言語処理学会第24回年次大会（NLP2018）招待論文

大規模コーパスに基づく日本語二重目的語構文の基本語順の分析

笹野遼平（名大）　奥村学（東工大）

研究の背景と着想

背景 二重目的語構文の基本語順の分析

- 人手分析や脳活動・読み時間計測に基づく研究が中心
- 分析対象とした用例については高信頼度な分析が可能
- しかし、コストが大きく多くの仮説の網羅的検証に不向き

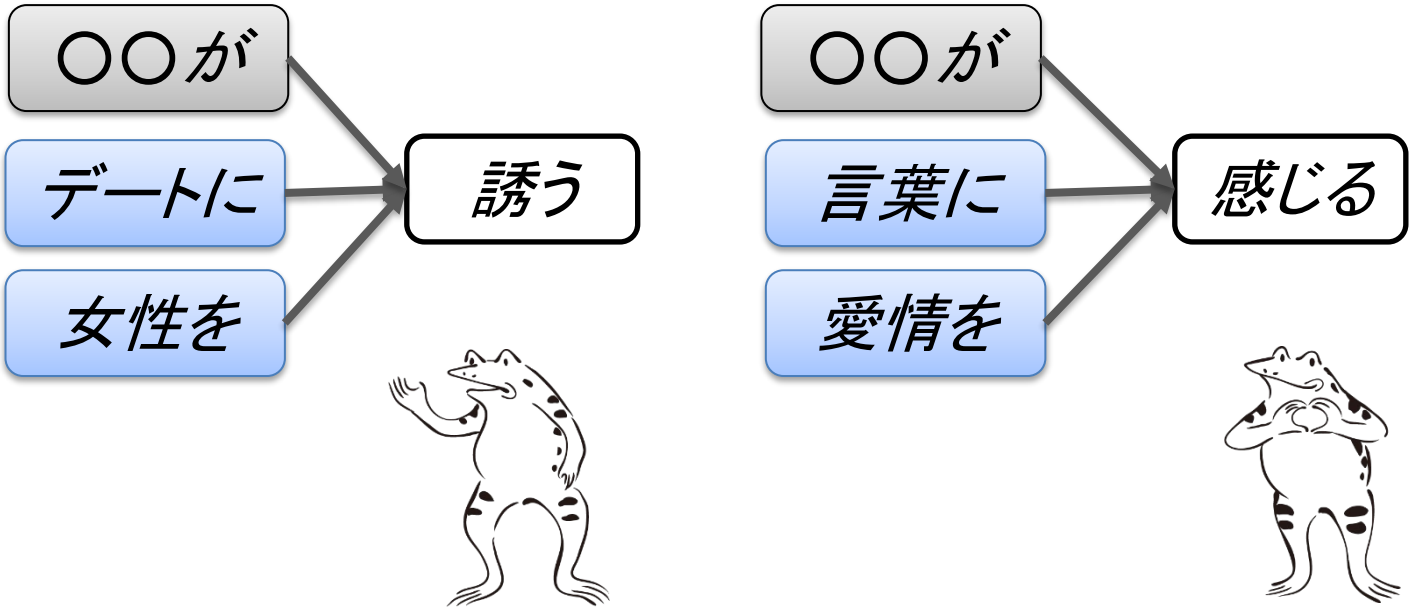
着想 大規模コーパスを用いて分析

- 大規模なコーパスを使うことで多少のノイズを無視し、低コストで網羅的に仮説を検証することが可能
- 間接的な分析ではあるものの、心理実験や脳科学等のコストの掛かる検証を行う前段階の分析として有用

もくじ

二重目的語構文とは

- 項として2つの目的語を取る構文
  - 日本語の場合、各目的語はヲ格とニ格で表される
  - ガ格も含めると項は3つ ⇒ 三項動詞文とも呼ばれる



二重目的語構文の基本語順

- 「～に～を」なのか？

コーパス中の出現数/割合

言葉に 愛情を 感じる　（118 / 97.5%）

愛情を 言葉に 感じる　（3 / 2.5%）

- 「～を～に」なのか？

デートに 女性を 誘う　（4 / 0.4%）

女性を デートに 誘う　（923 / 99.6%）

⇒ 基本語順を決める要因を明らかにしたい

本研究における“基本語順”

1. 日本語話者にとって最も自然で理解しやすい語順

2. コーパス中の出現率と強く関係すると仮定
   - 2つの目的語が与えられた場合に、ある語順が大半を占めるならばその語順を基本語順とみなす

3. 動詞/項の組合せにより異なるが組合せごとに1つ
   - 多くの先行研究で採用[Matsuoka’03, Miyagawa+’04, 中本+’06, etc.]
   - 語彙によらず基本語順は構文ごとに1つという立場[Hoji’85]、1つの文における基本語順が複数存在しうるという立場[Miyagawa’97]も存在

