LING 450: Data Science for Linguistics

Final Project Report

Submitted by: Akash Chaudhari, Divya Munot, Nanda Kishore Korrapolu, Satvik Garg, Sreyoshi Basu and Suhasini Patni

4th May 2023

A Statistical Analysis of the Syntactic Dependencies between Governing Verbs and their Argument Gerunds

Abstract:

This research is an investigation of the syntactic context within which gerunds of various types may occur. More specifically, this project uses statistical analysis to understand the dependency relation between controlling verbs and their argument gerunds. The Python library, Matplotlib, was used in various stages of the analysis to visualize the frequency with which a certain verb type co-occurs with a certain gerund type in each syntactic context. Dependency parsing was used to identify instances of verbs that have governing relations on the dependent gerunds. Given the nature of the data frame, it was predicted that Verbs of Predicative Complements govern most types of gerunds.

Project Description:

In English, the ing-form, also known as the gerund, has been shown to display both verbal and nominal properties (Seiss, 2008). Included in the gerund's verbal properties are "the governing of a direct object or the possibility of modification by adverbs." Contrastively, when the -ing forms appear as "subjects, objects and complements of prepositions," they demonstrate nominal properties. (Seiss, 2008).

A previous study on gerunds by the University of Rochester's Departments of Computer Science and Linguistics used computational methods to identify gerunds in a sentence and categorize them into six types, namely POSS-ing of, -ing of, POSS-ing, DET-ing, ACC-ing, and, VP-ing. Among these, POSS-ing of and -ing of gerund types display nominal properties, while the others are classified for their verbal nature. Their investigation focused on "the syntactic and semantic properties of -ing nominalizations." (Gerund Nominalizations, 2022). This prior research serves as the foundation for the present study on the context governing dependent gerunds. The following data from the previous study was utilized to expand our current research on gerund behavior: parts-of-speech tagging, identification of the position of every word in a sentence, classification of the type of gerund and dependency parsing on each sentence in the dataframe.

Current research on gerunds focuses on the syntactic context that governs dependent gerunds. A study of the syntax context includes an investigation of numerous governing lexical and functional elements such as nouns, verbs, adjectives and prepositions, and how they influence the selection of the specific gerund types with which they combine. This will provide a more precise syntactic perspective on the contexts in which a specific gerund type can occur. This project, more specifically, concentrates on deciphering the nature of the governing verbs and their co-occurrence with dependent gerund types, to make larger analyses about the syntactic context of a gerund. Dependency parsing on each sentence in the data frame provides the list of verbs that have governing relations on gerunds. Given the nature of the data frame, it is likely that most gerund types co-occur with similar verb-types. This hypothesis was proved in the statistical analysis which revealed that Verbs with Predicative Complements co-occur with all gerund types in a majority distribution, with the exception of POSSing-of gerunds, which portray an interesting case in that they mostly co-occur with Verbs of Change of State.

Methodology:

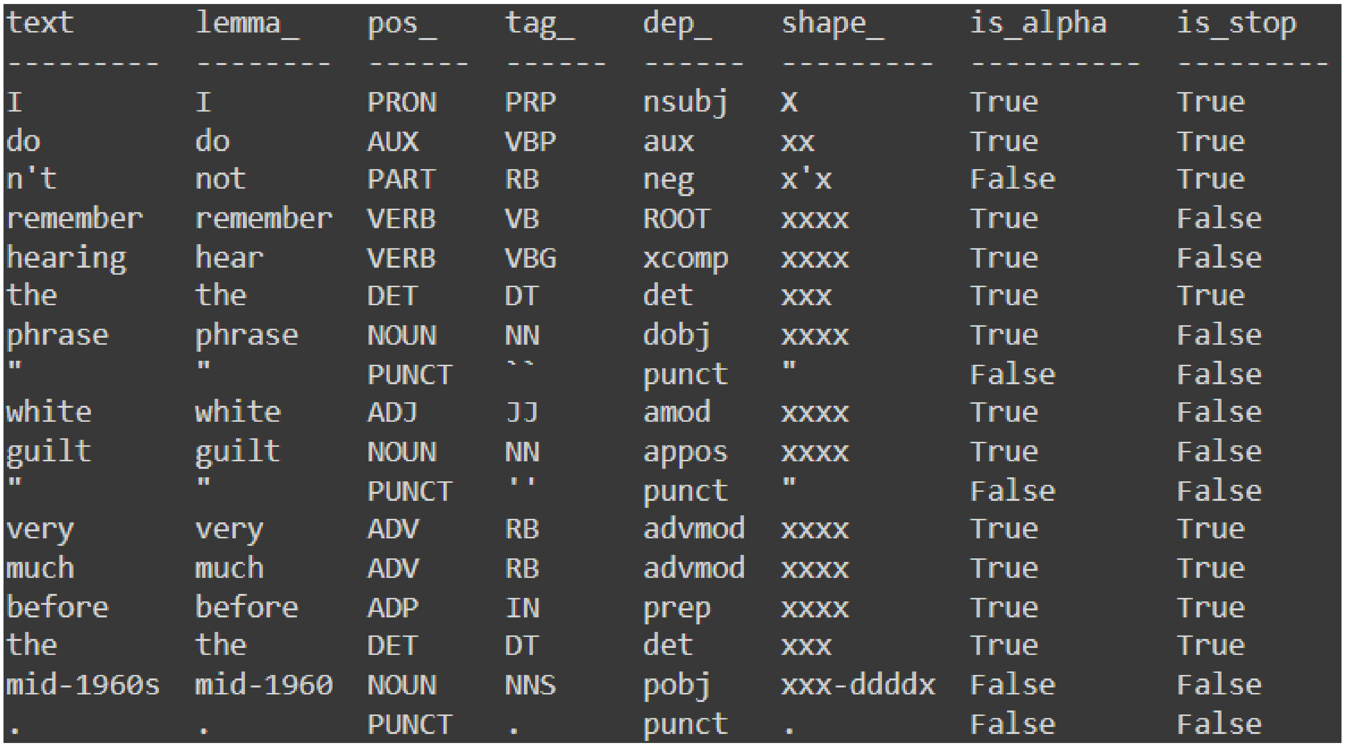
Step 1:

As the initial step in the research, verbs and their placement in the sentence needed to be identified. spaCy, a natural language processing tool that has “built in visualizations for syntax” was used to tag parts of speech and learn their placement in the sentence (spaCy, n.d). spaCy was also utilized to get the root form of the verbs being used in the sentence. This was crucial for verb classification in further steps.

The process of extracting the root form of a word from a tagged sentence is called lemmatization. This process is used because verbs in verbnet are stored in their lemmatized or root format.

Examples of this process are provided below.

Sentence: “I don't remember hearing the phrase " white guilt " very much before the mid-1960s.”



```
text       lemma_    pos_    tag_    dep_    shape_     is_alpha    is_stop
---------  --------  ------  ------  ------  ---------  ----------  ---------
I          I         PRON    PRP     nsubj   X          True        True
do         do        AUX     VBP     aux     xx         True        True
n't        not       PART    RB      neg     x'x        False       True
remember   remember  VERB    VB      ROOT    xxxx       True        False
hearing    hear      VERB    VBG     xcomp   xxxx       True        False
the        the       DET     DT      det     xxx        True        True
phrase     phrase    NOUN    NN      dobj    xxxx       True        False
"          "         PUNCT   ``      punct   "          False       False
white      white     ADJ     JJ      amod    xxxx       True        False
guilt      guilt     NOUN    NN      appos   xxxx       True        False
"          "         PUNCT   ''      punct   "          False       False
very       very      ADV     RB      advmod  xxxx       True        True
much       much      ADV     RB      advmod  xxxx       True        True
before     before    ADP     IN      prep    xxxx       True        True
the        the       DET     DT      det     xxx        True        True
mid-1960s  mid-1960  NOUN    NNS     pobj    xxx-ddddx  False       False
.          .         PUNCT   .       punct   .          False       False
```

Step 2:

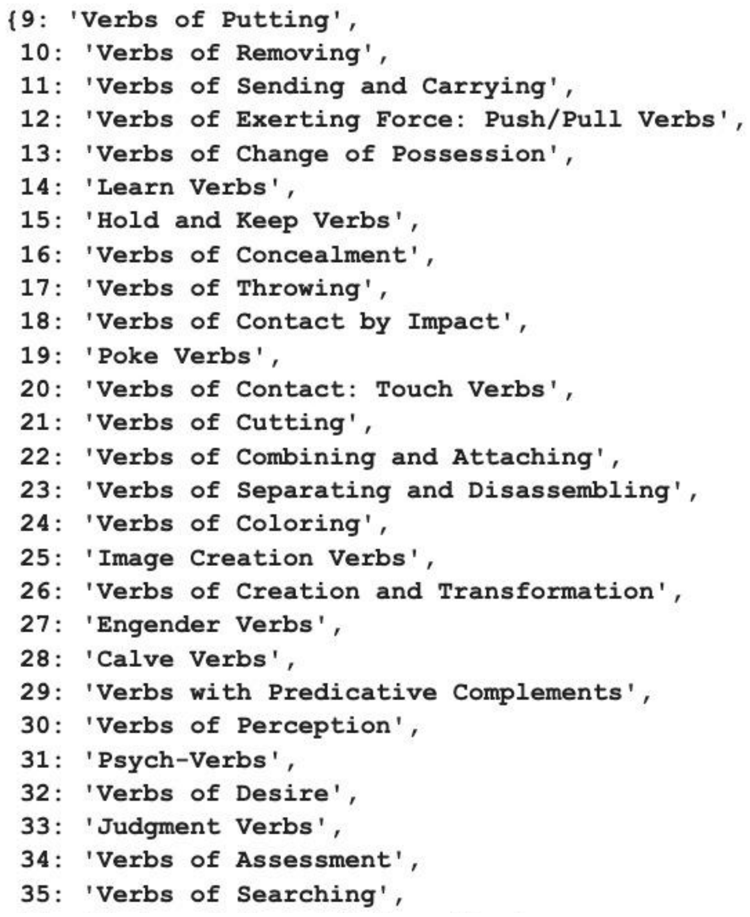
In order to execute the hypothesis, *VerbNet*, an open-source library, and the "largest online network of English verbs," was used to categorize verbs into type buckets (Green et al, 2017). VerbNet uses semantic and syntactic knowledge encoded in verbs to first categorize them into classes, and then further group those classes into 109 verb types.

According to their website, the list of verbs on VerbNet are:

> "described by thematic roles, selectional preferences of the arguments, and frames consisting of a syntactic description and a semantic representation with subevent structure patterned on the Dynamic Event Model of Pustejovsky and Moszkowicz (2011) and Pustejovsky (2013)" (Green et al, 2017).

