

言語研究資料としての The Movie Corpus の可能性と留意点*

Possibilities and Limitations of the Movie Corpus for Linguistic Research

山内 昇

YAMAUCHI, Noboru

摘要

Today, corpora with more than 100 million words can be accessed and searched online instantaneously on English-Corpora.org, thereby becoming more accessible for use in linguistic research. However, caveats to the use of corpora have been insufficiently shared amongst researchers, thus resulting in some corpus-based studies drawing invalid conclusions from inadequate analyses. This is partly due to the lack of opportunities to share caveats to prevent such circumstances. Therefore, this study attempts to present the knowledge and skills relevant to the use of the Movie Corpus (Davies, 2019) which is one of the corpora provided in English-Corpora.org. Firstly, a brief overview of the Movie Corpus is provided. Secondly, the study discusses the possibilities that the corpus affords linguistic research and how it benefits studies that focus on conventional meanings and uses, discourse functions, and pragmatic licensing conditions of linguistic phenomena in spoken language. The study is based on three key qualities of film dialogues: (i) because the intention and purpose of utterances in movies are designed to be understandable by audiences, it is easier for linguists/researchers to grasp the functional motivation of the linguistic expressions used; (ii) film characters tend to be less complex and more predictable than real people which makes it easier for linguists/researchers to assess their background and the knowledge shared between speakers; and (iii) because utterances in films are validated by native speakers of a language as part of the filmmaking process, they provide highly reliable data for linguistic research. Finally, the phrase *speaking of which*—a discourse marker indicating a change of topic—is discussed as a case study on how to extract data from the corpus using its search interface. The following important usage information will be included: (i) phrases that appear in search results might differ from spoken dialogue; (ii) the information displayed in search results may differ from the source information of the actual files that include subtitles; and (iii) the corpus may contain typographical errors (e.g., *speakig of which*, *speaking of wich*, and *speakingofwhich*). The issues raised in this paper can help improve future corpus-based studies, especially those using the Movie Corpus.

キーワード : コーパス *The Movie Corpus* 話し言葉 映画の会話 談話標識

Keywords: *corpus, The Movie Corpus, spoken language, film dialogue, discourse marker*

1. はじめに

近年、English-Corpora.org (<https://www.english-corpora.org>)において、数億語規模のコーパスがオンラインで使用可能な検索環境と共に公開されており、言語研究におけるコーパスの利用が拡大している。コーパスの利用方法に関しては概説書等で詳細な情報が公開されているが、コーパス利用時の注意点に関してはほとんど扱われておらず、研究者コミュニティ内で十分に共有されているとは言い難い。実際、コーパスの構成や検索ツールの仕様の理解不足により、誤った分析結果を導いてしまった研究も存在している。⁽¹⁾このような状況を作り出している要因の一つとして、コーパス利用時の注意点に関する情報公開の場が非常に少ないことが挙げられる。そこで本論では、コーパスを利用した言語研究の現状を改善するために、コーパスの初学者から専門家まで広く役立つよう、コーパス利用時における注意点について論じる。既存のコーパスを網羅的に扱うことはできないため、English-Corpora.org で公開されている The Movie Corpus (Movies) (Davies, 2019) を取り上げ、言語研究資料としての可能性を明確にした上で、利用時における注意点を明示する。本論の構成は以下の通りである。第二節では、Movies がどのようなコーパスであるのかを確認する。第三節では、映画等の会話の特徴を踏まえ、Movies の言語研究資料としての可能性を示す。第四節では、Movies を利用したデータ検索の具体例として、談話標識 *speaking of which* の該当例を可能な限り多く抽出する方法とその際の注意点を示す。第五節は全体のまとめである。

2. The Movie Corpus の概要

Movies は 1930 年から 2018 年にわたる約 90 年間に公開された約 2 万 5 千作の映画作品の英語字幕が収録されたコーパスである (Davies, 2019)。その規模は約 2 億語であり、The British National Corpus の話し言葉パートの約 20 倍に相当する (ibid.)。⁽²⁾ また、The Corpus of Contemporary American English (COCA) におけるジャンル SPOKEN の約 1 億 2 千万語を上回る。いわゆる自然会話のテキストと Movies のテキストの違いについては第三節で述べるが、ここでは Movies の規模が従来の話し言葉のコーパスよりも大きいことを示しておきたい。Movies に収録されている英語字幕は映画やドラマの字幕提供サイトの OpenSubtitles (<https://www.opensubtitles.org>) から収集されたものである。各データは Internet Movie Database (IMDb) (<https://www.imdb.com>) と関連づけられており、作品名だけでなく、公開年、ジャンル、筋書き、観客指定 (PG-13 指定など)、IMDb における評価等の情報も参照できる。英語字幕の収集方法など、Movies の制作過程に関しては Davies (2021) を参照されたい。Movies の検索インターフェイスでは、English-Corpora.org に公開されている他のコーパスと同様に、レンマ検索、同義語検索、ワイルドカード検索、OR/NOT 検索、CLAWS7 の品詞タグを利用した検索等が可能である。⁽³⁾ Movies は購入版も入手できるが、購入版の場合、著作権の侵害を避けるために全体の 5% が伏せ字 (@) にされている (<https://www.corpusdata.org/limitations.asp>)。本論では、多くの人が利用しやすいオンライン公開版を使用する。以下、Movies という場合、オンライン公開版を指す。

3. 言語研究資料としての The Movie Corpus の可能性

いわゆる「コーパス言語学」と呼ばれる分野で著名な John Sinclair は、映画やドラマの言語は人工的な環境で会話を模倣したものであり、自然会話を真に反映したものではないと述べている (Sinclair, 1991, p. 16)。一方 Davies (2021) は、映画等の会話は自然会話の一部の側面を確かに反映しており、Movies は話し言葉に関する豊富なデータを提供すると共に米英等の方言差や言語変化に関する研究に有益であると述べている。両者の意見の違いは映画等の会話と自然会話の相違点と類似点のどちらに重きを置いているのかにある。自然会話とは異なる傾向を示す現象を扱う場合には、Davies の立場でも Movies を使用することは推奨しないと思われる。本節では、映画等の会話の特徴に関する研究を踏まえ、Sinclair (1991) と Davies (2021) とは異なる観点から、Movies がどのような研究・調査に有益であるのかを示す。

映画等の会話の言語学的特徴に関しては数多くの研究がある (Bednarek & Zago, 2021)。ここでは紙幅の都合上、Quaglio (2008) の調査のみを取り上げる。Quaglio (2008) は、Friends (TV Series 1994–2003) の会話と自然会話を比較し、自然会話に見られる特徴(一部の罵り語の使用を除く)は Friends の会話にも観察され、逆に Friends に観察される特徴は自然会話にも観察されると指摘している。しかし、一部の表現の使用頻度には差が見られ、Friends の会話では、ヘッジ表現 (e.g., kind of; sort of)、等位タグ (e.g., and stuff like that)、曖昧指示 (e.g., stuff; thing)、談話標識 you know の頻度が低く、一部の強意副詞 (e.g., so; really; totally) 等の頻度が高い (Quaglio, 2008, pp. 199–208)。また、Friends の場合には、発話の重なり、発話の中断、不明瞭な表現がほとんどなく、発話のターンも均等に分配される (Quaglio, 2008, p. 208, fn. 7)。同様の特徴は映画の会話にも見られるため、発話の重なりや話者交替など、会話分析の分野で問題とされる現象を扱う場合には、Movies よりも自然会話のコーパスを使用する方が妥当である。

