コーパスを教育利用するための要件としての適応学年指標 ― 文法コーパスと TED コーパス構築

田淵 龍二
ミント音声教育研究所
tabuchiryuji@nifty.ne.jp

1. はじめに

ウェブの普及とともに，ニュースやビデオが語学教育現場で使われることが増えてきた。こうした言語資源利用は，従来の教科書と何が異なるのか。本論文では，あらかじめ語学学習用に編集された素材を教材と呼び，それ以外の言語資源を自然素材と呼び分けて考察する。「不自然だ」との批判もあるほどに「教科書のための英作文」も見受けられるが，ウェブの記事がそのまま使われることも増えている。制作時には自然素材であった英文が，教材に掲載される事例から見ると，教材と自然素材の違いは編集方針にある。実際教科書にはニューワード，キーセンテンス，エクササイズなどが付加されている。そもそも教科書には対応学年（大学ではレベルや対象）が明記されている。

さてコーパスはどんどん大規模になっており，語学教育に豊かな多様性を与えてくれる。コーパスには，データドリブン（データ駆動型）学習としての利用もあるが，やはり研究者のためが主だと見受けられる。そこで，コーパスを教育，特に語学授業現場での教材とする要件について研究した。

読解及び聴解過程を扱う認知科学の知見としてのリーダビリティ（文の読みやすさ学年指標）と語彙レベル（語彙親密度学年指標），そして認知単位としてのチャンク（英単語にして５～７単語，時間にして２秒前後）と発話速度（単位時間当たりの言語情報速度）がある。これらの知見をコーパス構築に取り入れた検索サイト事例を紹介しつつ，言語教育と言語処理の接点としてのコーパスについて研究する。

2. 先行研究

英文の読みやすさを数値で表すリーダビリティ公式は百年以上に渡って数多く開発され利用され続けている。英語母語話者向けの公式を日本人向けにカスタマイズしたものとして，MGJP（Mint Grade Level for Japanese Readers；田淵・湯舟, 2016）がある。これは認知科学におけるワーキングメモリとチャンクの理論（Card, 1983）を土台として日本人英語学習者向けに作成されたもので，適応学年を算出する。式は，SyPP をフレーズごと母音数，CPP をフレーズごと子音数，PPS を文ごとフレーズ数，log を常用対数，a, b, c を定数として一般に，a×(3×SyPP+2×CPP)+b×log(PPS)+c と表記され，英語母語話者の場合は a= 0. 07662, b=19.554, c= 3.141 であり，日本人英語学習者の場合は a=0.07496, b=7.926, c=4.618 である。ここで言うフレーズとは，英数字以外の記号たとえばカンマなどで区切られた文字列である。ちなみに，母語話者の場合の公式（MGEN）は，教育でよく使われるフレッシュキン・ケイド公式と高い一致（相関係数にして 0.98）を示している。

英文の語彙難易度を数値で表す指標としては，頻度によるものと親密度によるものなどがあるが，適応学年を算出する公式としては，日本の学制をもとに作成された語彙レベル公式 VGL（田淵, 2017）がある。これは中高の教科書と大学受験問題を言語資源として統計処理したものである。式は，H をテキストの総単語数に占める大学受験語彙構成比率（%）として 30.371×H＋8.7914 である。この公式が算出する最低学年は 9，つまり中 3 であることがわかる。

MGJP と VGL が算出する学年指標はいずれも小学 1 年を 1 とするアメリカ式の積算法で，中学 1 年が 7，高校 1 年が 10，大学 1 年が 13 である。小数点以下は四捨五入して換算する。

TED Talks（https://www.ted.com/talks ）は世界的規模の講演会で，その模様は無料動画としてクリエイティブ・コモンズ・ライセンス（CCL）のもとに公開され，著作物の共有・使用・二次創作などを可能である。また数多くのボランティアにより世界中の言語に翻訳された字幕も公開されている。

SCoRE（http://www.score-corpus.org/ ）は約 1 万本の短文からなる文法項目用例コーパス（中條他, 2016）である。用例の難易度は初級中級上級の 3 段階に分類されている。これも CCL である。

Talkies（http://www.mintap.com/talkies/talkies.html ）は，2 言語字幕同時表示を特徴とする語学学習専用ウェブ・プレーヤーで，チャンク（字幕）反復再生や自動生成問題演習機能を備えている。

