

2017/10/31

コーパスから探る英語話者の世界の見方

1

講座: コーパスから見えることば・社会・文化 コーパスから探る英語話者の世界の見方

東京外国語大学総合国際学研究院
大谷直輝

2017/10/31

東京外国語大学 体験授業

2

自己紹介

- 講師: 大谷直輝
- 所属: 東京外国語大学大学院 総合国際学研究院
- 職階: 講師
- 専門: 認知言語学、コーパス言語学
- 関心があること:
 - 「ことばを通して見えてくる人間の心のありようと、心を通して見えてくる言葉の働きの背後にある原理」
 - 「上り坂」と「下り坂」の数ほどどちらが多いか。
 - The highway goes from Tokyo to Osaka.
 - The bird is a land bird./The bird is a ground bird.

2017/10/31

コーパスから探る英語話者の世界の見方

3

本日の内容

1. はじめに
2. コーパス研究の背景
 - 言葉を通して心を見る(認知言語学の考え方)
3. コーパスの種類と特徴
4. ウェブで使えるコーパスの紹介
 - (i) COCA, (ii) TCSE, (iii) Google Books Ngram Viewer
5. コーパスを用いた研究の紹介
 - BNCの調査
6. おわりに

2017/10/31

コーパスから探る英語話者の世界の見方

4

1. はじめに

- 本発表の目的
 - 人間の世界の捉え方には人間が所属する文化や社会の影響が色濃く反映されている。そこで、大量の言語データが収集されている英語のコーパスを用いることで、(i) 英語の機能や構造、(ii) 英語の話者の心のありよう、(iii) 話者を困む文化や社会に対する理解を深める。
- 本発表の背景
 - コーパスには、現実に使用されるありのままのことばの姿が現れる。そのため、コーパスに収集されている言語データは、言葉の使用者である私たち人間が世界をどのようなものとして捉えているかの手掛かりになる。
 - コーパスに現れることばの姿は、学校で教えられる規範的な文法とはかなり異なることもある。

2017/10/31

コーパスから探る英語話者の世界の見方

5

1. はじめに

- 身近に存在するコーパス
 - コーパスは電子化された言語データの総体
 - 電子化された言語データは私たちの身近にある。
 - 電子書籍や電子新聞
 - SNS
 - ウェブ
 - 現代はこれまでにないほど、大量かつ多種多様な言語データに短時間で、アクセスすることができる時代。
- ウェブをコーパスとして使用する
 - Google Books Ngram Viewer (<https://books.google.com/ngrams>)
 1. big, small, large, little
 2. flight attendant, stewardess
 3. although, though

2017/10/31

コーパスから探る英語話者の世界の見方

6

2. コーパス研究の背景

- 問 なぜ私たちはコーパスを見るのか。
 - コーパスにはどのような情報が載っているのか。
 - コーパスからの情報はどのようなことを実証できるか。
 - コーパスを通して、言葉を用いる人間の心のありようや、人間を取り囲み言語の基盤となる文化や社会に対する理解を深められる。
- 確認が必要な点
 1. 人間は世界をありのままに認識するのではなく、主観的に捉える。
 2. 言語には人間による世界の認識の仕方が反映されている。
 3. 認識の仕方の違いは、各言語に色濃く反映されている。
- 背景となる言語観: 世界⇒人間(知覚+概念化)⇒言語
- コーパス言語学的手法: 言語⇒人間(知覚+概念化)⇒世界

