【研究論文】

オープンデータによる英語構文事例検索システムの可能性：TED Corpus Search Engine を例として*

Perspectives on Using Open Data to Build an English Grammatical Construction Search System: Introducing the TED Corpus Search Engine

長谷部 陽一郎
同志社大学
HASEBE, Yoichiro
Doshisha University

Abstract

This article discusses the great potential of the current movement toward open data in the field of English language education. For this purpose, I introduce and illustrate an online system specially designed to aid instruction of constructional patterns in English. The system, the TED Corpus Search Engine (TCSE), uses a set of open data of English presentations from TED Conferences and provides many functionalities, including full-text utterance retrieval of both transcript segments and corresponding videos using various search operators. The article shows that the open data movement, alongside technological innovations, could greatly benefit English language educators if they increase their awareness and make necessary efforts. I also present and discuss the fact that by making use of the appropriate data, it is possible to bridge the apparent gap between work in theoretical linguistics and the practice of language education. The TCSE is equipped with several features that help the user to extract instances of constructional patterns of English from the TED Talks corpus. Their designs are based on observations and theorizations from the field of cognitive linguistics.

キーワード： TED Talks、コーパス、構文、認知言語学
Keywords： TED Talks, corpus, constructions, cognitive linguistics

はじめに

インターネットとそれを基盤にした技術の発達は社会に様々な変化を生じさせている。その中で近年、注目が集まっているのは、情報の「オープン化」を目指す種々の活動である。物理的な距離や時差を超えて情報をやりとりすることが可能になった今、集積されたデータを共有することで、多くの人がその価値を享受できるようになる。また、データの応用的利用を通じて、技術的イノベーションを含む様々な展開が期待できる。このオープン化の波は、当然、教育の分野にも及んでおり、昨今、様々な試みが行われている。大学を始めとする高等教育機関の講義をインターネット上で公開する枠組みであるオープンコースウェア（OpenCourseWare, OCW）などは代表的な例である[1]。このような潮流を背景に、本稿では英語教育の実践に役立つシステムの開発にオープンデータを用いることの意義について考えると共に、実際にオープンデータを用いて開発したシステムの事例を示す。

本稿の構成は以下の通りである。1 節では、オープンデータの定義と特徴について論じる。今日、オープンデータ活動は国際的に重要な取り組みとなっており、日本でも総務省が「オープンデータ戦略」を推進している[2]。ここでは「オープンデータと言えるための条件」を明らかにし、基本的な意義について考える。また、オープンデータを利用した英語表現事例検索システムの構築という具体的な目的の達成に求められるデータの仕様について検討する。2 節では、そのような仕様を満たす、優れたオープンデータの例として TED が提供する英語プレゼンテーションのデータを紹介する。TED は 2,000 以上の TED Talks と呼ばれるプレゼンテーションの動画・音声とトランスクリプトを公開しており、それらは一定の条件のもと、自由に利用・再配布することが認められている。3 節では、この TED Talks データを利用して筆者が開発した TED Corpus Search Engine（TCSE）と称する検索システムの基本的な設計と機能を紹介する。続く 4 節では、TCSE の機能と理論的な背景との関係について論じる。TCSE の設計にあたっては、認知言語学において重視されている「構文」の概念に着目し、構文事例の検索・抽出を効果的に行うための各種機能の実装を目指した。それは実用性を前面に打ち出したシステムの構築を通じ、理論と実践を結びつける試みでもあった。本稿の最後には、英語表現事例検索システムの開発、ひいては英語教育という大きな取り組みにおいて、データのオープン化の流れが今後どのような変化をもたらし得るかについて述べる。

1. 言語教育資源としてのオープンデータ

1.1 オープンデータとは何か

そもそもオープンデータ（open data）とは何か、ここではまずこの問題に触れておきたい。インターネットは本来的に情報共有のインフラストラクチャーであり、インター

ネット上で公開され､認証手続きなどを経ることなくアクセス可能なあらゆるデータは、ある意味ですでに「オープン」と言える。しかし、単にアクセス可能であることと、利用に際しての種々の制約は別の問題である｡誰でも閲覧できる文書ファイルなどを含め、全ての制作物は著作権で守られており、その二次利用に関しては細心の注意が必要となる。一方で、価値のあるデータは、それ自体有用であるだけでなく、他のデータとの組み合わせにより、あるいは何らかのシステムに組み込むことにより、さらに高い価値を生み出すことがある。このような観点から、近年、科学・技術の分野では様々な種類・規模のデータがオープン化され、新たな発見や革新を生み出す基盤となっている。行政においても、透明性・信頼性の向上、各種事業の効率化や経済活動の促進といった目的から、公共データのオープン化を行うようになっている。そのような中、日本の総務省は「オープンデータ戦略の推進」ウェブサイト上で、データが「オープン」と言えるためには、

1. 機械判読に適したデータ形式で、
2. 二次利用が可能な利用ルールで公開されたデータ

である必要があるとしている [3]。ここには 2 つの条件が挙げられており、いずれも、データが単に「公開される」だけではなく､「適切に利用される」ために必要な条件となっている。1 の「機械判読に適したデータ形式」とは、具体的には PDF や JPG といったファイル形式や、XML や CSV といったテキストフォーマット形式を指す。広く普及した形式でデータを提供することで、スムーズな利用が可能になる。また、2 にある「二次利用が可能な利用ルール」が示されることで、利用者は再配布や改変が可能な範囲を明確に理解し、安心して新たな取り組みにつなげていくことができる。

さて、このような性質を持つオープンデータの公開が促進されることで、英語教育のあり方も少なからず変化する可能性がある。教師は日常の業務として、授業の目的や学習者のレベルに応じて教材を選定する。典型的にそれは教科書や市販のテキストであるが、それらに加えて補助的な教材を自ら作成することも多い。従来、そのような教材の作成は個人レベル､もしくは組織レベルで行われ､必ずしも広範囲な協働ではなかった。しかし今後は、より大きな教育コミュニティの中で、教材作成のためのデータをオープンな形で共有していくことが可能となるだろう。それは、多くの教師が直面する課題の発見と解消につながる。また、優れた実践の方法を模索する基盤となる。さらに、既存の素材を活かした新たな取り組みや枠組みの開発を促すと考えられる。

1.2 英語表現事例検索システムに適したデータの要件

具体的な例として、オープンデータを用いて英語表現の事例を検索するシステムの開

発について考えてみたい。もちろん、英語教育のために役立つデータは必ずしも教材作成の素材に限られる訳ではない。例えば、学習者の動機、学習歴、到達度などに関する統計情報なども、興味深い取り組みを可能にするだろう。しかしここでは、今日利用可能なオープンデータを用いた実践的な取り組みの事例として、英語表現検索システムの構築というテーマを取り上げることにする。この目的を達成するためにはどのような課題があるだろうか。

英語表現の事例を検索する方法としては、既存の教科書、参考書、辞書などを利用する他に、コーパスを利用する方法がある。イギリス英語の大規模コーパスとしては British National Corpus（BNC）が有名であり[4]、アメリカ英語に関しては Corpus of Contemporary American English（COCA）がよく知られている[5]。これらは様々な分野の英語テキストをバランスよく集積した均衡コーパス（balanced corpus）であり、英語学・英語教育の分野で盛んに用いられている（cf. 石川 2008; 投野 2015）。しかし、教育の現場における利用を想定した場合、事例の検索・抽出という目的でこれらを利用することには、少なくとも 3 つの障壁がある。第 1 にコーパスを構成するデータが必ずしもオープンでないため､結果を共有したり二次利用したりするのが難しいことである。第 2 に、大規模コーパスから得られる事例の多くは、断片的なテキストの一部だということである。言語学の語法研究であれば、ある形式を持った表現の存在や頻度を確かめること自体が重要な場合もあるが、教育的な目的で事例を用いる際には、それらがどのような場面で発話されたかについての情報が重要である。したがって、抽出された事例を含むテキストは、できるだけ完全な形で取得可能なことが望ましい。第 3 に、BNC や COCA などは基本的にテキストデータから構成されるコーパスであり、書き言葉（written words）セクションのデータはもちろん、話し言葉（spoken words）セクションのデータであっても、発話時の動画や音声を得ることは難しい。これまで英語教育の現場でも文字データが中心的位置を占めていたことは否めない。しかし、総合的な技能を育成することを目指す教育活動の中では、音声などの非文字データが重要になってくる。

