会話とスピーチの映像による日英対訳コーパス構築
― 自律学習を促す適応学年レベルのあるコンコーダンス

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

1. はじめに

誰でも手軽に使える英語学習用対訳コーパス構築が求められている。研究・学校教育・自律学習の 3 つのコーパス利用面のうち，今回は自律学習に重心を据え，一斉授業も視野に入れた。初学者による自律学習には，操作の簡明性，学習者レベル適応性が要件となる。学習者レベルがコーパスに影響するのは，検索語句入力と結果提示である。検索語句入力を支援する仕組みを工夫し，共起語の認知性を高める工夫（KWIC）の他，結果英文を中高大レベルで絞り込むフィルターを設置する。

2. 先行研究

単語難易度研究では，「分野が調整されている均衡コーパス」の方が「性能が高い」（江原, 2017）との報告がある。また田淵（2017）は検定教科書などのコーパスから語彙適応学年指標（VGL）を作成した。これらは，大規模コーパス頻度指標よりも，学習者に寄り添った分野コーパスの方が，親密度を反映しやすいとする考えで，背景知識（長期記憶）を活用する認知過程の認知科学的知見（Card, 1983）と符合する。大規模均衡コーパスを長期記憶に蓄えた巨人は実在しない幻だからである。

対訳では文ごとに訳す「文訳」が一般的で，認知科学的視点を取り入れた「フレーズ訳」は普及途上のようだ。Voice Of America（VOA）ニュース音声にフレーズ訳をつけた神田・湯舟・田淵（2010）では音声の自然な区切りであり意味の塊（chunk）でもある呼気段落（breath group）ごとに訳文（チャンク訳）を付けている。呼気段落は 2±1 秒の音声連続（田淵・湯舟, 2015）である。他方 TED Talks の字幕表示時間は 2.7±1 秒（田淵・湯舟, 2017）である。前者は映画の呼気段落長実測値，後者は講演の字幕表示時間と素性の違いはあるものの，概ね似通った分布とされる。このことから，TED 字幕をチャンクとみなしてフレーズごとの対訳とすることが，人の認知過程に即していると判断される。

3. 目的

文脈を映像で学習できる会話とスピーチの対訳コーパスを構築し，検索結果の英文から学年レベルに適応したものを絞り込める機能をもったオープンサイトとしてのコンコーダンス公開を目的とする。

4. 方法

(1) 実用性と品質，(2) 対訳による意味理解，(3) 分野として会話とスピーチ，(4) 初学者でも検索語を見つけやすい工夫，(5) ヒットしたフレーズ（字幕）から元の文章（場面）や作品を文脈単位で

視聴可能。さらに，(6) 自動生成問題演習を可能とし，(7) 自律的学習をうながす作りとする。言語資源映像として，会話には映画映像コーパス・セリーフ（Seleaf: http://www.mintap.com/ ），スピーチには TED（http://www.ted.com/ ）を採用した。Seleaf コーパスは名作映画 22 本，TED コーパスは 2007 年 12 月までの Talks 2,526 本のうち日本語字幕が具備されている 2,415 本から，英語と日本語の字幕数が同じ 1,971 本を採用した。

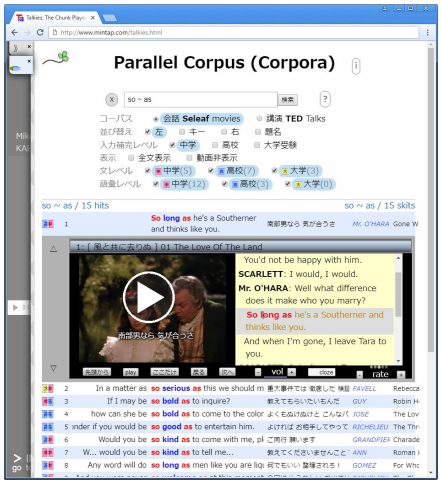

図 1. 対訳コーパス・コーポラの画面

5. 結果

構築したコーパス・コーポラ（Parallel Corpus / Corpora はオープンサイト Talkies をプラットホームとしたスライド窓で開く（図 1)。表題下のボックスに語句を入力して検索ボタンを押すと語句を含む字幕シーン一覧を出力。1 字幕 1 行。最大 500 項目。先頭行のみ自動的に再生画面表示。画面左がビデオと和文字幕，右が英字幕リスト。ヒット字幕前後 20～30 秒の文脈確認が可能。プレーヤーは語学教育専用仕様。再生箇所強調表示，字幕単位音声反復，再生速度調整，クローズテストに対応。TED では原作ビデオへのリンクがある。

