Grammar Patternsに基づく動詞項構造構文の自動抽出システムの構築 (A work-in-progress report)

福田航平（東京外国語大学大学院）

1. 背景と問題意識

動詞項構造構文（VACs）とは

VACs = Verb-argument constructions
動詞とその後ろに続く文法パターンのこと（動詞とその項構造パターン）

- **例：「give」という動詞の場合**
  - I **gave** him a book. → 「動詞 + 間接目的語 + 直接目的語」というパターン
  - I **gave** a book to him. → 「動詞 + 名詞 + 前置詞to + 名詞」というパターン

- **VACsの例**

構造	例文
動詞 + 間接目的語 + 直接目的語 (V n n-obj)	I **bought** him lunch. / I'll **show** you some pictures.
動詞 + to不定詞 (V to-inf)	She **agreed** to help. / Koji **wants** to read the book.
動詞 + that節 (V that)	I **think** that this is true. / I **hope** this information helps.
動詞 + 前置詞in + 名詞 (V in n)	We **participated** in the activities. / Miki **lives** in this town.

なぜVACsか

用法基盤・構文アプローチSLA研究において、VACsが分析ユニットの中心

（e.g Ellis & Ferreira-Junior, 2009a, b; Ellis et al., 2014; Ellis, et al., 2016; Römer et al., 2018; Römer, 2019: among others）

1. **動詞は文の構造と意味を決定する中心的役割を担う**ため、動詞を中心とした文法パターンであるVACsの習得は言語習得において重要
   - 日本の英語教育でも文法参考書で「動詞の語法」として重視
   - 用法基盤構文文法（Goldberg, 2006, 2019）では動詞を中心とする項構造構文が分析の中心

2. **VACsの知識が学習者の習熟度を予測できる因子**
   - 学習者が使用するVACsの種類やその数、動詞とVACsの結びつきが習熟度の予測に有効な指標となる（Kyle & Crossley, 2017; Hwang & Kim, 2022; Sung & Kyle, 2025）

用法基盤・構文アプローチ：「構文（construction）」の習得に関わる認知メカニズムを重視し、インプットの統計分布情報（frequency, contingency）が習得に重要であるとする立場（see Ellis & Wulff, 2025a, b）

本発表の目的

- **英語には動詞を中心とした文法パターン（動詞項構造構文；VACs）が様々ある**
  e.g.) 動詞 + to不定詞 / 動詞 + that節 / 動詞 + 名詞 + of + 名詞

- 大規模かつ体系的に、学習者の言語使用の分析や教材研究をするために、**一般的な英文法で扱われるVACsを網羅した分析**を行えるようにしたい

- 英語テキストから**網羅的に様々な種類のVACsを自動で特定するシステムの構築**が必要

- 網羅的な文法パターンのリソースであるGrammar Patternsに基づいて、VACsを網羅的に特定するシステムの構築を試みた

VACs抽出システムの必要性

コーパスからVACsを網羅的に自動抽出・分析できるツールがあれば、大規模かつ体系的に

1. 学習者コーパスを使ってL2英語学習者のVACs知識の習得・発達を分析できる
2. 学習教材が提供するインプットをVACsの観点から体系的に分析できる
3. 大規模コーパスから母語話者が使用する実際のVACs分布を体系的に記述・分析できる

本研究では、様々な英語のVACsを網羅的に自動抽出可能なシステムの構築を目指す

2. 先行研究とギャップ

先行研究：VACs抽出の方法

1. **マニュアル**（Ellis & Ferreira-Junior, 2009a, b; Park & Sung, 2023）
   - 手作業で構文パターンのラベルを付与
2. **統語依存フレーム**（Kyle & Crossley, 2017, Römer, 2019）
   - 動詞を中心とする統語依存関係（syntactic dependency）をのパターンをそのまま構文フレームとして利用する
3. **ルールベース**（Römer et al., 2015; Hwang & Kim, 2022）
   - 品詞タグや統語依存タグを利用して特定のVACsを識別するための一連の明示的ルールを作成する
4. **機械学習による分類**（Huang et al., 2021）
   - 特徴量から構文パターンを確率的に予測する機械学習分類器を作成する
5. **ファインチューニング**（Sung & Kyle, 2024a, b）
   - マニュアルで正解データを作成し、Transformerベースのファインチューニング

