

そうだったのか！ コーパスで知る コロケーションの世界

『日本語教育のための
文法コロケーションハンドブック』 活用法

中俣 尚己
(京都教育大学 講師)

2013年3月13日 (金)

本日の内容

1. 『文法コロケーションハンドブック』
とはなにか？
2. 『文法コロケーションハンドブック』
はどのようにして作られたか？
3. 『文法コロケーションハンドブック』
には何が書かれているのか？
4. 『文法コロケーションハンドブック』
はどう使うのがいいか？
5. 今後の展望

1. 『文法コロケーション ハンドブック』とはなにか？

- ・ コーパスとは
- ・ コロケーションとは
- ・ 例文作りとは
- ・ 『文法コロケーションハンドブック』とは

- ▶ コーパスを使って、
文法項目のコロケーション
をまとめた本（特に前接動詞）。
- ▶ 例文作りの手助けとして
作りました。

コーパスとは

実際に使用された言語データを
大規模に集め、
コンピュータで読めるようにしたもの。

主なコーパス：

現代日本語書き言葉均衡コーパス
(BCCWJ)

日本語話し言葉コーパス(CSJ)

日本語会話データベース

KYコーパス

コロケーションとは

- ▶ 「2語間の共起関係。一方の語が、他方の語が存在しない時よりも存在する時に生起しやすい場合、2語が共起している(collocate)と言う。」 McEnery & Hardie(2012)
- ▶ 語の意味を知っただけではその語が使えるようにはならず、どのような語とともに使われるのかを知らなければならない。Lewis(2000)
- ▶ 「条約」という語を使うには？
- ▶ 教育的には定義は緩いほうが良い。よって機能語と実質語の関係も含める。

例文作りとは

- ▶ 導入
- ▶ 練習（含・宿題）
- ▶ 評価（テスト問題）
- ▶ 全て広い意味で例文

ここで一句。

例文は
規則だけでは
作れない。

文法コロケーション ハンドブックとは？

- ▶ 日本語教師の仕事の中でかなりのウェイトを占める例文作りに「よく使うか」という視点を客観的なデータを元に導入するもの。
- ▶ 特に、言語学習で大きなウェイトを占める「コロケーション」を見ることが重要。

2. 『文法コロケーション ハンドブック』はどのように して作られたか？

- ・ 講師の来歴
- ・ 「てある」の研究からスタート
- ・ 「現場の声」と「コーパス」
- ・ データの収集方法
- ・ 効果の測定

中俣尚己 来歴

- ▶ 京都教育大学 日本言語文化専攻 卒業
- ▶ 大阪府立大学大学院で 博士号を取得
(日本語並列表現の体系)
大学院で学びながら非常勤講師として
初級・初中級の学生に日本語を教える。
- ▶ 京都外国語大学 嘱託研究員
日本語教員養成に関わる。
- ▶ 実践女子大学 国文学科 助教
- ▶ 京都教育大学 国文学科 講師

事の発端



事の発端

- ▶ 「中俣さんの企画、悪くはないんだけど、もうちょっと「真似できる」研究にしてほしい。なおかつ新しさのあるもの。来月までに新しい企画、よろしく！」
- ▶ 「わかりました。じゃあ、昔授業がうまくいかなかった「てある」と「ておく」にします。」
- ▶ (やけくそ)
- ▶ 実は、「てある」について考えたことなど1秒もなかった。

困ったので

- ▶ とりあえずコーパスを見ようと思った。
- ▶ 実は、これはCorpus-Driven Approachという研究手法であった。
- ▶ Corpus Based Approach = 仮説を立ててからコーパスを見る。
- ▶ Corpus Driven Approach = 仮説を立てずにコーパスを見る。

「てある」の2つの用法 (益岡1987ほか)