二重目的語構文の基本語順に関する先行研究

- 言語学や脳科学からのアプローチ
  - 言語学者/被験者の内省に基づく分析
    e.g. [Hoji'85, Matsuoka'03, Miyagawa+'04] / [中本+'06]
  - 読み時間/脳波等の計測に基づく分析
    e.g. [Koizumi+'04, Shigenaga'14 , 滝本+'15] / [高祖+'04, 犬伏+'09]
- 用例ごとに人手による分析・計測が必要
  - 分析対象とした用例については高い信頼度
  - 新しい用例に対しては改めて分析・計測が必要

⇒ 本研究: 大規模コーパスに基づく網羅的な分析

本研究で検証する仮説

- 代表的な仮説およびその類型として以下を検証

（詳細は後述）

A) 動詞によらず基本語順は「にを」である[Hoji'85]

B) 基本語順は動詞のタイプによって異なる[Matsuoka'03]

C) 省略されにくい格は基本語順において動詞の近くに位置する

D) 基本語順は二格名詞の意味役割や有生性によって異なる[Matsuoka'03,Ito'07]

E) 対象の動詞と高頻度に共起するヲ格、二格名詞は基本語順において動詞の近くに位置する

もくじ

用例収集におけるポイント

- 語順は直接観測可能な事象
  - コーパスから各語順の出現傾向を収集可能
  - 圧倒的に多くの用例に基づく網羅的検証

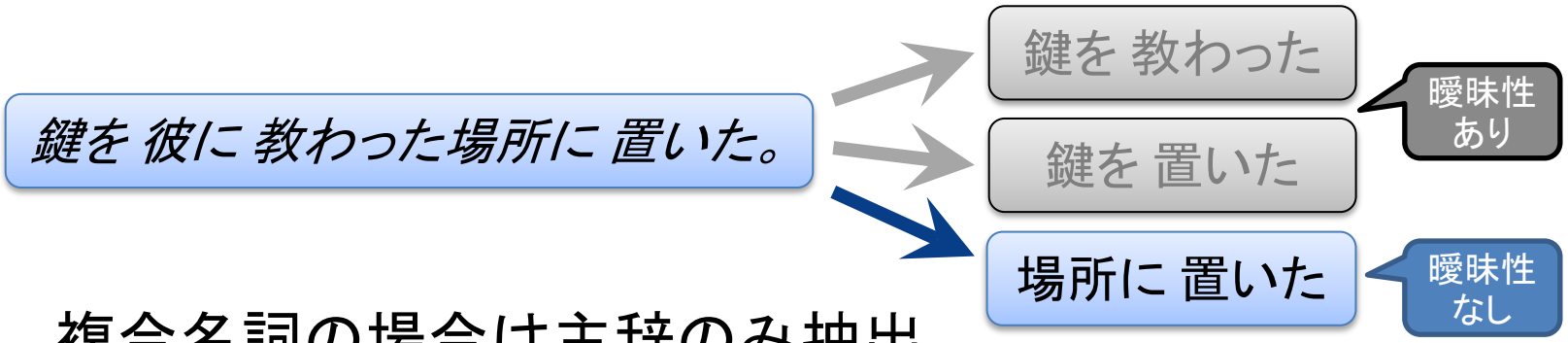
- 個別の事例の信頼性は低い
  - 個別事例から基本語順か判断するのは困難
  - 用例抽出の段階で誤りを含む可能性

夕飯を 先輩に 教わった 店で 食べた ✕ 夕飯を 先輩に 教わる

- 確からしい用例を大量に収集したい

用例の収集方針

- 非常に規模の大きいコーパスから（>100億文）
  - Web上のテキスト集合を句点等をもとに文に分割
  - コピーページ対策のため重複文を削除し使用
- 構文的曖昧性の少ない用例だけを収集
  - 河原ら[’02]の手法で信頼性の高い部分を抽出（精度98.3%）



  - 複合名詞の場合は主辞のみ抽出
  - ガ格や連体形等の扱いについては後述

分析対象とする動詞

- 以下の条件を満たす動詞の用例を収集
  a. 能動態で出現 & JUMANの基本辞書に含まれる
  b. ヲ格名詞、ニ格名詞がともにJUMAN辞書に含まれている用例の異なり数が500以上
  c. ヲ格名詞、ニ格名詞を両方持つ用例の割合が5%以上
- 条件を満たした動詞は648種類
  – １動詞あたりの出現数の平均は約35万、中央値は8.3万
  – ヲ格、ニ格を両方持つ用例数の平均は約3.8万、中央値は0.9万

もくじ

検証する仮説と分析方法

A) 動詞によらず基本語順は「がにを」である[Hoji'85]

B) 基本語順は動詞のタイプによって異なる[Matsuoka'03]

C) 省略されにくい格は動詞の近くに出現しやすい

昨日に比べ肌寒さを感じます。

直前要素⇒意味を決める⇒省略されにくい

～に … 感じる	＜	～を … 感じる	⇒ 基本語順:	*～に ～を … 感じる*
～に … 例える	＞	～を … 例える	⇒ 基本語順:	*～を ～に … 例える*