VerbNet documents all its verbs in their root forms, which made the process of lemmatization crucial to the success of this project. Further, VerbNet classifies each verb into classes using a numbering system, i.e, every verb listed on VerbNet has a unique number associated with it. An NLTK package was used to access the verbs classified on VerbNet, along with their specific class-numbers. The numbers associated with each verb-class indicated the larger verb-type they were associated with. However, NLTK did not provide a direct mapping of a verbs class to its corresponding verb-type. Therefore, the class-numbers, along with the associated verb-type classification, were added to a Python dictionary. This step was helpful in allowing easy access to the verb and its type as mentioned on VerbNet.

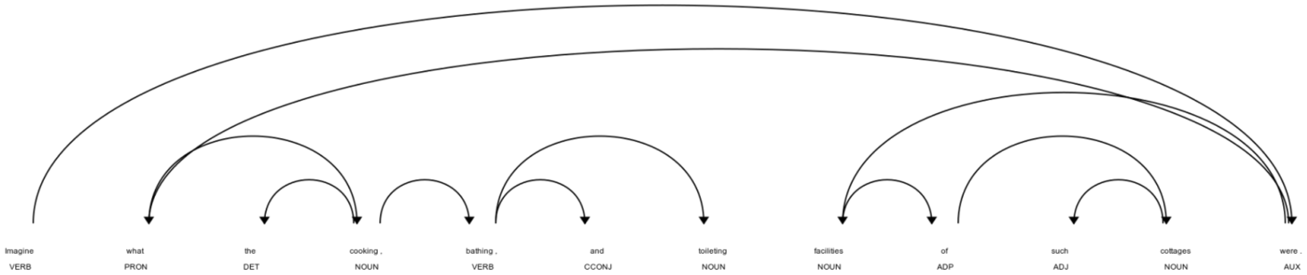
The following is a snippet of the Python dictionary that was manually compiled to map verbs in each class to their corresponding verb-type.



```
{9: 'Verbs of Putting',
 10: 'Verbs of Removing',
 11: 'Verbs of Sending and Carrying',
 12: 'Verbs of Exerting Force: Push/Pull Verbs',
 13: 'Verbs of Change of Possession',
 14: 'Learn Verbs',
 15: 'Hold and Keep Verbs',
 16: 'Verbs of Concealment',
 17: 'Verbs of Throwing',
 18: 'Verbs of Contact by Impact',
 19: 'Poke Verbs',
 20: 'Verbs of Contact: Touch Verbs',
 21: 'Verbs of Cutting',
 22: 'Verbs of Combining and Attaching',
 23: 'Verbs of Separating and Disassembling',
 24: 'Verbs of Coloring',
 25: 'Image Creation Verbs',
 26: 'Verbs of Creation and Transformation',
 27: 'Engender Verbs',
 28: 'Calve Verbs',
 29: 'Verbs with Predicative Complements',
 30: 'Verbs of Perception',
 31: 'Psych-Verbs',
 32: 'Verbs of Desire',
 33: 'Judgment Verbs',
 34: 'Verbs of Assessment',
 35: 'Verbs of Searching',
```

Step 3:

The next step was to understand how the verbs governed the gerund in each sentence. This required a process of elimination: to remove verbs that did not affect the gerund. The relationship crucial to the study of this hypothesis is the one between the verb and gerund. While there were several sentences in the data where multiple verbs co-occurred with a gerund, these didn't necessarily govern the gerund. Dependency parsing is an NLP tool that extracts a relationship in a sentence between a "head word" and words that modify the head (Dependency Parsing, n.d). Data generated in the previous study listed the dependency relations for each sentence in the data frame. The parsing stage isolates an accurate set of tokens that have a dependency on the gerund. These were matched to their class type and analyzed to understand the context of the gerund. An example of a dependency parsing visualization using spaCy is provided below.



Analysis:

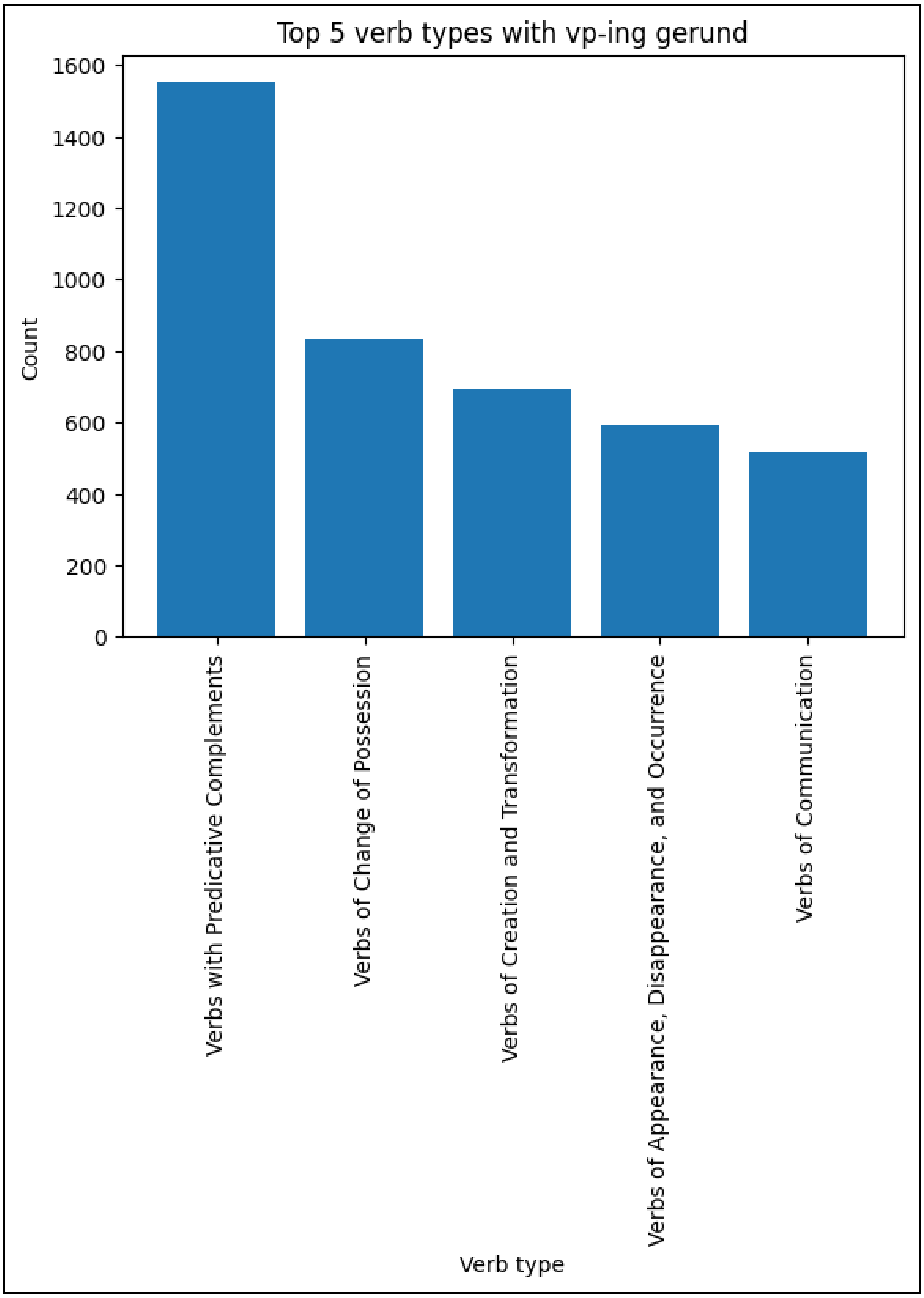
Out of 55059 tokens, a total of 17,586 token verbs that governed dependent gerunds were extracted. The following information was extracted from the data.

The 5 most frequently occurring verb-types with every gerund type are as follows.

1. Verbs of Predicative Complements
2. Verbs of Change of Possession
3. Verbs of Creation and Transformation
4. Verbs of Appearance, Disappearance, and Occurrence
5. Psych-Verbs

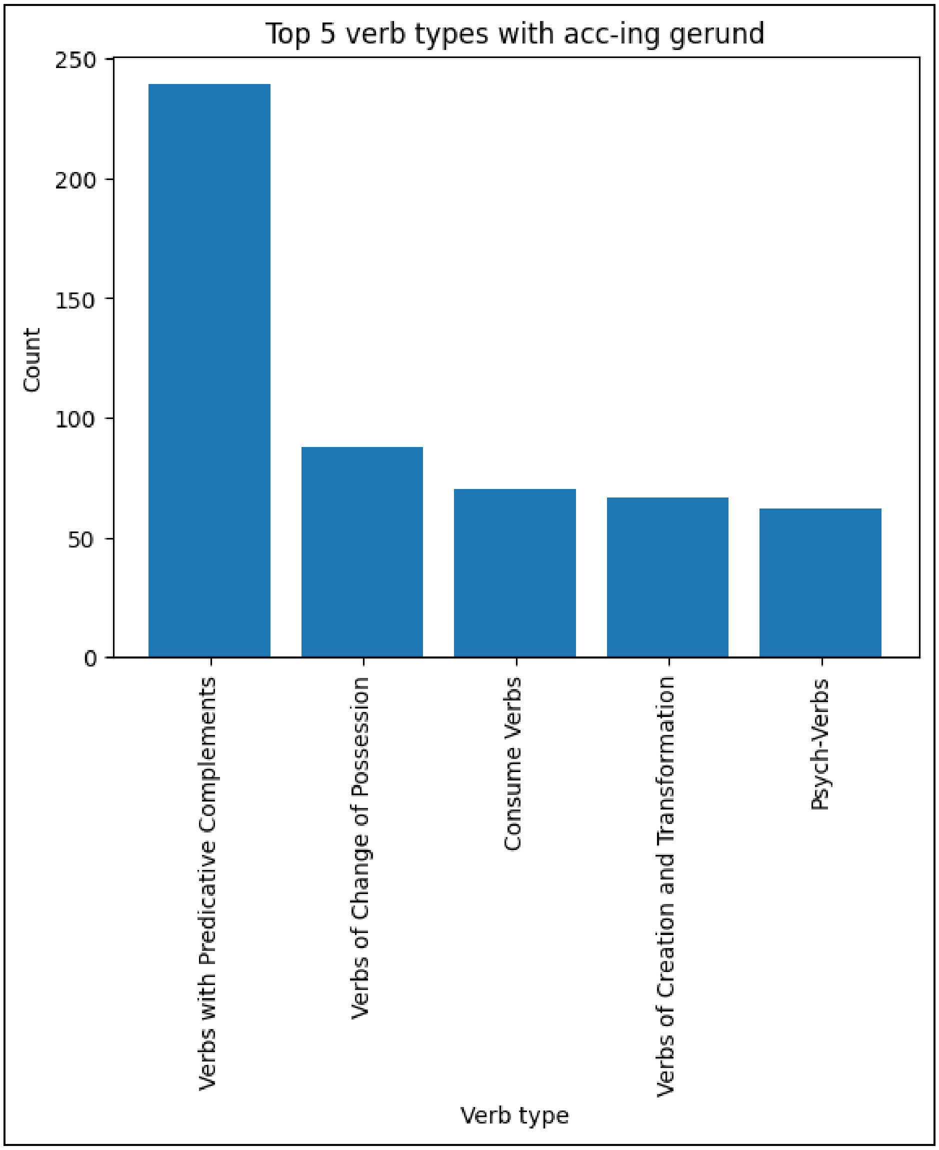
The relation between the verb type and their impact on dependent gerund types, was visualized using Matplotlib, a Python visualization library that helps create 2D graphs and plots. The graphs in this following section were made using this tool. These are graphs of the five most frequently occurring verb-types with each individual type of gerund.

