「シンポジウム」
TED Corpus Search Engine: TED Talks を研究と教育に活用するためのプラットフォーム

長谷部陽一郎

1．はじめに

本稿では TED（https://ted.com）で公開されている英語プレゼンテーションのトランスクリプト，翻訳テキスト，動画データをコーパスとして用いるための検索エンジンである TED Corpus Search Engine（TCSE）の概要とその可能性について論じる。まず 2 節では TCSE の主な機能を概観する。TED Talks という素材に特化した検索システムである TCSE には，そのことを生かした，いくつかの特徴的な機能が実装されている。次に 3 節では TCSE に含まれるデータの様々な統計値を示しながら，TED Talks の言語コーパスとしての特徴を素描する。また，TED Talks の言語がどれだけ「話し言葉」としての特徴を有しているか，あるいはそうでないのかという問題について，予備的な考察を示す。続く 4 節では TCSE の理論的な背景について述べる。ここでは TCSE が認知言語学の言語観に基づいて設計されており，その点において他のコーパス・システムと大きく異なっていることを示す。最後の 5 節では全体のまとめを行う。

2．TCSE の主な機能

TCSE（https://yohasebe.com/tcse）の最も重要な機能は TED コーパスを自在に検索し，その結果を様々な形式で表示することである。本節では，1）英語トランスクリプトの検索，2）多言語による対訳の表示・検索，3）検索対象となる言語単位と表示のバリエーションの 3 つに焦点を当て，システムの概要を示す。なお，TCSE はインターネットに接続された機器（PC，スマートフォン，タブレット）上のウェブ・ブラウザから利用可能である。個別の機能に関する詳しい解説は，サイト上で閲覧できるチュートリアル文書などを参照されたい。

2.1 英語トランスクリプトの検索

図 1 は TCSE のメイン検索パネルである。テキストボックスに英語表現を入力して SEARCH ボタンをクリックすると, 図2のように結果が一覧表示される。

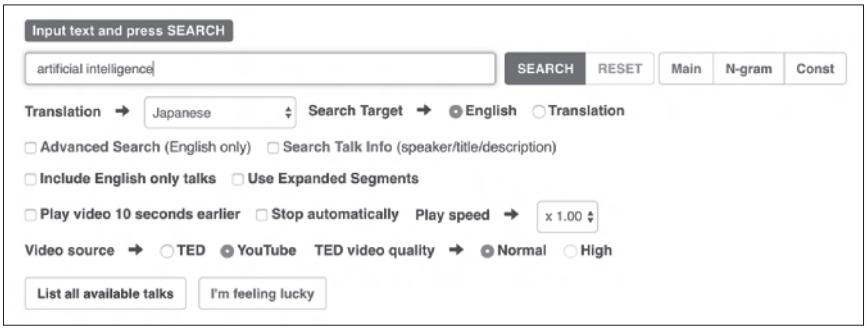


図 1　TCSE の検索パネル

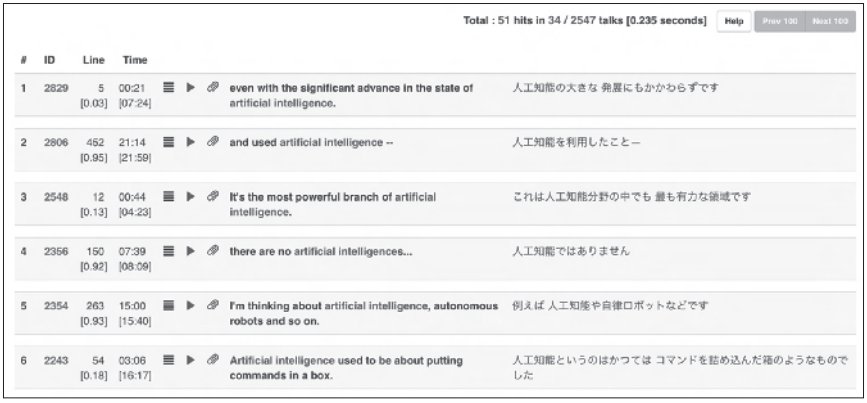


図 2　TCSE の検索結果の例

英語トランスクリプトの検索にあたっては，入力文字列を単純にマッチさせる通常検索に加えて, Advanced Search モードが存在する。Advanced Search モードでは，語の表層形，レンマ，品詞，あるいはこれらを組み合わせた文字列を用いた複合的な検索が可能である。