英語教育に資する表現事例を採取するには、BNC や COCA といった大規模均衡コーパスを利用すること以外に､Google に代表されるインターネット検索エンジンを利用する方法もある。大規模コーパスより手軽であり、なおかつ、検索対象となるデータがさらに大きいという利点がある。しかし、少なくとも英語教育に役立つ事例を採取するという目的において、理想的なデータソースにはなりにくい。先ほど、1）データのオープン性、2）データの完全性、3）動画・音声の利用可能性という 3 つのポイントについて述べたが、Google などの検索エンジンを用いる場合にも、この 3 点が障壁となる。

従来、これら 3 つのポイントを満たすようなデータを得ることは簡単ではなかった。しかし、近年、情報のオープン化が推進される中で、様々なデータが一定の条件のもと

に自由に利用できるようになった。それらは必ずしも教育的な目的で制作されたものではない。しかし、多様なデータの中には英語教育上の諸活動に有用な性質を持ったものがある。TED が公開する英語プレゼンテーションのデータは正にそのようなものであり、1）データのオープン性、2）データの完全性)、3）動画・音声データの利用可能性という 3 つのポイントを満たしている。そこで次節では TED Talks のデータの内容と形式について見ていきたい。

2. オープンデータとしての TED Talks

2.1 TED とは何か

TED（Technology, Entertainment, Design）は「世界に広める価値のある内容」を持ったスピーカーたちがプレゼンテーションを行うカンファレンスを定期的に開催する非営利団体である[6]。実施されたプレゼンテーションの動画（多くの場合 5 分〜20 分）は基本的にすべて公式ウェブサイト（http://ted.com）および YouTube 上で公開されており、現在、英語によるプレゼンテーションの数は約 2,200 となっている。また、トークの内容はボランティア・メンバーによりトランスクリプトとして書き起こされ、テキストデータの形で公開されている。同じくボランティアにより日本語を含む他言語への翻訳も精力的に行われている。

TED は「オープン」であることを重要な理念としており、トークの内容は Creative Commons ライセンス（BY-NC-ND）のもとに使用・再配布することができる[7]。商用利用はできないため、特別な許諾を得ない限り、TED のデータを利用して作成した教材を販売することはできないが、出典を明記し、改変を行わないという条件のもとで、様々な用途に利用することが認められている。なお、TED のオープンな姿勢は、かつてデータ利用のための Application Programming Interface（API）が公開されていたことにも現れている。API とはコンピュータ上のプログラムからデータにアクセスするための規約であり、TED では、データの有効で幅広い利用を促進するため、TED Talks を利用したソフトウェアを開発する技術者向けに、プレゼンテーションの詳細情報にアクセスする手段を公開していた。API の提供は 2016 年の半ばに終了してしまったが、教室内での利用を含め、データ自体の利用については従来と同様、オープンな姿勢が保たれている。

2.2 英語表現事例検索用データとしての TED Talks

このようなことから、TED Talks のデータは、1.2 節で述べた英語表現事例検索システムの構築に必要な 3 つの要件のうち、1 つ目の「データのオープン性」を満たしていると言える。では 2 つ目の「データの完全性」についてはどうか。検索結果として得ら

れた事例の文脈をどれだけ網羅的に把握することができるかという要件である。前述の通り、TED Talks はそれぞれが概ね 5 分から 20 分のプレゼンテーションで構成されている。内容は、科学、技術、政治、経済、芸術、福祉、生活など多岐にわたるが、各プレゼンテーションはそれぞれ内容的に独立している。また、プレゼンテーションの全文はもちろん、スピーカー名、タイトル、シノプシスなども公開されている。したがって、検索結果として得られた表現が「誰によって」「どのような目的で」「どのような話の流れの中で」発せられたのかといった情報が取得可能である。したがって、「データの完全性」についてもクリアしていると言ってよい。

それでは、第 3 の要件である「動画・音声データの利用可能性」についてはどうか。TED Talks は動画・音声データと共に公開されている。加えて、ボランティア・スタッフの手によって、英語のトランスクリプトが作成されている。それらは、動画ファイル再生時の字幕として利用することを想定したものであり、動画と同期させるため、一定のまとまりごとにプレゼンテーション開始時からの経過時間情報が付与されている。これにより、TED Talks から得られた英語表現事例は、それを含むプレゼンテーション全体の動画・音声ファイルが利用可能なだけでなく、表現が発話される様子を個別に確認することが可能となっている。したがって、「動画・音声データの利用可能性」の要件を完全に満たしている。

さて、英語表現事例検索システムの構築には、さらにもう 1 つの要件がある。それは「データの網羅性」である。大規模均衡コーパスやインターネット検索エンジンは、検索対象テキストの量が膨大であるため、個々のデータの信頼性と抽出の精度はともかく、必要となる表現事例を含んでいる可能性、すなわち「データの網羅性」が高い。一方で、TED Talks はあくまで約 2,200 のプレゼンテーションのトランスクリプトから成る、いわば特殊コーパス（specialized corpus）であり、大規模均衡コーパスやインターネット検索エンジンに比肩する網羅性を望むことはできない。また、すべてがプレゼンテーションのデータであることから、表現にある種の偏りがある。生活に密着した多くの日常的な表現や、あるいは書き言葉だけで用いられる表現の数々については、十分な事例を得ることができない可能性がある。しかしながら、約 2,200 のプレゼンテーションから得られる英語テキストの延べ語数はおよそ 500 万語であり、これは最初の本格英語コーパスとしてよく知られる Brown Corpus（Francis and Kučera 1964）の収録語数の 5 倍である[8]。また、異なり語数は 7 万 5 千を数える。この数字は一般的な第 2 言語としての英語学習者にとって習得が必要とされる語彙数を大きく超えており、部分的な漏れはあり得るものの、学ぶべき表現のかなりの部分を実質的に網羅していると考えられる。したがって、第 4 の要件「データの網羅性」についても、基本的にはこれを満たしていると言ってよいだろう。

以上のように、TED Talks のデータは、実際の英語教育に役立つ表現事例検索システ

ムを構築するのに適したデータの要件を満たしている。1 節で述べたように、今日、世界で活発に行われている「データのオープン化」は、価値のある情報を共有し、二次利用による展開を促すものである。そのようなオープン化の 1 つの形である TED Talks のデータを効果的に用いることで、英語教育の領域においても、ある種のイノベーションが期待できる。次節ではそのような試みとして開発されたシステム TED Corpus Search Engine の概要について述べる。