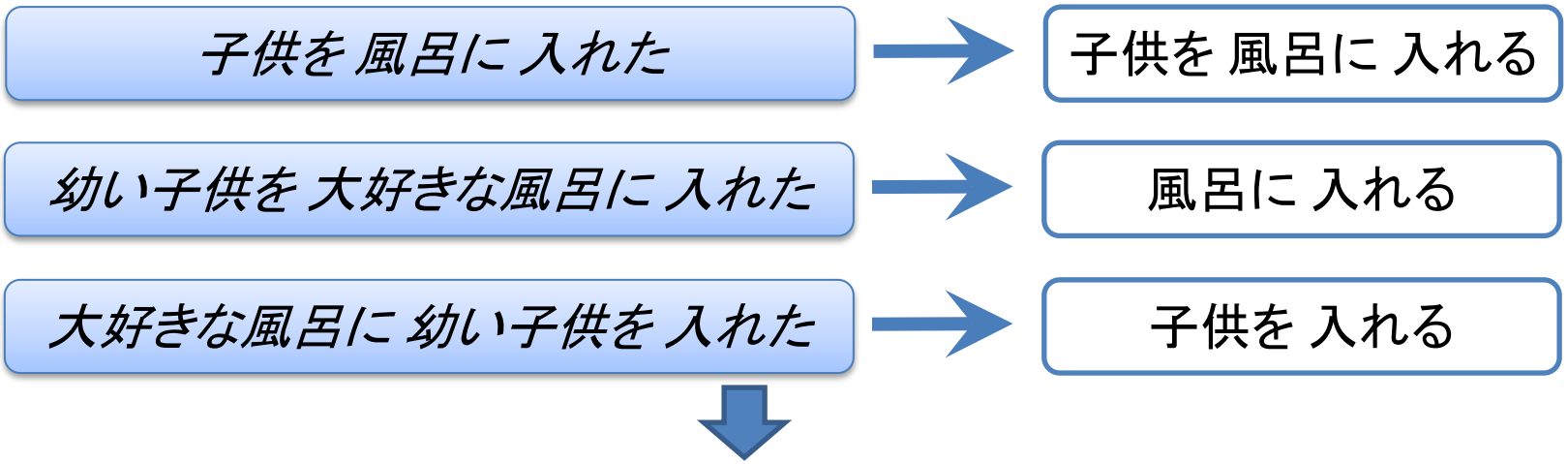
野生動物に例えると…

- 前述の条件を満たす動詞ごとに以下を算出
  1. ヲ格と二格の一方のみ出現した用例に占める二格の割合（横軸）※
  2. ヲ格と二格が両方出現した用例に占める「をに」語順の割合（縦軸）

一方のみ出現した用例の収集(※)