2.2 多言語による対訳の表示と検索

TED Talks のトランスクリプトは様々な言語への翻訳が進められており，TCSE ではそのうち，「1,000 件以上のトークが翻訳されている」という基準のもとに，日本語を含む 28 言語による対訳データを収録している。表示する対訳言語は検索パネル上で選択可能であるが，指定がない場合はウェブブラウザの設定に応じた言語が選択される仕様になっており，日本語環境では通常，図 2 のように日本語の対訳が英語のトランスクリプトと共に表示される（当該のトークが翻訳済みである場合）。また，対訳テキストを検索することも可能である。対訳言語を選択した上で，Search Target を Translation に設定すれば，テキストボックスに入力された内容をその言語の文字列と見なし，一致する表現をリストアップする。図 3 は日本語による検索結果の例である。

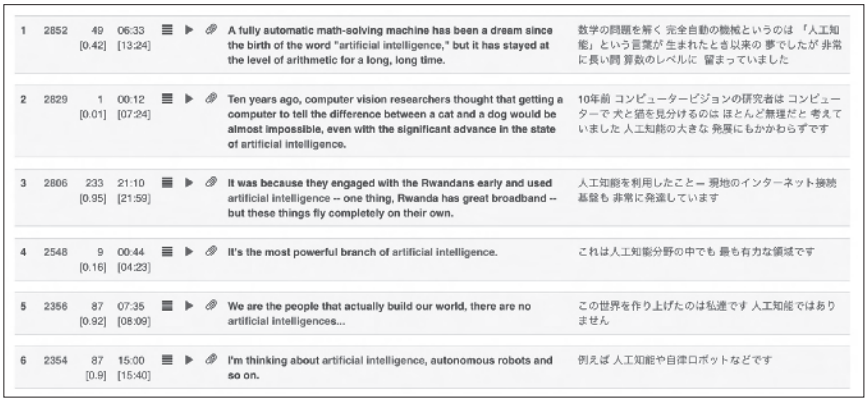
図３　日本語による検索結果の例

2.3 TCSE における検索対象の単位と表示のバリエーション

TED で公開されているトランスクリプトのデータは，1 画面に表示する字幕の幅を基準に構成されており，TCSE でも基本的にこれを採用している。つまり，TCSE において検索結果としてデフォルトで表示されるのは，入力テキストと一致する語句を含んだ 1 画面分の字幕文字列である。TCSE ではこの単位をセグメント（segment）と呼んでいる。

しかしながら，セグメントは往々にしてセンテンスに満たない単位であり，ユーザーとしては表示される個々の結果をより詳しい形で見たい場合があり得る。また，長めの文字列を入力したときなど，セグメント単位では期待された

表現が結果に含まれない場合がある。そこで TCSE では，拡張セグメント（expanded segment）という単位での検索・表示のオプションを設けている。拡張セグメントとは，セグメントがセンテンスの断片である場合，隣接するセグメントと連結し，センテンス全体が同じ要素内に収まるよう調整した単位である。図 4 は拡張セグメント・モードでの検索結果を示したものである。



1	2852	49 [0.42]	06:33 [13:24]	A fully automatic math-solving machine has been a dream since the birth of the word "artificial intelligence," but it has stayed at the level of arithmetic for a long, long time.	数学の問題を解く 完全自動の機械というのは 「人工知能」という言葉が 生まれたとき以来の 夢でしたが 非常に長い間 算数のレベルに 留まっていました
2	2829	1 [0.01]	00:12 [07:24]	Ten years ago, computer vision researchers thought that getting a computer to tell the difference between a cat and a dog would be almost impossible, even with the significant advance in the state of artificial intelligence.	10年前 コンピュータービジョンの研究者は コンピューターで 犬と猫を見分けるのは ほとんど無理だと 考えていました 人工知能の大きな 発展にもかかわらずです
3	2806	233 [0.95]	21:10 [21:59]	It was because they engaged with the Rwandans early and used artificial intelligence -- one thing, Rwanda has great broadband -- but these things fly completely on their own.	人工知能を利用したこと― 現地のインターネット接続基盤も 非常に発達しています
4	2548	9 [0.16]	00:44 [04:23]	It's the most powerful branch of artificial intelligence.	これは人工知能分野の中でも 最も有力な領域です
5	2356	87 [0.92]	07:35 [08:09]	We are the people that actually build our world, there are no artificial intelligences...	この世界を作り上げたのは私達です 人工知能ではありません
6	2354	87 [0.9]	15:00 [15:40]	I'm thinking about artificial intelligence, autonomous robots and so on.	例えば 人工知能や自律ロボットなどです