- (1) カレンダーに今月の予定が書いてあります。

×動作主

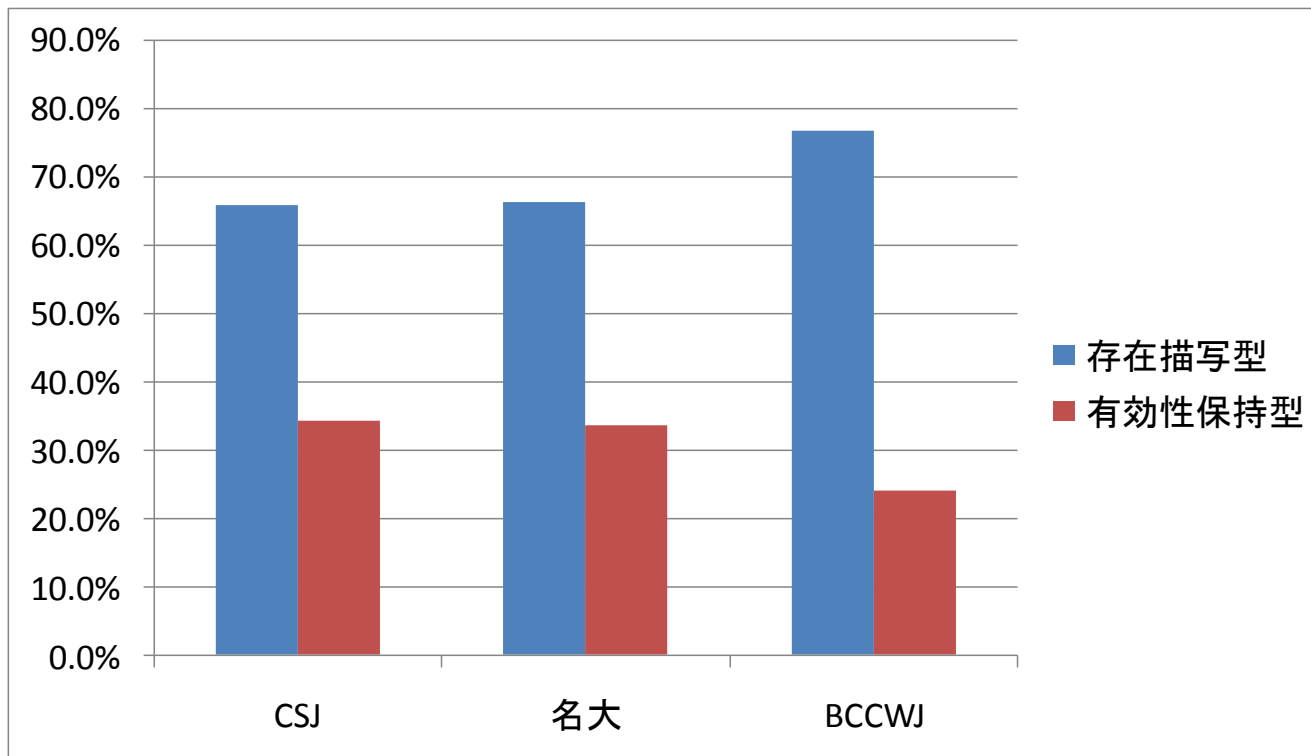
存在描写型

- (2) (私は) プレゼントはもう買ってあります。

存在描写型の方が数が多い。
存在描写型から導入すべき。
原沢(2007)

どのコーパスでも存在描写型が多い

中俣(2011)



「てある」に前接する頻度が高い語

中俣(2011)

順位	CSJ	名大	BCCWJ
1	書く (177/48.3%)	書く (278/50.8%)	書く (1864/38.4%)
2	置く (21/5.7%)	置く (57/10.4%)	置く (386/8.0%)
3	示す (16/4.4%)	する(31/5.7%)	する(203/4.2%)
4	する(14/3.8%)	作る(14/2.6%)	貼る (101/2.0%)
5	作る(7/1.9%)	言う(10/1.8%)	つける (74/1.5%)
6	拳げる(6/1.6%)	入れる(8/1.5%)	かける (69/1.4%)
7	振る(6/1.6%)	張る (7/1.3%)	示す (65/1.3%)
8	張る (6/1.6%)	揚げる(6/1.1%)	入れる(51/1.1%)
9	取る(5/1.3%)	取る(6/1.1%)	飾る (47/1.0%)

「である」の50%が書いてあるというのは、

- ▶ 言語学者としては興味を引かない事実であるが、言語教育者としては重要な事実。

疑問文とは共起しない

- ▶ 「てあるか」「てありますか」

といった形はほとんどない。

- ▶ 「みんなの日本語」本文にも疑問文はないが、練習問題や会話にちゃっかり紛れ込んでいる。

一方、その頃

- ▶ 京都外国語大学・日本語教員養成推進室
- ▶ 文法の説明は理解できても、良い例文を作るのが難しい学生が多かった。
- ▶ 例文は 規則だけでは 作れない。
- ▶ ベテランの先生も学生も
「この文法項目は、こういう動詞とよく使う、みたいなことが書いてあったらいいのに」

じゃあ、作りましょう！

- ▶ 当初は学会でセンセーショナルな発表をして仲間を募り、プロジェクトを立ち上げる予定だった。
- ▶ しかし、2011年8月に転機が！
- ▶ 1億語の書き言葉コーパスを、形態素情報を利用して検索できるシステム「中納言」が公開された。
- ▶ これを使えば1人でできる！

検索の方法など

- ▶ 見出し語93。 正確には「なければなら
ない」「なければいけない」
などを1項目にまとめているので
103項目。
- ▶ それぞれの「1つ前」または
「助動詞を挟んで2つ前」に
出現する動詞・イ形容詞・
ナ形容詞・名詞を検索した。
- ▶ 食べ ている。 ←収集
食べ られている。 ←収集
食べ られ つづけ ている。 ←収集しない

長単位モードで検索

▼ 前方共起条件の追加

キー (--- ▼ 10 ▼ 語 ▼)

品詞 ▼ の 大分類 ▼ が 動詞 ▼ ☐ ☒ 長単位の条件の追加

後方共起1 (キーから ▼ 1 ▼ 語 ▼)

