

日本語の研究と教育に役立つ コーパス調査の方法

砂川 有里子

日本台湾交流協会2018年度第2回日本語教育研修会[特別講演会]
2018年11月13日 台湾大学 / 11月15日 静宜大学

1

コーパスとは

- 実際に使用された言語データを大量に集めて電子化した言語研究のための言語資料

理想の言語ではなく、
現実に使用された言語を
分析する

2

現実の言語の分析例

- つかえや言いよどみにもルールがある。
- 非文法的であるとされた形式も無視できないほど多く使われていることがある。
- 言語の地域的あるいは時代的な多様性を捉えることができる。
- レジスターの違い(e.g.書き言葉と話し言葉)によって文法が異なっている。

3

日本語のコーパス

1960年代から 2000年代から
英語のコーパスに比べ、大きく遅れて構築された。
にわにはにわうらにわにはにわにわとりがいる
にわにはにわうらにわにはにわにわとりがいる
庭には二羽、裏庭には二羽、鶏がいる
学校にはいった

4

国立国語研究所



5

国立国語研究所

コーパスデータベース

<https://www.ninjal.ac.jp/database/>

6

BCCWJ

現代日本語書き言葉均衡コーパス
Balanced Corpus of Contemporary
Written Japanese

書籍、雑誌、新聞、白書、ブログなど
1億430万語のデータ

7

複数のレジスターから
日本語母語話者の書き言葉を
バランス良く集めたコーパス

現代日本語の書き言葉の多様性を把握し
その全体像を捉える

国立国語研究所
2011年に公開

8

BCCWJの用例検索システム

オンライン・無料

タイプ	概要	申込
少納言	文字列検索のみ 用例は500まで	不要
中納言	短単位／長単位検索 すべての用例をダウンロードできる	必要
NINJAL-LWP for BCCWJ	BCCWJの調査結果を示し、 語の特徴的な振る舞いを提示する	不要

9



前文語	外国語	序列	後文語	執筆者
とです。私たちの今までの英語学習は、本当にこの逆だったように思います。まず外國語			を読んで、その後に日本語に訳して理解してきた。英語、英語をマスター	松木 修三(著)
内容をつかむことが困難になっている。■ 20 08年6月29日	国語		国語学	「ベトナム自治地区」チーは中国四川省の一
である。「科」である。ちなみに、上田は「中」	中国語		学」、トハラ学、国語、童学均シ」とすることを「説り」、解説、「国語」と「学」に	安田 敏郎(著)
行く、というと彼女の夢は保育園だったはずがこの前聞いた「英語勉強したいから」	国語		大学近くで買ってきました。こういう、人の意見に流れやすい人はこのままでいい	
国語、を「英語」にすら他の別の仕方をみるとできる。このように基本的には、中国語入力のソフトをインストールする必要はありません。WindowsのSFCの中、中国語は表記による文字体系(すなアラファベット)であるために、中華中韓国語	国語		の表現で、そのまま書くのがむずかしい。昔の歌詞の歌詞がよくわからなくなる	ステイグン・ロジャー・フィッシャー(著) 鈴木 晃(訳)
の歌詞であり、そのまま間違えてしまうから、セタップすれば、すぐ日本語で文書を書くように				
10	国語			

I-JAS

多言語母語の日本語学習者横断コーパス

International Corpus of Japanese as a Second Language

国立国語研究所で構築中の学習者コーパス
2020年3月に完成予定(現在一部公開中)

13

多様なタスク

データの種類	タスク名
発話	ストーリーテリング (2種)
	対話(インタビュー)
	ロールプレイ (2種)
	絵の描写
作文	ストーリーライティング (2種)