図４　拡張セグメントを用いた結果表示

セグメントを用いるにせよ，拡張セグメントを用いるにせよ，個々の具体的な表現を詳しく確認する必要がある際には，トークの全文テキストを対訳と共に表示することが可能である。また，検索結果のそれぞれについて，発話箇所の映像をピンポイントで再生することができる。これは従来の多くの言語コーパスには見られない重要な特徴である。

以上，簡単にではあるが，コーパス検索システムとしての TCSE の機能を概観してきた。次節では，TCSE に収録されている TED Talks のデータがコーパスとしてどのような性質を持っているのかについて考える。

３．コーパスとしての TED Talks

TCSE は TED Talks のデータをコーパスとして用いるためのオンライン・システムであり，2014 年の 11 月に最初のバージョンが公開された（Hasebe, 2015）。その後，数多くの機能を実装すると共に，新たなデータの追加を定期的に行ってきた。本節ではこうした「TED コーパス」の特徴を概観する。

3.1 TED Talks のデータについて

TED では現在 2,600 以上の英語プレゼンテーションが公開されており，そのデータは Creative Commons ライセンス（CC BY-NC-ND）のもとで利用可能になっている。CC BY-NC-ND とは，再出典を明記し（BY），非商用（NC）で，内容の改変を行わないという条件のもとにデータの再配布（ND）を認めるタイプのライセンスであり, TCSE もこれに準拠した形でデータを利用している。

TED が主催するカンファレンスにはいくつかの種類があるが，公式のイベントの他に，TEDx と呼ばれる非公式のカンファレンスが存在する。TEDx は TED からライセンスを受けた各地のコミュニティーが独自に開催するもので，厳密には TED カンファレンスと区別される。しかし，その一部は TED の公式サイトでデータが公開されており，上記の「2,600 以上」というプレゼンテーション数にも含まれている。TCSE では，TED により公開されている全データのうち 2,547 件のデータを収録している（2018 年 2 月現在）。収録対象となっていないプレゼンテーションがあるのは，英語以外の言語で話されているものや，トランスクリプトの形式が TCSE のシステムに合致しないものを除外しているためである。

なお，TED ではプレゼンテーションの動画に加えて，発話内容のトランスクリプトを公開しているが，トランスクリプトの多言語への翻訳はボランディア・メンバーの作業によるものである。組織としての TED は翻訳の作業自体には携わらないが，コミュニティーとして翻訳・校閲作業が円滑に進められる仕組みを構築している。登録済みボランティアメンバーによる翻訳テキストは別の登録メンバーによる校閲を受け，これをクリアした翻訳のみが公開されることになっている[1]。

3.2 基本統計値

表 1 に TCSE が収録している TED Talks データの基本統計値を示す。

表 1　TCSE に収録された TED Talks の基本統計値

項　目	総　数
トーク	2,547
セグメント	691,788
拡張セグメント	308,720
エレメント（延べ）	6,166,375
エレメント（異なり）	83,909
語（延べ）	5,306,719
語（異なり）	83,891

それぞれの項目について簡単に解説したい。トーク（talk）とは個々のプレゼンテーションを指す。セグメント（segment）とは，TCSE における基本的な言語単位であり，トランスクリプトにおける「1 画面分の字幕文字列」に一致する。セグメントは「センテンス未満」の単位であることが多いが，そのようなセグメントを隣接するセグメントと結合し，センテンスを 1 つ以上含むセグメントに拡張したものが拡張セグメント（expanded segment）である。エレメント（element）は単語と類似した単位であるが，トランスクリプト内に含まれるメタ表示（Applause や Laughter など）や，句読点などの記号類も含んでいる。したがって，これらはその下の「語（延べ）」や「語（異なり）」よりも大きな値になっている。