先行研究：VACs抽出の方法

1. Römer et al. (2015)：ルールベース

- Grammar Patternsに基づいた動詞+前置詞パターン（e.g. “V about n”, “V in n”）約20種類をRASPパーザー解析済みBNC-XMLから抽出
- 後続するN.C. EllisやU. Römerの研究（e.g. Ellis et al., 2016）はこれをもとにしているため、分析対象が動詞+前置詞パターンに偏重している
- **全体精度：Precision 0.78%・Recall 0.53%・F1-score 0.612**

2. Huang et al. (2021)：機械学習分類

- BNCから6,133文の母語話者データとEFCAMDATから1,000文の学習者英語データを組み合わせ、Maximum Entropyモデルを用いて構文パターン分類器を作成
- 動詞トークンの特徴量（語・品詞・依存関係・語埋め込みなど）から各構文パターンの確率を計算し、49種類の構文タイプを識別可能なシステムを構築
- **全体精度：84.2%（accuracy）**

先行研究：VACs抽出の方法

3. Kyle (2016), Kyle & Crossley (2017) (TAASSC) ：統語依存フレーム

- VACsを利用した非常に多様な**統語的洗練性指標を算出するツール**
- 統語依存関係を利用して、動詞とそれに直接従属する要素をVACフレームとしている

  e.g.) I **think** that he **did** that.

  → `nsubj-v-ccomp` / `nsubj-v-dobj`

- 動詞とその直接従属要素を依存関係ラベル（ccomp／xcomp など）でひとまとめにするため、一般的な英文法で扱われる「V＋that節／V＋wh節／V＋to不定詞」のような補文パターンと一対一には対応しない

  e.g.) I **think** that he did that. → `nsubj-v-ccomp`

  I **wonder** whether he did that. → `nsubj-v-ccomp`

先行研究：VACs抽出の方法

4. Hwang and Kim (2022)：ルールベース

- 依存関係と語彙情報の両方を使って手動で作成したルールを使用して、**11種類の構文**を識別するシステムを開発
- **全体精度：Precision 0.86 / Recall 0.82 / F1-socre 0.82**

5. Sung and Kyle (2024a, b)：ファインチューニング

- **9種類の構文**をマニュアルアノテーションしたデータを作成し、 RoBERTa-baseでファインチューニング
- **全体精度：F1-score 0.912（L1） / 0.928（L2 Spoken） / 0.915（L2 Written）**

※ 上記で挙げている先行研究は、対象としている構文パターンやデータセットはそれぞれ異なるため単純な比較は不可

先行研究の問題点

- **特定の構文パターンのみに偏重**
  - Römer et al. (2015)とそれに後続する研究（e.g. Ellis et al., 2016）は、動詞+前置詞パターンに偏重
  - 逆にHuang et al. (2021)は前置詞は全て1つにまとめられ、前置詞タイプを区別しない
- **対象とする構文の種類が少ない**
  - Hwang and Kim (2022)：**11種類の構文が対象**
  - Sung and Kyle (2024a, b)：**9種類の構文が対象**
- **一般的な英文法で扱われるような文法パターンに一致しない部分がある**
  - 教育・教材研究にも活用するためには、一般的な英文法で扱われるパターンとの整合性も重要

以下の要件を満たすVACs抽出システムの作成を試みる

❶網羅性　❷一般的な英文法との整合　❸高精度

3. リソース：
Collins COBUILD Grammar Patterns

Collins COBUILD Grammar Patterns I: Verbs (Francis, et al., 1996)