3. 目的

学習利用する言語資源の適応学年を素材選択肢に加えたコーパスを構築し，授業や個別学習を現場で支援することを目指す。教材として利用する言語資源として，スーパープレゼンで知られる TED Talks（最多視聴ビデオ上位 360 本）と日本の中高生を対象とした SCoRE 文法例文集（10,113 本）を使う。TED コーパスは，閲覧したいビデオの絞込みを目的とし，文法コーパスは，学習に使う例文の選択を目的とする。TED Talks は 2,500 本を超えているが，全部を対象とすると結果表示の絞込みなどに困難が予想されたので，視聴数の多い上位 360 本と手ごろな数にした。

4. 方法

TED コーパスではビデオ 1 本ごとにテキスト解析をおこないリーダビリティ（MGJP）と語彙レベル（VGL）を求める。文法コーパスではすでに初級中級上級の 3 段階に分類済みなので，それぞれの学年レベルを求める。文法コーパスではリーダビリティ（MGJP）のみを採用する。

TED コーパスも文法コーパスもどちらも簡便性を高めるために，フィルターによる既定値選択方式とする。TED コーパスでは利用者が所望のビデオに素早くたどり着けるように，ヒットしたビデオごとに日本語表題はもちろん，アイコン画像や紹介記事，そして高頻度語彙を出現頻度に応じた大きさでの表示とする。さらに，TED Talks 全体におけるそのビデオの難易度が一目で分かるような等高線図を採用する。授業やスマホでの利便性を考慮し，コーパスと Talkies プレーヤーを同一のドメインに配置する。文法コーパスはテキストだけなので，読み上げ機能を使って音声を提供する。Talkies をプラットフォームとして，2 つのコーパスを利用する仕組みとなる。

5. 結果

構築したコーパスは，TED が http://www.mintap.com/talkies/talkies.html?ted360 ，文法が http://www.mintap.com/talkies/talkies.html?score である。

TED コーパスの画面を図 1 に示す。上部に選択フィルターが並び，下にヒットした動画へのリンクが並ぶ。拡大図は語彙レベルを選択中で，中 3 以下から大 2 以上までの選択肢があり，他に文レベルや速さなど言語情報系列と収録時間やトピックスなど動画情報系列のフィルターが見える。画像アイコンのすぐ上にある等高線図は縦横がそれぞれ語彙レベルとリーダビリティで右上ほど難易度が高い。円形の点がそのビデオの総合難易度の位置を示す。

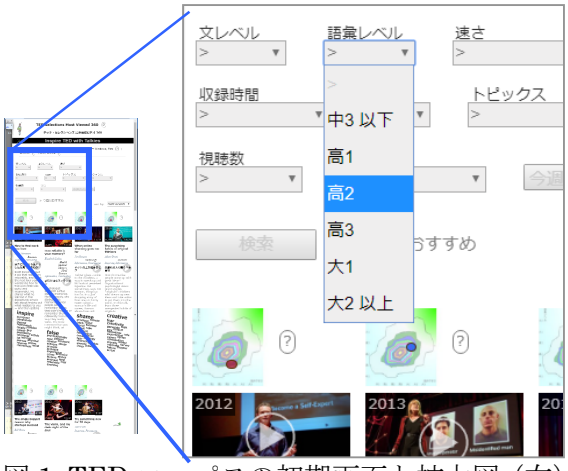

図 1. TED コーパスの初期画面と拡大図（右）

文法コーパスでは，難易度分類を学年に換算するため全文のリーダビリティ（MGJP）を求めた。結果を図 2 に示す。級毎の平均値は順に 8.2（中 2），10.3（高 1），13.4（大 1）で，隣接する級間の効果量（Cohen's *d*）は初級中級で 1.97，中級上級で 1.66 となり，明確に区分されていることが判明した。そこで初級・中級・上級をそれぞれ中学・高校・大学受験と名付けた。上級を大学受験としたのはピークが高 3 にあることを考慮したものである。