3.3 継続時間および発話速度

コーパスとしての TED の 1 つの特徴は，いずれもおよそ 10〜20 分程度でそれぞれが完結した内容のトークとなっていることである。また，スピーカーの多様さには目を見張るものがある。1 人のスピーカーが複数のトークを担当することもあるが，TCSE に収録された 2,547 のトークは，実に 2,166 組の異なるスピーカーによるものである。しかし，そのように多様なスピーカー達によるトークであっても，継続時間や発話速度は，全体を通じてある程度の統制がみられる。

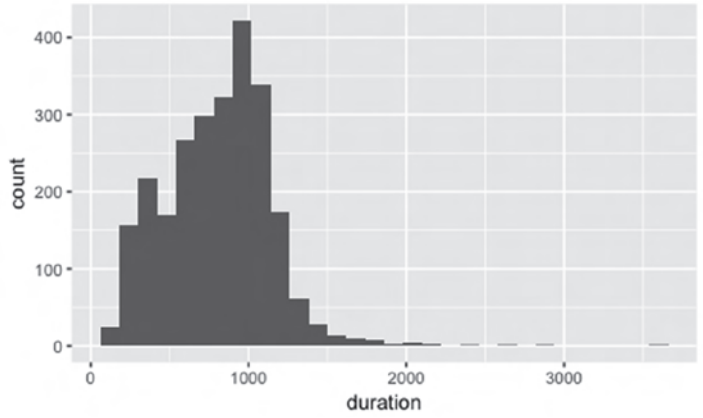


図 5　TCSE に収録されたトークの継続時間（秒）分布

図 5 は継続時間（秒）の分布を示したものである。いくらかの外れ値がみられるものの，それらを除けば 1,000 秒に満たない位置にピークを持つ分布を成していることが分かる。全トークの平均継続時間は 802.12 秒（SD = 330.57），すなわち約 13 分である。

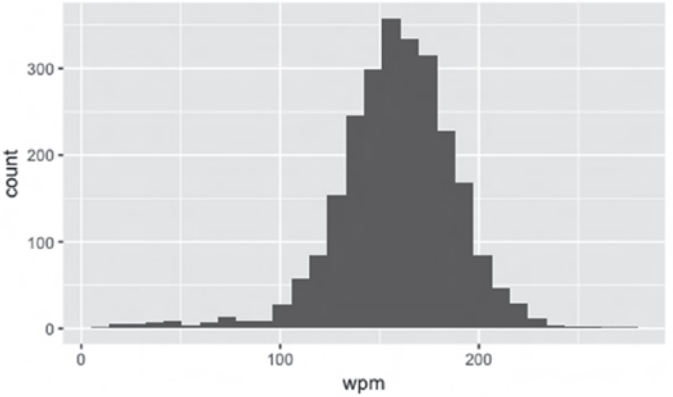


図６　TCSE に収録されたトークの平均発話速度分布

次に図 6 に発話速度（wpm）の分布を示す。平均発話速度は 158.20 wpm（SD = 30.47）である。なお，TED には舞踏や音楽を含むトークも多く含まれており，一部のトークについては正確な発話速度を計測することが難しい。図 6 で 100 wpm 以下に位置するデータは概ねそのようなものである。しかし，いずれにしても全体的には 160 wpm あたりをピークとした分布を示す。Carpenter and Just（1977）は英語の標準的な発話速度を 160 wpm としており，TED の平均発話速度はこれに一致する。

3.4　難易度

TCSE には言語教育に資する様々な機能が実装されているが，その 1 つはトークの相対的な難易度指標の表示である。この機能は Flesch-Kincaid Readability Ease（以下，Flesch-Kincaid 値とする）を計算することによって実現されている。Flesch-Kincaid 値は書かれた文章の「読みやすさ」を評価するための指標であるが，TCSE では多数のトークを教育的な目的で選り分ける際の参考になる要素の 1 つとして本指標を採用している。Flesch-Kincaid 値の計算式は次の通りである。

206.835−1.015（総語数 / 総文数）−84.6（総音節数 / 総語数）

図 7 は TCSE に収録された TED Talks のデータにおける Flesch-Kincaid 値の分布を示したもので，平均値は 57.27（SD = 10.16）である。なお，この平均値を伝統的な Flesch-Kincaid 値の評価の目安に照らすと，Fairly Difficult（高校レベル）と評価される（Gray, 2012）。