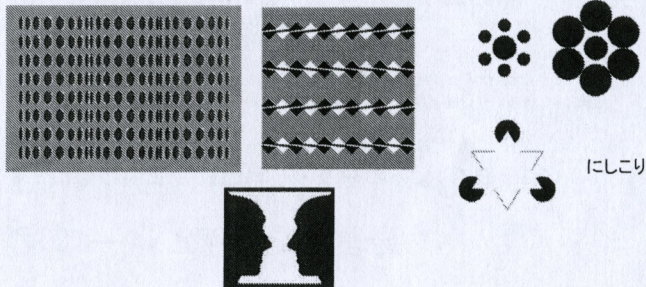
2017/10/31

コーパスから探る英語話者の世界の見方

7

2. コーパス研究の背景

1. 人間は世界を主観的に捉える。
 - 人間が認識した世界はありのままの客観的な世界ではなく、人間にとっての世界である。



2017/10/31

コーパスから探る英語話者の世界の見方

8

2. コーパス研究の背景

2. 言語には人間による世界の認識の仕方が反映されている。
 - 問題: 次の表現には人間による世界の認識がどのように表れているか。
 - (1a) The glass is half empty.
 - (1b) The glass is half full.
 - 外部世界のどこに注目するかで言語表現は異なる。
 - (2a) This highway goes from Tokyo to Osaka.
 - (2b) The mountain range runs from Canada to Mexico.
 - 静的な状態を動きがあるものとしてとらえている。
 - (3a) The bird spends its life on the land.
 - (3b) The bird spends its life on the ground.
 - 指示対象を背景的な七木と関連付けて記憶する。
 - (4a) John broke the vase (with the hammer).
 - (4b) The hammer broke the vase.
 - (4c) The vase broke.
 - 連続する事態のどこに注目するかで言語表現が異なる。

2. コーパス研究の背景

3. 認識の仕方の違いは、各言語に色濃く反映されている。

- 問題: How many fingers do you have?
 - finger: one of the four long thin parts on your hand, not including thumb
 - toe: one of five fingers movable parts at the end of your foot
- 日英語では<手の指>の切り分け方に違いがある。
- <手>と<指>を区別する点は日英語共通
- 世界の言語には<手>と<指>を区別せず同じ名前で呼ぶ言語もある。
- さらに、<手>と<腕>を区別せず同じ名前で呼ぶ言語の数はさらに多い。
- 一見すると、<手>と<腕>をしないなんて想像できないかもしれない。
- しかし、日本語ではlegとfootを区別せず、「足」と呼ぶ。

手				足	
finger	指	語1	語3	toe	指
thumb				foot	足
hand	手		語4	leg	
arm	腕	語2		英語	日本語
英語	日本語	言語α	言語β		

言語によって事態の切り取り方は異なる。

2. コーパス研究の背景

- 背景となる言語観:
 - 世界⇒人間(知覚+概念化)⇒言語
 - 人間は世界をありのままに認識するのではなく、主観的に捉えるため、言語には人間による世界の認識の仕方が反映されている。各言語には、言語に固有の文化や社会に根差した認識の仕方が現れる。
- コーパス言語学の手法:
 - 言語⇒人間(知覚+概念化)⇒世界
 - コーパスに収集されている言語データは、言葉の使用者である私たち人間が世界をどのようなものとして捉えているかが現れるので、コーパスを用いて言語の使用者である人間の心のありようを探る。

3. コーパスの種類と特徴

- コーパスの種類
 - コーパスは様々な目的や用途に応じて編纂される
 - 汎用コーパス/特殊目的コーパス
 - 総合的な目的か特定の言語研究のために編纂されるか
 - 共時コーパスと通時コーパス
 - 同時代の言語データか複数の時代区分からデータを使用するか。
 - 話し言葉コーパス/書き言葉コーパス
 - 伝達手段に応じた区別
 - ウェブコーパス
 - ウェブ上に撮りこむかに関する区別
 - 学習者コーパス
 - 学習者の負例を含むコーパス
 - タグ付きコーパス
 - 品詞や文法関係等の情報を含んだコーパス

3. コーパスの種類と特徴

- コーパスのデータとしての特徴
 - 電子化された言語データが収集されている。
 - 大量の言語データが収集されている。
 - 産出された言語データが収集されている。
 - 談話を構成する言語データが収録されている。(つまり、収集された言語データは独立した単文の集合ではない)。
 - このような特徴を利用して言語に関するどのようなことがわかるか。

3. コーパスの種類と特徴

1. 電子化された言語データが収集されている。

- 電子化され数値として扱えるため、コーパス内の語句の分布状況や使用頻度を定量的に表すことができる。
- 用例の検索や収集、コロケーションの調査等が容易に行える。

2. 大量の言語データが収集されている。

- コーパスに収集されている大量の言語データにおける言語分布を調べることで、人間が持っている言語知識のありようを調べられる(現在のコーパス言語学において最も注目されている特徴)。

3. コーパスの種類と特徴

3. 産出された言語データが収集されている。

- 収集されている用例が作例ではなく実例であることから、コーパスを用いることで、言語の使用実態に基づいた一般化を行うことが可能となる。

4. 談話を構成する言語データが収録されている(収集された言語データは独立した単文の集合ではない)。

- コーパスに収集された言語データは、単文の集合ではなく、談話の一部を構成する。そのためコーパス内の各文は、文の内部で働く意味論的な機能と同時に、単文を超えて、談話内部で働く語用論的な機能を持つ。
 - 次の文の意味の違いは何か。
 - (a) John lost his wallet. (b) As for John, he lost his wallet.
 - (c) What John did was lose his wallet. (d) It was John who lost his wallet.