16

中納言

前文語	キ一	後文語	キ二	語彙素	語彙素分類	語彙素	語彙素	活用形	活用形	サブ	サブ
等	が	挙げ	られ	る	名詞普通名詞一般	名詞普通名詞一般	名詞普通名詞一般	松	出版・書籍	松	
21	た	い	ほ	じ	「」(1) いじ(1)マレー(1)の(1)家庭に(1)香港(1)で(1)いる(1)といふ内容で(1)	コク	国語	松	出版・書籍	松	
70	た	い	ほ	じ	「」(1) いじ(1)この(1)「国語」(1)の(1)書(1)す(1)意味(1)を(1)明解(1)にする(1)ために(1)国語(1)	コク	国語	川	出版・書籍	川	
85	じ	か	う	じ	「」(1) いじ(1)この(1)「国語」(1)の(1)書(1)す(1)意味(1)を(1)明解(1)にする(1)ために(1)国語(1)	コク	国語	大	出版・書籍	大	

全ての用例がダウンロードできる。

学習者の母語

◆ 多言語母語話者(17ヶ国20地域12言語)

表1 調査対象の12言語		
インドネシア語	タイ語	ベトナム語
韓国語	トルコ語	中国語
ハンガリー語	英語	フランス語
スペイン語	ドイツ語	ロシア語

14

I-JASの特徴

- すべての音声データとテキストデータがダウンロードできる。
- さまざまな比較が可能。
 - 母語話者と学習者
 - 教室環境と自然環境
 - 母語別
 - 海外学習者と国内学習者、など
- 中納言による検索が可能。

17

N L B

中納言

N L B

頻度
統計値

前文語	キ一	後文語	キ二
を飲みます。おなかが	冷え	ているときは、腹巻きをしたり、	
りが、と食つた。からだが	冷え	かけていたから、それと、わざ	
を示す、食べ物の残りが	冷え	て固まった白い血が散らかって	
を帯びたが、それのが	冷え	て消えてしまうと、霍れ自身	
お行が悪くなり、カラダが	冷える	傾向に。できるだけ脚を組んだ	
つて、ますますカラダが	冷える	という寒覚理におちいることも、	
トではますますカラダが	冷え	てしまします。されば、ソック	
床が		0	8.24
やれ		4	5.93
中の		9	7.16
だ		13	8.17
語の特徴的な振る舞いが観察できる。			
BCCWJの調査結果が示される。			
…が冷える 115種類			
コロケーション			
コーパス全体			
頻度 MI LD			
体が冷える	48	8.66	5.04
身体が冷える	15	9.06	5.43
足が冷える	12	8.28	4.65
【人名】が冷える	8	1.31	-2.29
お腹が冷える	8	9.72	6.02
床が	10	8.24	
やれ	4	5.93	
中の	9	7.16	
だ	13	8.17	
語の特徴的な振る舞いが観察できる。	4	4.78	
BCCWJの調査結果が示される。	2.25		
…が冷める	5.71		

2020年完成時の人数

2018年現在660名公開

◆ 大規模(合計1,050人)

協力者の内訳	人数
海外の学習者数	850人
国内の学習者数	100人
	50人
日本国内の日本語母語話者数	50人

15

語彙的自他動詞対の研究

ナロック・パルデシ・赤瀬川(2015)

18

語彙的自他動詞対

形式上コンパクトな ほうが派生元	立つ→立てる
• 移る→移す	tat-u : tat-eru
• 開く→開く	sak-eru : sak-u
• 死ぬ→殺す	utur-u : uru
	裂く→裂ける
	hirak-u : hirak-u
	sin-u : koros-u

19

仮説: 形式的に派生された動詞のほうが頻度が低く、
形式的な派生元となる動詞のほうが頻度が高い。
ナロック・パルデシ・赤瀬川(2015)



語彙的自他動詞対でも成立する

言語は頻度の高いものを形式上コンパクトに表現する
傾向がある。 Zipf (1935)

Zipfの法則

20

他動化型

(自動詞) → (他動詞)

あ	あ
開く	→ 開ける
立つ	→ 立てる

サイズ(小) < サイズ(大)

21

自動化型

(他動詞) → (自動詞)
焼く → 焼ける
裂く → 裂ける

サイズ(小) > サイズ(大)