しかし、映画等の会話は、以下 3 点の特徴を有するため、話し言葉における言語現象の典型的な意味や談話機能、使用条件等に関する調査・研究において貴重な言語研究資料となる。第一に、映画の会話等の場合、登場人物が理由なく発話することはなく、登場人物の発話は慎重に選択されており、発話が繰り返されたり、曖昧になったりすることが少ない (Lucy, 1996, p. 167)。映画等の会話は発話の意図や目的を視聴者が理解できるように作られているため、分析者の視点から言語表現の使用動機を把握しやすいと言える。自然会話の場合、発話者自身に発話の意図や目的を確認しても説明できない場合がある。第二に、Wray (2008, p. 175) によれば、ドラマの場合には登場人物の思考や感情を描写するために言語や行動が使用されるが、現実の生活ではそれらを隠すために言語や行動が使用される場合がある。また、映画等の会話では、登場人物が現実の人物よりも複雑でなく予測可能である (Mattsson, 2009, p. 30)。映画等の会話の場合には、話者に関する背景知識や話者間の共有知識を分析者が把握しやすいと言える。第三に、映画等の会話は映画製作の過程で繰り返し母語話者による承認 (imprimatur) を受けるため、母語話者により妥当と認められた豊富な発話を提供する (Alvarez-Pereyre, 2011, p. 61)。自然会話コーパスの場合、発話の書き直しや修正は施されないため、発話の容認可能性に関しては使用者側で判断する必要がある。以上を踏まえると、Movies は話し言葉における言語現象の典型的な意味や談話機能、使用条件等に関する調査・研究に有益であると考えられる。

もちろん、映画等の会話には自然会話とは異なる言語変種が混在する場合があるため、その点には注意が必要である。例えば、Star Wars 作品に登場する Yoda は通常とは異なる語順の英語を話す (e.g., When 900 years you reach, look as good you will not. (*Return of the Jedi* (1983)) (Crystal, 1987, p. 98))。Doctor Who (TV Series 2005-) に登場する Chantho という異星人は発話を自身の名前の間に挿入する (e.g., Chan, that would be rude, tho! (*Doctor Who* (TV Series 2007) Season 3 Episode 11 (Utopia)))。これらは特定の登場人物に特有の言語変種であるが、特定の人物像を喚起するいわゆる「役割語・キャラクター言語」(金水, 2000) に相当するような言葉遣いが含まれる場合もある。例えば、Bones (TV Series 2005–2017) に登場する Finn Abernathy の発話には米国南部方言の特徴 (ain't や二重否定などの使用) が見られる (Mitchell, 2015, pp. 300–301)。

本節では、映画等の会話の特徴に関する先行研究を踏まえ、言語研究資料として Movies を使うことでどのようなことが可能になるのかを示した。次節では、Movies を使用したデータ検索の一例として、談話標識の一つである *speaking of which* という表現の検索を取り上げ、検索時の留意点を示す。当該表現に関する研究では、*which* の指示対象の有無が争点の一つとなっている (山内, 2020)。自然会話から得られたデータを使用した場合、当該表現の使用条件とは別の要因 (話者の注意力など) により指示対象が不明瞭になる可能性がある。映画等の会話の場合にはそのような要因が関わり難いため、映画等の会話から指示対象が不明瞭な使用例が見つかれば、当該表現の使用条件を検討する上で価値の高いデータとなる。

4. The Movie Corpus の留意点

コーパスを利用した言語研究は、検索結果の再現性が確保されており、データの客観性と信頼性が高いというイメージがある。しかし、実際には、使用したコーパスやツールの仕様等の様々な要因から検索結果に変動が生じるため、コーパスを使用すれば、自動的に客観性と信頼性の高いデータが得られるわけではない。客観性と信頼性の高いデータを抽出するには、検索手法や検索結果の取り扱いに注意する必要がある。本節では、Movies を使用したデータ検索の具体例として、話題転換の際に使用される談話標識の一つである *speaking of which* (e.g., We've been invited to Rachel and Jamie's wedding—speaking of which, did you know that they're moving to New York? (Cambridge University Press, n.d.)) の検索を取り上げる。以下では、検索インターフェイスの検索文字列の入力欄に “speaking of which” と入力し、検索を実行した場合にどのような点に注意すれば、当該表現のデータを可能な限り多く抽出できるのかを示す。

4.1. 二種類の検索結果

最初に、検索インターフェイスの検索結果は検索毎に変動する場合があることを示す。説明の都合上、検索インターフェイスにおける検索結果の表示画面の見方を説明する。検索結果の表示画面は FREQUENCY と CONTEXT に分かれる。FREQUENCY には、検索文字列の使用頻度 (FREQ) が表示される。CONTEXT には、通し番号、公開年、制作国、作品名、検索文字列に該当箇所とその前後文脈が表示される (図 1 参照)。⁽⁴⁾

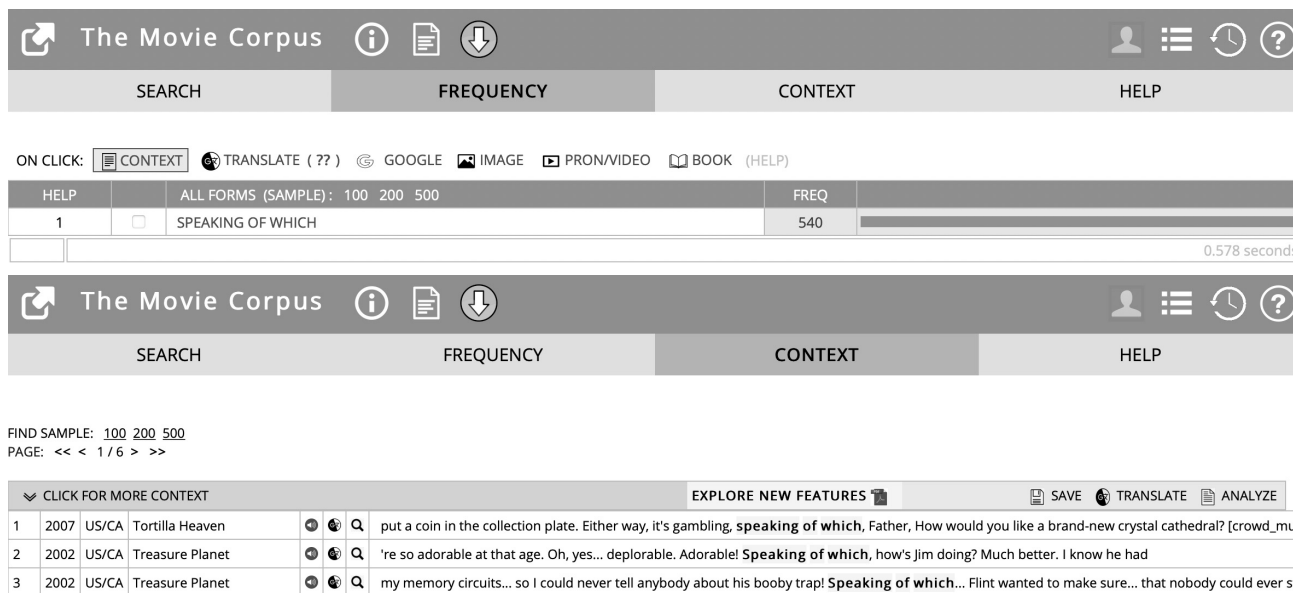


図1 FREQUENCY(上)とCONTEXT(下)の表示画面

CONTEXT の検索結果は 1 頁につき 100 行ずつ表示される。検索結果の文字列が一致する場合には重複例として扱われ、検索結果には表示されない。重複例のため非表示になったデータがある場合、直後に表示されるデータの作品名の後に、非表示となったデータの数が表示される。通し番号、公開年、制作国をクリックすると CONTEXT+ が表示され、作品の概要や広い前後文脈を確認できる。

