Gimme Gimme Gimme (A Distribution of Gerunds)

Amelia Ostrow, Simone Jackson, Tanvi Shroff, Ian Harding, Suganya Rajendran Schmura
LING 250 - CSC 250 - CSC 450 - CSC 250 - LING 450

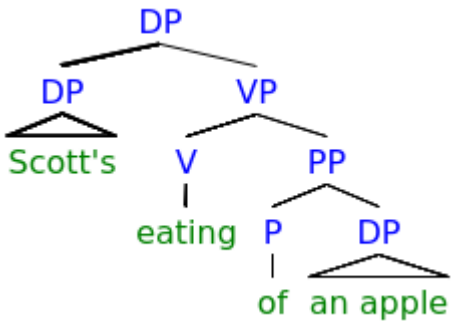
1. **Introduction**

Gerunds in English are the source of many linguistic studies due to their syntactic and semantic complexity as they can display varying mixtures of both verbal and nominal properties. There are six different types of gerunds that we will be focusing on in this paper: possessive gerunds using 'of' prepositional phrases (POSS-ing_{of}), gerunds using 'of' prepositional phrases (-ing_{of}), possessive gerunds (POSS-*ing*), determiner gerunds (DET-*ing*), accusative gerunds (ACC-*ing*), and verb phrase gerunds (VP-*ing*) (Grimm & McNally, 2015; Malouf, R; Seiss, 2008). Below, an example of each type of gerund phrase (the bolded phrase) and its representation in a syntax tree will be provided to illuminate the syntactic and semantic properties of each gerund type.

POSS-ing_{of} gerunds act as genitive subjects with a preposition phrase headed by 'of' as exemplified in (1).

(1) Some thought **Scott's eating of an apple** in class was disruptive.

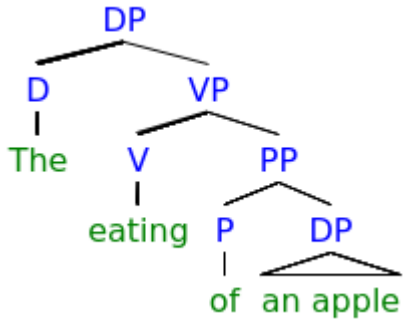
(1a)



Gerunds that employ the -*ing*of construction include a preposition phrase headed by 'of', as shown in (2).

(2) **The eating of an apple** happened in class.

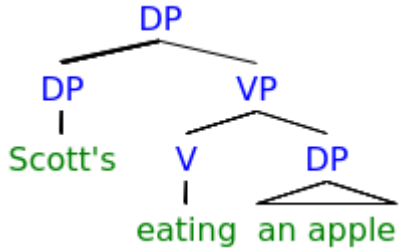
(2a)



POSS-*ing* constructions act as genitive subjects, as shown in (3).

(3) We thought **Scott's eating an apple** was humorous.

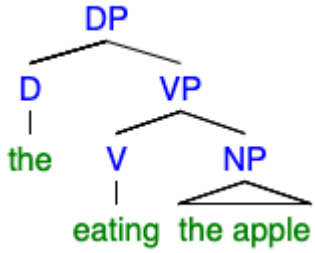
(3a)



DET-*ing* constructions have a similar syntactic construction as POSS-*ing* constructions, but they use a determiner head instead as shown in (4) and (4a).

(4) It was **the eating the apple** that annoyed me.

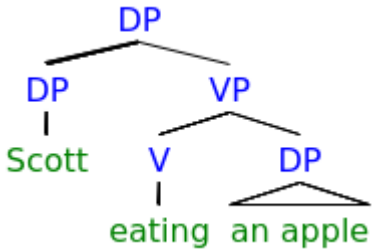
(4a)



ACC-*ing* constructions have an accusative subject, as shown in (5).

(5) **Scott eating an apple** in class made us hungry.

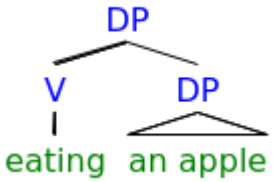
(5a)



VP-*ing* constructions are subjectless verbal gerund phrases, as shown in (6).

(6) **Eating an apple** is what Scott did in class.

(6a)



Linguistically, this question is interesting because there is such wide variation in how these nominalizations function in a sentence, but they do not seem to, on the surface, have consistent meanings or distributions within or across types.

The previous class created a program that parses a corpus of gerund constructions for dependency relations and summarized the frequency of each gerund type which was determined by this dependency parsing. They found that (VP-*ing*) constructions occur the most frequently, followed by ACC-*ing*, DET-*ing*, -ing_{of}, POSS-*ing*, and finally POSS-ing_{of}. They had the most trouble with the ACC-*ing* category which shows the least accuracy in their automatic labeler due to the oversimplified criteria that was used to classify it.