Graph 1:



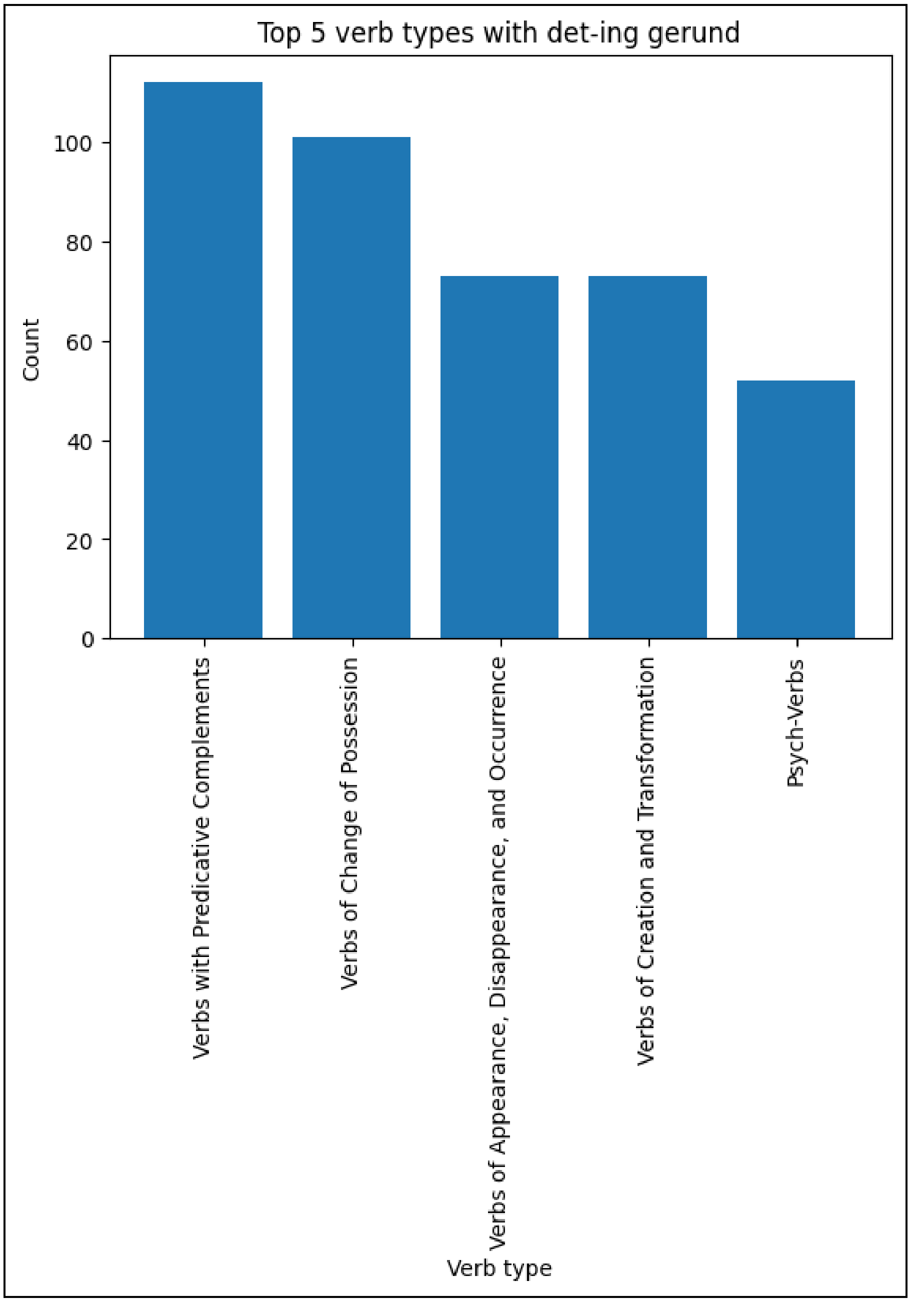
In the data, VP-ing gerunds occur most frequently, comprising almost 90% of all gerunds. As shown in the graph above, these gerunds most frequently co-occur with verbs of predicative complements, appearing in over 1500 dependencies.