Grammar Patterns = コーパスデータに基づいた文法パターン辞書

- COBUILD Grammar Patternsは、コーパスデータからボトムアップで文法パターンとそのパターンで使用される語彙を特定し、**網羅的にリストアップ**したもの（Hunston, 2000）
- COBUILD Grammar Patternsを**英語学習に役立てる**ことのできるリソースとして開発
- VACsの網羅的リソースとして利用している研究がある（e.g. Römer et al., 2015; Hunston & Su, 2017; Hunston, 2019, 2025; Perek & Patten, 2019)

例) “V n *of* n”パターン

	Verb group	noun group	*of*	noun group/-ing clause
Subject	**Verb**	**Object**	**Adjunct**	
The settlement	absolved	the company	of	all criminal responsibility.
	Clear	your mind	of	other thoughts.
They	suspected	him	of	doing away with Beryl.

パターン内の動詞を意味グループに分類

- **The ‘rob’ and ‘free’ group:** *defraud, denude, deprive, strip, cure, divest, rid, relieve* etc.
- **The ‘inform’ group**: *advise, assure, convince, inform, persuade, notify, remind, warn* etc.
- **The ‘acquit’ and ‘convict’ group**: *accuse, acquit, convict, suspect* etc.

4. VACs抽出システムの構築

VACs抽出システムの構築

Grammar Patternsの例文を正解データとし、RoBERTa-baseのFine-tuningでVAC抽出モデルを構築した

1. **Grammar Patternsに基づいて68種類の動詞パターンを抽出対象に採用**
2. **Grammar Patternsの例文データに手作業で動詞パターンのラベルをアノテーション**
   - 1文につき1つではなく、文中に出てくる抽出対象となる動詞全てにラベルを付与
   - CONLL-U Formatで動詞トークンのMISC欄にラベルを付与
3. **RoBERTa-baseでNERモデルとしてFine-tuningを実行**（方法は基本的にSung & Kyle, 2024bに倣った）

CONLL-U Formatでのラベル付与の例

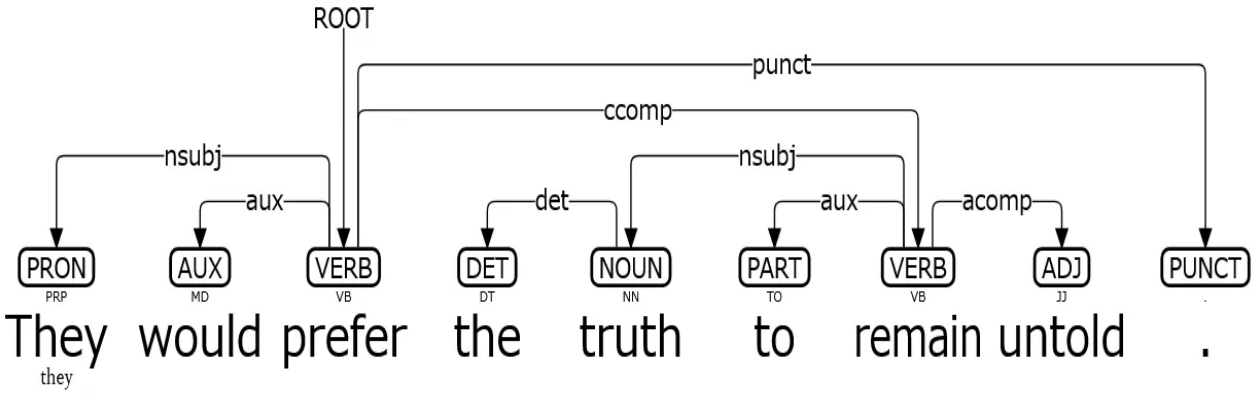
ID	FORM	LEMMA	UPOS	XPOS	FEATS	HEAD	DEPREL	MISC
1	He	he	PRON	PRP	Case=Nom | Gender=Masc | Number=Sing | Person=3 | PronType=Prs	2	nsubj	–
2	pleaded	plead	VERB	VBD	Tense=Past | VerbForm=Fin	0	ROOT	Pattern=V_to-inf
3	to	to	PART	TO	–	4	aux	–
4	speak	speak	VERB	VB	VerbForm=Inf	2	xcomp	Pattern=V_with_n
5	with	with	ADP	IN	–	4	prep	–
6	me	I	PRON	PRP	Case=Acc | Number=Sing | Person=1 | PronType=Prs	5	pobj	–
7	privately	privately	ADV	RB	–	4	advmod	–
8	.	.	PUNCT	.	PunctType=Peri	2	punct	–