While examining the work on gerund nominalizations done by the previous class, we noticed that different gerund types seemed to have different functions in a sentence. In order to take the next steps from this previous project, we decided to formally investigate if different controlling words have different distributions of gerund types based on how those controlling words function in a sentence. If it is true that different nominalization types have different functions in a sentence, then we would expect to see that different controlling words will have different distributions of gerund nominalization types depending on the function of the controlling word.

2. **Methodology**

In order to test our prediction, we used the data frame that the previous class made to extract the controlling word for each nominalization phrase, the dependency between the controlling word and the nominalization, and the part of speech of the nominalization. In this analysis, we use part of speech and dependency type as stand-ins for the controlling word's

function in a sentence; words with the same part of speech have the same syntactic distributions by definition, and we hoped that the dependency relations indicate a semantic relationship between the controlling word and the gerund. We then grouped all the controlling words that were identical strings (we did not lemmatize them) and quantified how many of each type of nominalization phrase each word controlled. In order to ensure that these distributions had enough data, we excluded from further analysis any controlling words that appeared fewer than 68 times in the dataset.

In order to look for trends solely between controlling words and their respective distributions across gerund types, we performed latent class analysis (LCA) on the distribution data. Our goal with this was to potentially find classifications of distributions throughout the data, and then examine the associated controlling word classes to identify semantic distinction between classes.

The analysis initially involved only the distributions of controlling words with more than 68 occurrences. However, results for all of 2-6 latent classes produced Bayesian Information Criterion (BIC) values far over 10,000, revealing the poor fit of the models.[1] That being said, the lowest BIC occurred when the data was fitted for 2 classes, indicating that the best models out of the 6 were with 2 latent classes. Afterwards, this method was repeated with the full dataset in hopes of better-fitting classification. 2-4 latent classes were tested and produced similarly ill-fitting models with unacceptably high BIC values. The best fit was, again, with 2 classes.

The analysis did not reveal any existing latent classes in the data, which prompted us to conclude that there are no underlying trends in the distributions of controlling words across gerund types. This directed us back to controlling word characteristics and functions as potential sources of information on gerund types. The characteristic we subsequently looked to was the dependency relation between a controlling word and an associated gerund.

We separated the dependency relations found in the data into the nine categories outlined by the Universal Dependencies website, as seen below, with the details of what types of relations was in each category:

Core dependents of clausal predicates	nominal subjects (nsubj), passive nominal subjects (nsubj:pass), direct object (obj), indirect object (iobj), clausal subject (csubj), clausal passive subject (csubj:pass), clausal complement (ccomp), open clausal complement (xcomp)
Non-core dependents of clausal predicates	oblique nominal (obl), temporal modifier (obl:tmod), adverbial clause modifiers (advcl), adverbial modifier (advmod)

[1] For reference - while there is no ‘good’ BIC value that can be used as a standard for all analysis, the lower the BIC value, the better the resulting models for n classes. Thus, a BIC as high as 10,000 indicates the models will predict very poorly, in any case.

Special clausal dependents	vocative, expletive (expl), auxiliary (aux), passive auxiliary (aux:pass), copula (cop), marker (mark)
Noun dependents	appositional modifier (appos), nominal modifier (nmod), temporal nominal modifier (nmod:tmod), possessive nominal modifier (nmod:poss), clausal modifier of noun (acl), adnominal relative clause modifier (acl:relcl), adjectival modifier (amod)
Compounding and unanalyzed	compound, flat
Coordination	conjunct (conj), coordination (cc)
Case-marking, prepositions, possessive	case marking (case)
Loose joining relations	list, parataxis
Other	unspecified dependency (dep)

Table 1: Categorizations of dependency relations.

3. Results

To analyze the data, we began by graphing the distribution of gerund type by the part of speech of the controlling word, as shown in Figure 1. This is included for the sake of completeness; the dependency relations illustrate syntactic relationships more granularly, and the part of speech data was supposed to be the basis of a syntactic analysis. Therefore, this data was overlooked in favor of the insights we gleaned from the dependency data.