図 1 における FREQUENCY の表示画面には、上述の文字列検索に該当する箇所の頻度として「540」という数値が表示されている。FREQUENCY に表示される頻度と CONTEXT における最終行の通し番号は一致するはずである。しかし、FREQUENCY に表示される頻度は一定であるが、CONTEXT の最終行の通し番号は変動する場合があります、検索により「539」になる場合と「540」になる場合とがある(図 2 参照)。前者を「検索結果 A」と後者を「検索結果 B」と呼ぶことにする。⁽⁵⁾

537	1986	Misc	Circus of the Stars #11	there were so many broken marriages in this town. It's an epidemic. Speaking of which . Dr. Roberts is getting a divorce? You didn't hear
538	1971	Misc	Bakuto gajjin butai	It's called Kadesa. A dance to welcome visitors. Quite an honor. Speaking of which , may I ask what you hope for? That we form an
539	1970	Misc	May Morning	don't think I've got any hope? Discuss it with your tutor. Speaking of which , explain to me why you chose Finlake? They gave me a
538	1986	Misc	Circus of the Stars #11	there were so many broken marriages in this town. It's an epidemic. Speaking of which . Dr. Roberts is getting a divorce? You didn't hear
539	1971	Misc	Bakuto gajjin butai	It's called Kadesa. A dance to welcome visitors. Quite an honor. Speaking of which , may I ask what you hope for? That we form an
540	1970	Misc	May Morning	don't think I've got any hope? Discuss it with your tutor. Speaking of which , explain to me why you chose Finlake? They gave me a

図2 検索結果 A(上)と検索結果 B(下)の最終行

検索結果 A と B の(重複例のため非表示になったものも含め)全データを数えると、どちらの場合でもデータの総数は 540 例であることが分かる。調査・研究に検索インターフェイスを使用するには、その仕組みを可能な限り理解しておく必要がある。以下では、使用頻度と最終行の通し番号が一致しない理由を示す。

検索結果 A において通し番号が 539 となる理由は通し番の重複である。CONTEXT の検索結果は 1 頁につき 100 行ずつ表示されるため、1 頁目における最終データの通し番号は 100 になるはずである。しかし、検

索結果 A の場合、1 頁目の最後に表示されるデータ (\$21 a Day—(Once a Month)(1941)) の通し番号が 101 となっており、2 頁目最初のデータ (The Perfect Sleep (2009)) の通し番号も 101 となっている (図 3 参照)。

100	1946	US/CA	The Dark Corner	🔍🔍🔍	pick you up at a rummage sale. I'm a sucker for bargains. Speaking of which , if you can't get nines in those nylons, I'll
101	1941	US/CA	\$21 a Day - (Once a ...	🔍🔍🔍	passed on his information, this whole thing could blow up in our faces. Speaking of which , when are we going to pick up the explosive? Oh,
101	2009	US/CA	The Perfect Sleep	🔍🔍🔍	is that sick, burning desire to get it back. The perfect sleep. Speaking of which , you probably think this is one of those stories, a study
102	2009	US/CA	Paper Man	🔍🔍🔍	at the very beginning. That's a very good place to start. - Speaking of which ... Close your eyes. Okay, you can open them. -

図 3 検索結果 A における 1 頁目の最後(上)と 2 頁目の最初(下)

つまり、検索結果 A の場合、通し番号 101 が 2 回割り当てられているため、最終データの通し番号が 539 になるということである。一方、検索結果 B の場合、通し番号 101 が重複していないため、最終データの通し番号は 540 になる (図 4 参照)。

99	1946	US/CA	The Dark Corner	🔍🔍🔍	pick you up at a rummage sale. I'm a sucker for bargains. Speaking of which , if you can't get nines in those nylons, I'll
100	1941	US/CA	\$21 a Day - (Once a ...	🔍🔍🔍	passed on his information, this whole thing could blow up in our faces. Speaking of which , when are we going to pick up the explosive? Oh,
101	2009	US/CA	The Perfect Sleep	🔍🔍🔍	is that sick, burning desire to get it back. The perfect sleep. Speaking of which , you probably think this is one of those stories, a study
102	2009	US/CA	Paper Man	🔍🔍🔍	at the very beginning. That's a very good place to start. - Speaking of which ... Close your eyes. Okay, you can open them. -

図 4 検索結果 B における 1 頁目の最後(上)と 2 頁目の最初(下)

検索結果 A の場合、非表示の重複例を入れると 1 頁目に 101 例のデータが含まれていることになる。検索結果 A の 1 頁目に 101 例が含まれる理由は不明である。現時点では、検索インターフェイスの仕様と考えるしかない。

それでは検索結果 B の場合に \$21 a Day—(Once a Month) の通し番号が 100 になるのはなぜだろうか。実は検索結果 A と B では、データの表示順序が一部異なる。例えば、検索結果 A の場合、Maid in Manhattan (2000) を出典とするデータは 7 番目に表示されるが、検索結果 B の場合には、7 番目ではなく 187 番目に表示される (図 5-6 参照)。

6	2002	US/CA	Highway	🔍🔍🔍	the world being... a fucking blender... and me being a wild strawberry. Speaking of which ... you all look a little torqued... and I am the discoverer
7	2002	US/CA	Maid in Manhattan	🔍🔍🔍	- And you're what? - I'm working hard for the money. Speaking of which , you hand in your application? Management? Yeah. What are
8	2001	US/CA	Bones	🔍🔍🔍	sleep in a spook house... surround me with chicks more scared than me. Speaking of which ... Why would you say that about my sister, man? -

図 5 検索結果 A における通し番号 6-8

6	2002	US/CA	Highway	🔍🔍🔍	the world being... a fucking blender... and me being a wild strawberry. Speaking of which ... you all look a little torqued... and I am the discoverer
7	2001	US/CA	Bones	🔍🔍🔍	sleep in a spook house... surround me with chicks more scared than me. Speaking of which ... Why would you say that about my sister, man? -
8	2001	US/CA	Antitrust	🔍🔍🔍	he'll get back to work now. - Yeah. Copy that. - Speaking of which ... - Okay. I'm on my way back. - Did
186	2003	US/CA	Open Water	🔍🔍🔍	'm sorry, honey. I threw up. Fish got ta eat too. Speaking of which , the thing patting at your leg, is just a little cleaner
187	2002	US/CA	Maid in Manhattan	🔍🔍🔍	- And you're what? - I'm working hard for the money. Speaking of which , you hand in your application? Management? Yeah. What are
188	2002	US/CA	The Skulls II	🔍🔍🔍	. Well, maybe you don't know me as well as you think. Speaking of which , rumor has it you've been tapped. It's just a

図 6 検索結果 B における通し番号 6-8(上)と 186-188(下)

検索結果 B の場合、検索結果 A よりも 1 頁目のデータ数が 1 例少ないため、\$21 a Day—(Once a Month) が通し番号 100 として表示される。検索結果 A と B には、このように表示される箇所が異なるデータが数多く存在している(表 1 参照; 網掛け部分は頁を跨いで表示位置が変わるデータである)。