河原ら[’02]の手法で収集した用例そのままではバイアスが存在

- 基本的に曖昧性のない係り受けから収集



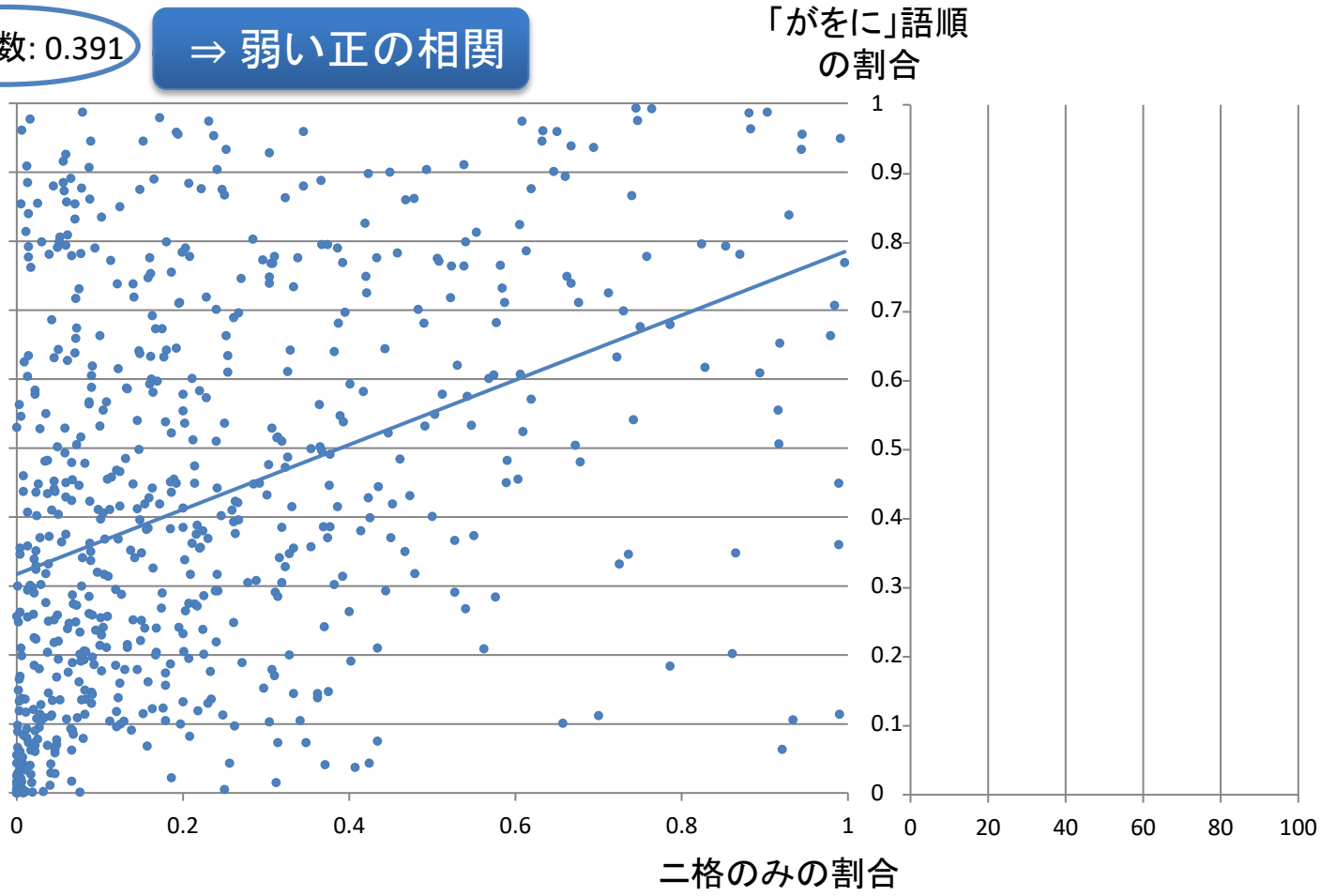
近い位置に出現する用例の方が集まりやすい！

- [解決策] ヲ格と二格の一方のみ出現した用例（＝横軸）として先行するガ格も収集された用例のみを使用
  - 「は」「も」などの副助詞を伴って出現する項がある場合、および、連体形の動詞も収集対象から外す（∵ヲ格or二格項である可能性があるため）

実験1: 格の出現率と語順

相関係数: 0.391

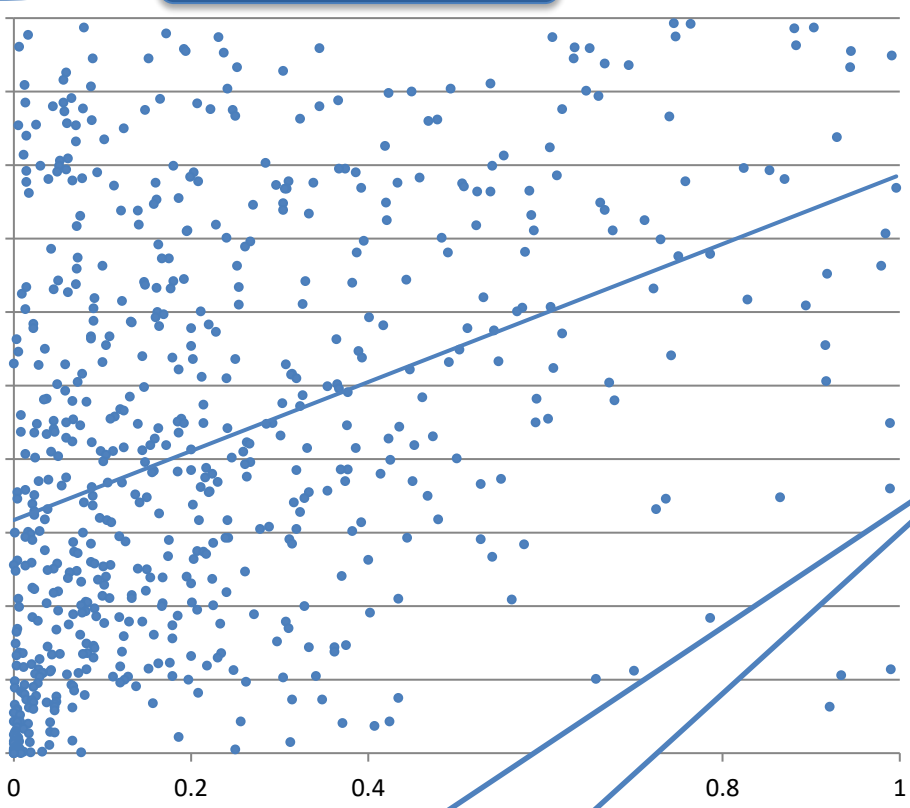
⇒ 弱い正の相関



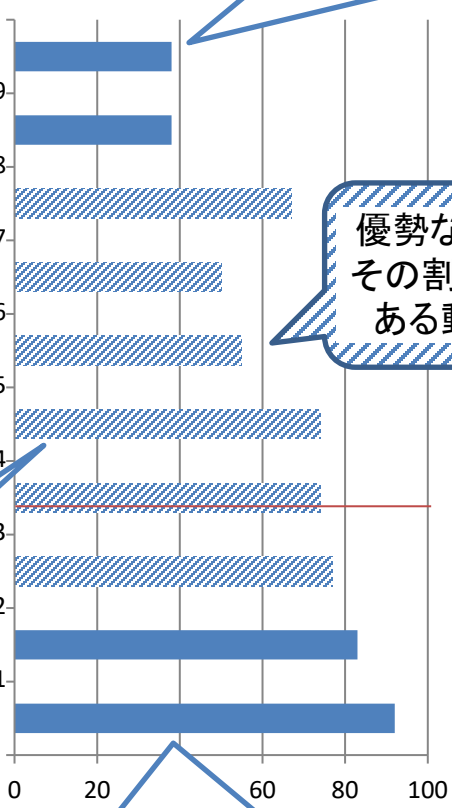
実験1: 格の出現率と語順

相関係数: 0.391

⇒ 弱い正の相関



「がをに」語順の割合



ニ格のみの割合

動詞の例と図中の縦軸,横軸の値
例える:0.993,0.745, 見立てる:0.994,0.657
代える:0.933,0.252, 生かす:0.916:0,056