Graph 2:



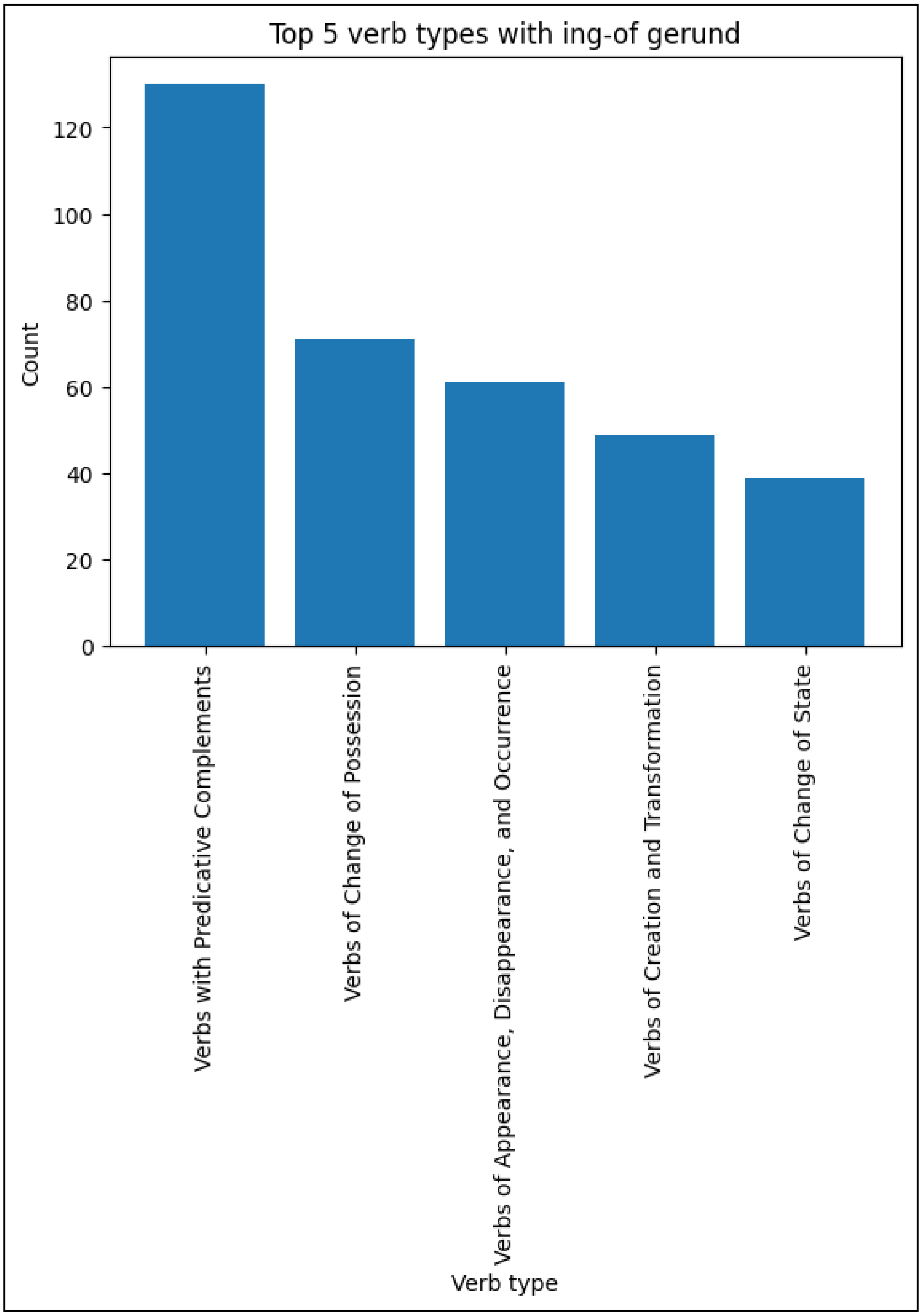
Similar to VP-ing gerunds, ACC-ing gerunds also co-occur most frequently with verbs with predicative complements, followed by verbs of change of possession.

Graph 3:



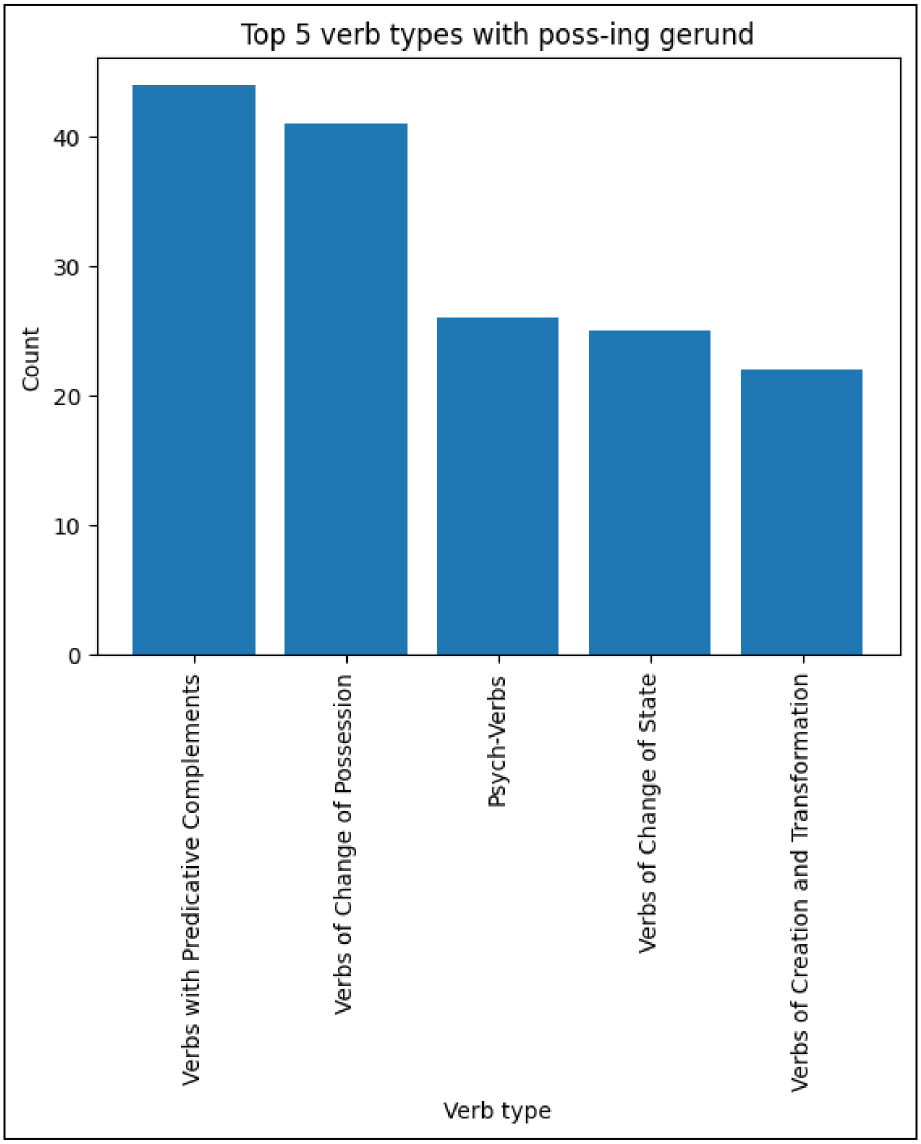
Similar to VP-ing and ACC-ing gerunds, DET-ing gerunds too co-occur most frequently with verbs with predicative complements, followed by verbs of change of possession.