3. TED Corpus Search Engine について

3.1 基本的な機能と使用方法

筆者は TED Talks のプレゼンテーション・データを用いて、特定の英語表現の実例を検索し、結果を一覧表示するとともに、当該箇所の動画・音声をピンポイントで再生できるシステム TED Corpus Search Engine（TCSE）を開発した。下記 URL で公開しており、ウェブ・ブラウザから利用可能である。

http://yohasebe.com/tcse

2014 年 11 月に最初のバージョンを公開した後、様々な機能を加えながら、検索対象となるプレゼンテーションの追加・更新を継続的に行ってきた。現在（2016 年 9 月 25 日）、2,172 件の英語プレゼンテーションから表現事例を検索し、該当する部分の動画をピンポイントで再生することが可能となっている。2,172 のプレゼンテーションの延べ語数は 5,340,652、異なり語数は 77,039 である（記号類を含む）。動画の総時間は約 485 時間、プレゼンテーションあたりの平均時間は約 13 分である。

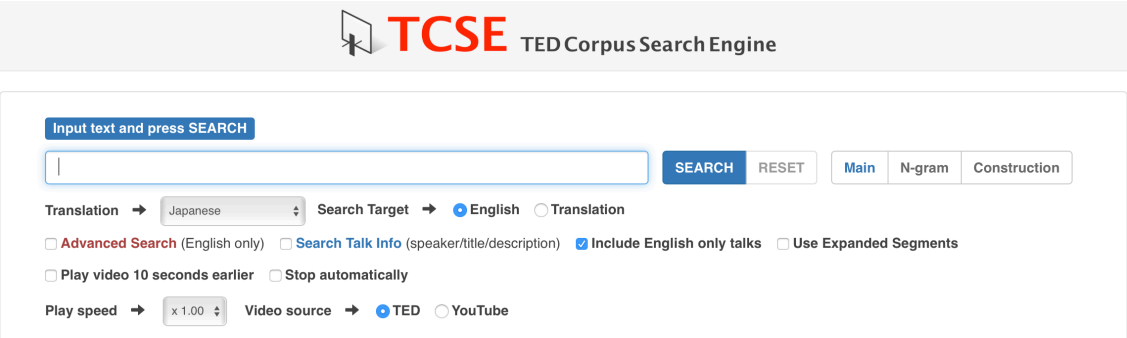
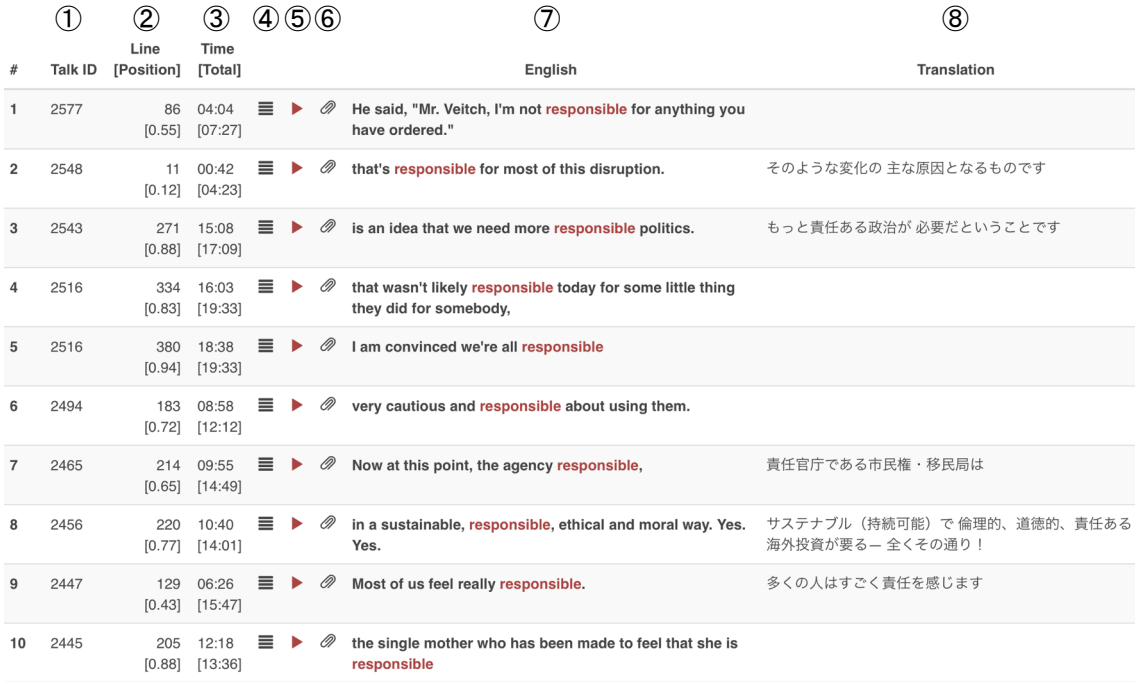


図 1　TCSE の入力パネル



	①	②	③	④⑤⑥	⑦	⑧
#	Talk ID	Line [Position]	Time [Total]		English	Translation
1	2577	86 [0.55]	04:04 [07:27]		He said, "Mr. Veitch, I'm not responsible for anything you have ordered."	
2	2548	11 [0.12]	00:42 [04:23]		that's responsible for most of this disruption.	そのような変化の 主な原因となるものです
3	2543	271 [0.88]	15:08 [17:09]		is an idea that we need more responsible politics.	もっと責任ある政治が 必要だということです
4	2516	334 [0.83]	16:03 [19:33]		that wasn't likely responsible today for some little thing they did for somebody,	
5	2516	380 [0.94]	18:38 [19:33]		I am convinced we're all responsible	
6	2494	183 [0.72]	08:58 [12:12]		very cautious and responsible about using them.	
7	2465	214 [0.65]	09:55 [14:49]		Now at this point, the agency responsible,	責任官庁である市民権・移民局は
8	2456	220 [0.77]	10:40 [14:01]		in a sustainable, responsible, ethical and moral way. Yes. Yes.	サステナブル（持続可能）で 倫理的、道徳的、責任ある海外投資が要る— 全くその通り！
9	2447	129 [0.43]	06:26 [15:47]		Most of us feel really responsible.	多くの人はすごく責任を感じます
10	2445	205 [0.88]	12:18 [13:36]		the single mother who has been made to feel that she is responsible	

図 2　TCSE の検索結果表示画面

TCSE の基本的な使用は簡単である。画面上部の検索ボックスに検索したい文字列を入力して、SEARCH ボタンをクリックすると（図 1)、検索対象となっているプレゼンテーションに含まれるすべての英語表現から合致するものが抽出され、一覧表示される（図 2)。例えば *responsible* という語を検索すると、245 件の事例が得られる。各事例について表示される項目と可能な操作は次の通りである。

① Talk ID : プレゼンテーションの識別 ID である。マウスカーソルを近づけると、スピーカー名とタイトルが表示される。

② Line [Position] : プレゼンテーション内での行番号および相対位置を示す。例えば 90 行あるプレゼンテーションの 11 行目の表現であれば、11 [0.12]と表示される（相対位置は 0 から 1 の値をとる)。クリックすると、プレゼンテーション全体を行単位で表示するウィンドウが立ち上がる。

③ Time [Total] : 表現が発話される箇所（プレゼンテーション開始時を 00:00 とする）とプレゼンテーション全体の所要時間を示す。クリックすると、プレゼンテーション全体を行単位で表示するウィンドウが立ち上がる。

④ パラグラフ表示 : アイコンをクリックすると、プレゼンテーションのトランスクリプトがパラグラフ情報を含んでいる場合には、全体をパラグラフ単位で表示するウィンドウが立ち上がる。