語彙素 ▼ が ている ☐ ☒ 長単位の条件の追加

▲ 後方共起条件の追加

キー = 「品詞」が「動詞」
後方共起1 = 「語彙素」が「ている」
で検索！！

検索の方法など

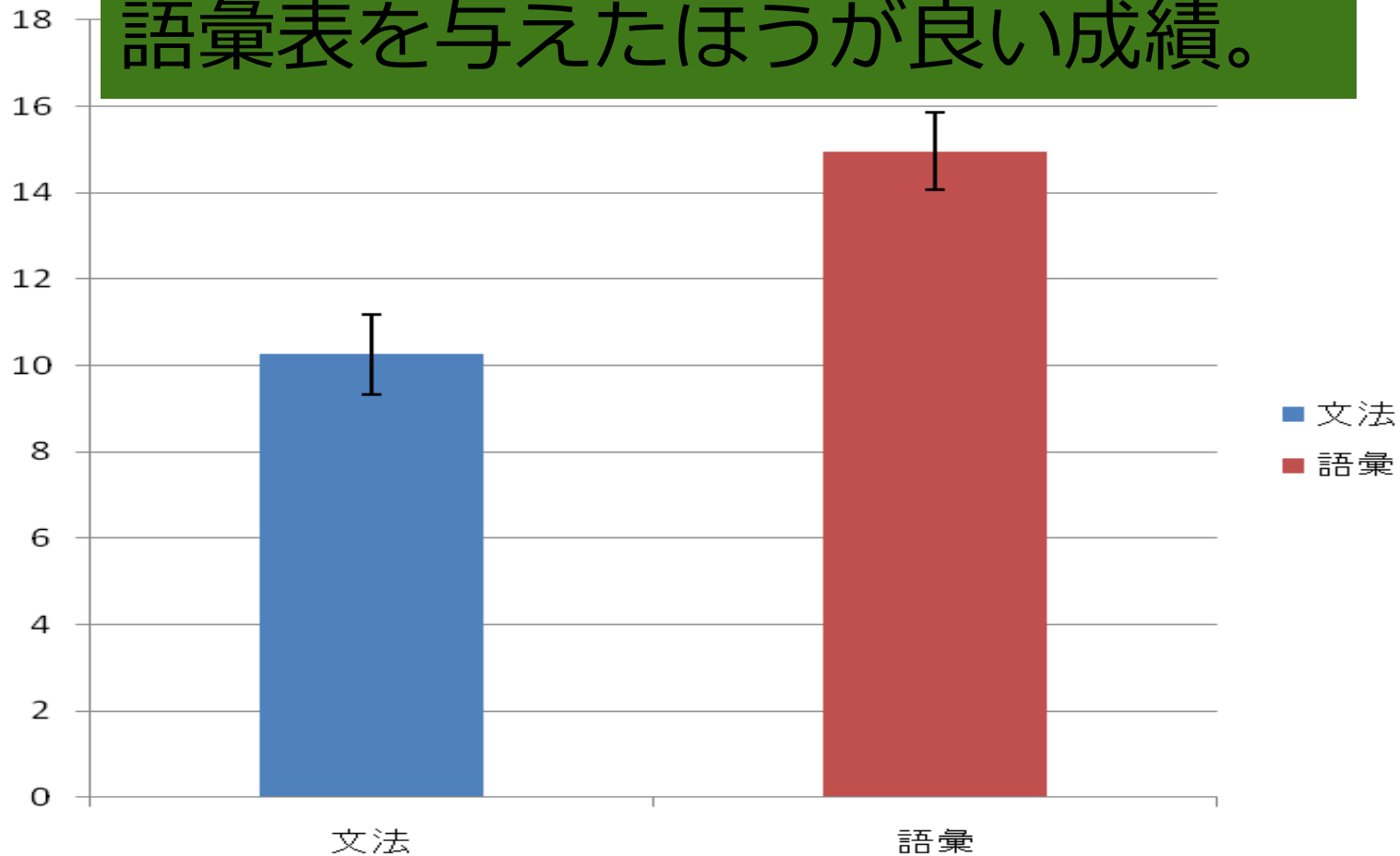
- ▶ 検索条件は項目ごとに異なり、できるだけ多くの例が得られるようにした。
- ▶ データは主として2011年に収集。当時とはシステム・内部データが変更されているため、今の数字とは微妙に異なる。
- ▶ ただし、上位のランキングが変わるほどの影響はない。

様々な苦労を経て.....