Graph 4:



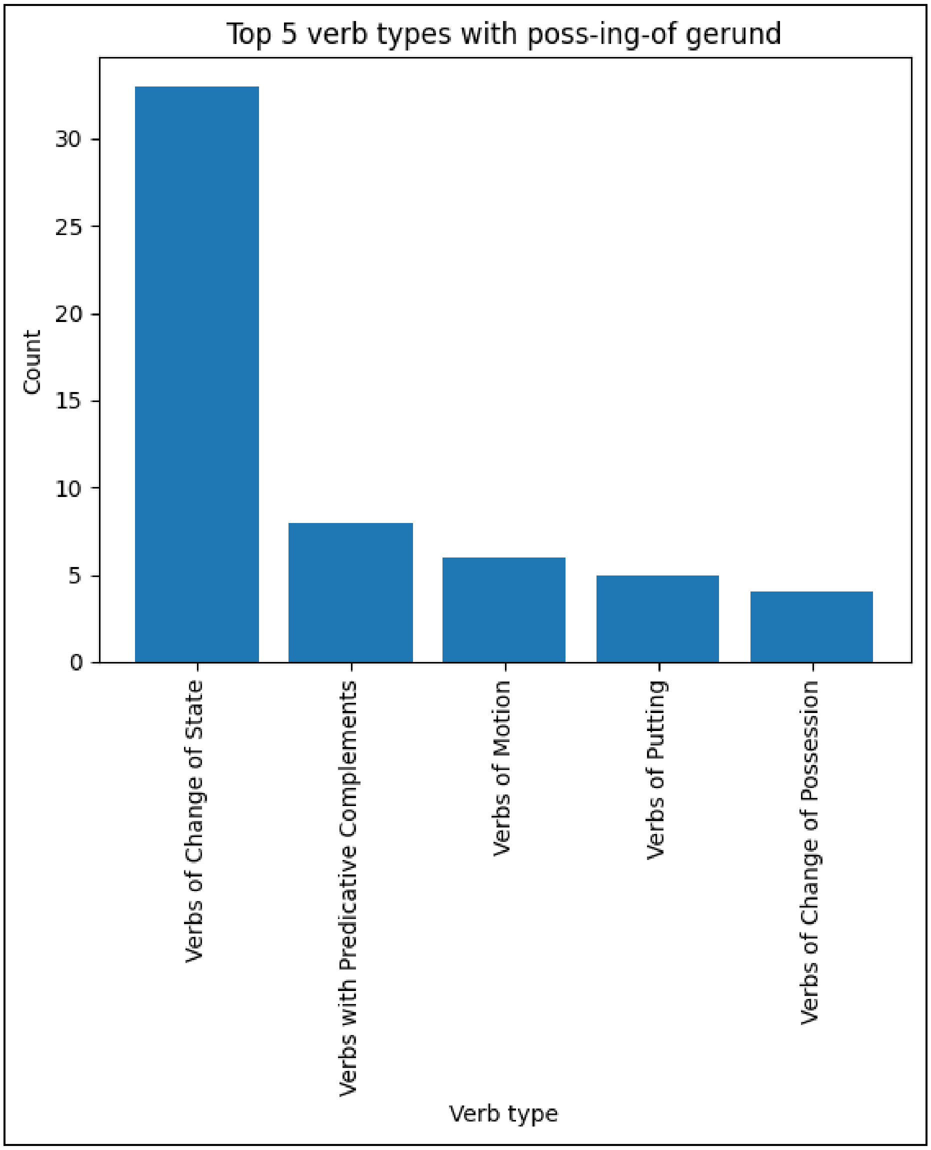
Similar to VP-ing, ACCing, and DET-ing gerunds, ing-of gerunds are seen to co-occur most frequently with verbs with predicative complements, followed by verbs of change of possession.

Graph 5:



Similar to VP-ing, ACCing, DET-ing, and ing-of gerunds, POSS-ing gerunds too co-occur most frequently with verbs with predicative complements, followed by verbs of change of possession.