⑤ 動画再生 : アイコンをクリックすると、動画ウィンドウが立ち上がり、当該の表現

を含む箇所の動画再生が開始される。また、これと同期する形でテキストが表示される（図 3）。

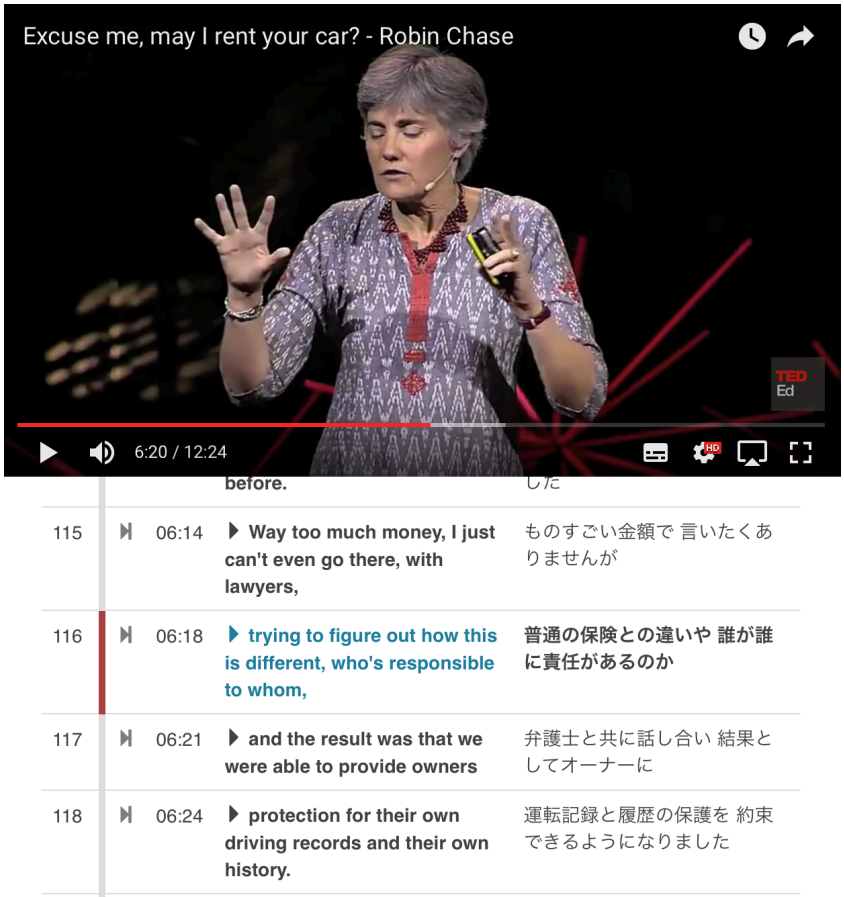
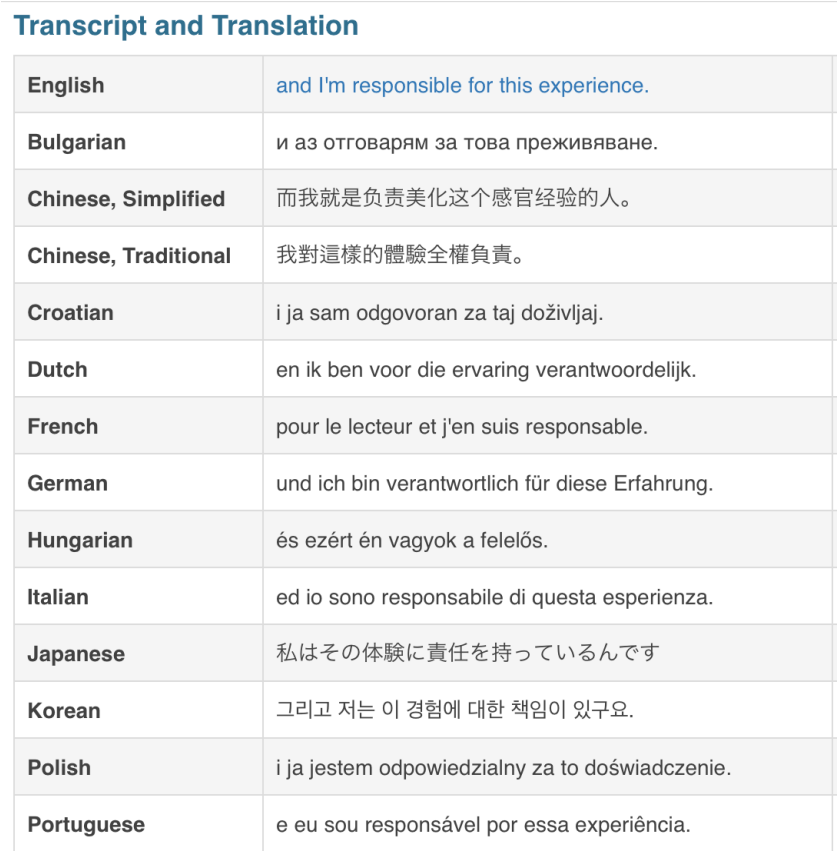


図 3　動画再生画面

⑥　URL のコピー：アイコンをクリックすると動画再生ウィンドウを表示するための URL が表示される。これをコピーすることで、特定の表現事例だけを後からテキストと共にピンポイント再生することができる。

⑦　検索結果（English）：検索文字列に合致する英語表現のテキストが表示される。クリックすると、その表現の品詞構成および統語解析データを表示するウィンドウが立ち上がる。他言語による翻訳データが利用可能な場合は、その表現の翻訳結果の一覧が表示される[9]（図 4）。



Transcript and Translation

English	and I'm responsible for this experience.
Bulgarian	и аз отговарям за това преживяване.
Chinese, Simplified	而我就是负责美化这个感官经验的人。
Chinese, Traditional	我對這樣的體驗全權負責。
Croatian	i ja sam odgovoran za taj doživljaj.
Dutch	en ik ben voor die ervaring verantwoordelijk.
French	pour le lecteur et j'en suis responsable.
German	und ich bin verantwortlich für diese Erfahrung.
Hungarian	és ezért én vagyok a felelős.
Italian	ed io sono responsabile di questa esperienza.
Japanese	私はその体験に責任を持っているんです
Korean	그리고 저는 이 경험에 대한 책임이 있구요.
Polish	i ja jestem odpowiedzialny za to doświadczenie.
Portuguese	e eu sou responsável por essa experiência.

図 4　翻訳データの表示例

⑧　検索結果 (Translation) : 当該のプレゼンテーションに翻訳データが存在する場合、表現に対応する翻訳テキストが表示される。翻訳言語は検索時に入力パネル上で指定することが可能だが、指定がない場合、ウェブ・ブラウザの設定情報を取得して、可能な限り自動的に翻訳言語を決定する。例えば、日本語環境では（そのプレゼンテーションが翻訳済みならば）自動的に日本語の翻訳テキストが表示される。

3.2 高度な検索

画面上部のチェックボックスやセレクターを切り替えることで、様々な設定が可能だが、ここでは重要な 2 つの設定項目についてのみ述べる。まず 1 つ目は高度な検索のモードである。Advanced Search と書かれたチェックボックスをクリックすると、単純な文字列検索ではなく、各種の演算子を用いた検索が可能になる。高度な検索を指定しない場合、検索ボックスに入力された文字列がそのまま検索される。したがって、*regular* という文字列を指定すると、*irregular* や *regularly* といった語を含む表現も結果一覧に含まれる。一方、高度な検索を指定した場合は、入力した文字列の両端や空白文字が語境界として認識される。したがって regular という文字列をそのまま指定しても、*irregular* や *regularly* などは結果に含まれない。その他、表 1 と表 2 に示すように、語

の表層形（surface form）、基本形（lemma）、品詞（part of speech）、およびこれらの組み合わせによる検索が可能になっている[10]。

表 1

高度な検索で利用可能なシンタクス

検索対象	形式	備考
表層形（SURFACE）	`SURFACE`	
基本形（LEMMA）	`[LEMMA]`	
品詞（POS）	`{POS}`	
表層形（品詞を指定）	`SURFACE{POS}`	途中に空白は挿入しない
基本形（品詞を指定）	`[LEMMA]{POS}`	途中に空白は挿入しない
論理和（A or B）	`A|B`	
ワイルドカード	`*`	1 個以上の要素に対応
セグメント開始点	`^`	