優勢な語順であっても
その割合が80%以上で
ある動詞は4割のみ

平均: 0.328

動詞の例と図中の縦軸,横軸の値
近付ける:0.487,0.326, 詰める:0.474,0.214
挙げる:0.448,0.026, 適用する:0.421,0.266

動詞の例と図中の縦軸、横軸の値
発揮する:0.098,0.001, 感じる:0.043,0.256
及ぼす:0.015,0.312, もたらす:0.000,0.000

実験1が示唆する内容

A) 動詞によらず二重目的語構文の“基本語順”は「がにを」である ⇒ *No!*
   - ただし、全体で多い語順はガニヲ語順（=67.2%）[Hoji'85]
   - 優勢な語順であってもその割合が80%以上である動詞は4割のみ

B) 基本語順は動詞タイプによって異なる[Matsuoka'03] ⇒ *No?*

C) 省略されにくい格は動詞の近くに出現しやすい ⇒ *Yes!*

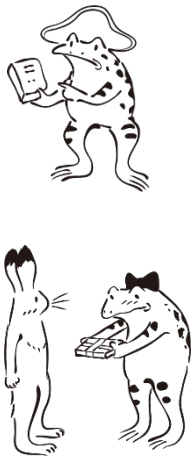
基本語順と動詞タイプ

- 動詞タイプ: 使役・起動交替を適用した時に
  - Showタイプ: ニ格名詞が主語となる

    彼に 本を 見せた ⇒ *彼が 見た*

  - Passタイプ: ヲ格名詞が主語となる

    本を 彼に 渡した ⇒ *本が 渡った*



- 基本語順との関係に関する仮説[Matsuoka'03]
  - Showタイプの動詞は「がにを」が基本語順
  - Passタイプの動詞は「がをに」が基本語順

実験2: 動詞タイプごとの語順の割合

Show タイプ			Pass タイプ					
動詞	用例数	「がをに」率	動詞	用例数	「がをに」率	動詞	用例数	「がをに」率
知らせる	372927	0.522	戻す	146145	0.771	落とす	129406	0.351
預ける	77671	0.399	泊める	2551	0.748	漏らす	15193	0.332
言付ける	176	0.386	くるむ	1277	0.603	浮かべる	52038	0.255
諭す	329	0.325	伝える	216113	0.522	向ける	319614	0.251
見せる	163172	0.301	乗せる	111757	0.496	残す	212591	0.238
被せる	23324	0.256	届ける	71428	0.491	埋める	50615	0.223
教える	144282	0.235	並べる	54949	0.481	混ぜる	39220	0.200
授ける	9294	0.186	返す	43045	0.448	当てる	203360	0.185
浴びせる	17714	0.177	ぶつける	66620	0.436	掛ける	164960	0.108
貸す	54359	0.118	付ける	529550	0.368	重ねる	74584	0.084
着せる	26791	0.113	渡す	154272	0.362	建てる	32203	0.069
		平均 0.274						平均 0.365

- 基本的に各タイプの動詞は[Koizumi+'04]が心理実験で使用した動詞を使用（ただし、out: はかせる, in: 知らせる&言付ける）
- 平均値の差を並べ替え検定で検定⇒有意差なし（p=0.177）
- 仮説Bを支持しない（cf. [Miyamoto+'02,Koizumi+'04]も同様の結論）

検証する仮説(続)

D) 基本語順は二格名詞の意味役割や有生性により異なる[Matsuoka'03, Ito'07]

– 二格名詞が有生性をもつ所有者(着点)を表す場合、有生性を持たない場所(着点)を表す場合よりも「がにを」語順をとりやすい

先生に 本を 返却した ⇔ *本を 学校に 返却した*

E) 省略されにくいヲ格名詞、ニ格名詞は動詞の近くに出現しやすい

脳に … 与える ＜ 影響を … 与える ⇒ 基本語順: *脳に 影響を 与える*

身に … 付ける ＞ 能力を … 付ける ⇒ 基本語順: *能力を 身に つける*

実験3: ニ格のカテゴリと語順

- カテゴリはJUMAN 辞書のカテゴリ情報をもとに決定
  - 用例数が100万を超える8つのカテゴリが対象

カテゴリ	用例数	「がをに」率	出現頻度の高い名詞の例
場所-機能	1376990	0.499	下、横、外、中、方向、...
動物-部位	1483885	0.441	手、身、頭、肌、胸、...
人	5511281	0.387	友達、人、市長、私、先生、...
人工物-その他	2751008	0.372	パソコン、ファイル、風呂、写真、本、...
場所-施設	1618690	0.342	部屋、店、所、冷蔵庫、学校、...
場所-その他	2439188	0.341	場所、世界、サイト、位置、前面、...
数量	1100222	0.308	図、何、表、半分、値、...
抽象物	10219318	0.307	ブログ、心、リスト、視野、元、...
合計	26500582	0.353	