表 1 検索結果 A と B における通し番号の相違点一覧

作品名	検索結果 A		検索結果 B	
	頁番号	通し番号	頁番号	通し番号
(Tortilla Heaven ~ Highway)	(1)	(1~6)	(1)	(1~6)
Maid in Manhattan	1	7	2	187
(Bones ~ The Dark Corner)	(1)	(8~100)	(1)	(7~99)
\$21 a Day - (Once a Month)	1	101	1	100
The Perfect Sleep	2	101	2	101
(Paper Man ~ Pistol Whipped)	(2)	(102~116)	(2)	(102~116)
Lost Boys: The Tribe	2	117	3	283
(Grindhouse ~ Open Water)	(2)	(118~187)	(2)	(117~188)
(The Skull II ~ The Immigrant)	(2)	(188~218)	(2)	(119~218)
The Hang Over Part III	3	219	4	361
(A Thousand Words ~ Take My Wife)	(3)	(220~283)	(3)	(219~282)
(Impulse ~ The Device)	(3~4)	(284~360)	(3~4)	(284~360)
Wrong Turn 6: Last Resort	4	361	5	457
(Screwed ~ 12 Gifts of Christmas)	(4)	(362~456)	(4)	(362~456)
(Wishin' and Hopin' ~ Rogue)	(4)	(457~484)	(4)	(458~485)
Noise	5	485	6	525
(The Tinger's Tail ~ Murder Dot Com)	(5~6)	(486~524)	(5~6)	(486~524)
(Mrs. Ratclif's Rev. ~ Bakuto Gaijin Butai)	(6)	(525~538)	(6)	(526~539)
May Morning	6	539	6	540

以上、検索結果が二種類生じる理由を検索インターフェイスの仕様を踏まえて説明した。検索結果が二種類になり得ることを事前に把握していないと、例えば Microsoft Excel のシートにデータをコピーして処理を行う際に、検索結果 A と B が混在し、気付かないうちにデータの重複や欠落が発生する可能性がある。⁽⁶⁾

なお、この現象は English-Corpora.org の検索インターフェイスを使用する限り、別のコーパスを使用する場合でも発生する可能性があるため、Movies の利用を避ければ済むわけではない。検索インターフェイスを使用する場合には、検索毎に検索結果における最終行の変動の有無を確認する必要がある。

4.2. “speaking of which” により余分に検索されるもの

“speaking of which” の検索結果には重複例が含まれるため、FREQUENCY に表示される 540 という頻度を Movies における当該表現の使用頻度と見なすことはできない。検索結果に重複例等の余分なデータがどの程度含まれているのかを把握することは分析データの抽出過程で必要な作業である。余分なデータは結果の解釈に影響を与える場合があるため、検索結果から手作業で取り除く必要がある。以下では、除外が必要となるケースとしてどのようなものがあるのかを示す。

第一に、検索インターフェイスの仕様で非表示になっていない重複例を除外する必要がある。検索インターフェイスでは、検索結果に表示される文字列が一致する場合、重複例として扱われ、検索結果に表示されない。例えば、通し番号 23 のデータは通し番号 22 のデータと重複するため非表示になっている(図 7 参照)。

21	2000	US/CA	Loser	🔍	🔍	🔍	. Come on, guys. I just washed all these towels. Oh, speaking of which , Paul, next time don't use so much starch. Hey
22	2000	US/CA	The Broken Hearts Cl...	🔍	🔍	🔍	's over... and you won't catch me going through any mourning cycle. Speaking of which , I haven't seen Princess Taylor all week. He's been
24	1999	US/CA	Dogma (1)	🔍	🔍	🔍	I? m gon na yank your sac off like a paper towel. Speaking of which , you? re awfully nude. - Rufus, is it?

図7 検索結果に非表示となる重複例(検索結果A)

検索インターフェイスの仕様により重複例が非表示となる場合には、直後のデータの作品名の後に非表示となっているデータ数が明示されるため、重複例の存在を容易に把握できる。図7の場合、直後のデータの作品名に「(1)」と表示されている。しかし、検索結果に表示される文字列が一致しない場合には、実際には重複例であっても、重複例として認識されずに検索結果に表示される(図8参照)。

324	2015	US/CA	Girl on the Edge	🔍	🔍	🔍	uh... - we thought we confiscated her phone. - Uh-huh. Uh... speaking of which , uh... what about communication? Uh... speaking of which,
325	2015	US/CA	Girl on the Edge	🔍	🔍	🔍	. Uh... speaking of which, uh... what about communication? Uh... speaking of which , uh... what about communication? Well, she gets a family
379	2013	US/CA	Slink	🔍	🔍	🔍	'D BRING HER BACK SOMETHING REALLY NICE. BRING HER BACK SOMETHING REALLY NICE. SPEAKING OF WHICH , I'LL PROBABLY NICE. SPEAKING OF
380	2013	US/CA	Slink	🔍	🔍	🔍	HER BACK SOMETHING REALLY NICE. SPEAKING OF WHICH, I'LL PROBABLY NICE. SPEAKING OF WHICH , I'LL PROBABLY JUST GO AND BUY HER SOMET
381	2013	US/CA	Slink	🔍	🔍	🔍	NICE. SPEAKING OF WHICH, I'LL PROBABLY JUST GO AND BUY HER SOMETHING SPEAKING OF WHICH , I'LL PROBABLY JUST GO AND BUY HER SOMETH

図8 検索結果に表示される重複例①(検索結果A)

図8における2つの作品の実際の映像を確認すると当該表現の発話は一度のみであるため、それぞれの最初のデータ以外は重複例である。Slink (2013) の字幕ファイルを確認すると、同一の文字列が繰り返し入力されており、機械的に重複例として処理することは難しいことが分かる(図9参照)。

```

90
00:05:04,338 --> 00:05:05,704
BRING HER BACK SOMETHING REALLY
NICE.
SPEAKING OF WHICH, I'LL PROBABLY

91
00:05:05,706 --> 00:05:07,005
NICE.
```

```

SPEAKING OF WHICH, I'LL PROBABLY
JUST GO AND BUY HER SOMETHING

92
00:05:07,007 --> 00:05:08,407
SPEAKING OF WHICH, I'LL PROBABLY
JUST GO AND BUY HER SOMETHING
BECAUSE HE'S NOT GONNA HAVE
```

図9 Slink (2013) の字幕ファイル

図8のように重複例が連番で表示されている場合、重複例を容易に発見できるが、同一の作品が異なる名称でも収録されている場合には連番で表示されないため発見が難しい(図10参照)。

118	2007	US/CA	Grindhouse	🔍	🔍	🔍	be waiting for us to join at the Texas Chilli Parlor. Oh shit. Speaking of which? What happened with you and Nate last night? Well you know
130	2007	US/CA	Death Proof	🔍	🔍	🔍	'll be waiting for us at the Texas Chili Parlor. Oh, shit. Speaking of which ... what happened with you and Nate last night? Well, you

図10 検索結果に表示される重複例②(検索結果A)

図10におけるDeath Proof (2007) はGrindhouse (2007) の一部を作品化したものである。これら2作品には、全く同じ発話が収録されているため、一方を重複例として扱う必要がある。⁽⁷⁾ 発話の内容は同じでも、字幕フ

ファイルの作成方法により字幕の細部が異なるため、重複例として認識されていない。図 7 における重複例 (1 例) も含めて、540 例から 5 例を重複例として除外する必要がある。

第二に、検索結果には当該表現の言い直しが含まれている場合がある (図 11-12 参照)。

367	2013	US/CA	Hansel & Gretel Get ...	🔍	line of business, I like to move pretty sporadically... So, um, <u>speaking of which</u> ... Witch? Yeah, yeah speaking of which, like, how
368	2013	US/CA	Hansel & Gretel Get ...	🔍	sporadically... So, um, speaking of which... Witch? Yeah, yeah <u>speaking of which</u> , like, how did you get into the growing business? I

図 11 検索結果における通し番号 367-368 (検索結果 A)

ASHTON: So, um, did you just move here? AGNES: Well, actually very recently. In my line of business, I like to move pretty sporadically... ASHTON: So, um, speaking of which... AGNES: Witch? ASHTON: Yeah, yeah speaking of which, like, how did you get into the growing business?