表 2

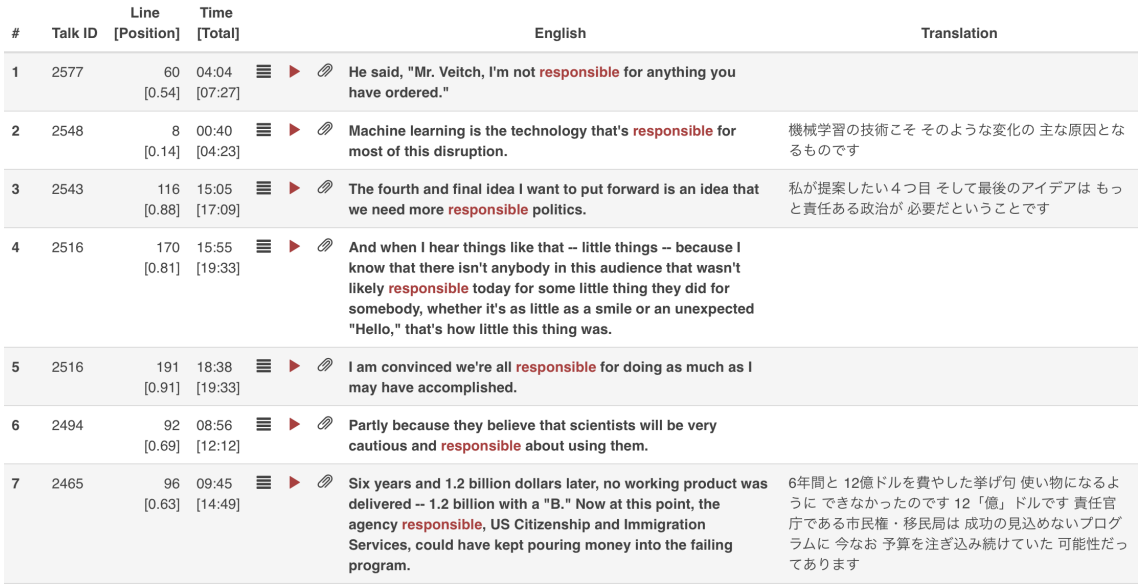
高度な検索の例

検索文字列	マッチする表現の例
`[excite]`	*excite, excites, excited, exciting*
`{n}`	すべての名詞（noun）
`{v}`	すべての動詞（verb）
`to * surprise`	*to our surprise, to his surprise,* etc.
`[read] {DT} [news|paper|article]`	*they read these articles,* *reading the paper or something,* *I was reading the news at six*, etc.
`^ having {v}`	*Having started the process,* *Having said that,* etc.
`[help] {n}`	*an aunt offered financial help,* *we called people for help,* etc.

3.3 セグメントと拡張セグメント

検索時に設定が可能なもう 1 つの重要な項目は、テキストデータの分割単位である。前述の通り、TED Talks のトランスクリプトは字幕表示を主な目的として作成されている。TCSE では 1 画面分の字幕テキストをセグメント（segment）と呼ぶが、4 節で論

じるように、セグメントは理論的な観点から見ても妥当な処理単位である。そこで、TCSE ではこのセグメントを検索と表示の基本的単位として採用している。しかし、多くのセグメントは文の部分要素であるため、文全体を見渡したいようなときには、前後のセグメントを参照しなくてはならない。また、1 つの文の中で隣接した要素であっても、異なるセグメントに属している場合には、1 つの検索文字列にマッチさせることができない。そこで TCSE では、内部に文の開始点と終点を少なくとも 1 つずつ含むよう、1 個以上のセグメントを結合する形で再構成した拡張セグメント（expanded segment）の単位を設けており、必要に応じて切り替えられるようにしている。



#	Talk ID	Line [Position]	Time [Total]	English	Translation
1	2577	60 [0.54]	04:04 [07:27]	He said, "Mr. Veitch, I'm not responsible for anything you have ordered."	
2	2548	8 [0.14]	00:40 [04:23]	Machine learning is the technology that's responsible for most of this disruption.	機械学習の技術こそ そのような変化の 主な原因となるものです
3	2543	116 [0.88]	15:05 [17:09]	The fourth and final idea I want to put forward is an idea that we need more responsible politics.	私が提案したい４つ目 そして最後のアイデアは もっと責任ある政治が 必要だということです
4	2516	170 [0.81]	15:55 [19:33]	And when I hear things like that -- little things -- because I know that there isn't anybody in this audience that wasn't likely responsible today for some little thing they did for somebody, whether it's as little as a smile or an unexpected "Hello," that's how little this thing was.	
5	2516	191 [0.91]	18:38 [19:33]	I am convinced we're all responsible for doing as much as I may have accomplished.	
6	2494	92 [0.69]	08:56 [12:12]	Partly because they believe that scientists will be very cautious and responsible about using them.	
7	2465	96 [0.63]	09:45 [14:49]	Six years and 1.2 billion dollars later, no working product was delivered -- 1.2 billion with a "B." Now at this point, the agency responsible, US Citizenship and Immigration Services, could have kept pouring money into the failing program.	6年間と 12億ドルを費やした挙げ句 使い物になるように できなかったのです 12「億」ドルです 責任官庁である市民権・移民局は 成功の見込めないプログラムに 今なお 予算を注ぎ込み続けていた 可能性だってあります

図 5　拡張セグメント表示モードでの検索結果の例

以上のように、TCSE には、TED Talks のデータを言語教育資源として有効活用するための機能が実装されている。TCSE の開発では、教師あるいは学習者自身にとって容易に使いこなせることが重要な指針となっている。しかし、それと同時に、TCSE の設計は認知言語学の理論的な考え方に根ざしており、様々な機能や特徴が、これを直接的または間接的に反映している。次の 4 節では、そのような TCSE の認知言語学的側面について述べておきたい。

4. TCSE の認知言語学的側面

4.1 認知言語学における構文の概念

TCSE は TED Talks のトランスクリプトに含まれる文字列を網羅的に検索する。検索対象は、単独の語でも、あるいは複数の語から成る句でもよい。しかし本稿のタイト

ルにもあるように、TCSE は「英語構文事例検索システム」として役立つことを目指している。TED Talks のデータを得て、認知言語学において論じられるところの「構文」の検索と抽出が、部分的にではあれ可能になると期待されるのである。本節では、TCSE においてこの目論見がどれだけ達成されているかを示したい。

構文という概念は、認知言語学において、「語」や「文」といった形式的で伝統的な言語単位に増して重要なものと認識されている。認知言語学の考え方は、近年、英語教育の世界においても多くの関心を集めており、構文についても、その重要性が次第に認識されるようになってきた（Littlemore 2009; Tyler 2012; 谷口 2011; 長 2016）。しかし、構文という概念を言語教育に応用するにあたっては、常にある種の困難が生じる。その要因は第一に構文という概念自体の定義が抽象的であること、第二に個々の表現を特定の構文に属する事例として認定することが必ずしも容易でないことである。