- カテゴリにより「がをに」語順の割合が異なることを確認
  - ただし、ニ格が有生名詞（cf. 人）の場合「がにを」語順をとりやすいという仮説と合致しない
  - 意味役割を限定できていないことが原因である可能性

実験4: ニ格の意味役割と語順

- [滝本+'15]を参考に下記を抽出(e.g.「本を学校に返却」「先生に本を返却」)
  - ヲ格名詞のJUMAN辞書におけるカテゴリが『人工物-その他』
  - ニ格名詞のカテゴリが『人』である用例の異なり数と、『場所-施設』である用例の異なり数がいずれも100以上である126動詞
- ニ格名詞のカテゴリごとに語順の割合を調査
  - 語順に有意な差があった動詞94個中、『人』である場合の方が「がにを」語順をとりやすい動詞は64個、その逆は動詞は30個
  - 仮説Dを支持する結果

動詞	「がをに」率 ニ格が『人』|『場所-施設』	関連する出現頻度の高い用例の抜粋とその頻度
据える	0.852 > 0.203	アクセルを主役に-:8、ターゲットを女性に-:6、ホテルにカメラを-:4
展示する	0.851 > 0.359	館に美術品を-:46、館に資料を-:46、製品を一同に-:39、商品を一同に-:26
返還する	0.915 > 0.491	許可証を市長に-:253、登録証を市長に-:149、営業店にレンタカーを-:51
運ぶ	0.244 < 0.577	荷物を部屋に-:250、荷物を部屋に-:188、子供に餌を-:92、客に料理を-:73
落とす	0.173 < 0.595	携帯をトイレに-:486、電話をトイレに-:282、トイレに携帯を-:168、私に爆弾を-:16
郵送する	0.224 < 0.664	合格者に書類を-:180、書類を自宅に-:156、用紙を客に-:61、合格者に証書を-:48

検証する仮説(続)

D) 基本語順は二格名詞の意味役割や有生性により異なる [Matsuoka'03, Ito'07]

– 二格名詞が有生性をもつ所有者(着点)を表す場合、有生性を持たない場所(着点)を表す場合よりも「がにを」語順をとりやすい

先生に 本を 返却した ⇔ *本を 学校に 返却した*

E) 省略されにくいヲ格名詞、ニ格名詞は動詞の近くに出現しやすい

脳に … 与える ＜ 影響を … 与える ⇒ 基本語順: *脳に 影響を 与える*

身に … 付ける ＞ 能力を … 付ける ⇒ 基本語順: *能力を 身に つける*

ヲ格名詞、ニ格名詞、動詞の組み合わせごとの「がをに」語順の割合の調査

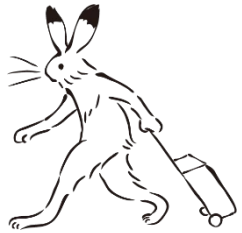
- ヲ格名詞と動詞、ニ格名詞と動詞、それぞれの共起度合いと、「がをに」語順の関係を調査
  - 名詞と動詞の共起度合いの尺度には、以下で定義される正規化自己相互情報量(NPMI) を使用（常に共起⇒1、独立⇒0、共起しない⇒-1）

$$\mathrm{NPMI}(n_{ニ}, v) = \frac{\mathrm{PMI}(n_{ニ}, v)}{-\log(p(n_{ニ}, v))} \quad \text{ただし、} \mathrm{PMI}(n_{ニ}, v) = \log\frac{p(n_{ニ}, v)}{p(n_{ニ})p(v)}$$

  - ニ格名詞と動詞のNPMIとヲ格名詞と動詞のNPMIの差を算出
  - 仮説Eが正しいならばニ格名詞と動詞のNPMIの方が大きな値となる場合、「がをに」語順の割合が大きくなるはず
- 500回以上出現した組み合わせ2417個を調査