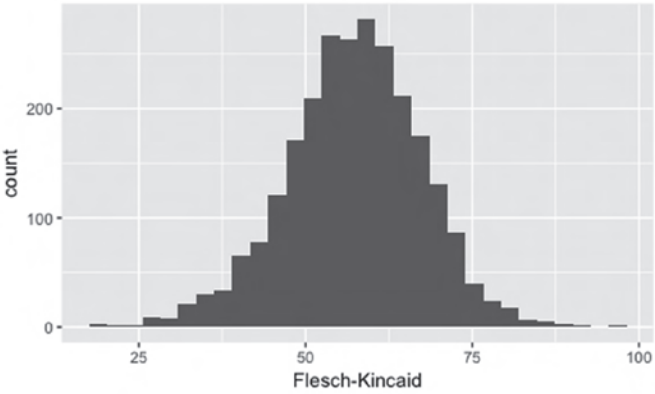


図 7　TCSE に収録されたトークの Flesch-Kincaid 値分布

3.5　TED コーパスの言語的特徴

ここでは，特殊コーパス（specialized corpus）としての TED コーパスの言語的特徴について予備的な考察を行う。一般的に言語コーパスは書き言葉（written language）から構成されたものと話し言葉（spoken language）で構成されたものとに大別される。TED Talks は英語の話し言葉によるプレゼンテーションであり，第一義的には話し言葉コーパスの一種として分類されるべきである。しかし，日常会話の言語を話し言葉の典型とするなら，聴衆の前で準備された内容を披露するプレゼンテーションの言語はそのような典型からはいくらか離れたバリエーションと考えるのが妥当であろう。では，TED Talks の言語は典型的な話し言葉から「どれだけ」また「どのように」離れているのだろうか。現在のところ，この問題に即答することは難しい。ここでは予備的な見通しを得るために，大規模コーパスとの定量的な比較調査の結果に基づいた考察を示す。

Corpus of Contemporary American English（COCA）は異なる使用域の現代アメリカ英語から成る大規模コーパスである（https://corpus.byu.edu/coca）。COCA では，話し言葉（spoken），フィクション（fiction），一般雑誌（popular

magazines），新聞（newspapers），学術雑誌（academic journals）の 5 つの使用域のデータがほぼ均等に（各 20 パーセント）含まれるよう調整されている（Davies, 2010）。そこで，これら 5 つの使用域のサブコーパスから抽出した COCA の語彙データ（レンマ＋頻度）と，TCSE から抽出した TED Talks の語彙データ（レンマ＋頻度）とで順位相関係数（Kendall's *tau*）を用いた分析を実施した[2]。なお，分析を実施するにあたっては，比較の対象となる 2 つのコーパスのどちらかだけに生起する語彙項目が大量に出てくることが予想される。事実，TED Talks においては特定のトークだけに頻出するような語が固有名詞を中心として大量に存在する。こうした事情による影響を最低限にとどめ，より一般的な語彙項目の相対頻度に基づいた分析を実施するため，TED Talks だけに出現する語，および COCA だけに出現する語をすべて除外し，結果として残った 28,180 種のレンマと各コーパス内での頻度を入力値とした[3]。その結果を表 2 に示す。なお, 数値計算と後に示すグラフの出力には R（Version 3.4.3）を使用した[4]。

表 2　TED コーパスと COCA［使用域別］の比較

コーパス	順位相関係数（Kendall's *tau*）	*p* 値
TED ⇔ COCA [Popular Magazine]	0.61	< 0.001
TED ⇔ COCA [Spoken]	0.59	< 0.001
TED ⇔ COCA [Academic Journal]	0.57	< 0.001
TED ⇔ COCA [News]	0.57	< 0.001
TED ⇔ COCA [Fiction]	0.51	< 0.001