4. ウェブで使えるコーパスの紹介

• オンラインで使用できる無料コーパス

- COCA (Corpus of Contemporary American English)
 - <https://corpus.byu.edu/coca/>
 - アメリカ英語5億語からなる均衡コーパス
 - 言語使用域や類義語の比較に便利
- TCSE (TED Corpus Search Engine)
 - <https://yohasebe.com/tcse/>
 - TED Talkのコーパス
 - 話し言葉やスピーチに特化した特徴が見つかる。
- Google Books Ngram Viewer
 - <https://books.google.com/ngrams>
 - 語彙の歴史的な変遷の調査が行える。

4.1 COCA

• COCAを用いた言語調査

- 検索の種類
 1. 語や句の検索
 - mysterious, "break", more * than, rough NOUN
 2. 言語使用域別の語や句の検索
 - you know, I know, she knows
- コロケーションの検索
 - look into
- 語と語の振る舞いの比較
 - warm and hot, small and little, boy and girl.

4.2 TCSE

・ TCSEを用いた言語調査

- ・ TCSE: TED Corpus Search Engine (TCSE) は TED (<https://ted.com>) が公開している 2,400 件 以上のプレゼンテーションの英語トランスクリプトと多言語翻訳テキストを解析してデータベースに格納し、全文検索を可能にしたシステム
- ・ 検索語からTEDの該当場面への移動
 - ・ Tokyo, by the way
- ・ 日本語からの翻訳を用いた検索が可能
 - ・ 人工知能, もしかして
- ・ 英語学習にも使える

4.3 Google Books Ngram Viewer

・ Google Books Ngram Viewerを用いた言語調査

- ・ 語の使用頻度を通して、世界の情勢を知る。
 - ・ male, female, man, woman
 - ・ disabled, challenged, handicapped, crippled
 - ・ bisexual, gay, lesbian, transgender
 - ・ green house effect
- ・ アメリカ英語とイギリス英語の違いが分かる。
 - ・ favor, favour, favourite, favorite
 - ・ center, centre

5. コーパス研究の紹介

・ 使用するコーパス

- ・ British national Corpus
- ・ 1億語のコーパス (書き言葉9割、話し言葉1割)

・ 分析する対象

1. arrive {in/at/on}
2. SVCの {go/come/fall/turn/grow}
3. 類義的な句動詞 {burn up/burn down}

5.1 事例研究1: arrive in

・ 問 arrive {at, in on}に意味の違いはあるか。

	arrive at (5202)		arrive in (1856)		arrive on (394)	
1	house	55	London	73	scene	69
2	airport	49	Britain	58	time	40
3	station	35	England	56	earth	8
4	scene	30	time	40	island	8
5	hospital	28	Paris	30	morning	7
6	destination	25	country	27	site	7
7	hotel	25	New	26	doorstep	6
8	office	25	city	19	foot	6
9	school	21	France	19	September	6
10	figure	20	America	18	Sunday	6
11	Heathrow	19	York	17	desk	5
12	point	19	Hong	16	may	5
13	work	19	Kong	16	Monday	5
14	conclusion	17	town	16	planet	5
15	decision	17	Moscow	14	stage	5

2017/10/31 コーパスから探る英語話者の世界の見方 21

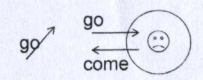
5.2 事例研究2: SVC

- 問2 英語の第二文型
 - SVCで用いられる {go/come/fall/turn/grow}
- (a) The science experiment **went wrong**.
- (b) His dream **came true**.
- (c) The baby **fell asleep**.
- (d) The wine **turned sour**.
- (e) I **grew strong**.