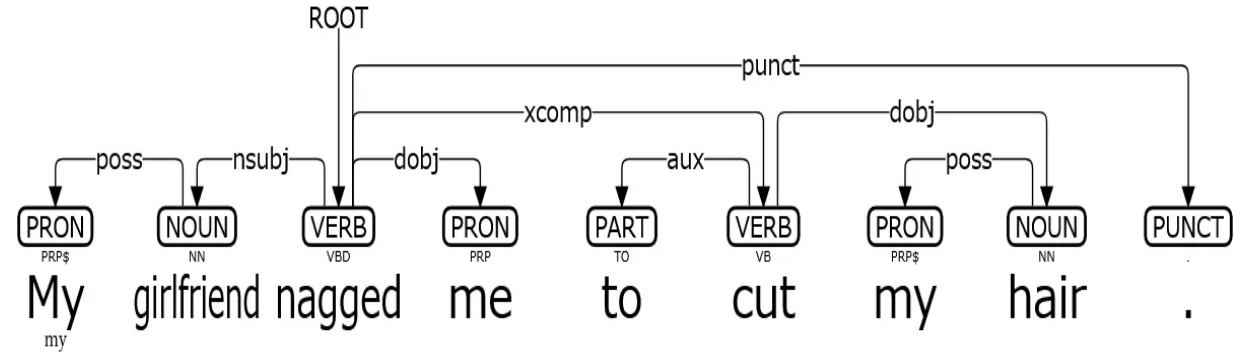
VACs抽出システムの構築

比較として品詞・統語依存タグをベースにVACsを識別するルールベースのシステムも作成

spaCyの`en-core-web-trf`モデルの解析結果を利用し、統語依存関係に基づく`DependencyMatcher`を使ってマッチングルールを様々なVACsに対し定義した

マッチングルールの例：V_n_to-infパターン





マッチングルール１：

① 動詞がある（アンカー動詞とする）
② アンカー動詞に ccompの依存関係で依存している動詞トークンがある（ comp_token）
③ comp_tokenにnsubjの依存関係で依存している名詞・代名詞がある
④ comp_tokenに依存している不定詞マーカーのtoがある（TAG = TO)

マッチングルール２：

① 動詞がある（アンカー動詞とする）
② アンカー動詞に xcompの依存関係で依存している動詞トークンがある（ comp_token）
③ アンカー動詞に dobjの依存関係で依存している名詞・代名詞がある
④ comp_tokenに依存している不定詞マーカーのtoがある（TAG = TO)

VACs抽出システムの構築：抽出例（イメージ）

テキスト例

The new employee seemed nervous during his first presentation. He talked about his previous experience and described what he wanted to achieve. The manager found his ideas interesting and encouraged him to develop them further.

ラベル付与

The new employee **seemed** nervous during his first presentation. He **talked** about his previous experience and
V_ADJ **V_ABOUT_N**
described what he **wanted** to **achieve**. The manager **found** his ideas interesting and **encouraged** him to **develop**
V_WH **V_TO_INF** **V_N_OBJ** **V_N_ADJ** **V_N_TO_INF** **V_N_OBJ**
them further.