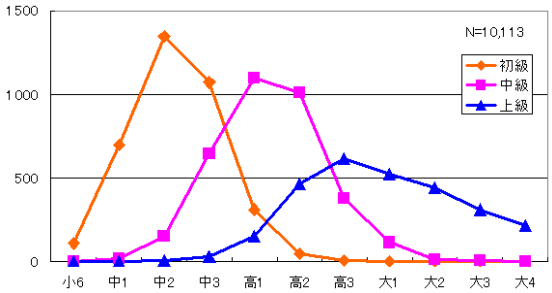

図 2. 文法用例の級別リーダビリティ分布

文法コーパスの画面を図 3 に示す。SCoRE との操作親密性を考慮してレイアウトを共有させた。縦に 4 つの欄があり，左から文法項目，キー語彙，例文と並び最右翼が利用者が選んだ Talkies プレーヤーへのリンクで，再生候補である。図では中高だけを選んでいる。中学は赤いアイコンで，高校は青い。

図 3. 文法コーパスの画面

TED コーパスも文法コーパスも再生ボタンをクリックすると Talkies で再生がはじまる。その様子を図 4 に示す。左がコンピュータによる文法コーパスで，順にテキストが読み上げられていく。このときキー(文法表現の要所）は自動的に空所となっているので，聞き取るよう努めることになるが，わからない箇所はマウスでポイント（あるいはタッチ）することで単語を覗き見る（赤字）ことができるようにしている。図 4 の右はスマホでの TED 再生画面で，映像の下に日本

語字幕が付き，下段の英語リストでは音声に合わせて該当部分がハイライトされていく。どちらのコーパスも再生時には右側に畳み込まれているので，タブをクリック（あるいはタップ）して引き出す。

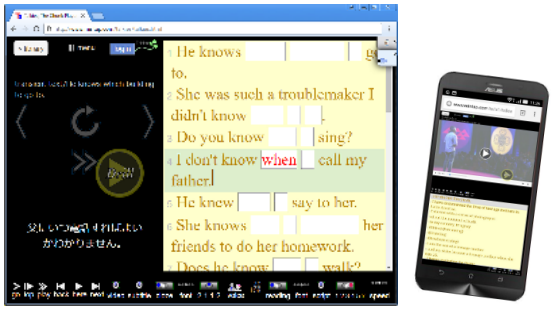

図 4. Talkies による再生画面（右はスマホ）

6. 考察

TED コーパスは 2017 年 9 月，文法コーパスは 12 月に公開したばかりなので，実証研究はこれからであるが概ね好評で，授業で毎週利用しているとの報告を得ている。特に，教材の選定から再生利用までの流れが円滑かつ迅速であること，音声に合わせて文字が同期的に提示されること，苦手な音声部分だけを反復して聴取できること，そして本研究のテーマである学年表示が e-ラーニング学習の敷居を下げ動機づけを強めると期待できるなどの声が寄せられ始めている。

7. 議論

文法コーパスの原作は中条清美氏をリーダーとする科研費研究チームが制作した SCoRE である。氏らはこれをクリエイティブ・コモンズ（CC）として公開しており，かつ，それを Talkies 対応に翻案することに賛同していただけた。また TED も TED Talks Usage Policy（https://www.ted.com/bout/our-organization/our-policies-terms/ted-talks-usage-policy ）で CC であることを明記している。このような最新の著作権ルールがコーパスを活用した学習機会の増大をもたらし，文化の発展に結びつくと期待されるが，他方では異論もあることからも，法改正も見据えた議論が望まれる。

8. おわりに

研究者による研究者のためのコーパスが，これからは，先生と生徒のための教具という方向に枝を茂らせていくと思われる。その際，素材の適応学年表示は，ビデオやニュース記事や文献などの言語資源を教材として提供するサイト（コーパス検索エンジン）の必須要件となるであろう。

参考文献

Card, S. K., & Moran, T. P., & Newell, A., (1983). *The psychology of human-computer interaction*. Lawrence Erlbaum Associates, Inc.

田淵龍二・湯舟英一（2016）．「音韻符号化の予測時間に基づく日本人英語学習者向けリーダビリティ公式の開発」. *Language Education & Technology, 52*, 359-388.

田淵龍二（2017）．「日本人英語学習者向け語彙レベル適応学年算出公式の試験的開発」. *LET Kanto Journal, 1*, 25-35.

中條清美，内山将夫，赤瀬川史朗，西垣知佳子（2016）．「データ駆動型英語学習における教育用例文コーパス SCoRE の活用」．言語処理学会第 22 回年次大会発表論文集 2016_P19-3, 1081-1084.