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

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

現実の言語の分析例

- つっ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。

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。

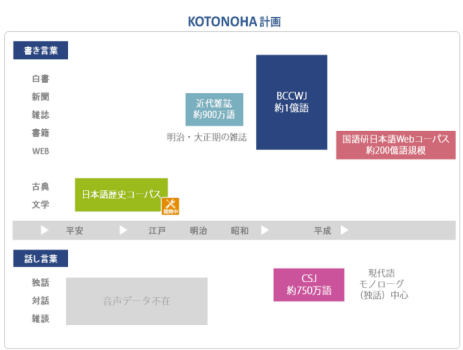
にわにはにわうらにわにはにわにわとりがいる

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

国立国語研究所



国立国語研究所

コーパスデータベース

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

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

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

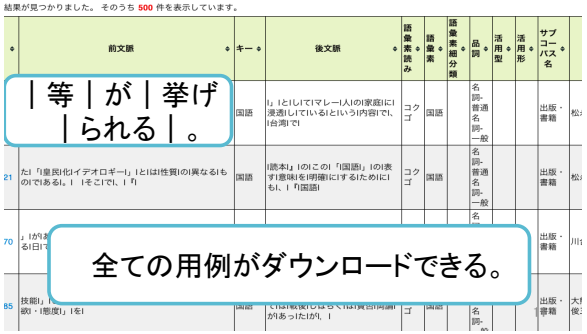
国立国語研究所
2011年に公開

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

中納言



NLB

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

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

2020年完成時の人数

2018年現在660名分公開

◆ 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	教室環境	100人
	自然環境	50人
日本国内の日本語母語話者数		50人

15

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

I-JASの特徴

1. 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。
2. 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
3. 中納言による検索が可能。

17

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

立つ→立てる

形式上コンパクトなほうが派生元

- 移る―移す　tat-u : tat-eru
- 開く―開く　sak-eru : sak-u
- 死ぬ―殺す　utur-u : [illegible]
- 　　hirak-u : hirak-u
- 　　sin-u : koros-u

裂く→裂ける

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）

立つ→立てる　　裂く→裂ける

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

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）

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）

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	裂く	74	裂ける

結論

他動化型も自動化型も形式的に短い動詞のほうが形式的に長い動詞よりも高い頻度で使用されている。

ZIPFの法則が日本語でも成立する

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）

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

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）

（χ^2=8.57；df=1；p<.005）

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）

（χ^2=3.98；df=1；p<.05）

考察

1. 中国人学習者は，自然発話を行う際に，母語である中国語の漢字音が介在してしまう（負の転移）。
2. 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。
3. 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。
4. 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。

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）

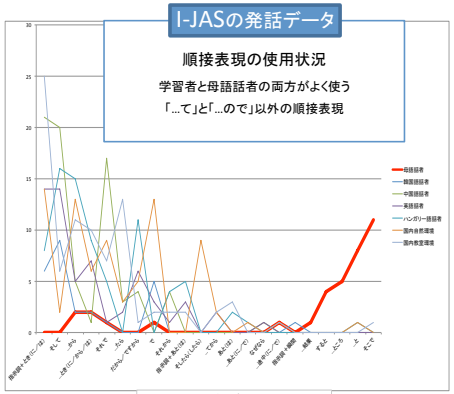


図2 母語話者と学習者グループの順接形式の使用状況

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）」

「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。

「…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いました。

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

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。

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。

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。

テシマウの使用状況（母語話者との比較）

砂川（2018）

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							
		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	平均
てしまう	104	39	31	36	39	14	25	34	31
ちゃう	115	3	6	21	13	15	34	20	18
合計	219	42	37	57	52	29	59	54	47
母語話者との比較	100	19	17	26	24	13	27	25	22

過去形と非過去形

	母語話者	学習者							
		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	合計
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。

母語話者

- ま多分僕は、ま企業に、言っちゃうんだろうな。
- 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、の皮も入れちゃったりして。
- 年寄りになったら動けなくなっちゃうから。

学習者

- ちょっとえーけがをしちゃったのでー。
- 重い病気なのではい亡くなってしまいました。
- 今はちょっと怠け者になってしまいました。
- 知っていたけど忘れてしまいました。