図 12 通し番号 367-368 の実際の発話 (検索結果 A)

図 12 は Ashton と Agnes (魔女) の会話である。Ashton が speaking of which を使用したところ、Agnes は Witch? (魔女だって?) と聞き返している。それに対し、Ashton は再度 speaking of which を使用している。このような言い直しの事例は 1 例としてカウントできる。

第三に、検索結果には当該表現のデータとして表示されていても、実際の映像を確認してみると当該表現が使用されていない場合がある。“speaking of which” の検索結果には実際の発話と異なる事例が 13 例含まれている (表 2 参照)。これらは非該当例として検索結果から除外する必要がある。⁽⁸⁾⁽⁹⁾

表 2 検索結果と実際の発話の不一致 (検索結果 A)


作品名	実際の発話
48 The Associate (1996)	speaking of realism
92 Support Your Local Gunfighter (1971)	speaking of business
94 Where Were You When the Lights Went Out (1968)	since you mention it
97 The Tingler (1959)	speaking of weird effects
100 The Dark Corner (1946)	speaking of bargains
241 Born Bad (2011)	by the way
277 Make the Yuletide Gay (2009)	speaking of
285 Stag Night (2008)	n/a
291 3-Day Weekend (2008)	speaking of
303 Me You and Five Bucks (2016)	speaking of
331 American Muscle (2014)	by the way
444 Dancer and the Dame (2015)	which reminds me
539 May Morning (1970)	by the way

このように実際の発話と字幕が異なる場合があるため、データの質を高めるには一つ一つ実際の映像を確認する必要がある。検索結果と実際の発話の不一致は OpenSubtitles 上の精度の低い英語字幕が Movies に収録されているために起こる。実際の映像を確認することなく、字幕の精度を判断することは難しい。この問題を改善するには、Movies のデータに OpenSubtitles 上の英語字幕を使用するのではなく、手間はかかるが Subtitle Extractor 等のソフトウェアにより DVD から直接抽出した英語字幕を使用する必要がある。

第四に、検索結果の出典情報と字幕ファイルの出典情報が一致しない場合がある。例えば、検索結果には、Tunnel Vision (1976) を出典とするデータが含まれているが、全く同じデータが The TV Corpus (TV) において CSI: Miami (TV Series 2008) のデータとしても収録されている (図 13 参照)。



Source information:


	Title Tunnel Vision (IMDB) (Open Subtitles) Year: 1976 / Genre: Comedy
	Plot A committee investigating TV's first uncensored network examines a typical day's programming, which includes shows, commercials, news programs, you name it. What they discover will surely ...
	More info Length: 70 min / Rating: R / IMDB rating: 5.1 (513 votes)

Expanded context:

. Now sell it like you did on tv. This is Amanda Brighton for Amanda's Orchards. Squeeze out every drop of goodness. Oh, you know you will. Little, dirty girl. So Carlos hid this camera in his safe deposit box. That's what the entire robbery is about. Someone went to a lot of trouble to get this tape. I better tell Horatio about this, A.S.A.P. Be careful with that. That's my mother's favorite painting. Lieutenant Caine. I didn't know you were an art lover. **Speaking of which**, Carlos, you and Amanda Brighton on tape has become an instant classic. Little, dirty girl. That's not meant for public viewing. Yes, blackmail never is, is it? Blackmail? Brighton's never paid me a cent. There are many forms of currency. See, now you're reaching, Caine. And I'm not alone. I'm impounding your vehicle. You can't do that. I just did, Carlos. Should I call you a cab



Source information:

	Series CSI: Miami (IMDB) (Years: 2002-2012: 232 episodes) Country: USA Genre: Action, Crime, Drama
	Series info The cases of the Miami-Dade, Florida police department's Crime Scene Investigations unit.
	Episode Tunnel Vision (2008) (IMDB) (Open Subtitles)
	Episode info Length: 44 min / Rating: TV-14 / IMDB rating: 1093887 (141 votes)
	Episode plot The team finds a body in a sinkhole, which turns out to be related to an underground bank robbery. The case also involves an incriminating video of a wealthy businessman's daughter with a drug dealer.

Expanded context:

. Now sell it like you did on tv. this is amanda brighton for amanda's orchards. Squeeze out every drop of goodness. Oh, you know you will. Little, dirty girl. so carlos hid this camera in his safe deposit box. That's what the entire robbery is about. Someone went to a lot of trouble to get this tape. I better tell h about this, A.S.A.P. Be careful with that. That's my mother's favorite painting. Lieutenant caine. I didn't know you were an art lover. **Speaking of which**, carlos, you and amanda brighton on tape has become an instant classic. Little, dirty girl. that's not meant for public viewing. Yes, blackmail never is, is it? Blackmail? Brighton's never paid me a cent. There are many forms of currency. see, now you're reaching, caine. And I'm not alone. I'm impounding your vehicle. You can't do that. I just did, carlos. Should I call you a cab

図 13 Movies の Tunnel Vision (上) と TV の CSI: Miami (下)

Tunnel Vision の実際の映像では検索結果にある発話を確認できないが、CSI: Miami の映像には同じ発話を確認できるため、前者の字幕として後者の字幕が Movies に収録されていることになる。後者のエピソード名が Tunnel Vision であることから、OpenSubtitles 上に前者の字幕として後者の字幕が誤ってアップロードされることが原因であると思われる。その他にも Movies には出典情報の間違いが数多く含まれている (表 3 参照)。

表3 検索結果における出典情報の誤り(検索結果A)

	誤	正
16	Doomsday Man (2000)	Superman: Doomsday (2005)
32	The Day I Ran into All ... (1998)	My Boyfriend is Type B (2005)
39	When He Didn't Come ... (1998)	Wonder Woman (1978); s02e15
42	Chasing Andy (1998)	Chasing Amy (1997)
47	UFC 15: Collision Course (1997)	CSI: Miami (2006); s04e17
63	David Copperfield: ... (1994)	The O.C. (2005); s02e15
91	Tunnel Vision (1976)	CSI: Miami (2008); s01e17
95	A Question of Identity: ... (1966)	CSI: Miami (2005); s06e16
96	Specials for United ... (1963)	Wonder Woman (1979); s03e15
101	\$21 a Day--(Once a Month) (1941)	MacGyver (1987); s02e21
109	Breaking Point (2009)	Impact Point (2008)
128	10 Days to a New York ... (2007)	CSI: Crime Scene ... (2005); s02e10
158	Big Lebowski: ... (2005)	The Big Lebowski (1998)
160	NTSB: The Crush of ... (2004)	The O.C. (2005); s03e03
168	Polly World (2004)	Polly World (2006)
169	Family Sins (2004)	MacGyver (1987); s02e12
172	The Dog Days of Summer (2004)	Dog Days of Summer (2007)
174	The World Man Does Las ... (2004)	Las Vegas (2007); s05e05
192	Out in the Cold (2002)	MacGyver (1987); s02e16
283	Take My Wife (2009)	Punchline (2016); s01e02
298	Jack Turner and the ... (2008)	Tales from the ... (1991); s03e07
396	Detention (2012)	Detention (2011)
430	Bull (2016)	Bull (2016); s01e01
431	Bull (2016)	Bull (2016); s01e01
495	Agatha Christie's ... (1987)	Agatha Christie's ... (2010); s03e01
496	A Simple Man (1987)	CSI: Miami (2008); s01e17
519	Black Gold (2011)	Day of the Falcon (aka Black Gold) (2011)
529	3 Blind Mice (2003)	Three Blind Mice (2008)
535	Circus of the Stars and ... (1992)	The O.C. (2005); s02e17
537	Circus of the Stars #11 (1986)	The O.C. (2005); s02e11