慣用表現等への対処

- 基本的に慣用表現はまとまって出現
  - 全用例を人手でチェック
  - e.g.「～を手に入れる」、「世界を股にかける」



- 機能動詞構文も影響する可能性もあるが保留
  - yes:「実行に移す」、???:「影響を与える」

- その他、不適切な用例を削除（2417⇒2302）
  - 「～を友達に知らせる」: SNSの定型表現
  - 「～を最大限に活用する」

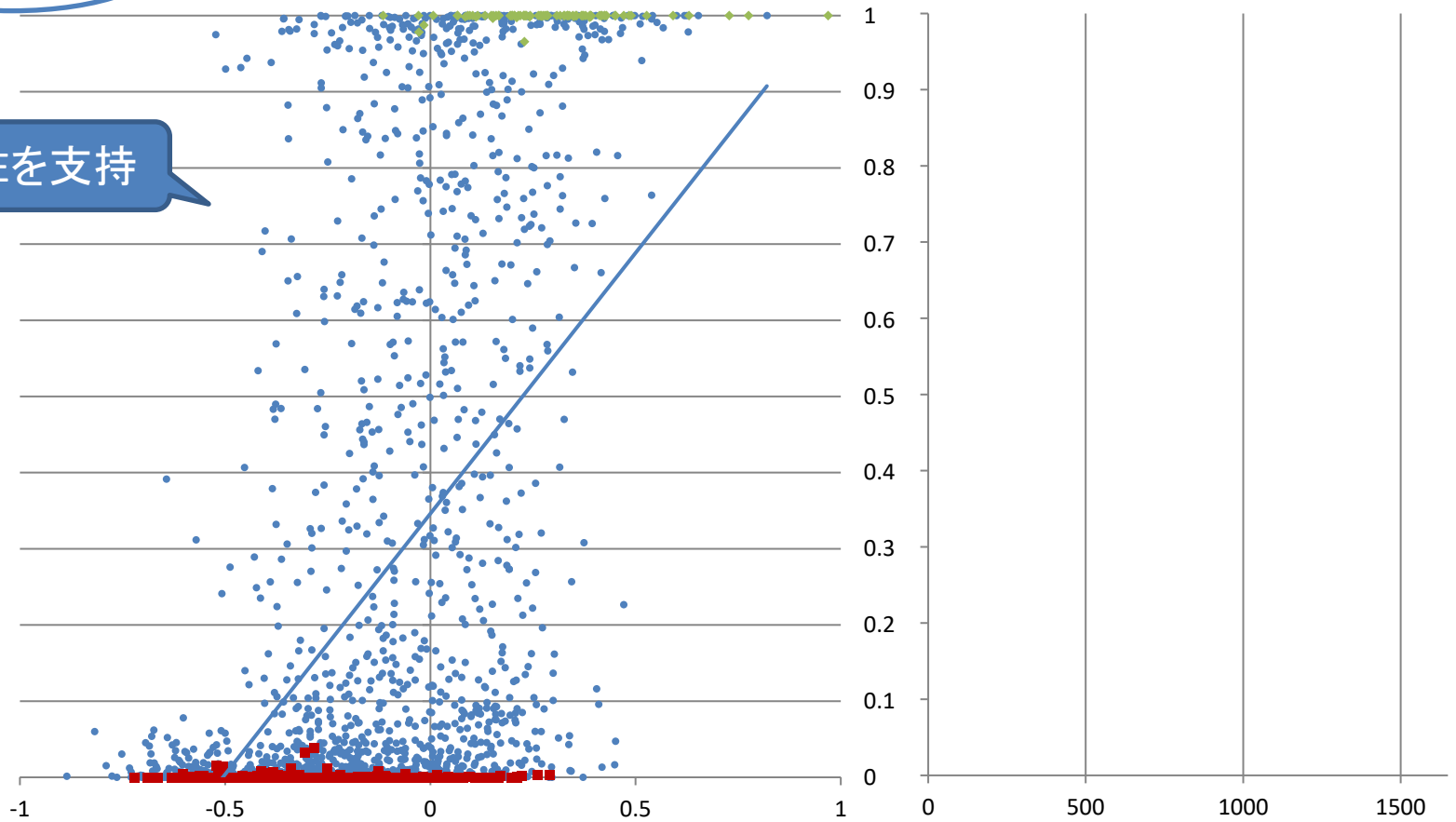
実験5: ヲ格名詞、ニ格名詞、動詞の組み合わせごとのNPMIの差と語順の関係

相関係数: 0.567
(0.513)

⇒ 正の相関

1814例 (+慣用表現: ニヲ: 404例、ヲニ: 84例)