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ゃないか。

先輩におごってもらっちゃった！

遺憾以外の気持ちを表すテシマウ

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

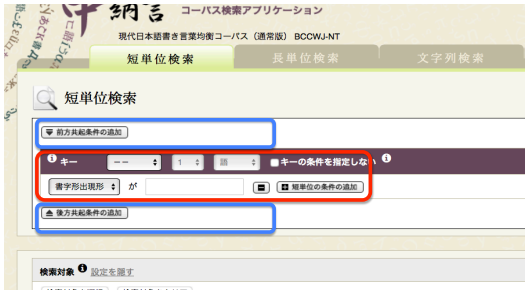
中納言の使い方

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

短単位検索



キーの前　　を

•キー　　読む

キーの後　　せる

ある程度のまとまった分量の	読ま	(せ)、クライマックスに至るまでのせることの方が大切と
実はお釈迦様のお経(を)	読ま	(せ)ていただいたら、精進を守らなを徹底してお
の場合、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)ていただきますと、いかに如信違いなくお受け取りに
った。＃そんなこんなで、今初	読ま	(せ)てもらいました。＃（３通分）たので
字し、次に音を消しせりふ(を)	読ま	(せ)再生する。＃大喜びで観ているまである。＃北京
も根本さんの『小さい目のフラ	読ま	(せ)ていただいて、自分であり続けか、可能な社会
学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父母に教え込まれ
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、しかも苦しゅうない
このような伝統的な立場からの	読ま	(せ)ていただいて、とても勉強になれが正しいかと

い。＃人質にして己を誘い出	割ら	(せ)て、茉屋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響	響か	(せ)ている。＃またそれぞれの表現れる。＃　ところ
みの神」とはいったい誰のこアマテラス(を)	怒ら	(せ)て天岩戸事件を引き起こし、高あるスサノオに好意
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（十）
たる。＃一定の審議時間を確票(を)	投じ	(させる)ことも、国会の空転を回避ある。
色の恥毛に覆われた淑女の秘	触れ	(させ)た。＃　ちろっ。＃　「はう、その
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

接続助詞「から」の検索

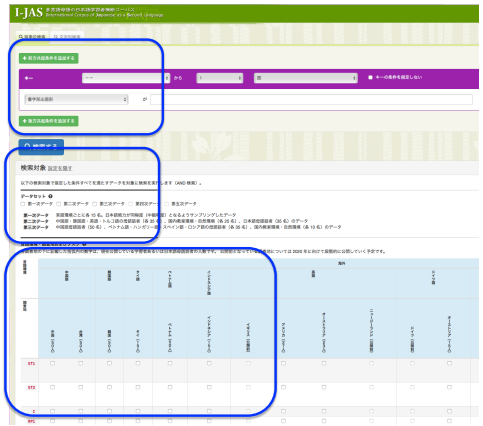
区切り記号の設定

文脈の語数の設定

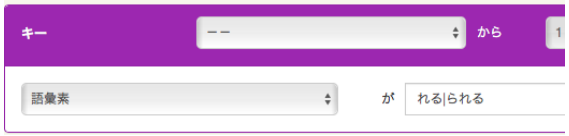
I-JAS

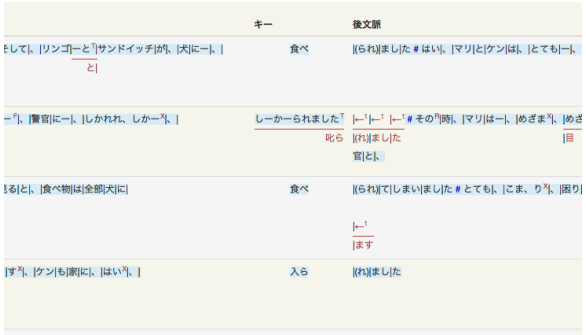
多言語母語の日本語学習者
横断コーパス

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

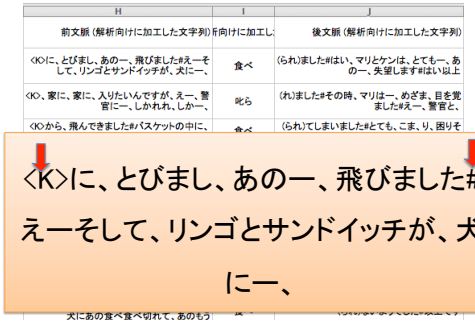


受身形の検索



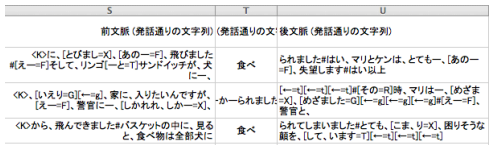


ダウンロードしたデータ
解析向けに加工した文字列



<K>に、とびまし、あのー、飛びました#えーそして、リンゴとサンドイッチが、犬にー、

発話通りの文字列

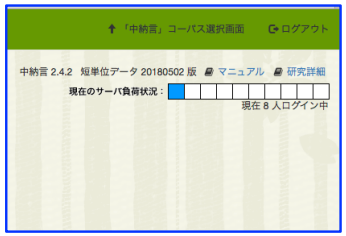


<K>に、[とびまし=X]、[あのー=F]、飛びました#[えー=F]そして、リンゴ[ーと=T]サンドイッチが、犬にー、

音声データと文字データ

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。



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://lsaj.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/

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）「中国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–中国語の字音の影響を中心に–」『日中言語研究と日本語教育』10，64-79.
- 砂川有里子（2018）「中級以降で指導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・丸山岳彦編『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）「フランス語と中国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、パルデシ・プラシャント、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」プラシャント・パルデシ、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.