抽出

```
[('seemed', 'V_ADJ'), ('talked', 'V_ABOUT_N'), ('described', 'V_WH'),
('wanted', 'V_TO_INF'), ('achieve', 'V_N_OBJ'), ('found', 'V_N_ADJ'),
        ('encouraged', 'V_N_TO_INF'), ('develop', 'V_N_OBJ')]
```

5.システムの評価と比較

VACs抽出システムの構築

正解データのデータ数

センテンス数	2867
トークン数	42352
ラベルつきトークン数	5771

80%を訓練データ
10%を検証データ
10%をテストデータ

全体的な精度

ファインチューニングモデル

Metrics	Score
Precision	0.925
Recall	0.947
F1-score	**0.936**

ルールベースモデル

Metrics	Score
Precision	0.931
Recall	0.909
F1-score	0.920

先行研究との比較

	Römer et al. (2015)	Huang et al. (2021)	Hwang & Kim (2022)	Sung & Kyle (2024, a, b)	本研究
手法	ルールベース	機械学習分類	ルールベース	ファインチューニング	**ファインチューニング**
精度	F1 = 0.612	0.842 (accuracy)	F1 = 0.82	F1 = 0.912 (L1データ)	**F1 = 0.936**
構文種類数	20種類	49種類	11種類	9種類	**68種類**