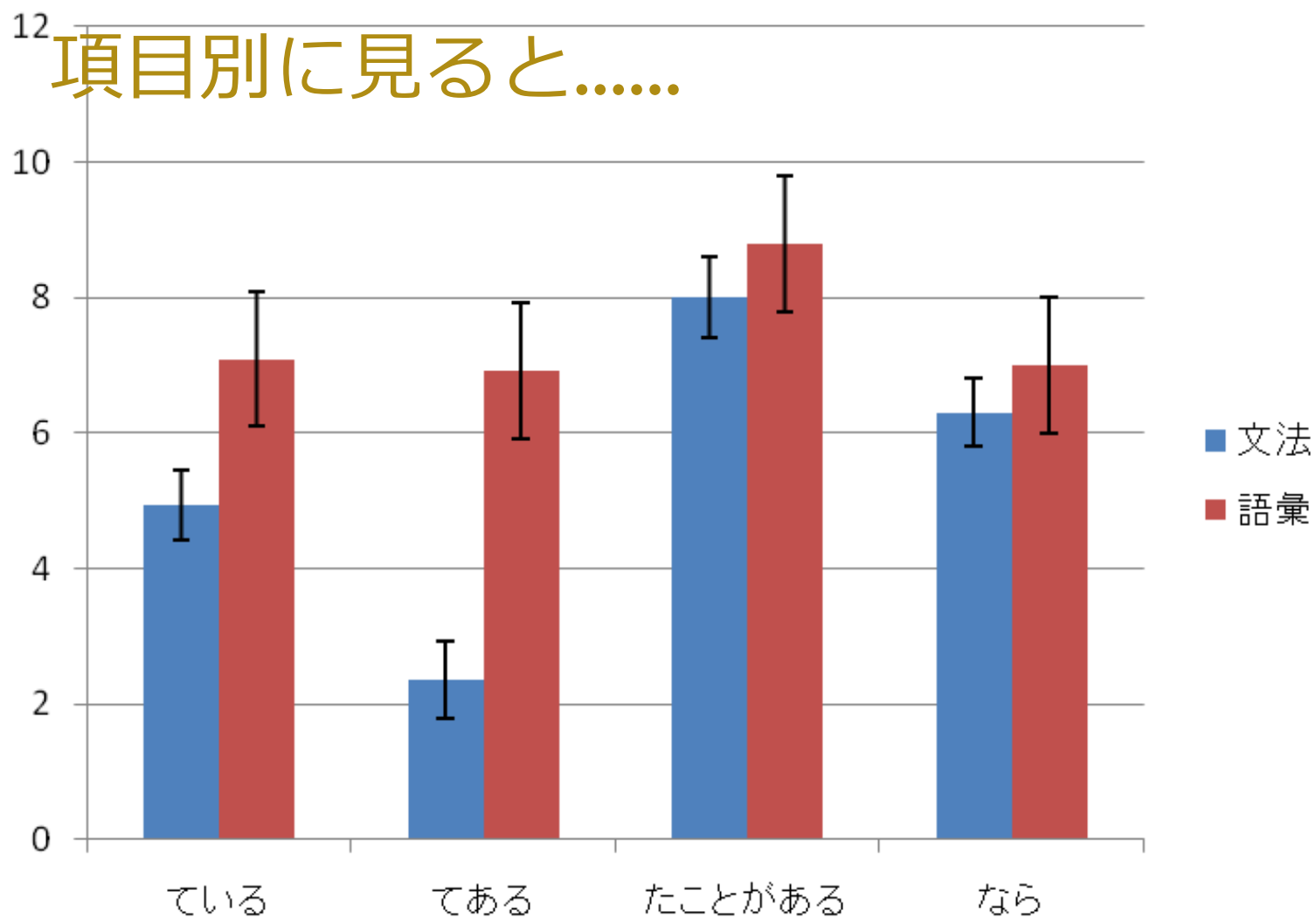
- ▶ 中納言は一度に10万件しかダウンロードできないが、「ている」は90万件以上ある。
- ▶ 3日かけてダウンロード、結合した結果、終止形は「なっている。」が最も多いということがわかり、「大変な事が起こりました！」と学科主任の部屋に駆け込むという事案が発生。

例文作り実験の結果。4問の合計。 中俣(2013)

語彙表を与えたほうが良い成績。



項目別に見ると.....



$p < .05$

$p < .01$

n.s.

n.s.

3. 『文法コロケーション ハンドブック』には何が書かれて いるか？

- ・ 実例を見ながら解説

4. 『文法コロケーション ハンドブック』はどう使うか？

- ・ 予想されるメリット
- ・ 現場の声
- ・ Q & A

ハンドブックのメリット (中俣2012)

(1) 教師にとっては、毎回の授業で準備する**例文作り**に有効である。特に、日本語の内省がきかないノンネイティブ教師にとっては、教科書にとらわれず、母語話者がよく使う例文を提示することが容易になる。

ハンドブックのメリット (中俣2012)

- (2) 学習者にとっては、日本語でよく使われるコロケーションを学ぶことができる。
- (3) 教科書作成において、導入すべき語彙の選定に必要な情報を提供できる。
- (4) 言語研究において、新たな角度からの知見を提供できる。

発売後の実際の声 (amazonのレビューより)

- ▶ 教える項目が実際にはどのような形で、どのような文脈で使われているのかイメージできる。
- ▶ 本書のように数字ではっきりと使用頻度のランクやどんな場面で使われているか(会話か？文書上か？ネット上か？)がわかると安心して「これ教える」「これ要らない」と決めやすいです。

海外実習での実習生に 使わせている例

- ▶ 文型の導入、文例、練習などの作成、
提示するカードの取捨選択。
- ▶ 文型シラバス以外で教えるときも、
教師が予想する使用文型と頻度、
文型とモードのずれを確認できます。

順位はどう考えればよいか？

▶ とりあえず50位
以内なら
「よく使う」と
判断してOK。

1 位から順番に例文を作れば よいか？

- ▶ いいえ。
- ▶ 上位の動詞が固い文章に多く見られる例になっていることも多い。
- ▶ 「ている」は「問題になっている」が非常に多い。
- ▶ 「ている」の「導入」には適さない。
- ▶ ただし、中・上級では「～が問題になっている」というようなコロケーションが多く、アカデミックな文章の書き出しによく用いられるという情報は重要になってくる。
- ▶ 「初級項目」というカテゴリが意味するものとは？

研究に使えるかも？

- ▶ 多分、93本の論文が容易に書けると思います。
- ▶ 伝聞の「そうだ」だけやりました。(中俣2014)