図 2. 選択ボタン群

検索語句は赤と青で強調表示される。一覧表では検索語句を縦に整列させる方式を採用。ヒット語句配置切替は，入力欄下ツールボタン「左・キー・右・題名」(図 2)。たとえば「キー」を選ぶとヒット語句がアルファベット順となる（図 3)。検索語句は「so * as」で，指定単語の so と as が赤，ワイルドカード（*）でヒットした任意語句が青。他に「コーパス，入力補完レベル，表示，文レベル，語彙レベル」の選択がある。入力補完（auto complete）は入力文字列を含む候補を選択可能にする機能で，たとえば a.c.c.o…と打ち込んでいくと according to，account for，hold * accountable for などがアルファベット順に提示される（図 4)。表示オプションの全文表示は，1 行で表示しきれない部分を折り返して可視化する。項目クリックで当該映像が視聴できる。動画非表示をオンにすると動画シーンの自動読み込みをしない。

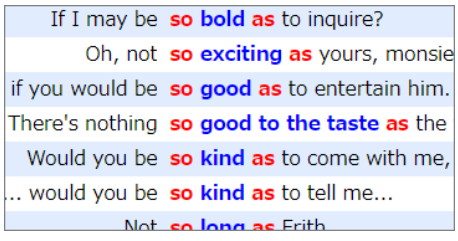

図 3. キー昇順の KWIC 方式

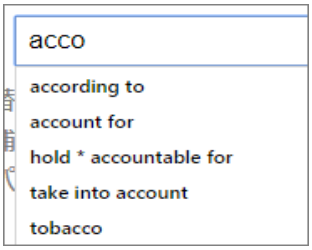

図 4. acco 入力時の補完候補

見出し語検索を基本とし，go で goes，went などがヒット。文字大小の区別はしない。書式には正規表現を採用。任意単語や文字検索，綴り字検索，論理和検索，疑問文検索が可能。検索は字幕単位で行う句検索仕様とした。入力補完辞書は中学・高校・大学受験の 3 種。中学が慣用句 371 個，単語

1,974 個の計 2,285 語句，高校が慣用句 185 個，単語 573 個の計 758 語句，大学受験が慣用句 1,461 個，単語 4817 個の計 6,278 語句，総計 9,321 語句を搭載。分類は学制別単語リスト（田淵，2017）などに従った。クローズテストでは，ヒットした語句が単語ごとに伏字となる。

検索結果の英文には日本の学制（中学・高校・大学）に応じた文レベルと語彙レベルを示すアイコンが示されていて，学制ごとの絞込みができる。

6. 考察

(1) 品質，(2) 対訳，(3) 分野（会話とスピーチ），(4) 初学者でも検索語を入力しやすい工夫，(5) 元の文章（場面）や作品を視聴，(6) 演習，(7) 自律的学習をうながす作りの観点から考察する。

(1) 世代を超えて定評のある名作映画と，世界中のさまざまな専門家によるプレゼンテーションで評価の高い TED Talks を採用することで学習素材としての品質が確保された。(2) 英文だけが並ぶコーパスでは意味理解面に困難があるが，対訳が意味理解を支援する。(3) 会話とスピーチの 2 分野採用で学習目的に応じた活用が期待できる。(4)「検索語が思い浮かばない」「綴りが不正確」との困難に対し，入力補完機能導入で，①タイプ省略，②うろ覚えでも検索語句発見，③打ち込んだ単語の慣用表現発見の機会増大などが期待される。(5) 音声と映像による場面表現の視覚化は，文字だけのコーパスを超えた臨場感ある言語体験を可能とした。(6) クローズテストではワンクリックで検索語句が隠され，反復学習による定着が期待される。(7) 一流の素材，対訳補助，学制別入力補完，検索語句ハイライト，文レベルと語彙レベルでの絞込み，各種並び替え，動画視聴，クローズテスト，原作ビデオへのリンクなどの仕掛けが初級者の動機づけとなり，自律的学習につながること（データ駆動型学習）が期待される。また HTML5 をベースとしたことで，スマホでも利用でき，自律学習機会の拡大に貢献すると思われる。