これらの動詞は、動詞の後に現れる形容詞が主語の結果状態を表す。
 BNCは品詞情報が付いているコーパスなので、5つの動詞go, come, fell, turn, growに形容詞が続く形を検索できる。

2017/10/31 コーパスから探る英語話者の世界の見方 22

5.2 事例研究2: SVC



- “S go C”と“S come C”におけるCの特徴

	go (200,562)		come (144,685)	
1	wrong	1,550	true	314
2	mad	383	alive	173
3	bust	162	complete	120
4	bankrupt	147	clean	106
5	unnoticed	130	close	96
6	crazy	115	unstuck	75
7	wild	109	cheap	69
8	cold	99	full	50
9	berserk	89	good	48
10	white	78	free	47
11	free	76	loose	35
12	straight	76	thick	31
13	hungry	72	easy	29
14	left	71	clear	27
15	red	71	early	16

2017/10/31 コーパスから探る英語話者の世界の見方 23

5.2 事例研究2: SVC

- “S {fall/grow/turn} C”におけるCの特徴

	fall (26275)		grow (18228)		turn (43928)	
1	asleep	644	old	350	sour	78
2	short	301	strong	99	professional	55
3	silent	221	big	82	white	51
4	foul	156	large	77	red	50
5	ill	131	tired	65	blue	39
6	due	88	loud	64	green	39
7	open	56	accustomed	56	black	37
8	vacant	35	tall	45	cold	35
9	flat	31	cold	37	nasty	34
10	sick	31	hot	36	full	32
11	dead	17	rich	36	pale	32
12	pregnant	14	fond	33	brown	27
13	free	11	small	33	pink	23
14	fast	7	bad	32	grey	21
15	full	7	warm	29	yellow	21

2017/10/31 コーパスから探る英語話者の世界の見方 24

5.3 事例研究3: burn up/burn down

- 類義的な句動詞 {burn up/burn down}
 - (1) (a) He burned up his house. [他動詞: VPO型]
 - (1) (b) He burned his house up. [他動詞: VOP型]
 - (2) (a) The fire burnt up. [自動詞(VP型): 主体主語]
 - (2) (b) The house burned down. [自動詞(VP型): 対象主語]
 - (3) The house was burned down. [受動態]
- 動詞burnはBNC全体で5400例その中で、upかdown現れるのは460例

	先行文脈	検索語	後続文脈	前置詞	文型	下位分類
1	... the sun will	burn	up their fuel much more rapidly.	up	他動詞	VPO型
2	... some hooligans	burnt	the house down and ...	down	他動詞	VOP型
3	The fire	burnt	up without much trouble.	up	自動詞	主体主語型
4	But the old pub had	burnt	down in the 1960s and ...	down	自動詞	対象主語型
5	Several trees were	burned	down on the Splash premises.	down	受動態	-

5.3 事例研究3

・結果

burn	VPO型	VOP型	VP型	受動態	総計
up	77 (55%)	16 (11%)	38 (27%)	10 (7%)	141 (100%)
down	68 (21%)	47 (15%)	92 (29%)	112 (35%)	319 (100%)
総計	145 (32%)	63 (14%)	130 (28%)	122 (27%)	460 (100%)

表1 burn {up/down}が現れる文法形式

burn	主体主語型	対象主語型	総計
up	25 (66%)	13 (34%)	38 (100%)
down	14 (15%)	78 (85%)	92 (100%)
総計	39 (30%)	91 (70%)	130 (100%)

表2 burn {up/down}の自動詞用法の分類

	コーパスに頻出する文型	⇒	コーパス分析から得られる示唆
burn up	VPO型・主体主語型	⇒	燃やす行為に注目
burn down	受動態、対象主語型	⇒	燃やされる対象に注目

表3 burn {up/down}の特徴

6. 終わりに

・まとめ

1. 人間は事態をありのままに客観的に見るわけではなく、自分なりの視点で見るので、言語表現には人間の認識の仕方が現れる。
2. コーパスから取り出したデータは、認知言語学的な視点を用いることで、言語を話す人間の心のありようや人間が属する文化や社会を映し出す鏡となる。
3. 大量の言語データが収集されているコーパスは、人間による世界の見方を考察するうえで、多くのヒントを与えてくれる。