表3における30件中、網かけ部分の10作品は、正しい出典情報でも映画作品が出典であるため、出典を修正した上でMoviesのデータとして扱うことができる。その他20作品はテレビドラマが出典である。本来であればTVの方に収録されるデータであるため、検索結果からは除外する。MoviesとTVを合わせて使う場合には、各データの正確な出典情報を確認しないと重複例が生じるため、注意が必要である。

第五に、検索結果には英語以外の言語による映画の英語字幕も含まれている。That Girl in Pinafore (2013)は中国語の映画であり、Bakuto Gaijin Butai (1971)とDetective Conan: The Time Bomber Skyscraper (1997)は日本語の映画である。また、上記表3で正しい出典情報として挙げたMy Boyfriend is Type B (2005)は韓国語の映画である。異なる言語の英語字幕の場合、翻訳の質などの問題も考慮する必要があるため、本論では上記4例を該当例から除外する。

以上を踏まえると、speaking of whichの検索の場合、①重複例(5例)、②言い直しの事例(1例)、③実際の発話では当該表現が使用されていない事例(13例)、④映画作品が出典ではない事例(20例)、⑤英語以外の言語の英語字幕の事例(4例)の合計43例(全体の約8.0%)が除外対象となる。これらを漏れなく発見するには、実際の映像を一つ一つ確認するしかない。しかし、例えば、検索結果の頻度が数十万規模になる場合には、全データの映像を確認することは実質的に不可能である。その場合、500例程度を無作為に抽出し、それらの映像を確認した上で使用するという方法が考えられる。

4.3. “speaking of which” により検索されないもの

“speaking of which” という文字列検索の検索結果として表示されるデータだけが speaking of which の該当例であるとは限らない。使用頻度が少ない場合には 1 例毎の重みが大きくなるため、該当例を可能な限り多く抽出する必要がある。ここではどのような事例が上記の文字列検索の検索対象から外れてしまうのかを示す。第一に、“speaking of” という文字列検索を実行した場合、検索結果には図 14-15 に示す 5 例が含まれている。

366	1958	US/CA	Imitation General	🔍	🔍	🔍	as stupid as you and that moron buddy of yours... Tsk tsk tsk. Speaking of whi... [Door_opening] Where is that dear boy? Ain't you two together
1059	2002	US/CA	Bang Bang You're Dead	🔍	🔍	🔍	should just be quiet. - Mr. Duncan makes an important point. - Speaking of wich , what about this " Bang, bang, I'm going to

図 14 “speaking of” の検索結果①

958	2001	US/CA	Thirteen Conversatio...	🔍	🔍	🔍	, also happens a way to me of being name Father of the Year. Speaking of that... guesses with whom I ran into. the world is a handkerchief
1452	2008	US/CA	Another Gay Sequel: ...	🔍	🔍	🔍	Golf Well, that have a great summer sir, Muffster. What the hell Speaking of that , do not have to take the plane I guess not want to
1808	2015	US/CA	Entourage	🔍	🔍	🔍	about to have a baby, asshole. I thought you were Sloan. - Speaking of , you were in my dream last night. - I hope I was

図 15 “speaking of” の検索結果②

図 14 の Imitation General (1958) における speaking of whi... は発話の途中で突然扉が開き、発話が遮られた事例である。一方、Bang Bang You're Dead (2002) の speaking of wich は which の誤字である。図 15 における 3 例は speaking of which の検索結果ではないが、実際の映像では speaking of which が使用されている。以上の 5 例は speaking of which の該当例として扱うことにする。

第二に、“of which” という文字列検索を行った場合、検索結果には図 16-18 に示す事例が含まれている。

372	1969	US/CA	Alice's Restaurant	🔍	🔍	🔍	gon na live in it. Hey, knock it off. Speakin' of which , can I crash here tonight? Yeah. Oh, yeah. - We
585	1990	US/CA	Postcards from the Edge	🔍	🔍	🔍	on, back to one. Into position, come on. Speakin' of which , do I have time to go to the men's room before we go
753	2000	US/CA	The Tigger Movie	🔍	🔍	🔍	hoo, hoo! It's all right in my letter. Speakin' of which , you wan na hear me read it once or thrice again? - Uh
806	2004	US/CA	Home on the Range	🔍	🔍	🔍	Well, it's no use cryin' over spilled milk. Speakin' of which , that's me. I'm the cow. Yeah, they're real
1122	2008	US/CA	Bachelor Party 2: Th...	🔍	🔍	🔍	for my new bro. - [All] right. - Mmm. Speakin' of which , have you ever had two girls at once? Uh, actually, sadly
1134	2007	US/CA	The Killing Floor	🔍	🔍	🔍	is a mistake. One which I will correct in time. Speakin' of which , yours is up. Hello. Man You really think that cop's gon

図 16 “of which” の検索結果①

1267	2008	US/CA	Misconceptions	🔍	🔍	🔍	comprehend what I'm about to do. Well, yeah, speaking' of which , how is your own " one man, one woman " marriage doin'
1755	2013	US/CA	Dead in Tombstone	🔍	🔍	🔍	n't handle their affairs like a couple of men. And, speaking' of which , a gang of gunfighters just rode into town. Reckon they're headed for
2374	1998	UK/IE	Little Voice	🔍	🔍	🔍	day without some dribbling' fat, can I? Eh, speaking' of which , I can call Sadie on me new instrument. That will freak her flabby

図 17 “of which” の検索結果②

1303	2014	US/CA	Dracula Untold	🔍	🔍	🔍	Q	u talked like us, prayed like us, fought like us... speaking of which... I am owed 1,000 boys. Why have I not seen them? Mehmed
1334	2009	US/CA	The Trotsky	🔍	🔍	🔍	Q	Uh, and I figured since you used to be married to Commissioner Archambault-Speaking of which, I saw the pictures from that wedding. - Hey. Leon. -
2782	1991	Misc	Millions	🔍	🔍	🔍	Q	them. What'll do you do then, take it back? #Speaking of which... Right mystery activity... champagne. #... follow me. A woman opens

図 18 “of which” の検索結果③

図 16 における 6 例は speaking に縮約が起きた事例である。図 17 における 3 例は speaking' という誤字の事例である。図 18 における Dracula Untold (2014) の事例は speaking の n が落ちた事例であり、The Trotsky (2009) と Millions (1991) の事例は speaking の直前のスペースが落ちた事例である。以上の 12 例も speaking of which の該当例に含められる。^{10) 11)}