構文の考え方は、認知言語学という大きな枠組みの一部を構成する構文文法理論として発展してきた（Fillmore, et al. 1988; Goldberg 1995; Hilpert 2014; 山梨 2009）。構文文法において構文は単なる連語ないしは熟語のみを指すのではない。形式と意味との記号的関係が、慣例的な定着と共に 1 つのパターンとしての心理的実在性を得たもの、これらを構文として総称する [11]。したがって、連語や熟語などは、当然、構文の一種であるが、構文という概念が外延として含む対象はさらに広い。その抽象度は様々であり、*let alone* のような定型句も構文であるが、V + NP *to* NP のようなパターンも構文である（*to*-与格構文）。また表現としての規模も固定的ではなく、上記の V + NP *to* NP というパターンが 1 つの構文であると同時に、その部分構造である V + NP や *to* NP もそれぞれ構文と言える。したがって、具体的な表現 e_1 を取り上げて、それを c_1 という構文の事例であると述べることは間違いではないが、e_1 に c_1 以外の構文 c_2, c_3, … が含まれる可能性は排除されない。認知言語学および構文文法の考え方に基づくと、文法とは様々な抽象度と規模を持った構文知識の集合であり、それらが同時並行的に実現形を得ることにより、発話の中で一定の機能を果たす言語表現が成立する。したがって、「構文」の観点から言語表現を考えるには、常に複層的な構造として対象を捉える必要がある。

そして、この複層性ゆえに、表現の中で実現している構文の事例を網羅的に示すのは容易でない。しかし、認知言語学的の考え方に基づいて各種の構文の指導・習得を目指すならば、必ずしも語や文といった伝統的な単位にこだわるのではなく、様々な抽象度と規模を持った「意味的・機能的まとまり」に焦点を当て、それらが集合的に構成するスキーマとしての構文の存在と働きの理解を目指すべきだろう。

このような観点から、具体的な表現の事例を通して学習者が個々の構文を把握する一助となるよう、TCSE はいくつかの特徴的な機能を備えている。いずれの機能も、あらゆる構文事例を抽出して網羅的に提示するようなものではない。あくまで、構文の重要性を踏まえ、その輪郭を断片的に浮き彫りにするための試みとして設計・実装された機

能である。多分に概念実証的ではあるが、理論的な考察が実際の言語教育に役立てられる可能性を示すものである。

4.2 イントネーション・ユニットとウィンドウ・オブ・アテンション

先に述べた通り、構文であることの条件に形式的な規模に関する規定はない。構文には、語よりも小さい形態素レベルの要素やパターンも含まれる。また、節や文の境界を越えるパターンもある。しかし、実際の談話の中で、話者と聞き手が「意識の対象」として一時に把握できる音声的・文字列的パターンの幅や概念の複雑性には限りがある。そのため、多くの典型的な構文は、形式的にも内容的にも、ある程度制限された規模で実現することになる。

Chafe（1987, 1994）は、イントネーション・ユニット（intonation unit）という単位が談話の構成に関わっていると論じる。発話は必ずしも文を基本的な単位として構成される訳ではなく、むしろ、時間軸上の一点で意識に上らせることができる情報のまとまりが重要な役割を果たす。それは韻律上のまとまりに対応しており、したがってイントネーション・ユニットと呼ぶことができる。Chafe は、英語においてイントネーション・ユニットの再頻値（mode）は 4 語であると述べている。

Langacker（2001, 2012）も同様の見解を示している。Langacker は、一時に活性化される概念構造の枠をウィンドウ・オブ・アテンション（windows of attention）と呼び、そのような、言わば「注意のフレーム」によって、話者と聞き手は複雑な概念内容を容易に取り扱い可能なまとまりとしてパッケージ化していると論じる（Langacker 2012: 562-563）。

(1) *// Since Joe hates salmon // we'll have to serve steak. //*
(2) *// Since Joe hates salmon // with such a passion // we'll have to serve the steak // we've been keeping in the freezer // for a special occasion. //*
(3) *// If he likes it / he'll eat it. //*

上記の例において「//」はウィンドウ・オブ・アテンションの境界を表している。(1) の文は 2 つの節を含んでおり、各々がウィンドウ・オブ・アテンションを構成する。Langacker によると、ウィンドウ・オブ・アテンションはこのように「節」の幅と一致することが多い。節という文法的なまとまりは、概念的なまとまりと連関しており、また韻律的なまとまりにも関わっている。しかし、節とウィンドウ・オブ・アテンションの一致はあくまで 1 つの典型に過ぎず、逸脱は決して珍しくない。(2) のように、節以下の単位がウィンドウ・オブ・アテンションを構成することも多い。また、(3) のようなある種の「決まり文句」では、複数の節が 1 つのウィンドウ・オブ・アテンションを

構成することもある（「/」は節の切れ目を表す）。

Chafe がイントネーション・ユニットとして、Langacker がウィンドウ・オブ・アテンションとして、それぞれ論じている発話の単位は、話者にとっての表現構成と聞き手にとっての表現解釈に大きく関わっていると考えられる。したがって、様々な規模の構文の中でも、とりわけ重要なのは、この種のまとまりの幅に一致する構文である。英語構文の習得にあたっては、また、英語構文の事例を検索・抽出することを目的としたシステムの設計においては、こうした「注意のフレーム」に着目することが必要であろう。

しかし、そのようなフレームは、通常のテキスト上に明確な境界線を伴った形で現れるものではない。また、韻律、姿勢、表情などから機械的に検知することも容易ではない。そこで、TCSE の開発においては、プレゼンテーションのトランスクリプトに含まれるセグメント情報に着目して、擬似的なイントネーション・ユニットおよびウィンドウ・オブ・アテンションの検出を目指した。3 節で示したように、TED Talks のトランスクリプトは動画にスーパーインポーズする字幕として用いることを前提に制作されている。したがって、1 つのセグメントは 1 画面上に収まる幅になる。しかし、単に画面幅だけに合わせる形で機械的に分節化される訳ではない。字幕は話者の発話のタイミングと同期させなければ意味がなく、必然的に、セグメントの切れ目はイントネーション・ユニットの境界上に設けられる。例えば、次の図 6 は、TCSE で *regardless of* の検索を行った結果の一部であるが、*regardless of* NP という構文の事例を概ね適切な幅で抽出できている。

English	Translation
Regardless of your time and place, there are some things that are constant.	時代や場所の違いはあっても 変わらないものがあります
But regardless of the format,	ただ 形式に関わらず
regardless of our intent.	ほとんど認識していません
regardless of whether it's in our bodies	それが私たちの体内にあろうが
every day, regardless of your own belief.	毎日 自身の信条に関わらず

図 6　セグメント検索の結果例

Chafe（1987, 1994）はイントネーション・ユニットの幅の再頻値は 4 語であると述べた。しかし、TCSE におけるセグメントの幅（＝1 画面分の字幕テキスト）は、4 語より大きいことが多い。2,172 のトークのデータを調査した結果、セグメントに含まれる語数の平均は約 8.6、再頻値は 8 であった。したがって、Chafe の論が正しいと仮定する

なら、TCSE のセグメントは概ね 2 つのイントネーション・ユニットないしはウィンドウ・オブ・アテンションに対応していると言えるだろう。

3 節でも示した通り、セグメントは多くの場合、文の一部を構成する要素であるため、文レベルでの観察が必要な場合には、検索および表示の幅を広げる必要がある。そこで TCSE ではそのような目的のために、少なくとも 1 つの文の開始点と終了点を含むまとまりを拡張セグメントとして定義し、これを基準とした操作を可能にしている。先に用いた *regardless of* を拡張セグメント表示モードで検索した結果の一部を図 7 に示す。