最も大きい順位相関係数を示したのは COCA［Spoken］との組み合わせではなく，COCA［Popular Magazine］との組み合わせであった。その後 COCA［Academic Journal］および COCA［News］がほぼ同じ値で続き，最も小さい係数を示したのは COCA［Fiction］であった。TED Talks の言語は基本的に話し言葉ではあるが，いわゆるダイアログではなく，聴衆に向けてなされるプレゼンテーションの言語である。したがって，典型的な話し言葉コーパスには見られない，「書き言葉」的な特徴が生じているとしても不思議ではない。プレゼンテーションでは基本的に，相手の返答によって共有知識の有無や理解度を確認して発話の内容を微調整することは難しい。したがって，情報の自然な流れをあらかじめ考慮した上で，適切な語彙，構文，そしてレトリックを用いた「語

り」を展開することが求められる。これは通常，「書き言葉」に求められるような特徴である。表 2 の結果は，TED コーパスが基本的には話し言葉でありながらも，説得的・説明的なプレゼンテーションの言語であるために，書き言葉の特徴をいくぶん備えていることを示唆している[5]。

上記の観察は TED コーパスの特殊性に着目したものであるが，TED コーパスが英語という言語の基本的（あるいは普遍的）な特性を抽出するのに役立つ資源であることについても触れておきたい。図 8 は TCSE のデータを X 軸，COCA［All］のデータを Y 軸に，レンマごとの頻度をプロットしたものである（Kendall's *tau* = 0.62, $p < 0.001$）[6]。コーパス間で規模が大きく異なるため，プロットの際に対数変換を施しているが，付加された近似曲線の形状からわかるとおり，両コーパスには線形に近い相関がみとめられる。このことは，TED コーパスがプレゼンテーションという固有の使用域に属し，その特徴を多分に備えたデータである一方で，同時に英語という言語の基本的な特徴を保持したものであることを示唆している。

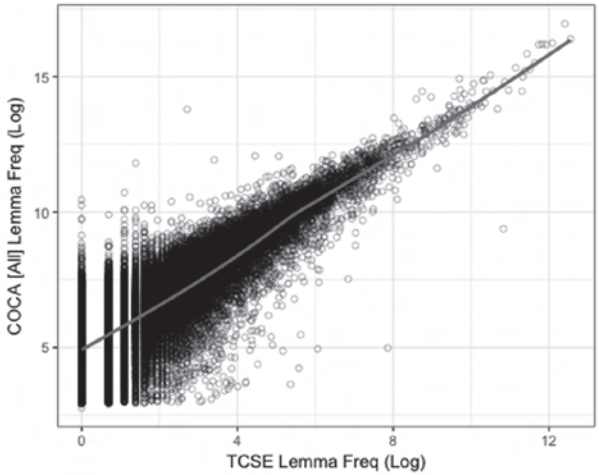


図 8　TCSE と COCA［All］のレンマ相対頻度プロット

以上，本節では TED Talks のコーパスとしての特徴を概観した。次節ではこのような特徴を持つ TCSE が，いくつかの認知言語学上の概念と強く結びついていることを論じる。

4．TCSE の認知言語学的特徴

本節では，TED コーパスを用いた言語研究／言語教育が，認知言語学（cognitive linguistics）の言語観と一致したものとなり得ることについて述べる。具体的には 1）用法基盤モデル，2）注意の枠，3）グラウンディングという 3 つの理論的概念を取り上げ，TCSE との関係について簡潔に論じていく。

4.1 用法基盤モデルの考え方

言語の用法基盤モデル（usage-based model of language）とは，認知言語学が掲げる言語観の基礎をなす考え方であり，言語を現実の使用から切り離された静的な体系として見るのではなく，個々の発話における状況や話者の一人称的体験と結び付いたスキーマ・ネットワークの動的な構造として見なすことを重視した言語観である（cf. Langacker, 2000; Bybee, 2010）。McEnery and Hardie（2011）が強調するように，用法基盤モデルの考え方はコーパス言語学と高い親和性を示す。一般にコーパスは語彙項目や構文の豊富な事例を提供し，言語研究や教育を現実の使用と結びつけると考えられている。しかし，多くのコーパスは豊富な用例・用法を提供するものの，それらが「現実の使用」の側面をどれだけ保持しているかについては考察の余地がある。多くのデータは文脈情報を欠いており，また話者の一人称的視点に基づく発話時の事態認識のあり方を再構成する材料にも乏しい。

