパスを構築すると，字幕だけでなく，ウェブ上で映像・音声もいっしょに確認することができるので，マルチモーダルな研究を行うことも可能になると考えられる。

本稿では，パラレルコーパスの構築方法に重きを置いていたので，構築したパラレルコーパスを実際の言語研究に活用する方法まで示すことはできなかった。本稿で構築したコーパスを利用して行われた研究として，韓国語の「P-kinun P」と日本語の「PことはP」という反復構文の韓日対照研究をした徐 (2022) が挙げられる。この研究は，とりわけ韓国語の反復構文が準口語においてどのような振る舞いを見せているのか，構文的な特徴を分析し，さらに，日本語の翻訳字幕を通じて韓国語の反復構文が日本語の反復構文とどのような対応関係を成しているのかを明らかにしたものである。本稿で構築したパラレルコーパスを利用することで，他にはどのような対照言語学的な研究ができるのかを示すことは，今後の課題としたい。

参考文献

磐崎弘貞 (2002)「「連結」と映画CCデータベース」城生佰太郎（編）『映像の言語学』81-114. 東京：おうふう.

上野達弘 (2021)「情報解析と著作権――「機械学習パラダイス」としての日本」『人工知能』36(6): 745-749.

篠原有子 (2018)『映画字幕の翻訳学 : 日本映画と英語字幕』京都：晃洋書房.

柴田厚 (2016)「既存の放送メディアを揺さぶるアメリカのOTTサービス」『放送研究と調査』66(3): 2-13.

末吉亙 (2012)「情報解析と著作権」『情報管理』55(6): 434-437.

徐敏徹 (2022)「パラレルコーパスを利用した述語反復構文の韓日対照研究―「Pㄱ]는 P」構文と「PことはP」構文を中心に―」『第73回朝鮮学会大会』[2022年10月2日. 於：Zoom].

曺英南 (2004)「字幕付き映画における韓日の言いさし表現の対応関係―「述部有り」の言いさし表現を中心として―」『言語文化と日本語教育』27: 102-115.

野原佳代子 (2014)『ディスカッションから学ぶ翻訳学 : トランスレーション・スタディーズ入門』東京：三省堂.

モナ・ベイカー，ガブリエラ・サルダーニャ（編） (2013)『翻訳研究のキーワード』（藤濤文子（監修・編訳）・伊原紀子・田辺希久子（訳）） 東京：研究社.

尹盛熙 (2016)「日本語の翻訳字幕における省略・縮約の実現―韓国語との対照分析―」『社会言語科学』18(2): 19-36.

渡邉里菜・山田優 (2020)「英日字幕翻訳のコーパスベース研究―プロ字幕とファンサブの比較分析―」『MITIS Journal』1(2): 1-23.

キム ＝ ボエ 김보애 (2013)「지시어 'こ, そ, あ'의 자막번역 분석-영화 '러브레터'와 '쉘 위 댄스' 일한 자막번역에서-」『일본어학연구』(38): 3-19. [指示語「こ・そ・あ」の字幕翻訳分析-映画「Love Letter」と「Shall we ダンス?」日韓字幕翻訳から-].

ミン ＝ ギョンモ 민경모 (2020)「다국어 병렬 말뭉치의 구축과 한국어교육 연구에의 활용」『한국학논집』(78): 187-220. [多国語パラレルコーパスの構築と韓国語教育研究への活用].

ソ ＝ サンギュ 서상규 (1999)「언어 연구의 도구로서의 컴퓨터 - 국어정보학과 사전편찬학의 응용 -」『언어정보의 탐구』1: 242-276. [言語研究の道具としてのコンピュータ-国語情報学と辞典編纂学の応用-].

オ ＝ ギョンハン・ノ ＝ ヨンヒ 오경한·노영희 (2021)「Netflix와 Youtube 플랫폼 내의 영화 자막오역 분석을 통한 영상번역 실태와 개선점: 한국어 번역본을 중심으로」『디지털융복합연구』19(3): 25-35. [NetflixとYoutubeプラットホーム内の映画字幕の誤訳分析を通じた映像翻訳の実態と改善点：韓国語翻訳本を中心に].

ユ ＝ ヒョンギョン・ファン ＝ ウンハ 유현경·황은하 (2010)「병렬말뭉치 구축과

응용」『언어사실과 관점』25: 5-40. [並列コーパスの構築と応用].

韓国放送通信電波振興院 (2021)「디즈니 플러스의 아시아 태평양 시장 진출 동향과 경쟁력」『미디어 이슈 & 트렌드』(47): 53-60. [ディズニープラスのアジア太平洋市場進出動向と競争力].

ファン ＝ ウンハ 황은하 (2021)「대조분석을 위한 말뭉치의 타당성 연구 -한중 대조분석을 중심으로-」『이중언어학』82: 259-286. [対照分析のためのコーパスの妥当性研究-韓中対照分析を中心に-].

Aijmer, Karin (2008) Parallel and comparable corpora, In Lüdeling, Anke and Kytö Merja (eds.) *Corpus linguistics : An international handbook*, 275-292. Berlin: W. de Gruyter.

Heo, Hoon, Hyunwoong Ko, Soohwan Kim, Gunsoo Han, Jiwoo Park, and Kyubyong Park (2021) PORORO: Platform Of neuRal mOdels for natuRal language prOcessing, `https://github.com/kakaobrain/pororo`.

Mochizuki, Hajime and Kohji Shibano (2015) Re-mining topics popular in the recent past from a large-scale closed caption TV corpus, *International Journal of Future Computer and Communication*, 4(2): 98-103.

Moreau, Eponine (2021) The subtitling of taboo language terms in the French version of Orange is the New Black: A corpus-based analysis, *Proceedings of the Using Corpora in Contrastive and Translation Studies Conference (6th edition),* 114-117.

A Method for Constructing Parallel Corpus by Using Over-the-Top Media Service

Abstract

The present paper aims to introduce a method for constructing Japanese-Korean and Korean-Japanese parallel corpora using subtitles from an over-the-top (OTT) media service, and it highlights what points need to be considered in doing so. In order to build a parallel corpus using an OTT service, an OTT service that offers high-quality content and a sufficient number of subtitles must be chosen. In this paper, Netflix was selected as the OTT service, and Japanese and Korean (translated) subtitles were simultaneously collected using Language Reactor. Subtitles collected from OTT services may have problems such as repetition, correction, and relay translation, so caution is required when using them for linguistic research.

Keywords: Netflix, subtitles, copyright, quasi-spoken language, language resource

受領日　2022年10月7日
受理日　2022年12月11日