とにかく表が大事！

- ▶ よく使う動詞リストを見て、例文を作ってみる。（中俣の例文はアテにしない）
- ▶ 一例一例が大事なのは言うまでもないが、本質的な機能が見えてくることもある。
- ▶ 新米の教師はこれで言語感覚を磨く。ベテランの方は直感の確認にどうぞ。

5. 今後の展望

- ・「生産性指数」という考え方
- ・ホップ・ステップ・ジャンプ！

頻度が増えれば、前の動詞の
数が増える.....わけではない！

	前の動詞の種類	頻度
た後	2,409	9,711
ておく	2,425	31,801

物質によって比重が決まっているように、
文法項目も「**どれだけの語と結びつくか？**」
=**生産性**が決まっているのではないか？
それを数値化できないか？

1年かけて7種類もの計算式
を試した結果.....

エントロピー

ジニ係数

HardenのC

Guiraud Index

修正Perplexity

標準化TTR

上位10語の割合

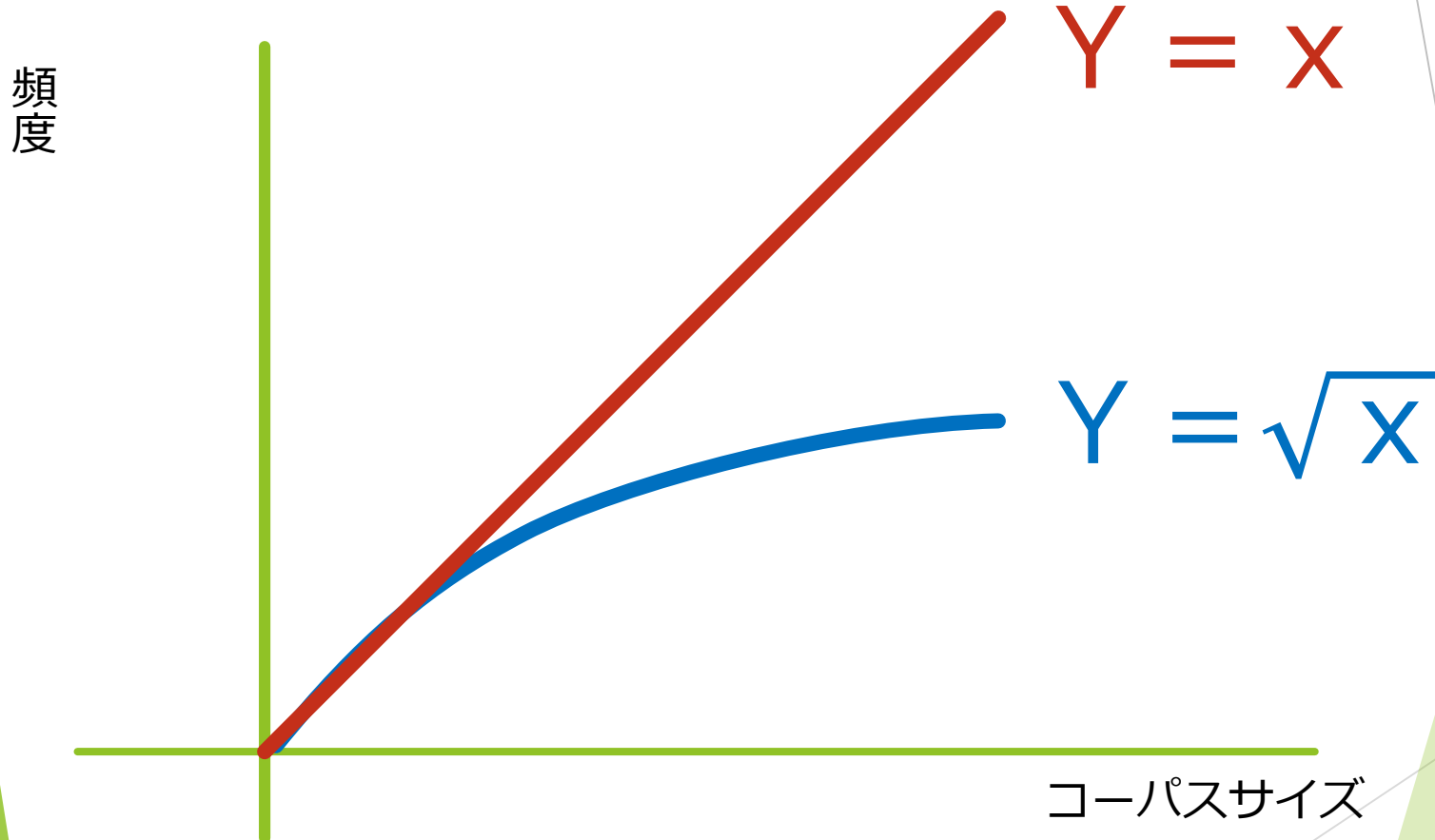
生産性指数(中俣 2015)

$$p = \frac{\text{Type}}{\sqrt{\text{Token}}}$$

これで確定。

$$\text{生産性指数} = \frac{\text{前接動詞の種類数}}{\sqrt{\text{項目の出現数}}}$$

何でルートが出てくるかというと。



のべ語数 = Token頻度
異なり語数 = Type頻度

生産性の高い項目

項目	生産性	項目	生産性	項目	生産性
たり	27.10	るとき	22.40	ようになる	21.91
た後	24.45	たら	22.29	ている	21.06
のに	23.98	から	22.22	てしまう	20.58
ことになる	23.09	させる	22.13	ながら	20.40
ようにする	22.93	たとき	22.03	と	20.36