一方で TED コーパスの場合はデータのすべてが個々に独立したトークであり，発話者の情報も開示されているため，取り出された表現は，それがいかに断片的であっても，固有の文脈や談話的意図と結びつけることが比較的容易である。もちろん，TED カンファレンスというきわめて限定された発話環境（＝聴衆を前にした 10〜20 分程度の英語によるプレゼンテーション）のみが対象となってしまうが，現実の使用と結びついた言語研究と教育を追求していくための貴重なリファレンスとなり得る。

4.2 注意の枠

注意の枠（windows of attention）という用語は Langacker（2001, 2012）によるもので，Chafe（1994）のイントネーション・ユニット（intonation unit）と共通した概念である。言語学や言語教育において，センテンスという単位は確固たる地位を有しており，多くの研究が，「あらゆる表現はセンテンスごとに

発話・理解されている」という前提のもとになされている。しかし，Langacker によると実際の談話における言語産出や理解は必ずしもその通りではない。言語使用者の注意の枠はより限定されており，多くの場合，言語産出や理解は主要な要素（trajector）とそれに次ぐ重要度を持つ要素（landmark）を中心とした局地的な構図の連続として展開する。このような観点から見ると，センテンスの区切りとは多分に恣意的なものであり，これを過度に重視することは，注意の枠の自然なつながりを適切に捉えることを阻害する。2.3 節で示したように TCSE において基本となっている言語単位はセンテンスではなくセグメントである。セグメントは TED のトランスクリプトが「字幕 1 画面分」を基準に構成されていることを利用した単位であり，形式的なセンテンスの幅にとらわれない。TCSE の提供する言語データ単位が Langacker の論じる注意の枠と完全に一致するわけではないが，話し言葉においてとりわけ重要な「注意の枠に基づいた言語の発話と理解」を考えるにあたり，興味深い材料になると思われる。

4.3 グラウンディング

グラウンディングとは Langacker（2002, 2008）の用語であり，談話の参与者間で共有された知識基盤のもとで事物を認識様態的（epistemically）に同定する認知的プロセスを指す。英語ではあらゆる名詞句（nominal）と定形節（finite clause）はグラウンディングのプロセスを経てはじめて成立する。グラウンディングは明示的なマーカーを伴わずに実現することもあるが，英語の場合，グラウンディング叙述詞（grounding predicate）と呼ばれる特定の語／形態素がこれを起動する役割を担う。名詞に付加する冠詞や，節を構成する際の法助動詞や時制接辞は，グラウンディング叙述詞の典型である。4.1 で論じた用例基盤の考え方とも重なるが，グラウンディングは言葉によって表される内容を現実世界に位置付ける役割を果たす。こうした考え方に基づく言語研究／言語教育を実施するにあたっては，コーパスの豊富な事例が有効であるが，従来的なコーパスでは，必要な文脈的・談話的情報のすべてを再現することが難しい。その点，トークごとに完結しており，かつ発話の状況や文脈情報が完備されている TED コーパスであれば，様々な関連要素を考慮に入れた言語研究／言語教育が可能になる。

5．まとめ

本稿では TED Corpus Search Engine（TCSE）の概要と可能性について論じた。2 節では TCSE の主な機能を概観した。3 節では TCSE に含まれるデータの様々

な統計値と共に，TED Talks の言語コーパスとしての特徴を示した。また，TED Talks の言語が単なる「話し言葉」ではなく，ある種の「書き言葉」的性質を有していること，しかしその一方で，大規模な均衡コーパスが示すような普遍的特徴をある程度備えていることについて述べた。そして 4 節では，1）用例基盤モデル，2）注意の枠，3）グラウンディングという 3 つの理論的概念を取り上げ，TCSE が認知言語学の考え方と一致する特徴を持つことについて論じた。

謝　辞

本稿の内容は英語コーパス学会第 43 回大会（関西学院大学）シンポジウム「話し言葉コーパスの構築と利用」における発表に基づいている。当日の発表とそれに先立つワークショップでは数々の有益なコメントを頂いた。また，『英語コーパス研究』の査読者からは詳細かつ的確な指摘と示唆を頂いた。ここに記して感謝したい。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。