仮説Eを支持



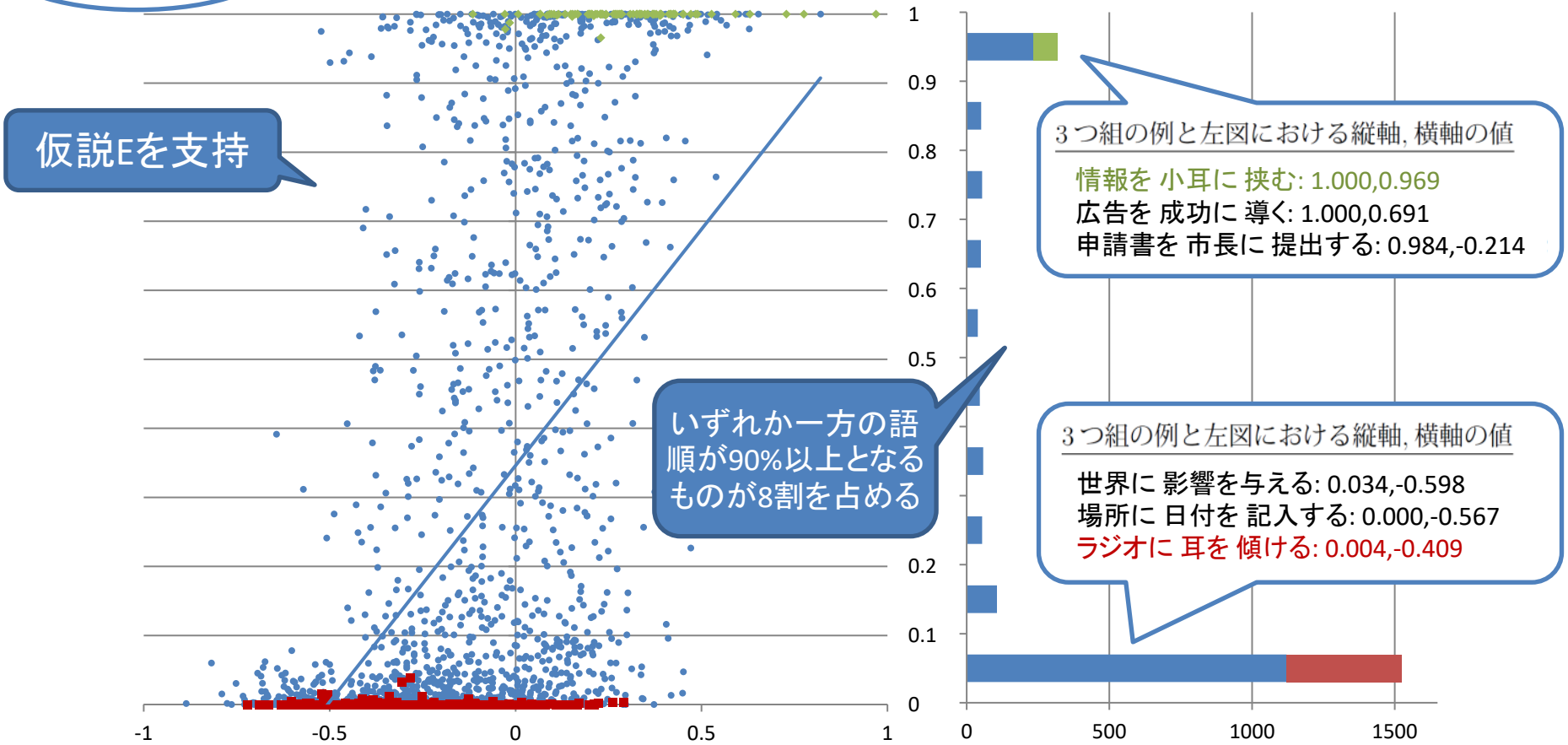
実験5: ヲ格名詞、ニ格名詞、動詞の組み合わせごとのNPMIの差と語順の関係

相関係数: 0.567
(0.513)

⇒ 正の相関

1814例 (+慣用表現: ヲニ: 404例、ニヲ: 84例)

「がをに」語順の割合



3つ組の例と左図における縦軸, 横軸の値

情報を 小耳に 挟む: 1.000,0.969
広告を 成功に 導く: 1.000,0.691
申請書を 市長に 提出する: 0.984,-0.214

3つ組の例と左図における縦軸, 横軸の値

世界に 影響を与える: 0.034,-0.598
場所に 日付を 記入する: 0.000,-0.567
ラジオに 耳を 傾ける: 0.004,-0.409

もくじ

1. 研究の背景と動機
2. 日本語二重目的語構文と語順
3. 分析に使用する用例の収集
4. 大規模コーパスに基づく基本語順の分析
5. まとめと今後の展望

実験結果が示唆する結論

- 動詞の6割は「にを」、4割は「をに」語順が優勢[≠A]
- ただし、優勢な語順の割合が80%以上である動詞は4割のみ
- Pass/Showタイプの違いは基本語順と無関係[≠B]
- 省略されにくい格は動詞の近くに出現する傾向[=C]
- 二格名詞が着点を表す場合、有生性を持つ場合の方が「にを」語順をとりやすい[=D]
- 対象の動詞と高頻度に共起するヲ格名詞、二格名詞は動詞の近くに出現しやすい[=E]
- ヲ格名詞、二格名詞、動詞の3つ組が与えられた場合、一方の語順が90%以上を占める組合せが8割

手法の限界と今後の展望

- 一部の動詞について正しく分析できていない可能性
  - ヲ格が副助詞「は」や「も」を用いて表現されやすい動詞
  - 連体節などを含む長い項を取りやすく項と動詞の間の係り受け関係に曖昧性が存在することが多い動詞
- コーパス中の出現割合と強く関係するとの仮定
  - 関係はあると考えられるもののどの程度強く関係するかは不明
  - 脳科学などのより直接的な検証も行うことが望ましい
- 語順と意味役割の関係の分析
  - 意味役割の分析において語順が手掛かりとなる可能性
- 大規模データに基づくテキスト中の語順分析
  - 語順に影響を与える既知の要因（項の長さ、新情報か否か、etc.）が実際の語順にどのように寄与しているかの分析等