生産性が高い項目の特徴

- ▶ 意味的制約を持たない「たり」が1位。
- ▶ 「～る時」「～た後」「たら」「ながら」など時間に関係する表現も生産性が高い。
- ▶ 「と」よりも「たら」の方が汎用的であることを書き言葉コーパスで示せたのは画期的。
- ▶ 「のに」は逆接の「のに」と「目的」の「のに」が混ざっているので本当は低い。

生産性の低い項目

項目	生産性	項目	生産性	項目	生産性
てあげる	11.65	ないか	9.82	ませんか	7.25
しょうか	11.64	たこと がない	9.52	終わる	6.50
てください	10.58	ましょう か	8.88	お~になる	5.94
ないでく ださい	10.08	すぎる	8.05	お~する	2.46
である	9.91	終わる	7.91	やむ	0.53

生産性が低い項目の特徴

- ▶ 「終わる」「終える」「やむ」など**終了を表すもの**。
- ▶ 「ませんか」「てください」「お～する」など**話し言葉でよく使うものは生産性が低い**。
- ▶ この傾向は「から」>「ので」など**他の項目も同じ**。
- ▶ 書き言葉にこれらの項目が出現しないから、**というわけではな**

4. 生産性と難易度

- ▶ 山 中俣さん, これは作っただけ?
- ▶ 中 (えっ!? 他に何をしろと...) は, はい, 作っただけです。
- ▶ 山 並び替えたりはしてない?
- ▶ 中 あ, やってませんけど, それは
すぐ
す
に出来ます。(その場でソート
る)

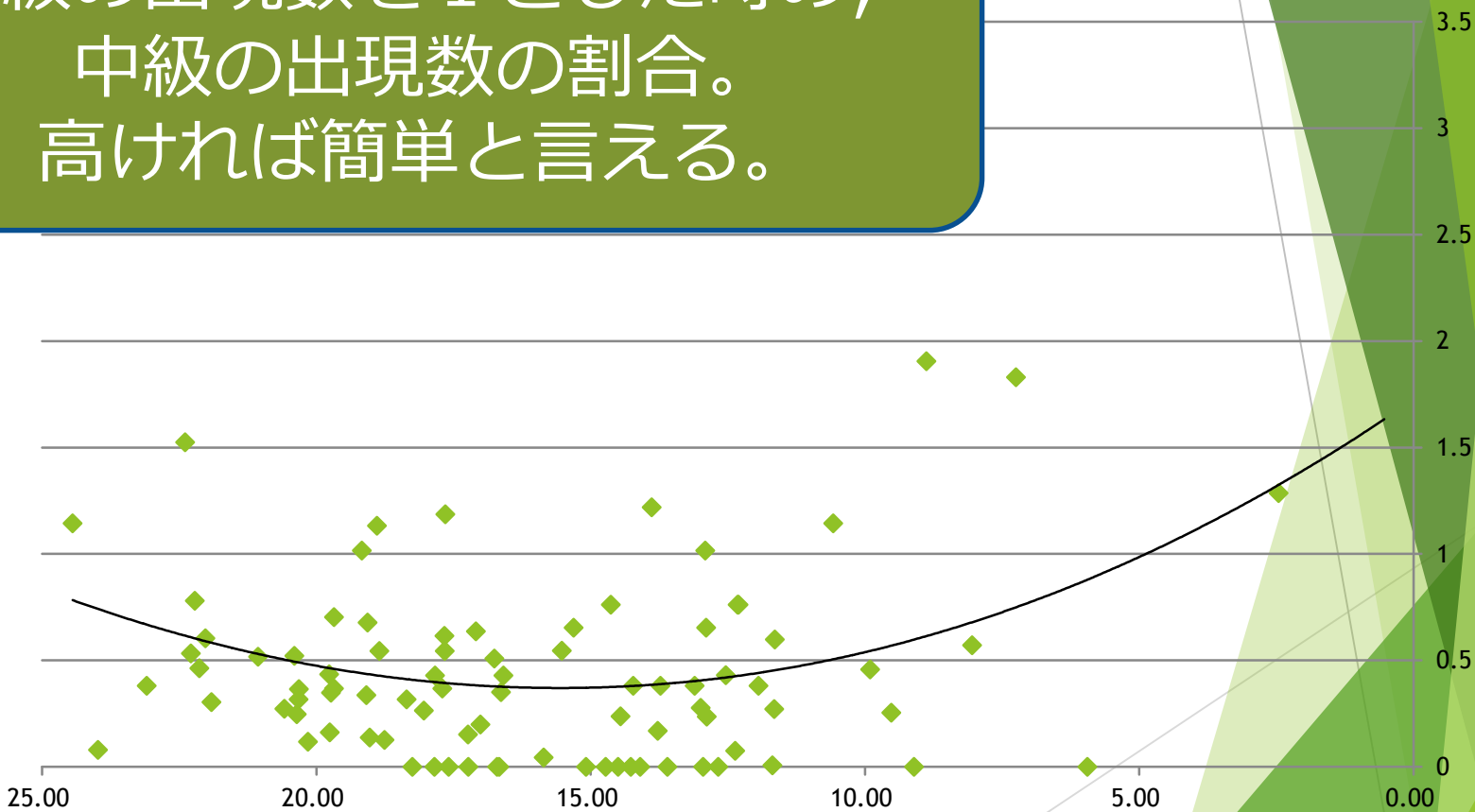
- これね、
生産性が高いのは簡単で、
真ん中が難しくて、
低いのがまた簡単にな
ってる気がする。