※ 上記で挙げている先行研究は、対象としている構文パターンやデータセットがそれぞれ異なるため、単純な精度の比較はできない

6. 使用例：教科書のVACs頻度分析

分析の内容

分析対象

- ある1つのシリーズの現行検定英語教科書を中1~高3までを通して分析（6冊）
- メインレッスンの本文とキーセンテンスの部分のみを抽出し、分析対象とした

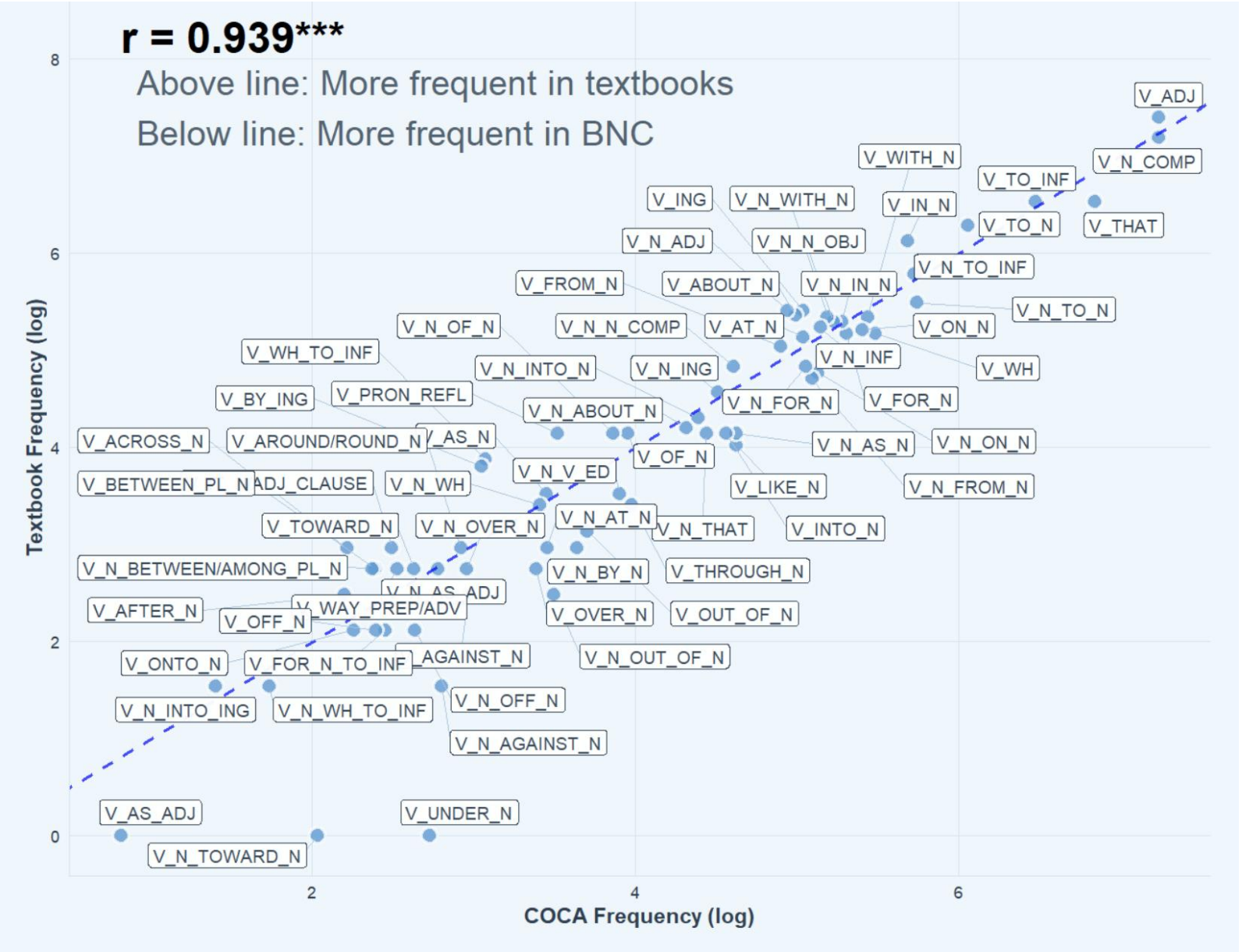
分析内容

- VACsの使用頻度の点で、教科書の英語は自然な英語使用と比較してどうか。
  - 教科書のVACsの頻度は参照コーパス（BNC）と相関するか
  - 教科書で有意に過剰／過少使用されているVACsはあるか

対象教科書データの語数

VACs	中1	中2	中3	高1	高2	高3	合計
総語数	2080	3090	3797	8151	8881	11232	37231
対象VACs総頻度	265	375	461	971	1119	1330	4520