7. 議論

対訳コーパスは，学習面では語学における作文支援，研究面では機械翻訳への応用も期待されている。それらの観点から今回構築したコーパスを振り返ると，(Ⅰ) 検索語句は英語のみで日本語での検索には未対応，(Ⅱ) 映画コーパスは字幕ごとに和訳が正確に対応するフレーズ訳としてあるが，TED コーパスでは字幕とその訳文が意味的にずれているセンテンス訳となっている場合が多い（表 1）。表 1 の右端は筆者による対訳例である。英語字幕ごとに黄青緑灰で配色した。TED の日本語字幕部分で

表 1. TED 対訳コーパスにおける日英字幕のズレとフレーズ訳（右）

#	英語字幕	日本語字幕（文訳）	対訳例（フレーズ訳）
80	And having said to your dad, Nic,	君のお父さんのニックに	お父さんと約束したから
81	that I would try to teach you, I was then slightly confused	君にピアノの演奏を教えると約束しておきながらも	君に教えようと思ったけど 困ったことに
82	as to how I might go about that	ピアノに近づくことができないのに	どうしたらいいかわかんないよ
83	if I wasn't allowed near the piano.	どうやって教えるんだと困ってしまいました	だってピアノに近寄れないんだから

memo: from TED Talks: *In the key of genius* by Derek Paravicini and Adam Ockelford

のずれが目立つ。フレーズ訳採用で英語の語順による直読直解効果が期待できる。作文支援や機械翻訳に利用するにはセンテンス単位に編集すればよいが相当な作業量となる。そこで TED 翻訳ボランティアに始めからフレーズ訳をお願いしたいが，センテンス訳の方が慣れていることもあり，簡単には移行できないと思われる。語学学習用にフレーズ訳を専門にするチームや文化を育成することも課題であろう。（Ⅲ）均衡コーパスと分野コーパスの考えがある。語学学習用としては今回の分野コーパスが適していると筆者は考えている。この視点から見ると分野としてニュース・コーパスを加えたい。しかし，ニュース映像利用には著作権処理の壁が予想される。(Ⅳ) 映画コーパスは著作権の保護期間が終了した名作映画に限られている。TED コーパスは，TED の理念（ideas worth spreading; 価値ある考えを広めよう）及び利用規約に明記されているクリエイティブ・コモンズ・ライセンスにも支えられてかなり自由度の高い活用が可能となっている。しかし，アメリカ（フェアユース）と日本（例外規定）の法体系の違い，さらに日本では情報通信技術の進歩に法改正が追いつかず，過度にリスクを恐れる風潮を踏まえると，まずは学術団体や研究者間での合意形成が必要である。また，環太平洋経済パートナシップ協定（TPP）における著作権保護期間の 70 年への延長や非親告罪化の動きも，コーパス構築に与える影響の観点から注目が必要であろう。(Ⅴ) 語学学習では，語彙表現の定着に反復練習が欠かせないとされることから，構築したコーパスに簡単なドリル（クローズテスト）を装備し，より本格的にはプラットホームである Talkies の演習機能に譲る仕様にしてある。

8. おわりに

学習者の学力に即したユーザーフレンドリーなコーパスとなっていれば幸いである。

謝辞

50 万フレーズに及ぶ日英対訳コーパスを可能にしてくれた TED の理念に感謝する。また，地道な翻訳作業に携わり，現在もなお週 3～4 本の翻訳をアップしている数多くのボランティアに感謝する。

参考文献・参考資料

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

江原遥（2017) E5-3 「生コーパスからの単語難易度関連指標の予測」. 言語処理学会 第 23 回年次大会発表論文集.

神田明延・湯舟英一・田淵龍二, (2010). 「チャンクで速読トレーニング」. 国際語学社.

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.

Parallel Corpus / Corpora (2018) http://www.mintap.com/talkies/talkies.html?corpora