ほんまかいな
～？

KYコーパスとの比較

中級以上

超級の出現数を1とした時の、
中級の出現数の割合。
高ければ簡単と言える。



生産性高

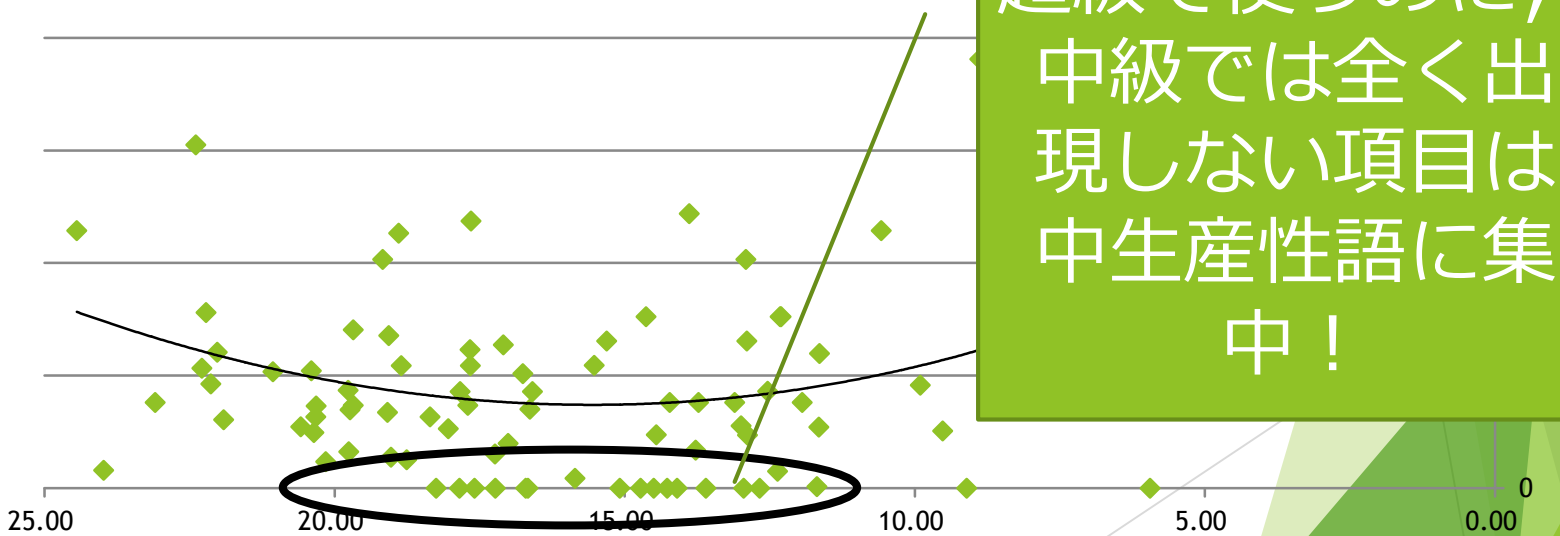
生産性低

KYコーパスとの比較

中級時点

超級でも使わないような項目の難易度を論じても仕方ないので、除外してある。

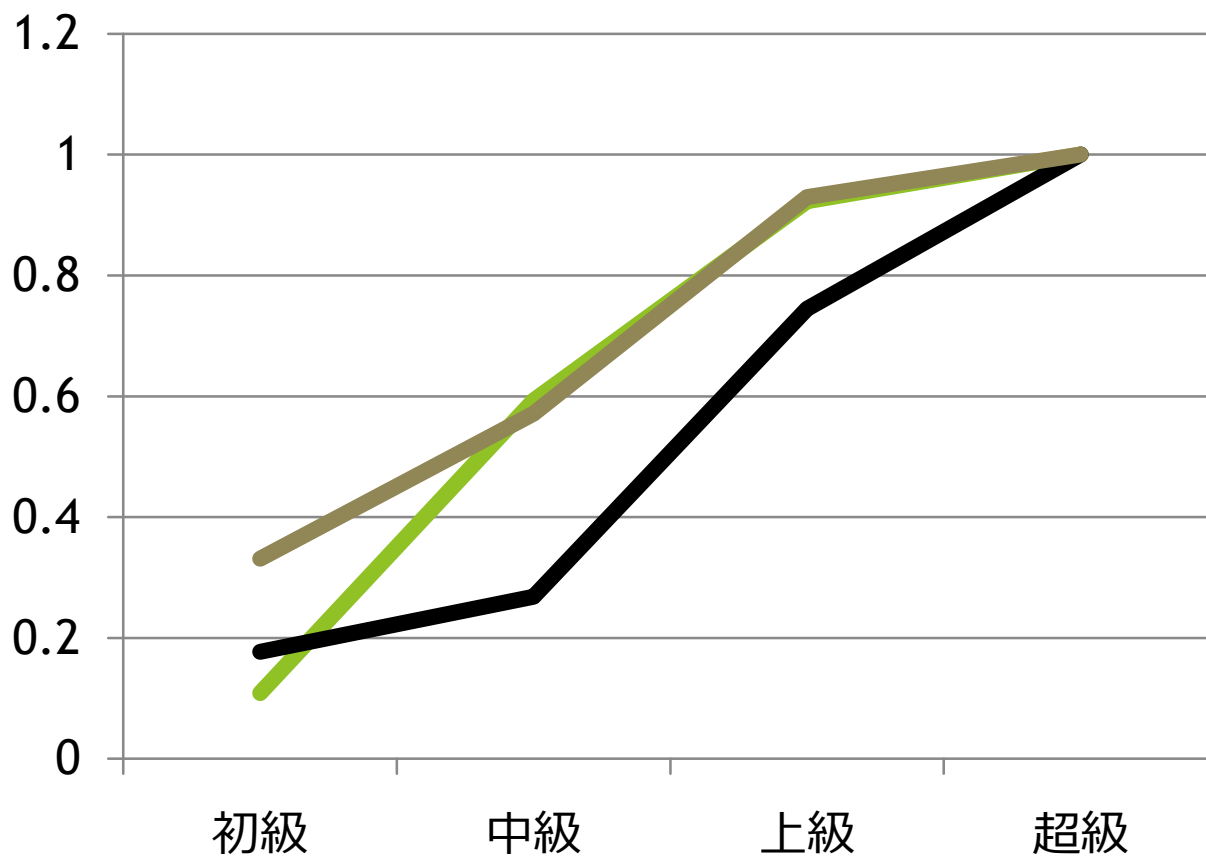
超級で使うのに、
中級では全く出現しない項目は
中生産性語に集中！



生産性高

生産性低

生産性高・生産性中・生産性 低に3分割。



超級で使うのに
中級で使わない
項目

— 高い **1**
— 中 **10**
— 低い **5**

なぜ中生産性語が難しいのか？

	文法項目の生産性	降水確率
高い	制約がないので簡単	90%なら迷わない
中	中途半端に制約が多いので迷う	50%が一番困る
低い	制約が強すぎて、逆に限られた組み合わせだけ覚えればOK	10%なら迷わない

宣伝

- ▶ 全ての項目の生産性の正確な数値は
今度『計量国語学』29巻8号の論文
に掲載しますが.....。
- ▶ 5月に出版される
**『シリーズ 現場に役立つ
日本語教育研究 1
データに基づく文法シラバス』**
にも解説つきで掲載します！

生産性指数の真価は
中上級項目で

▶ どう～ても

▶ いくら～ても

中上級項目に関しては.....

- ▶ 2016年
- ▶ 中上級で、主にアカデミックライティングや検定対策で必要になる項目を中心に、『文法コロケーションハンドブック』と同じアプローチで取り組んだ論文集を刊行！
- ▶ 「条件を表す表現」「対比を表す表現」など**10の機能**ごとに、**100を超える形式**を扱う、ハンドブック寄りの論文集。
- ▶ **中俣尚己(編)『シリーズ 現場に役立つ 日本語教育研究 5 コーパスから始まる例文作り』**

さらにその先へ

▶ 1000項目レベルの
Corpus-Basedの
ハンドブック。