第三に、“ofwhich” および“speakingofwhich” という文字列検索を実行すると、それぞれ図 19-20 に示すような事例を発見できる。

2	2009	US/CA	Across the Hall	🔍	🔍	🔍	Q	game at Rucker Park. You don't need to tell me that. Speaking ofwhich, we called your ass. What, you can't call people back?
3	2008	US/CA	One, Two, Many	🔍	🔍	🔍	Q	That's funny - - a guy who made love to his dog. Speaking ofwhich, my dog passed away when I was seven. Oh, really? I
11	1998	US/CA	Firestorm	🔍	🔍	🔍	Q	I'm goin' to the Baia, do some fishin'. Speakin' ofwhich, I was wasting my time up on the Chilkoote yesterday, where it runs
16	1987	US/CA	Cold Steel	🔍	🔍	🔍	Q	He probably stopped somewhere... to pick up some eggnog for the brandy. Speaking ofwhich, you smell like a distilley. We had a little party at the office

図 19 “ofwhich” の検索結果

1	2012	US/CA	Hayride	🔍	🔍	🔍	Q	KIND OF A TOUCHY SUBJECT. IT'S STILL KIND OF A TOUCHY SUBJECT. SPEAKINGOFWHICH, WHEREIS SUBJECT. SPEAKINGOFWHICH, WHEREIS YOUR OLD MAN? SPEAKINGOF
2	2012	US/CA	Hayride	🔍	🔍	🔍	Q	. IT'S STILL KIND OF A TOUCHY SUBJECT. SPEAKINGOFWHICH, WHEREIS SUBJECT. SPEAKINGOFWHICH, WHEREIS YOUR OLD MAN? SPEAKINGOFWHICH, WHEREIS YOUR OLD
3	2012	US/CA	Hayride	🔍	🔍	🔍	Q	TOUCHY SUBJECT. SPEAKINGOFWHICH, WHEREIS SUBJECT. SPEAKINGOFWHICH, WHEREIS YOUR OLD MAN? SPEAKINGOFWHICH, WHEREIS YOUR OLD MAN? HE'SUPAT THEE

図 20 “speakingofwhich” の検索結果

図 19-20 の 7 例は、字幕ファイルの質に問題があり、speaking/of/which の各語を分かち書きするスペースが落ちている事例である。図 20 の 3 例中の 2 例は 1 例目の重複例であるため除外する必要がある。その他 5 例は、字幕作成の過程で発生した間違いであるため、speaking of which の該当例に含められる。

以上をまとめると、Movies には“speaking of which” の文字列検索では抽出できない当該表現の事例として、①発話の途中で遮られている事例 (1 例)、②他の言語表現の検索結果であっても実際の発話では speaking of which が使用されている場合 (3 例)、③縮約が起きている事例 (6 例)、④誤字・脱字の事例 (12 例) が含まれている。これらの 22 例は該当例として追加可能である。

対象外となったデータの把握に加えて、その他の当該表現に可能な形式が使用するコーパスに含まれていないことの確認も大切である。当該表現に起こり得る縮約の可能性の一つとして、speakin' o' which という形式も考えられる。“speakin' o' which” という検索を行えば、少なくとも Movies にそのような形式が含まれていないことが分かる。さらに、speaking と of which の間には副詞句が介在する可能性もあるかもしれない。“speaking * of which” という検索を行えば、そのような事例は含まれていないことが分かる。

4.4. 第4節のまとめ

本節では、Movies から *speaking of which* の該当例を抽出する際の注意点を示した。検索インターフェイスに“speaking of which”という文字列検索を実行すると 540 という頻度が表示される。その内の 43 例は余分に検索された事例として除外する必要があり、検索文字列の検索対象から外れている事例として 22 例を追加できる。最終的な当該表現の使用頻度は 519 例となる。540 と 519 という数値はそれほど違わないようにも見えるが、除外された事例と追加された事例があるため、その中身は大きく異なる。データの中身が変わると分析結果も変わる可能性がある。例えば、非該当例の除外を行わない場合、Movies 内の当該表現の初出は \$21 a Day—(Once a Month) となる。しかし、4.2 節で示したように、\$21 a Day—(Once a Month) の事例の出典情報には誤りがあり、実際には 1941 年ではなく 1987 年の使用例であった(表 3 参照)。なお、本節で指摘した注意点は、Movies を使用する場合に限られるものではなく、別の言語現象を調べる場合や English-Corpora.org の別コーパスを使用する際にも問題になる。¹²⁾ 検索インターフェイスの使用を避け、購入版を使用すれば解決するわけでもない。

5. おわりに




代表的な英語のコーパスである通称 Brown Corpus の場合、コーパスの構成に関する詳細なマニュアルが公開されており、それを読めばデータ抽出時の注意点をある程度把握可能である (Francis & Kucera, 1979)。しかし、English-Corpora.org で公開されているコーパスの場合、直感的に使用可能な検索インターフェイスが既に整備されているためか、Brown Corpus に用意されていたようなマニュアルは公開されていない。“English-Corpora.org: a guide tour” (<https://www.english-corpora.org/pdf/english-corpora.pdf>) という文書が 2020 年に公開されたが、そこに記載されている情報はどのような検索が可能なのかといった内容に限られており、本論で指摘したような注意点は示されていない。コーパス利用の注意点に関する情報共有が十分に進んでいない中でコーパスを使用した言語研究が増加すると、誤った使い方をする研究がこれまでよりも増加することが予想される。そのような状況に鑑み、本論では English-Corpora.org で公開されている Movies を取り上げ、その言語研究資料としての可能性を明確にした上で、利用時における注意点を示した。

本論で指摘した注意点は、一部のコーパスに関する瑣末な問題ではないかと考える読者もいるかもしれないが、(他の言語によるコーパスも含め) あらゆるコーパスに共通する問題である。コーパスから得られたデータを使い、仮説の構築と検証を行う場合、主張の根拠となるデータが実際には存在していない可能性がある。また、コーパスから抽出した数値を使い、定量的な研究を行う場合には、検索インターフェイスから機械的に出力された数値を盲信してしまうと、言語事実と合わない分析結果を導いてしまう危険性がある。

コーパス利用時の注意点は利用前に網羅的に把握できるものではなく、最終的には使用者側が注意点を察知する感覚を身に付ける必要がある。本論で行ってきたような、具体的な注意点とその原因の指摘があれば、そのような感覚を身に付ける際の足掛かりになるだろう。それらの情報を共有していくことが、コーパスを利用した言語研究における現状の改善に有効であると思われる。

- * 本稿における第四節は、山内 (2021) におけるデータ収集の概要部分の一部を抜粋し、大幅な修正を加えたものである。本稿の執筆に際し、査読委員の先生方を含め、数多くの方々に貴重なコメントをいただきました。特に大名力先生には、山内 (2021) の段階から、貴重なご指摘とご助言をいただくだけでなく、本稿の執筆に際しても継続的にご指導をいただきました。この場をお借りし、全ての方々に心より感謝の意を申し上げます。本稿における不備や誤りは全て筆者の責任によります。

