

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）

補足資料やスライドやコードを載せた
GitHubページのアクセスQRコード



01

背景と問題意識

06

使用例：教科書分析

02

先行研究とギャップ

07

今後・まとめ

03

Collins COBUILD Grammar
Patterns

04

VACs抽出システムの構築

05

システムの評価と比較

補足資料やスライドやコードを載せた
GitHubページのアクセスQRコード



1. 背景と問題意識

補足資料やスライドやコードを載せた
GitHubページのアクセスQRコード



本発表の目的・概要

- 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 を特定するシステムの構築を試みた

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

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; Huang et al., 2021; 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）

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) 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パーサー解析済み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）

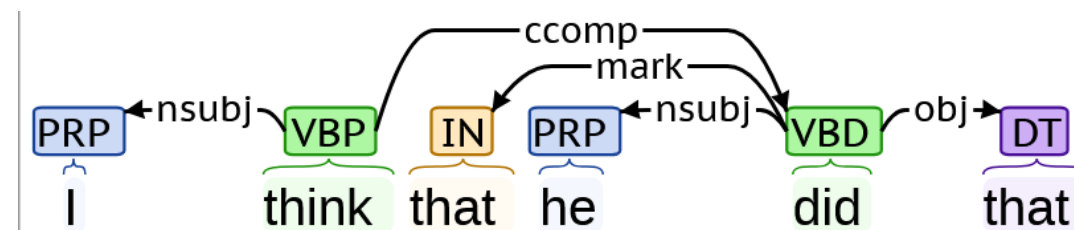
先行研究：VACs抽出の方法

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 / mark-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-score 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

①網羅性 ②一般的な英文法との整合 ③高精度

3. リソース： Collins COBUILD Grammar Patterns

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

15

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われる語彙を特定し、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の例文データを使用し、ファインチューニング（RoBERTa-base）でモデルを構築

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でファインチューニングを実行（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

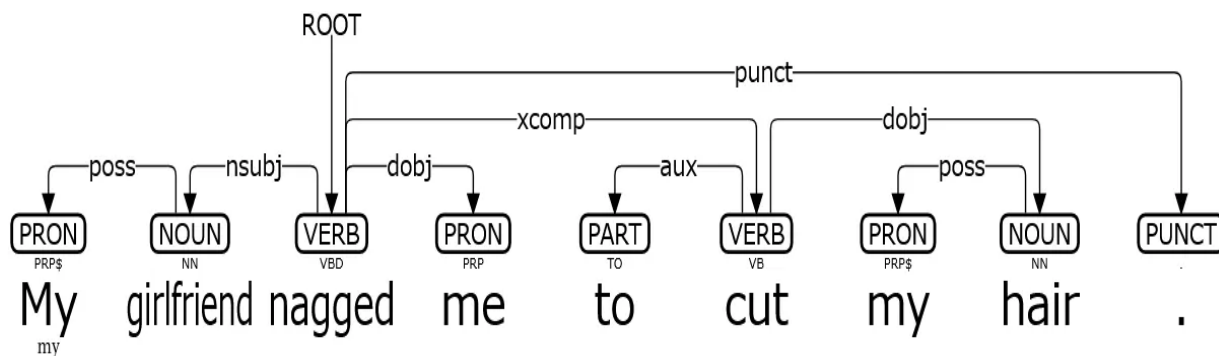
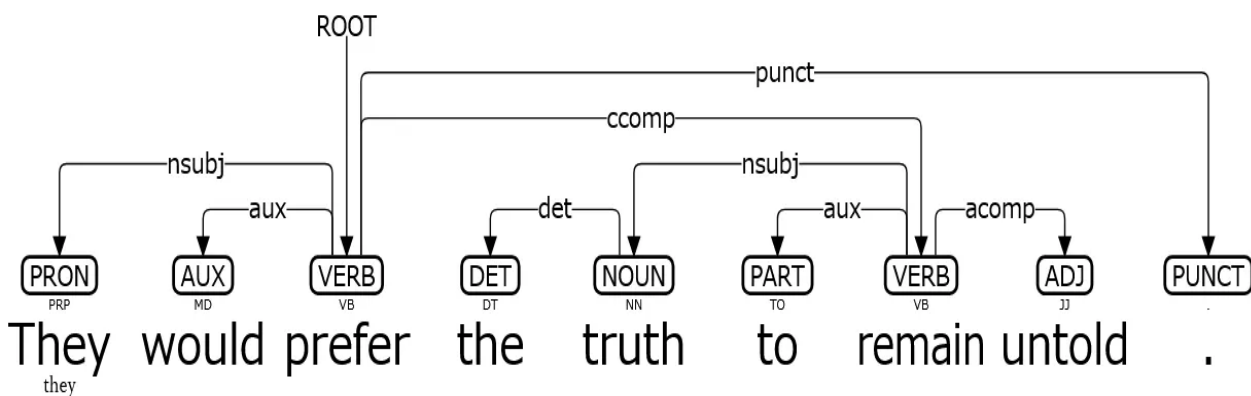
ID	FORM	LEMMA	UPOS	XPOS	FEATS	HEAD	DEPREL	MISC
1	He	he	PRON	PRP	Case=Nom Gender=Masc Number=Sing Person=3 PronType=Prs	2	nsubj	–
2	pleaded	plead	VERB	VBD	Tense=Past VerbForm=Fin	0	ROOT	Pattern=V_to-inf
3	to	to	PART	TO	–	4	aux	–
4	speak	speak	VERB	VB	VerbForm=Inf	2	xcomp	Pattern=V_with_n
5	with	with	ADP	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 doobj 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の to がある（TAG = TO）