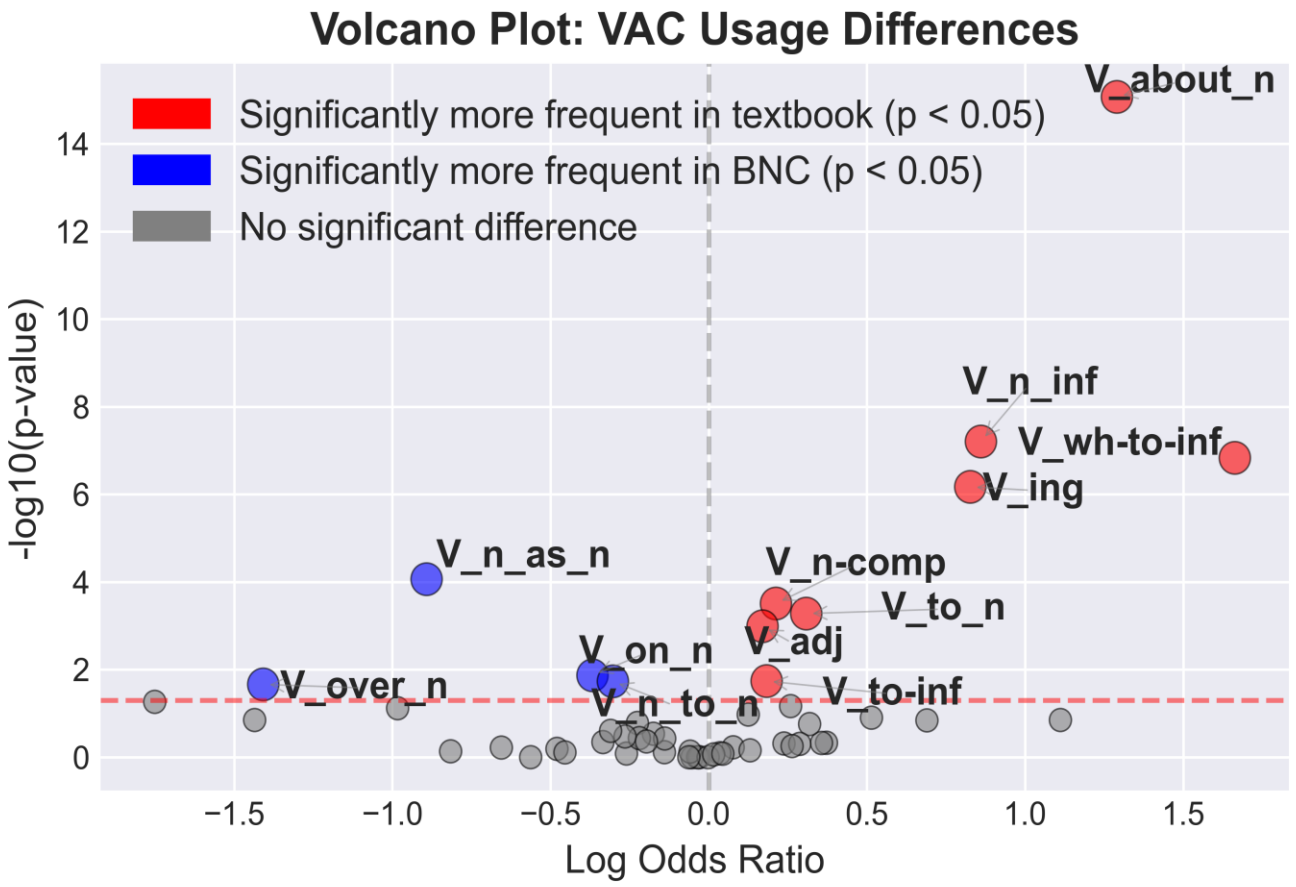
教科書と参照コーパスのVACs頻度の比較



教科書におけるVACsの頻度と参照コーパス（BNC）におけるVACsの頻度は強い相関を示している

VACsの使用頻度の点から見て、教科書は自然な英語使用を反映しているといえる

教科書で過剰／過少使用されているVACs



Notes:

- 縦軸：対数化したp値（赤線より上が $p < 0.05$）
- 横軸：効果量（Log Odds Ratio）
- Fisher's exact testで検定

過剰使用（overuse）

- **V_wh-to-inf** (e.g.) Now we **know how to** negotiate.
- **V_about_n** (e.g.) Let's **think about** it together.
- **V_n_inf** (e.g.) Miki **helped me cook** lunch.
- **V_ing** (e.g.) Did you **enjoy hiking** this morning?
- **V_to_n** (e.g.) I **go to dance lessons.**
- **V_n-comp** (e.g.) I **am a dancer.**
- **V_adj** (e.g.) It **was great.**
- **V_to-inf** (e.g.) I **want to go** with you.

過少使用（underuse）

- **V_over_n** (e.g.) We **argued over household chores.**
- **V_n_as_n** (e.g.) I **consider him as a friend.**
- **V_on_n** (e.g.) He is **concentrating on his studies.**
- **V_n_to_n** (e.g.) I **lent some money to** my father.

7. 今後・まとめ

今後

1. L1コーパスデータを使って、もっと**自然な英語に対するVACs抽出システムの精度を検証**する

2. **学習者データにVACsのアノテーションを付した正解データを用意**し、
   - モデル訓練に学習者データを含める
   - 学習者データに対する精度を検証する

3. **どのような場合に精度が高い／低いのか検証する**

4. **上のような検証を経たのちに、誰にでも利用可能なパッケージとして公開する**

今後

5. 大規模母語話者コーパスに適用し、VACsデータベースを作成する

- 教育・研究に利用な可能なデータベースを作成・公開

6. 学習者コーパスに適用し、学習者のVACs使用を分析する

- 構文多様性（Constructional diversity）の測定（既存のツールとの比較）
- 習熟度が上がるにつれて使用頻度が変わるVACsの特定
- 各VACs内の動詞分布の観点から学習者のVACs使用の特徴を分析

まとめ

- **今回作成を試みたVACs抽出システムの特徴**
  1. **網羅性**：68種類を対象
  2. **一般的な英文法との整合**：Grammar Patterns準拠
  3. **高精度**：F1 = 0.936

- **例文データをモデル学習に使用しているため、もっと一般的なテキストに汎化できるか検証の必要あり**