御清聴
ありがとうございました。

今後の活動のため、
ご意見・ご質問を
お待ちしております。

参考文献

- ▶ 中俣尚己(2011)「コーパス・ドライブン・アプローチによる日本語教育文法研究－「てある」と「ておく」を例として－」森篤嗣・庵功雄(編)『日本語教育文法のための多様なアプローチ』ひつじ書房, pp.215-233
- ▶ 中俣尚己(2012)「コーパスに基づいた語彙情報つき文法ハンドブックの構想」『実践国文学』81, pp.71(24) - 60(35)
- ▶ 中俣尚己(2013)「日本語教育における例文づくりのためのコーパス研究――条件の「なら」を例として――」『京都教育大学国文学会誌』40, pp.90(1)- 80(11)
- ▶ 中俣尚己(2014)「伝聞の「そうだ」が伝えるもの―機能語と実質語のコロケーション研究―」『京都教育大学国文学会誌』41, pp.130(1)- 114(17)
- ▶ 中俣尚己(2015)「初級文法項目の生産性の可視化―動詞に接続する文法項目の場合―」『計量国語学』29-8, pp.275-295

参考文献

- ▶ 原沢伊都夫(2007)「テアル2形式（能動型と受動型）における話者の認識」『静岡大学国際交流センター紀要』1, 静岡大学国際交流センター, pp.1-8
- ▶ 益岡隆史(1987)『命題の文法』くろしお出版.
- ▶ Lewis, M. (2000). There is nothing as practical as a good theory. Lewis, M. (Ed) Teaching Collocation: Further Development in the Lexical Approach. pp.10-27. Heinle: Cengage Learning.
- ▶ McEnery, Tony & Hardy, Andrew. (2012) Corpus Linguistics: Method, Theory and Practice. Cambridge University Press.