VACs抽出システムの構築：抽出例

```
1 text = """
2 The new employee seemed nervous during his first presentation.
3 He talked about his previous experience and described what he wanted to achieve.
4 The manager found his ideas interesting and encouraged him to develop them further.
5 """
6
7 df, data_list = vac_identification(text, vac_nlp)
8
9 display(df)
```

	Token	VAC_label
0	The	
1	new	
2	employee	
3	seemed	V_ADJ
4	nervous	
5	during	
6	his	
7	first	
8	presentation	
9	.	

10	He	
11	talked	V_ABOUT_N
12	about	
13	his	
14	previous	
15	experience	
16	and	
17	described	V_WH
18	what	
19	he	
20	wanted	V_TO_INF
21	to	
22	achieve	V_N_OBJ
23	.	

24	The	
25	manager	
26	found	V_N_ADJ
27	his	
28	ideas	
29	interesting	
30	and	
31	encouraged	V_N_TO_INF
32	him	
33	to	
34	develop	V_N_OBJ
35	them	
36	further	
37	.	

対象となるトークンに
VACsのラベルが自動
で付与される

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	80%を訓練データ 10%を検証データ 10%をテストデータ
トークン数	42352	
ラベルつきトークン数	5771	

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

VACs抽出システムの構築


```
1 text = """
2 He has been playing the guitar for 10 years.
3 Don't take me for a fool.
4 """
5
6 df, data_list = vac_identification(text, vac_nlp)
7 display(df)
```



0	He	
1	has	
2	been	
3	playing	V_N_OBJ
4	the	
5	guitar	
6	for	
7	10	
8	years	
9	.	

10	Do	
11	n't	
12	take	V_N_FOR_N
13	me	
14	for	
15	a	
16	fool	
17	.	

```
1 text = """
2 I saw the man crossing the street.
3 She chose a dress matching your tie.
4 """
5
6 df, data_list = vac_identification(text, vac_nlp)
7 display(df)
```