English	Translation
Regardless of your time and place, there are some things that are constant.	時代や場所の違いはあっても 変わらないものがあります
But regardless of the format, two of my favorite materials are history and dialogue.	ただ 形式に関わらず 好きな題材は 「歴史」と「対話」です
For the most part, prosecutors step onto the job with little appreciation of the impact of our decisions, regardless of our intent.	ほとんどの人は 検察官になるとき 自分の決断が 意図に関係なく持つ影響力を ほとんど認識していません 広い裁量を持つにもかかわらず
It seemed to us that DNA, the most fundamental structure of life, that codes for the production of all of our proteins, is both a product of nature and a law of nature, regardless of whether it's in our bodies or sitting in the bottom of a test tube.	そして生命の基本構造であるDNAは 私たちを形作る すべてのたんぱく質をコードしているので 私たちは それを天然物であり 自然法則だと考えたのです それが私たちの体内にあろうが 試験管の中にあろうが関係ありません
And that influence affects all of us, every day, regardless of your own belief.	私達 全員に影響を与えます 毎日 自身の信条に関わらず

図 7　拡張セグメント表示の例

なお、図 7 に示された 5 つの例はいずれも単一の文で構成された拡張セグメントであるが、TED Talks のトランスクリプトでは、1 画面中に文の切れ目が生じることもある。したがって、拡張セグメントが常に 1 文で構成されるとは限らない。セグメントがイントネーション・ユニット／ウィンドウ・オブ・アテンションの近似であるのと同様に、拡張セグメントは文レベルのまとまりの近似である。繰り返しになるが、認知言語学的な観点から見ると、談話の中で文は必ずしも中心的な単位ではない。例えば図 7 の 4 つ目の例は形式的には 1 文であるが、内部に様々なレベルの構文が含まれている。このような例を扱うためにも、構文という概念に着目した英語教育や事例検索システムの設計においては、イントネーション・ユニット／ウィンドウ・オブ・アテンションのように、文の概念にとらわれない単位を想定する必要がある。

4.3 構文とコロケーション

一般的に構文は、一定の慣例的な定着度を持った記号構造パターンと定義される（Goldberg 2006）。また、そのようなパターンは、基本的に意味的あるいは談話的な要

請のもとに構築されるというのが認知言語学における考え方である。そうであるならば、構文が取り得る様々なバリエーションについて考える際はもちろん、構文の成り立ちを考えるにあたっても、コロケーション（collocation）、すなわち語同士の結びつきやすさの傾向を知ることが重要である。このような考え方に基づいて、TCSE には N-gram 検索機能を実装している。N-gram とは、ある記号列から N 個の隣接した要素を取り出したとき、特定の組み合わせがどの程度出現するかを表す指標であり、これによって、ある語の前後にどのような語が生じやすいかが一目瞭然となる。TCSE では TED Talks のトランスクリプトに含まれる全ての語について 2-gram から 4-gram までを抽出・集計し、検索可能なデータとして提供している。

TCSE の画面右側にあるボタンをクリックして N-gram 検索モードに移行し、語を入力して実行すると、合計出現頻度と出現プレゼンテーション数に基づいて順位付けされた N-gram 情報（2-gram から 4-gram）が表示される。図 8 と図 9 は、*species* という語の N-gram を求め、この語が右端に現れる事例のみを抽出・集計した結果の一部を示したものである [12]。

	Word1	Word2	Freq	Num of Talks
1	a	species	65	54
2	our	species	69	54
3	the	species	71	49
4	other	species	59	44
5	different	species	36	32
6	of	species	38	32
7	human	species	20	18
8	this	species	29	20
9	new	species	42	19
10	one	species	26	20
11	endangered	species	23	18
12	these	species	20	16
13	all	species	18	16
14	many	species	14	13
15	another	species	14	11

図 8　2-gram の表示例

	Word1	Word2	Word3	Word4	Freq	Num of Talks
1	the	origin	of	species	11	9
2	us	as	a	species	5	5
3	of	the	human	species	5	5
4	evolution	of	our	species	5	5
5	history	of	our	species	5	5
6	we	as	a	species	6	5
7	percent	of	all	species	4	4
8	the	number	of	species	3	3
9	a	recently	arrived	species	3	3
10	are	the	only	species	2	2

図 9　4-gram の表示例

図 8 と図 9 から、*species* が *human, origin, evolution, endangered* といった語と複数回共起していることがわかる。例えば *human species* という組み合わせは 18 のプレゼンテーションにおいて合計 20 回使用されている。また、*endangered species* という組み合わせは 18 のプレゼンテーションにおいて合計 23 回使用されている。このような共起情報は、教科書や参考書にも一部示されてはいる [13]。しかし、それらは本来、豊富な言語経験の中で個々の話者が自ら蓄積していくべき「言語感覚」の一部である。したがって、英語教育の実践においては、語と語の結びつきに関する情報を取り出して与える場合でも、単に「語 w_1 は語 w_2 と結びつきやすい」と伝えるだけでなく、確かな事例と共に確認させる、あるいは「経験させる」ことが理想であろう。TCSE 上では、N-gram 表示をクリックすることで、当該の組み合わせ事例を検索し、前後の表現を含めた形で表示したり、動画・音声をピンポイントで再生したりすることができる。これにより、コロケーション情報を単なる数値としてではなく、実例と共に提示することが可能になっている。

4.4 構文パターンのリスト

構文は認知言語学における理論的な概念であるが、英語教育の領域では伝統的に「熟語」「連語」「群動詞」「定形表現」などを総称的に表す用語としても用いられてきた。両者は基本的に矛盾しない。伝統的（かつ一般的）な意味での構文は認知言語学的な意味における構文の一種であり、それらが英語教育において「構文」として特別にカテゴライズされるのは、その多くが教育的な重要性を帯びていることの表れだろう。そこで TCSE では、既存の英語学習参考書やその他のリファレンスで扱われている 1,160 の構文パターンのリストを格納し、それぞれの事例を TED Talks のトランスクリプトから

抽出して表示できるようにしている。TCSE のメイン画面の右側にある Construction ボタンをクリックするとリストが一覧表示されるが、さらにリンクをクリックすると、検索文字列の例と、TED Talks のトランスクリプトとは別に新たに作成された例文が示される。学習者は、作例を見て基本的な用法を確認した上で、TED Talks からの実例を検索・参照することにより、各構文のより深い理解を目指すことができる（図 10）。

1,160 の構文パターンの多くは、*all the more* や *as it were* といった、英語学習における「頻出表現」である。したがって、到達レベルにもよるが、学習者はそれらを何らかの形ですでに学んでいる可能性がある。しかし、認知言語学が前提とする言語の用例基盤モデル（usage-based model of language）に従うならば、真の意味での言語知識は、文脈やその他の付随的情報を伴った「事例」と共に習得される（cf. Langacker 1988; Barlow and Kemmer 2000）。したがって、既習の表現であっても、理解が単なる説明的記述のレベルに留まっているのであれば、実際の事例に触れることによって、ネットワーク的言語知識の中で位置付ける作業が必要である。TCSE に収録された 1,160 のパターンは、このような想定のもとに、学習者の構文理解を動的に活性化することを目指したものである。

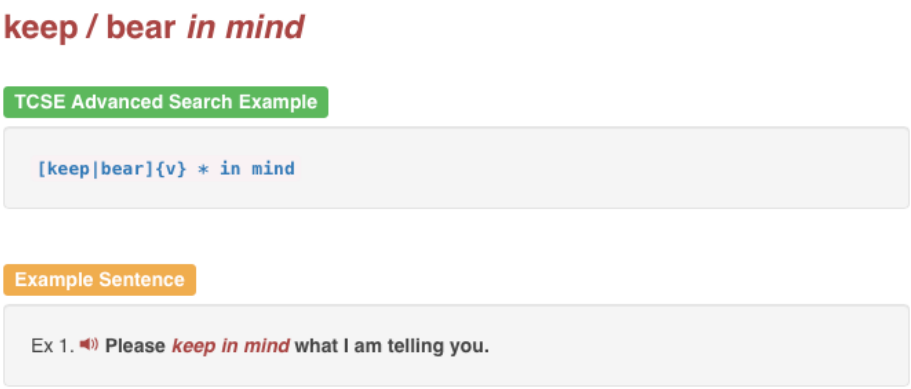


図 10　構文パターンの表示例

以上のように、TCSE は認知言語学における理論的な考察を背景に、構文事例を効果的に検索・抽出するための様々な機能を備えている。これらはいずれも TED Talks の内容的にも形式的にも優れたデータに依拠しており、その意味において、TCSE の開発は、データのオープン化という世界的な潮流の中で可能になった試みと言える。