22

他動化型と自動化型の頻度

統語論的派生パターン	派生元の動詞 (形態的に短い)		派生された動詞 (形態的に長い)	
	頻度	立つ	頻度	立てる
他動化型	182	82%	38	17%
自動化型	103	74%	36	26%
合計	285	74	74	43

23

中国語母語話者の漢字の発音 (非漢字圏学習者との比較)

砂川・黒沢(2017、近刊)

25

同形語と非同形語

- ◆ 同形語:
 - 中国語にも同じ形の漢字単語があるもの
 - 学校, 経済, 政治, 地図, 能力, 歴史
 - 緊張, 専門, 得意, 無理, 試験, 先生, 人間, 勉強
- ◆ 非同形語:
 - 中国語にはその形の漢字単語がないもの
 - 映画, 会社, 授業, 自分, 卒業

電影, 公司, 課業, 自己, 畢業

26

中国人の発話データ
発音の誤用は同形語のほうが多い

正用と誤用の頻度(頻度2以上の漢語名詞)

	正用頻度	誤用頻度	合計
同形語	1,943 (91.4%)	183 (8.6%)	2,126 (100%)
非同形語	965 (94.4%)	57 (5.6%)	1,022 (100%)

砂川・黒沢(2017)

($\chi^2=8.57$; $df=1$; $p<.005$)

27

フランス語母語話者との比較 -正用と誤用の頻度と比率-

母語	フランス語		中国語	
頻度／比率	頻度	比率	頻度	比率
正用	2,735	93.3%	3,115	92.0%
誤用	195	6.7%	271	8.0%
計	2,930	100%	3,386	100%

砂川・黒沢(近刊)

($\chi^2=3.98; df=1; p<.05$)

28

考察

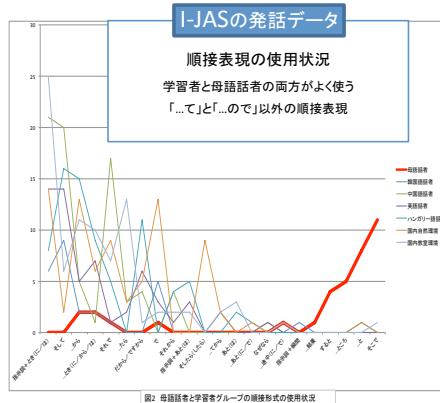
- 中国人学習者は、自然発話をを行う際に、母語である中国語の漢字音が介在してしまう（負の転移）。
- フランス人学習者に比べて中国人学習者のほうが漢字単語の使用頻度が高い。
- 使用に対する誤用の割合は中国人学習者のほうが大きい。
- 漢字圏の学習者の漢字単語の習得は、発音に関しては、非漢字圏の学習者に比べて不利である。

29

ストーリーテリングにおける接続詞の使用状況 (母語話者との比較)

砂川(2018)

30



31

調査結果

△「梯子をのぼる」場面

母語話者:「そこで」
学習者:「そして」「それで」「だから／ですから」

△「警官に見つかる」場面

母語話者:「…と」「…ところ」「すると」
学習者:「指示詞＋とき」「…とき(に)」

32

「そこで」と「そして」

【母語話者】 外から大きな声で呼んでも、マリは起きません。**そこで**ケンは梯子を持ってきて、二階の窓から、家の中に入ろうとしました。

【学習者】 ケンは、あー、マリは、ぐっすり、寝ました。**そして**、あー梯子、ケンは梯子を持ってきて、あー二階の窓から一家の中に入ろうと思いました。

33

「…と」と「…とき」

【母語話者】 ケンが梯子を使って二階の窓から家に入ろうとすると、警官に見つかってしましました。

【学習者】 ケンは、近くにある梯子、を使って、そのまま梯子に上がろうとした**とき**、警官が、見つかりました。

34

考察

「そこで」「…と」「すると」「…ところ」

- 時間関係や因果関係を表しているだけではない。
- 談話の重要な展開場面をマークしたり、次に起こることの意外性や期待感を表現する機能を果たす。

35

場面の展開

- 外から大きな声で呼んでも、マリは起きません。**そこで**ケンは梯子を持ってきて、二階の窓から、家の中に入ろうとしました。
- どうしようか迷った結果えとケンは梯子を上り、二階の窓から、入ろうとしました。**すると**警官の方に見つかってしまい、注意をされました。