0	I	
1	saw	V_N_ING
2	the	
3	man	
4	crossing	V_N_OBJ
5	the	
6	street	
7	.	

8	She	
9	chose	V_N_OBJ
10	a	
11	dress	
12	matching	V_N_OBJ
13	your	
14	tie	
15	.	

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

分析対象

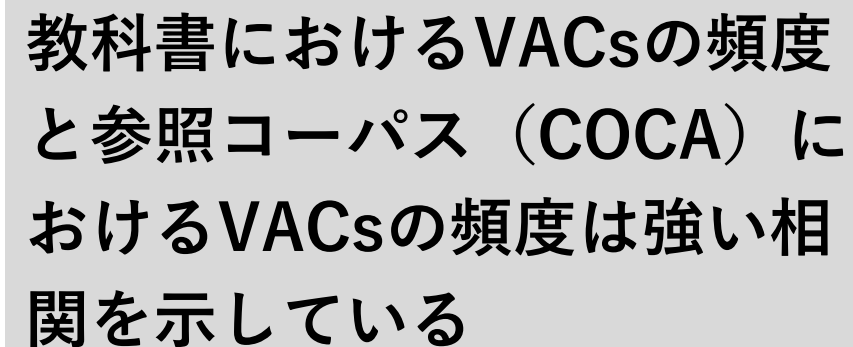
- ある1つのシリーズの現行検定英語教科書の中1～高3までを通して分析（6冊）
- メインレッスンの本文とキーセンテンスの部分のみを抽出し、分析対象とした

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（COCA）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

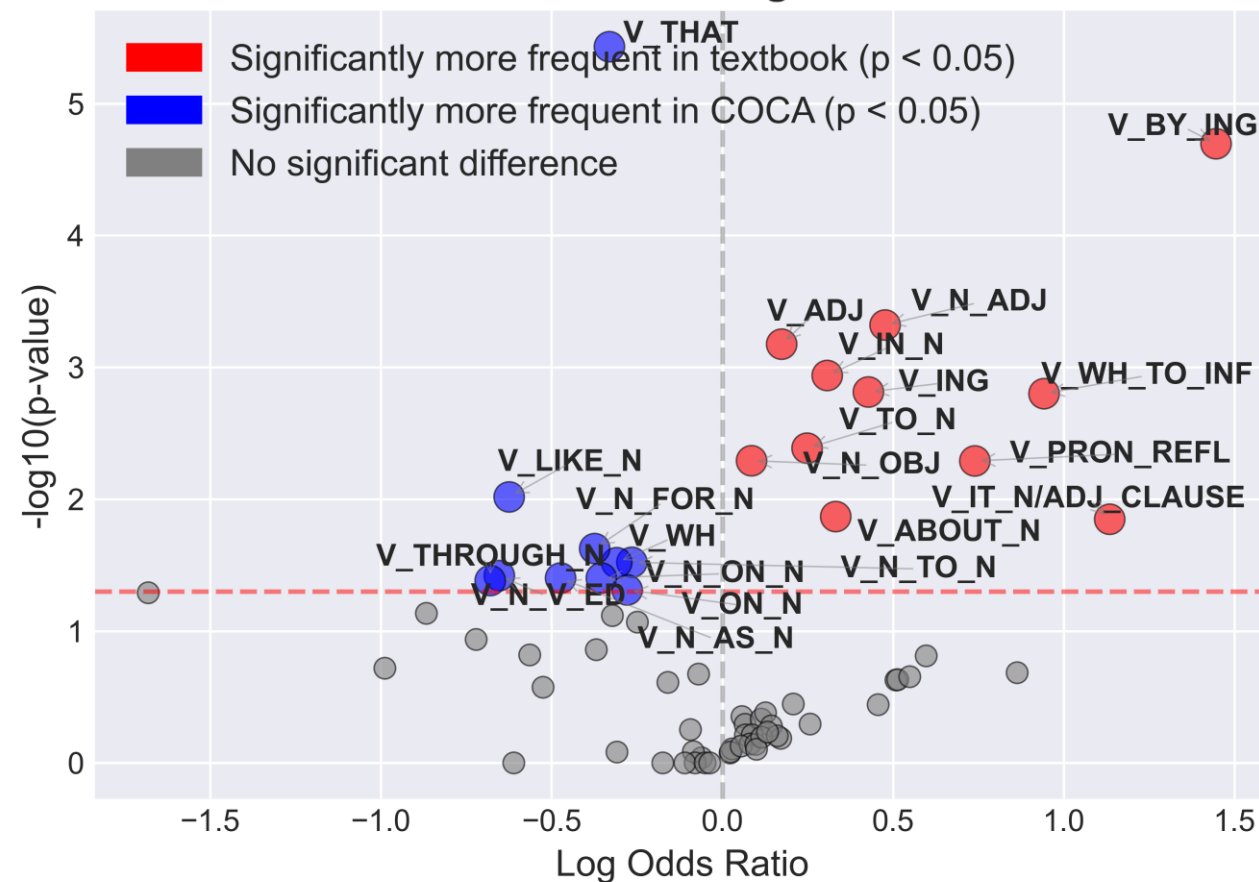
VACs	中1	中2	中3	高1	高2	高3	合計
総語数	2080	3090	3797	8151	8881	11232	37231
対象VACs総頻度	265	375	461	971	1119	1330	4520