Graph 6:



While verbs of predicative complements co-occur in a maximum frequency with most gerund types, in POSS-ing of gerunds, this trend shifts. Instead, verbs of change of state co-occur most frequently (over 30 times) with POSS-ing of gerunds, while verbs with predicative complements co-occur less than ten times. This is in-line with our hypothesis, where we state that while all gerund types co-occur with similar verb types, POSS-ing of gerunds have their own unique co-occurrence.

The co-occurence of the second highest frequency also displays a similar trend as above. Verbs of change of possession hold second position across all gerund types except POSS-ing of gerunds. However, it is interesting to note that the remaining top three positions display variations in the verb-type and gerund-type co-occurrence. Further research into this variation might potentially reveal more nuanced insights into why certain verb types co-occur with certain gerund types. Moreover, a thorough understanding of the characteristics of each gerund type is also important in drawing conclusions about them being selected by specific verb-types. This will also provide an explanation as to why POSS-ing of gerunds behave differently as compared to the rest. Since the study on various gerund types is still new and there is little literature discussing these variations, it is beyond the scope of this present study.

Limitations and Failures:

To understand the syntactic context of gerunds, it is prudent to analyze the environment in which they occur. They can be done by looking at all parts of speech. In this study, verbs have been utilized to understand the behavior and context of gerunds. However, testing the hypothesis did not come without a few failures and/or limitations.

While the aim was to analyze the relationship between verb types and gerund types in specific contexts,, the analysis initially took into consideration every instance of a verb in each sentence. It began by lemmatizing and classifying all the verbs that occurred in each sentence, using verb-net to tag all the verbs. The graphs generated in this stage were not insightful, as they took into account a massive amount of data, most of which was not relevant to the analysis.

In the second iteration of the analysis, an attempt was made to reduce and focus the list of verbs that had governing relations on argument gerunds. spaCy, which provided information on the position of the verbs in each sentence, was used to extract only the verbs that occurred in positions before the gerund. While this was more in line with the hypothesis and produced similar results to those presented in the study, it was still incorrect because not all verbs with governing relations on the gerund had to occur before the gerund.

Only in the third iteration was the use of universal dependencies to extract only verbs having dependence relations to gerunds executed. This revealed that such verbs might appear at any position in the sentence, not just before the gerund. The use of dependency parsing to extract only the governing verbs resulted in a significantly smaller and more precise dataset for analysis. In a further study, it may be relevant to look at specific tags of the dependency relation between the verb and its gerund, for instance, whether the gerund occurs as the verb's direct object, or in a clausal relation to it. This might provide a more nuanced insight into the relationship between a governing verb and its dependent gerund.

As with any statistical study, some of the tools used also presented with certain limitations. Most significantly, the lack of a specific API within the NLTK of VerbNet Class that allows easy access to verb-type from the verb-class necessitated a significant amount of manual code to extract this data. Following a closer inspection of the documentation, it was found that the number following the verb-class accurately represents the index of its verb-type. This issue was solved by manually building a Python dictionary that mapped the verb-class index number to the corresponding verb-type. This was achieved after an exhaustive search through the github of NLTK that has VerbNet stored in an XML format.

It is also worth noting that the data frame used for this study has substantial discrepancies in the number of each gerund type. For example, the data frame contains approximately 55,000 phrases, with VP-ing gerunds accounting for nearly 41,660 tokens. -ing of, POSS-ing, ACC-ing, and DET-ing gerunds have a token count ranging from 2000 to 5000, while POSS-ing of gerunds has just 255 occurrences. Since VP-ing accounts for nearly 90% of the data, the data was not normalized. This is because normalizing a difference of this magnitude would result in significant data loss and inaccurate results.

Currently, searching a certain verb's lemmatised form on VerbNet returns several different verb-class categorizations for the same verb. For instance, the verb 'hear' returns the classes ['discover-84-1-1] and ['see-30.1-1-1']. The class-ids on each verb are relevant to mapping them to their corresponding verb-type. For the purpose of this study, the verb-class to verb-type mapping has been achieved by selection of the first class provided on VerbNet. However, for future study, it may be worth exploring which verb-class is more relevant to the context of the gerund it occurs with and selecting the more appropriate verb-class.

Resources:

https://verbs.colorado.edu/verbnet/

Dependency parsing. (n.d.). NLP-Progress. Retrieved May 4, 2023, from http://nlpprogress.com/english/dependency_parsing.html

SpaCy · Industrial-strength Natural Language Processing in Python. (n.d.). Retrieved May 4, 2023, from https://spacy.io/

Seiss, M. (2008). "The English -ing Form." http://csli-publications.stanford.edu/.

Scott Grimm and Louise McNally. "Nominalization and Natural Language Ontology". In: Annual Review of Linguistics 8.1 (2022), pp. 257–277. doi: 10 . 1146 / annurev-linguistics-031120-020110.

Scott Grimm and Louise McNally. "The -ing dynasty: Rebuilding the semantics of nominalizations". In: Semantics and Linguistic Theory 25 (2015), pp. 82–102.

"A Large Scale Corpus of Gerund Nominalization" (2022), Department of Linguistics & Computer Science, University of Rochester

Kara Passmore. "POSSESSIVE-ING and ACCUSATIVE-ING Constructions in English". Linguistics Senior Thesis. Swarthmore College, 2003.

"Gerund Study." *GitHub*, Department of Computer Science and Department of Linguistics. University of Rochester, 2022, https://github.com/hwang42/Gerund-Study.