36

考察

「そこで」「…と」「すると」「…ところ」

- 中級学習者はこれらの表現を使えないため、ダイナミックな談話展開がうまくできない。

中級以降で使えるようになるのかどうか、追跡調査が必要。

37

テシマウの使用状況 (母語話者との比較)

砂川(2018)

38

テシマウ(チャウ)の使用頻度

I-JASの発話データ

表2 テシマウの使用頻度と使用比率

母語話者	学習者							平均
	スペイン	フランス	ベトナム	ハンガリー	韓国	自然環境	教室環境	
てしまう	104	39	31	36	39	14	25	31
ちゃう	115	3	6	21	13	15	34	20
合計	219	42	37	57	52	29	59	47
母語話者との比較	100	19	17	26	24	13	27	22

39

過去形と非過去形

母語話者	学習者								合計
	スペイン	フランス	ベトナム	ハンガリー	韓国	自然環境	教室環境		
過去形	64 (29%)	36 (86%)	31 (84%)	54 (95%)	38 (73%)	15 (52%)	30 (51%)	38 (70%)	242 (73%)
非過去形	155 (71%)	6 (14%)	6 (16%)	3 (5%)	14 (27%)	14 (48%)	29 (49%)	16 (30%)	88 (27%)

母語話者は非過去形が7割。
学習者は過去形が7割。

40

母語話者

- ま 多分僕は、ま企業に、言っちやうんだろうな。
- あ ったかい所にばっかりいたくなってしまうので。
- ぶ どうとか、の皮も入れちゃつたりして。
- 年寄りになつたら動けなくなつちやうから。

学習者

- ち ょつとえーけがをしちやつたのでー。
- 重い病気なのでい亡くなつてしまひました。
- 今はちょっと急け者になつてしまひました。
- 知っていたけど忘れてしまひました。

41

頻度差の原因

- 中級学習者は、「完全に済ませることを強調する」、または「遺憾・後悔を表す」の用法だけ。
- 母語話者はそれ以外の用法も使う。

まあ害虫みたいなもので、まあ殺してしまえばいいんじやないか。

先輩におごってもらつちやつた！

42

遺憾以外の気持ちを表すテシマウ

- きれいな景色を見ると、ついカメラを向けてしまいます。
- 店内が静かなので、何時間でもいてしまいそう。
- なんと大企業に受かつちやいました。
- 宝くじ買って大金あてちゃおうっと！
- いっそのこと殺してしまおうか。

43

中納言の使い方

44

BCCWJ

現代日本語書き言葉均衡コーパス 通常版

- ユーザー登録は
<https://chunagon.ninjal.ac.jp/useraccount/register>

45

中納言 コーパス選択
コーパス検索アプリケーション

ご利用になりたいコーパスをクリックしてください。

コーパス名	略称	状態	関連サイト	番号
現代日本書き言葉均衡コーパス 通常版	BCCWJ-NT	利用可	関連データベース	
現代日本書き言葉均衡コーパス 非-nuTrans 版	BCCWJ-OT	利用可	非-NuTrans 版	
日本語歴史コーパス	CNU	利用できません	利用申請	
日本語訳し言葉コーパス	CSU	利用可		
多音語訳し言葉の日本語学習者側面コーパス	I-JAP	利用可	関連データベース	
名大会話コーパス	名大会話コーパス	利用可		