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

Volcano Plot: VAC Usage Differences



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_BY_ING (e.g. let's **start** by comparing ...)
- V_IT_N/ADJ_CLAUSE (e.g. I **thought** it best to leave ...)
- V_WH_TO_INF (e.g. ... and **knew** how to cook ...)
- V_PRON_REFL (e.g. How can they **prepare** themselves?)
- V_N_ADJ (e.g. It **makes** me happy.)
- V_ING (e.g. Did you **enjoy** hiking this morning?)
- V_ABOUT_N (e.g. Let's **think** about it together.)
- V_IN_N (e.g. They **live** in London now.)
- V_TO_N (e.g. I **go** to dance lessons.)

● 過少使用（underuse）

- V_THROUGH_N (e.g. I **went** through 17 operations ...)
- V_N_V_ED (e.g. I **had** three wisdom teeth extracted.)
- V_LIKE_N (e.g. It **looks** like a mountain ...)
- V_N_AS_N (e.g. I **consider** him as a friend.)
- V_N_FOR_N (e.g. I don't **blame** you for being upset.)
- V_N_ON_N (e.g. You can **put** it on your window.)
- V_THAT (e.g. I **know** that he is honest.)
- V_WH (e.g. I don't **know** why Miki is sad.)
- V_ON_N (e.g. ... children can **focus** on the story.)
- V_N_TO_N (e.g. Can you **send** a video to me?)