おわりに

本稿では、「機械判読に適した形式で、二次利用が可能な利用ルールと共に公開されたデータ」と定義されるオープンデータの意義について論じると共に、英語教育における

実践的な目的に合致するデータの例として、TED が公開する英語プレゼンテーションのトランスクリプトおよび動画・音声の紹介を行った。また、このデータを活用して構築した英語構文事例検索システムである TCSE の基本的な機能と使用方法について論じた。さらに、TCSE の設計が認知言語学における「構文」の概念に根ざしたものであることを示し、各種の機能がそれぞれ理論的な視点に基づいて実装されていることを明らかにした。

インターネットの発展、そしてオープン化の潮流によって、様々な、また大量のデータが一定の条件のもと、自由に利用できるようになってきた。教育のために利用可能な素材には、これまで多くの質的・量的制約があった。私たち教師の多くは、そのような制約を殊更に意識することもなかったかもしれない。しかし、データのオープン化の流れは、ある意味で世界的な社会変革の過程であり、教育における前提や方法は社会の変化に大きく影響を受ける。本稿では、英語教育にとって今後起こりうる変化が、教師にとっても学習者にとっても歓迎すべきものであることを示した。しかし今回は、データのオープン化に関連する、ある重要なプロセスについて十分に考察を行うことができなかった。それは、教材やシステムを用いることで得られた知見を共有し、コミュニティの中で新たな展開を呼び起こすプロセスである。TED Talks を始め、優れたデータの数々が利用可能になるのは言うまでもなく素晴らしいことであるが、それを十分に活かすためには、実践から得られた知見をバランスよく組み入れた方法論が必要となる。オープン化の流れが、コンテンツとしてのデータの共有のみならず、それを活かす実践的方法の共有にまで発展するなら、さらに大きな社会的イノベーションとなるだろう。英語教育とは、そのようなイノベーションから最も多くの恩恵を受けることができる分野だと思われる。

注

* 本稿は 2016 年 6 月 4 日に JACET 中部支部大会において行われたシンポジウム「英語力向上のための多様なリソース活用の新展開」での研究発表に基づいている。

1 OpenCourseWare (Wikipedia)
https://en.wikipedia.org/wiki/OpenCourseWare

2 総務省ホームページ
http://www.soumu.go.jp

3 オープンデータと言えるための条件（総務省）
http://www.soumu.go.jp/menu_seisaku/ictseisaku/ictriyou/opendata

4 British National Corpus
http://www.natcorp.ox.ac.uk

[5] Corpus of Contemporary American English
http://corpus.byu.edu/coca

[6] Our Organization (TED)
https://www.ted.com/about/our-organization

[7] Usage Policy (TED)
https://www.ted.com/about/our-organization/our-policies-terms

[8] Brown Corpus Manual
http://icame.uib.no/brown/bcm.html

[9] TCSE では、翻訳済みプレゼンテーションの数を主な基準として選定した 20 の言語による翻訳テキスト表示が可能になっている。

[10] 現時点での実装では、通常検索と高度な検索の両方で文字種別（大文字・小文字）を区別しない。なお、品詞指定文字列の詳細については TCSE ホームページを参照されたい。

[11] 「構文」という語が construction の訳語であることに注意されたい。「文」の文字が含まれてはいるが、construction の概念自体が文の単位と特別な結びつきを持っている訳ではない。

[12] 紙幅の制約から、ここでは 3-gram の結果表示を省略する。なお、図の中で Frequency（総頻度）と Number of Talks（出現プレゼンテーション数）の値が必ずしも降順になっていないのは、TCSE の現在の実装では N-gram の順位を決定するのに、粗頻度ではなく、統計的ばらつき（dispersion）の値を用いているからである（cf. Gries 2008）。

[13] 事実、『ウィズダム英和辞典』（三省堂）の *species* の項には *human species* と *endangered species* が例として登場している。

謝辞

JACET 中部支部大会シンポジウムの企画と司会を担当され、本稿の元になった研究発表をまとめるきっかけを作ってくださった大森裕實先生、JACET 中部支部事務局幹事としてお世話くださった佐藤雄大先生、TCSE の開発と改良の過程で有益なコメントと励ましをくださった宮浦国江先生に感謝申し上げます。

付記　本研究の一部は科学研究費（若手研究 B：25870898）の補助を受けて行われた。

参考文献

石川慎一郎（2008）『英語コーパスと言語教育―データとしてのテクスト』大修館書店.

谷口一美（2011）「応用認知言語学と語彙学習―文法理論を英語教育に活用する(2)」『大阪教育大学紀要（第Ⅰ部門）』59(2), 63-74.

長加奈子（2016）『認知言語学を英語教育に生かす』金星堂.

投野由紀夫（2015）「教育利用のためのコーパス情報とツールの活用」投野由紀夫（編）『コーパスと英語教育』（pp. 181-206）ひつじ書房.

山梨正明（2009）『認知構文論―文法のゲシュタルト性』 大修館書店.

Barlow, M and S. Kemmer (Eds.). (2000) *Usage Based Models of Language*. Stanford: CSLI.

Chafe, W. L. (1987) Cognitive constraints on information flow. In R. S. Tomlin (Ed.), *Coherence and Grounding in Discourse* (pp. 21-51). Amsterdam: John Benjamins.

Chafe, W. L. (1994) *Discourse, Consciousness, and Time: The Flow and Displacement of Conscious Experience in Speaking and Writing*. Chicago: University of Chicago Press.

Fillmore, C. J., P. Kay, and C. O'Connor. (1988) Regularity and idiomaticity in grammatical constructions: The case of *let alone*. *Language* 64, 501-538.

Francis, W. N. and H. Kučera. (1964) Brown Corpus Manual [Revised and Amplified, 1979]. Department of Linguistics, Brown University.

Goldberg, A. E. (1995) *Constructions: A Construction Grammar Approach to Argument Structure*. Chicago: University of Chicago Press.

Goldberg, A. E. (2006) *Constructions at Work: The Nature of Generalization in Language*. Oxford: Oxford University Press.

Gries, S. Th. (2008) Dispersions and adjusted frequencies in corpora. *International Journal of Corpus Linguistics* 13(4), 403-437.

Hilpert, M. (2014) *Construction Grammar and Its Application to English*. Edinburgh: Edinburgh University Press.

Langacker, R. W. (1988) A usage-based model. In B. Rudzka-Ostyn (Ed.), *Topics in Cognitive Linguistics* (pp. 127-161). Amsterdam: John Benjamins.

Langacker, R. W. (2001) Discourse in cognitive grammar. *Cognitive Linguistics* 12(2), 143-188.

Langacker, R. W. (2012) Elliptic coordination. *Cognitive Linguistics* 23(3), 555-599.

Littlemore, J. (2009) *Applying Cognitive Linguistics to Second Language Learning and Teaching*. Hampshire, UK: Palgrave.

Tyler, A. (2012) *Cognitive Linguistics and Second Language Learning: Theoretical Basics and Experimental Evidence*. London: Routledge.