パスワード変更 メールアドレス変更 コーザ情報の変更 コーザ相談 収集履歴インポート

46

現代日本書き言葉均衡コーパス（通常版） BCCWJ-NT

短単位検索 長単位検索 文字列検索

外国人旅行者数 外国人旅行者数

前方共起条件の追加

キー キーの条件を指定しない

書字形出現形 が

検索対象 設定を握る

検索対象を選択 検索対象をクリア 全て

検索動作 設定を握る

文頭中の区切り記号 文頭中の文区切り記号 文頭の区切り記号 文頭文類の語数 20 検索

ダウンロードオプション 設定を表示する

47

短単位検索

現代日本書き言葉均衡コーパス（通常版） BCCWJ-NT

短単位検索 長単位検索 文字列検索

短単位検索

前方共起条件の追加

キー キーの条件を指定しない

書字形出現形 が

検索対象 設定を握る

48

キーの前
・キー
キーの後
を
読む
せる

49

前方共起1 キーから 1 語 キーと結合して表示

書字形出現形 が を 短単位の条件の追加

キー キーの条件を指定しない

語彙素 が 読む 短単位の条件の追加

後方共起1 キーから 1 語 キーと結合して表示

語彙素 が せる 短単位の条件の追加

50

ある程度のまとまった分量の 読ま (せ)、グラマックスに至るまでの せることの方が大切と
実はお駆け様のお経(を) 読ま (せ)でいただいたら、精進を守らな を徹底してお
り場合、教科書の問題(を) 読ま (せたあと、次のように指示する。 ころを
新聞一枚があります。 # これ 読ま (せていたらきっと、いかに如信
した。 # そんなこんなで、今初 読ま (せてもらいました。 # (3通分)
すし、次に音を消せりふ(を) 読ま (せ)再生する。 # 大喜びで観ている たので
も根本さんの『小さい目のフラ 読ま (せ)ていたらいて、自分で取り組
声の再教育をするため、本(を) 読ま (せたり、講義をしたりすることも
は、東京からご着任早々八丈実 読ま (せ)て欲しいと申されての、しかも
このような伝統的な立場からの 読ま (せ)ていたらいて、とても勉強にな
れが正しかった

51

前方共起1 キーから 1 語 キーと結合して表示

書字形出現形 が を 短単位の条件の追加

キー キーの条件を指定しない

品詞 の 大分類 が 動詞 短単位の条件の追加

後方共起1 キーから 1 語 キーと結合して表示

語彙素 が せる|させる 短単位の条件の追加

52

い。 # 人質にして己を誘い出 割ら (せ)て、英層について知ろうとする
大きなことを考え
れぞれの詩が精一杯にある審 話か (せ)ている。 # またそれ表現
される。 # ところ
みの神」とはいってい誰のこ 怒ら (せ)て天岩戸事件を引き起こし、清
アマテラス(を)
旗』を用いてきたし、カラス 占わ (せる)『鳥神事』という行事が各地
(+) たる。 # 一定の審議時間を確 投じ (させる)ことも、国会の空転を回復
票(を)
色の筋毛に覆われた淑女の秘 触れ (させ)た。 # ちろっ。 # 「は
が刑務所を出てきた。 # 前 死な (せ)てしまったというのに、約三年
う、その

53

接続助詞「から」の検索

短単位検索 長単位検索 文字列検索

短単位検索

前方共起条件の追加

キー キーの条件を指定しない

書字形出現形 が から 短単位の条件の追加

AND 品詞 の 中分類 が 助詞-接続助詞 短単位の条件の追加

後方共起条件の追加

区切り記号の設定

文脈の語数の設定



55

I-JAS

多言語母語の日本語学習者 横断コーパス

・ユーザー登録は

<https://chunagon.ninjal.ac.jp/useraccount/register>

56

57

受身形の検索



58

キー	後文脈
そして、 リンゴーと サンドイッチが、犬 にー、	食べ [(られ)ました]# はい、 マリと ケン は、 ともー 、 と
ー F、 警官 にー、 しかれ、 しかー X、	しーかーられましたT ー ー ー ー # そのN 時、 マリ はー、 めざまX、 めざ 此ら [(れ)]ました 官 と 。
る と 、 食べ物 は 全部 犬 に	食べ [(られ)]で しま ました]# とでも、 こま、り 、 因り すX、 ケン も 家 に 、 はいY、
	入ら [(れ)]ました