注

- (1) 大名 (2012) は、基礎データの信頼性に問題が生じるケースを取り上げ、その原因を詳細に検討している。大名力氏によれば、大名 (2012) における指摘は、実際の論文名等に言及していない場合でも、現実起きた問題を踏まえて書かれたものがほとんどである (私信)。
- (2) **Movies** が公開されているサイト (<https://www.english-corpora.org/movies/>) には、年代と制作地域ごとの収録語数が記載されている。また、収録された各データの出典や語数などが明記されたファイルも公開されている (https://www.english-corpora.org/movies/files/sources_movies.zip)。上記のサイト上で示されている総語数は 199,479,302 語である。しかし、公開されているファイルから総語数を計算すると 195,717,877 語になり、199,479,302 語と一致しない (2021 年 5 月 12 日時点)。
- (3) 滝沢 (2021) は **English-Corpora.org** における **COCA** の品詞タグに誤りがあることを指摘している。**Movies** においても品詞タグを利用する場合には注意する必要がある。
- (4) 各データの作品名をクリックすると **IMDb.com** の該当ページにアクセスできる。**CONTEXT** に表示されている「」は **Google** 翻訳のサイトを利用した各データの音声読み上げ機能である。「」をクリックすると **Google** 翻訳のサイト上で検索結果の他言語における翻訳を表示できる。「」をクリックすると各データを構成する各語の意味や傾向等を確認できる。
- (5) 検索結果 A と検索結果 B がどのような条件下で出現するのかは明確ではない (2021 年 10 月 1 日時点)。少なくとも筆者の環境では、検索結果 A が表示された場合に、検索をやり直すと検索結果 B が表示されるようになる。
- (6) 検索結果が二種類ある場合には、両方の検索結果を **Excel** の同じシートにコピーし、**Sort & Filter** の機能と **EXACT** 関数を活用すれば、表示位置の変動するデータを機械的に発見できる。
- (7) 検索結果には **Kill Bill: The Whole Bloody Affair** (2011) を出典とするデータが含まれているが、同作品は **Kill Bill: Vol. 1** (2003) と **Kill Bill: Vol. 2** (2004) から構成される。両作品の字幕は **Movies** に収録されていないため、重複例は発生していない。公開年度が変わる点には注意する必要がある。
- (8) **Stag Night** (2008) の場合、実際の映像では **speaking of which** に対応する表現が使用されていない。
- (9) 絶版等の様々な事情により映像を **DVD** 等の媒体で入手困難な場合もある。本論の場合、**Certainly** (2011) と **Billy** (2011) は映像媒体を入手することができない。**Garfield's Pet Force** (2009) は映像媒体を入手できたが、**speaking of which** の使用場面がカットされていた。便宜的に、これらの作品の事例は実際の発話でも当該表現が使用されていると見なし、該当例として扱うことにする。
- (10) **Millions** (1991) を出典とする事例の正しい出典情報は **Millions** (2004) である。両作品は同名の異なる作品である。
- (11) ある表現に起き得る誤字・脱字のパターンを前もって網羅的に予測することは不可能であるが、コーパスに含まれる誤字・脱字のパターンを部分的に把握することは可能である。例えば、“w*h”や“*f”という検索を実行すれば、**wich** や **off** (**of** の **f** が重複したもの) といった誤字・脱字を発見でき、使用した文字列検索の網羅性を確認できる。
- (12) **\$21 a Day—(Once a Month)** の事例は **The Corpus of Historical American English (COHA)** にも誤った出典情報で収録されている (2021 年 10 月 1 日時点)。COHA は 2021 年にアップデートされ、①**Movies** と **TV** のテキストの一部を追加、②2010–2019 年のテキストを追加、③1810–1819 年のテキストの除外、④テキストの修正 (Alatrash et al., 2020)、⑤メタデータの修正、⑥重複例の除外という変更が施された (<https://www.english-corpora.org/coha/>)。①の変更により、**Movies** における問題が **COHA** にも引き継がれている。

参考文献

- Alatrash, R., Schlechtweg, D., Kuhn, J., & Schulte im Walde, S. (2020). CCOHA: Clean Corpus of Historical American English. *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, 6958–6966. <https://aclanthology.org/2020.lrec-1.859>
- Alvarez-Pereyre, M. (2011). Using film as linguistic specimen: Theoretical and practical issues. In R. Piazza, M. Bednarek, & F. Rossi (Eds.), *Telecinematic discourse: Approaches to the language of films and television series* (pp. 47–67). John Benjamins. <https://doi.org/10.1075/pbns.211.05alv>
- Bednarek, M., & Zago, R. (2021). Bibliography of linguistic research on fictional (narrative, scripted) television series and films/movies, version 4 (January 2021). <https://unico.academia.edu/RaffaeleZago>
- Cambridge University Press. (n.d.). Speak. In *Cambridge dictionary*. Retrieved December 15, 2021, from <https://dictionary.cambridge.org/us/dictionary/english/speak>
- Crystal, D. (1987). *The Cambridge encyclopedia of language*. Cambridge University Press.
- Davies, M. (2004). *British National Corpus (from Oxford University Press)*. <https://www.english-corpora.org/bnc/>
- Davies, M. (2008). *The Corpus of Contemporary American English (COCA)*. <https://www.english-corpora.org/coca/>
- Davies, M. (2010). *The Corpus of Historical American English (COHA)*. <https://www.english-corpora.org/coha/>
- Davies, M. (2019). *The TV Corpus*. <https://www.english-corpora.org/tv/>
- Davies, M. (2019). *The Movie Corpus*. <https://www.english-corpora.org/movies/>
- Davies, M. (2021). The TV and Movies corpora: Design, construction, and use. *International Journal of Corpus Linguistics*, 26(1), 10–37. <https://doi.org/10.1075/ijcl.00035.dav>
- Francis, W. N., & Kucera, H. (1979). *Brown Corpus manual: Manual of information to accompany A Standard Corpus of Present-Day Edited American English for use with digital computers*. <http://licame.uib.no/brown/bcm.html>
- 金水敏. (2000). 「役割語探求の提案」佐藤喜代治 (編) 『国語史の新視点』(国語論究第 8 集) (pp. 311–351) 明治書院.
- Lucey, P. (1996). *Story sense: Writing story and scripts for feature films and television*. McGraw-Hill.
- Mattsson, J. (2009). *Subtitling of discourse particles: A corpus-based study of well, you know, I mean, and like, and their Swedish translations in ten American films* [Doctoral dissertation, University of Gothenburg]. Gothenburg University Publications Electronic Archive. https://gupea.ub.gu.se/bitstream/2077/21007/1/gupea_2077_21007_1.pdf
- Mitchell, J. G. (2015). Ain't no *Bones* about it: Dialect discrimination in primetime. In P. Donaher & S. Katz (Eds.), *Ain'thology: The history and life of a taboo word* (pp. 298–322). Cambridge Scholars Publishing.
- 大名力. (2012). 「コーパス利用の落とし穴」堀正広 (編) 『これからのコロケーション研究』 (pp. 227–264) ひつじ書房.
- Quaglio, P. (2008). Television dialogue and natural conversation: Linguistic similarities and functional differences. In A. Ädel & R. Reppen (Eds.), *Corpora and discourse: The challenges of different settings* (pp. 189–210). John Benjamins. <https://>

doi.org/10.1075/scl.31.12qua

Sinclair, J. M. (1991). *Corpus, concordance, collocation*. Oxford University Press.

滝沢直宏. (2021). 「ly 副詞の網羅的記述研究に向けたデータの処理方法: COCA (full text) の処理に関する覚え書き」『現代英語談話会論集』 16, 1-9.

Wray, A. (2008). *Formulaic language: Pushing the boundaries*. Oxford University Press.

山内昇. (2020). 「Speaking of which の構文化分析再考」『英語語法文法研究』 27, 103-118.

山内昇. (2021). 「字幕翻訳における談話標識の翻訳ストラテジーに関する語用論的研究: Speaking of which を事例として」『日本語用論学会第 23 回大会発表論文集』 16, 219-222.