7. 今後・まとめ

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**

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

1. Ellis, C. N., & Ferreira-Junior, F. (2009). Constructions and their acquisition: Islands and the distinctiveness of their occupancy. *Annual Review of Cognitive Linguistics*, 7(1), 188-221. <https://doi.org/10.1075/arcl.7.08ell>
2. Ellis, N. C., & Ferreira-Junior, F. (2009). Construction learning as a function of frequency, frequency distribution, and function. *The Modern Language Journal*, 93(3), 370-385. <https://doi.org/10.1111/j.1540-4781.2009.00896.x>
3. Ellis, N. C., Römer, U., & O'Donnell, M. B. (2016). VACs in L2 Knowledge and Processing. *Language Learning*, 66(S1), 123-153. https://doi.org/10.1111/lang.5_12177
4. Ellis, N. C., & Wulff, S. (2025). Usage-based Language Learning: Linguistic constructions and their learning. https://doi.org/10.31234/osf.io/fqtsx_v1
5. Ellis, N. C., & Wulff, S. (2025). Usage-based Second Language Acquisition. https://doi.org/10.31234/osf.io/bnq9w_v1
6. Francis, G., Hunston, S., & Manning, E. (1996). *Collins COBUILD grammar patterns I: Verbs*. Collins CoBUILD.
7. Goldberg, A. E. (2006). *Constructions at Work: The Nature of Generalization in Language*. Oxford University Press. <https://doi.org/10.1093/acprof:oso/9780199268511.001.0001>
8. Goldberg, A. E. (2019). *Explain Me This: Creativity, Competition, and the Partial Productivity of Constructions*. Princeton University Press. <https://doi.org/10.1515/flin-2019-2020>
9. Huang, Y., Murakami, A., Alexopoulou, T., & Korhonen, A. (2021). Subcategorization frame identification for learner English. *International Journal of Corpus Linguistics*, 26(2), 187-218. <https://doi.org/10.1075/ijcl.18097.hua>
10. Hunston, S. (2019). Patterns, constructions, and applied linguistics. *International Journal of Corpus Linguistics*, 24(3), 324-353. <https://doi.org/10.1075/ijcl.00015.hun>
11. Hunston, S. (2025). *Pattern, Construction, System: A Unified Approach to Grammar and Lexis*. Cambridge University Press. <https://www.cambridge.org/core/product/290B38732701BC5C991813CC63641D26>
12. Hunston, S., & Su, H. (2017). Patterns, Constructions, and Local Grammar: A Case Study of 'Evaluation'. *Applied Linguistics*, 40(4), 567-593. <https://doi.org/10.1093/applin/amx046>
13. Hwang, H., & Kim, H. (2022). Automatic Analysis of Constructional Diversity as a Predictor of EFL Students' Writing Proficiency. *Applied Linguistics*, 44(1), 127-147. <https://doi.org/10.1093/applin/amac046>
14. Kyle, K. (2016). *Measuring syntactic development in L2 writing: Fine grained indices of syntactic complexity and usage-based indices of syntactic sophistication* [Georgia State University].
15. Kyle, K., & Crossley, S. (2017). Assessing syntactic sophistication in L2 writing: A usage-based approach. *Language Testing*, 34(4), 513-535. <https://doi.org/10.1177/0265532217712554>
16. Park, J.-H., & Sung, M.-C. (2023). Expansion of verb-argument construction repertoires in L2 English writing. *International Review of Applied Linguistics in Language Teaching*, 62(2), 903-925. <https://doi.org/10.1515/iral-2022-0145>
17. Perek, F., & Patten, A. L. (2019). Towards an English Constructicon using patterns and frames. *International Journal of Corpus Linguistics*, 24(3), 354-384. <https://doi.org/10.1075/ijcl.00016.per>
18. Römer, U. (2019). A corpus perspective on the development of verb constructions in second language learners. *International Journal of Corpus Linguistics*, 24(3), 268-290. <https://doi.org/10.1075/ijcl.00013.roe>
19. Römer, U., O'Donnell, M. B., & Ellis, N. C. (2015). Using COBUILD grammar patterns for a large-scale analysis of verb-argument constructions: Exploring corpus data and speaker knowledge. In N. Groom, M. Charles, & S. John (Eds.), *Corpora, Grammar and Discourse: In honour of Susan Hunston* (pp. 43-72). John Benjamins. <https://doi.org/10.1075/scl.73.03rom>
20. Sung, H., & Kyle, K. (2024). Annotation Scheme for English Argument Structure Constructions Treebank. In S. Henning & M. Stede, *Proceedings of The 18th Linguistic Annotation Workshop (LAW-XVIII)* St. Julians, Malta.
21. Sung, H., & Kyle, K. (2024). Leveraging pre-trained language models for linguistic analysis: A case of argument structure constructions. In Y. Al-Onaizan, M. Bansal, & Y.-N. Chen, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing* Miami, Florida, USA.
22. Sung, H., & Kyle, K. (2025). Usage-based analysis of L2 oral proficiency: Characteristics of argument structure construction use. *Studies in Second Language Acquisition*, 1-27. <https://doi.org/10.1017/S0272263125100995>