注

1. TED トランスクリプトの翻訳については公式サイトで詳しい手順が公開されている。https://www.ted.com/participate/translate/get-started
2. 順位相関分析に広く用いられている Spearman's *rho* に比べ，Kendall's *tau* はとりわけ同順位の要素が多く存在するデータにおいて高い信頼性を示すと考えられている（cf. Howell, 1997）。
3. 今回の調査では語のレンマと頻度を入力値とし，品詞の区別を行っていない。したがって，品詞が異なっていてもレンマが同形の場合には同じ語として扱われる。これは，TCSE と COCA で用いている品詞解析システムが異なり，正確な比較ができないためである。TCSE で用いている Enju（http://www.nactem.ac.uk/enju）はテキストに Penn Treebank 形式の品詞タグを付与するが，COCA で用いている CLAWS（http://ucrel.lancs.ac.uk/claws）は CLAWS タグセットに基づいた出力を行う。品詞の違いを考慮したより精緻な調査は今後の課題である。
4. 本調査では 2014 年 3 月に取得した COCA の上位 6 万語の頻度データを使用した。COCA の総語数は 2018 年 2 月現在で約 5 億 7 千万語であるが，2014 年当時の総語数は約 4 億 4 千万語であり，当該のデータセットに含まれる上位 6 万語の総頻度は 3 億 7 千万語である。
5. TCSE と COCA［Popular Magazine］が最も高い相関係数を示す理由の 1 つとして、一般雑誌のテキストにはインタビューなど話し言葉の引用が多く含まれることから，話し言葉的特徴が備わっているということも考えられる（匿名の査読者から

の指摘に感謝する）。

6. TCSE と COCA の頻度データをプロットするにあたっては，1 万語や 10 万語などで正規化した調整頻度ではなく，粗頻度を対数変換した値を用いた。これは，各データセットにで極端な値を示す一部の語の影響を抑えるためである。なお，Johannessen and Guevara（2011）によるウェブ・コーパスに関する研究でも同様の手法が採用されている。

参考文献

Bybee, J. (2010) *Language, Usage and Cognition*. Cambridge: Cambridge University Press.

Carpenter, P. A. and M. A. Just (1977) “Reading Comprehension as Eyes See It.” In Just, M. A. and P. A. Carpenter (eds.), *Cognitive Processes in Comprehension*. New York: Psychology Press, pp. 109–140.

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

Davies, M. (2010) “The Corpus of Contemporary American English as the First Reliable Monitor Corpus of English.” *Literary and Linguistic Computing 25*, 4: 447–464.

Gray, C. J. (2012) “Readability: A Factor in Student Research?” *The Reference Librarian 53*, 2: 194–205.

Hasebe, Y. (2015) “Design and Implementation of an Online Corpus of Presentation Transcripts of TED Talks.” *Procedia: Social and Behavioral Sciences 198*, 24: 74–182.

Howell, D. C. (1997) Statistical Methods for Psychology. Belmont: Cengage Wadsworth.

Johannessen, J. B. and E. R. Guevara (2011) “What Kind of Corpus is a Web Corpus?” In B. S. Pedersen, G. Nešpore and I. Skadiņa (eds.), *NODALIDA 2011 Conference Proceedings*, pp. 122–129.

Langacker, R. W. (2000) “A Dynamic Usage-based Model.” *Cognitive Linguistics 12*, 2: 143–188.

Langacker, R. W. (2001) “Discourse in Cognitive Grammar.” In Barlow, M. and S. Kemmer (eds.), *Usage-based Models of Language*. Stanford: CSLI, 1–63.

Langacker, R. W. (2002) “Deixis and Subjectivity.” In Brisard, F. (ed.), *Grounding: The Epistemic Footing of Deixis and Reference*. Berlin: Mouton de Gruyter, pp. 1–28.

Langacker, R. W. (2008) *Cognitive Grammar: A Basic Introduction*. *Oxford*: Oxford University Press.［山梨正明（監訳）『認知文法序説』研究社］

Langacker, R. W. (2012) “Elliptic Coordination.” *Cognitive Linguistics 23*, 3: 555–599.

McEnery, T. and A. Hardie (2011) *Corpus Linguistics: Method, Theory and Practice*. Cambridge: Cambridge University Press.［石川慎一郎（訳）『コーパス言語学：手法・理論・実践』ひつじ書房］

（同志社グローバル・コミュニケーション学部　Email: yhasebe@mail.doshisha.ac.jp）