59

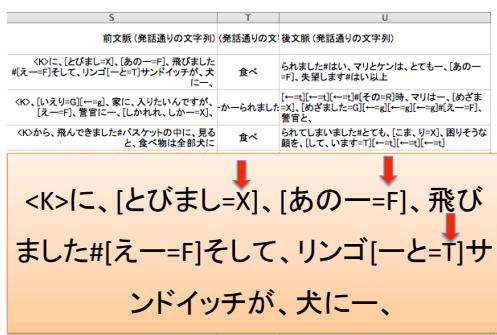
ダウンロードしたデータ 解析向けに加工した文字列

H	I	J
前文脈(解析向けに加工した文字列)	後文脈(解析向けに加工した文字列)	
<K>に、とびまし、あのー、飛びましたえーを して、リンゴとサンドイッチが、犬にー、	食べ (られ)ました# はい、マリケンは、とてもー、あ のー、失望します以上	
<K>、家に、家に、入りたいんですけど、えー、警 官にー、 しかれ、 しかー X、	此ら (れ)ました# の時、マリはー、めざま、目を覚 ました えー、警官と、	
<K>から、飛んできました# バケツの中に、	食べ (られ)しました# とても、こま、り、因りそ	

<K>に、とびまし、あのー、飛びました#
 えーそして、リンゴとサンドイッチが、犬
 にー、
 大にあの食べ様へ切れて、あのもうー

60

発話通りの文字列



61

音声データと文字データ

62

455 8fb4

ダッシュボード

- ナビゲーション
- ダッシュボード
- サイトホーム
- ▶ サイトページ
- ▼ マイコース
- ▶ I-JAS

- コース概要
- I-JAS

63

中納言の使い方について

- I-JAS の中納言検索画面右上にあるマニュアルを読んでください。



64

URL

- ◆ 国立国語研究所コーパスデータベース
<https://www.ninjal.ac.jp/database/>
- ◆ BCCWJ
http://pj.ninjal.ac.jp/corpus_center/bccwj/
- ◆ I-JAS
<http://ljsaj.ninjal.ac.jp>
- ◆ 少納言
<http://www.kotonoha.gr.jp/shonagon/>
- ◆ 中納言
<https://chunagon.ninjal.ac.jp>
- ◆ NINJAL-LWP for BCCWJ
<http://nlb.ninjal.ac.jp>
- ◆ 現代語自他対一覧表
<http://watp.ninjal.ac.jp/resources/>

65

参考文献

- 李在鎬・石川慎一郎・砂川有里子(2018)『新 日本語教育のためのコーパス調査入門』くろしお出版
- 砂川有里子・黒沢晶子(2017)「中國語を母語とする中級日本語学習者の発話に見られる日本語漢語名詞の使用状況—中國語の字音の影響を中心にして」『日中言語研究と日本語教育』10, 64-79.
- 砂川有里子(2018)「中級以降で指摘が必要なテシマウの用法について—学習者と母語話者の使用状況調査に基づく考察—」藤田保幸・山崎誠編『形式語研究の現在』和泉書院 479-499.
- 砂川有里子(2018)「ストーリーテリングにおける順接表現の談話展開機能」庵功雄・石黒圭・丸山信彦編『時間の流れと文章の組み立て—林言語学の再解釈』ひつじ書房 75-106.
- 砂川有里子・黒沢晶子(近刊)「フランス語と中國語を母語とする日本語学習者の漢語名詞の習得状況—自然発話にみられる発音の誤用分析」『フランス人日本語学習者の誤用から考える』ひつじ書房
- ナロック・ハイコ・バルデシ・ブランシャント、赤瀬川史朗(2015)「日本語自他動詞対のコード化の頻度論的動機付け: 大規模コーパスによる検証」ブランシャント・バルデシ・桐生和幸、ハイコ・ナロック(編)『有対動詞の通言語的研究—日本語と諸言語の対照から見えてくるもの』くろしお出版. 25-41.

66