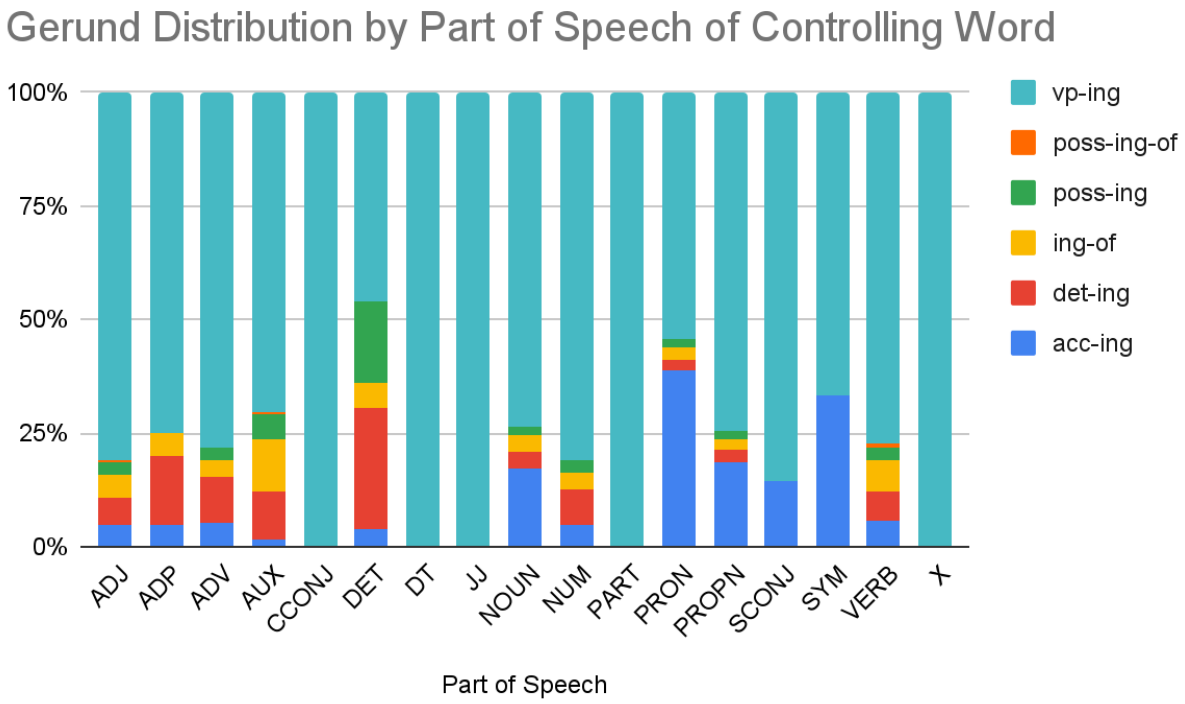


Figure 1. Distribution of gerund type by part of speech of controlling word.

We then graphed the gerund type by the dependency relation with the controlling word and present here both the graph with the raw data (Figure 2) and the normalized one (Figure 3). The raw data indicates that by far the most common dependency relations are *acl* and *advcl*; after that, the most common is *conj*; then the types that can contain any type of gerund (and some others); and finally the rest have very few data points. The normalized data shows the proportional distribution of gerund types for each dependency relation. As expected, most dependency relations are dominated by VP-*ing* and ACC-*ing*, though there are some that contain most or all of the gerund types.

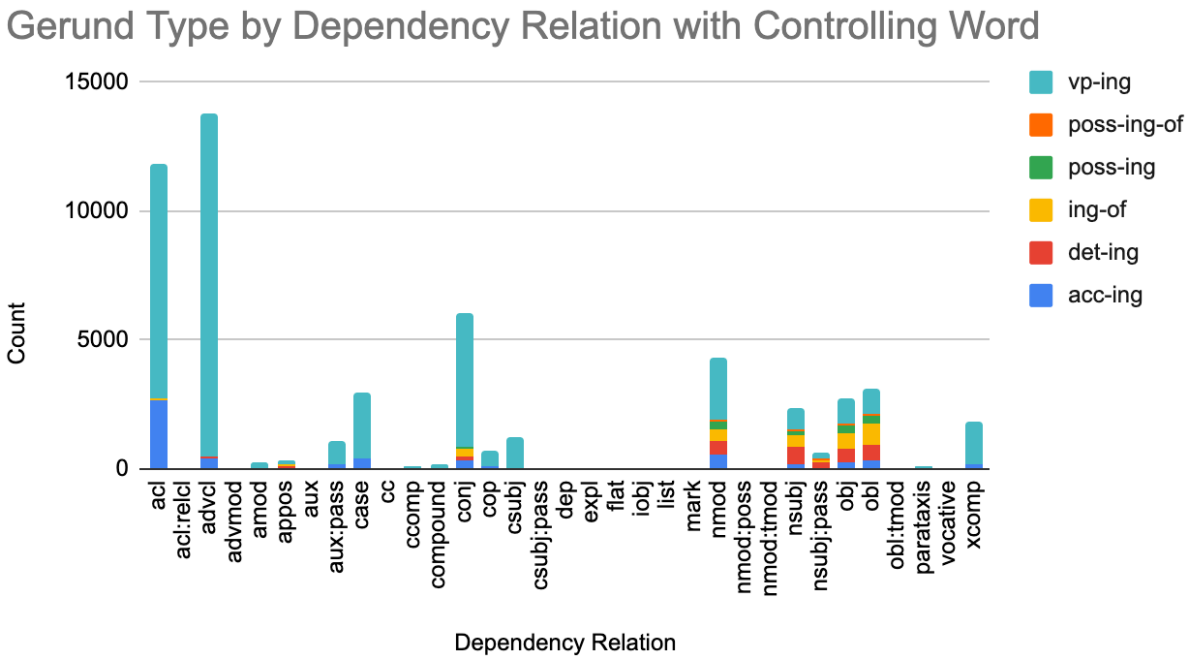


Figure 2. Distribution of gerund type by sub-categories of dependency between gerund and controlling word.

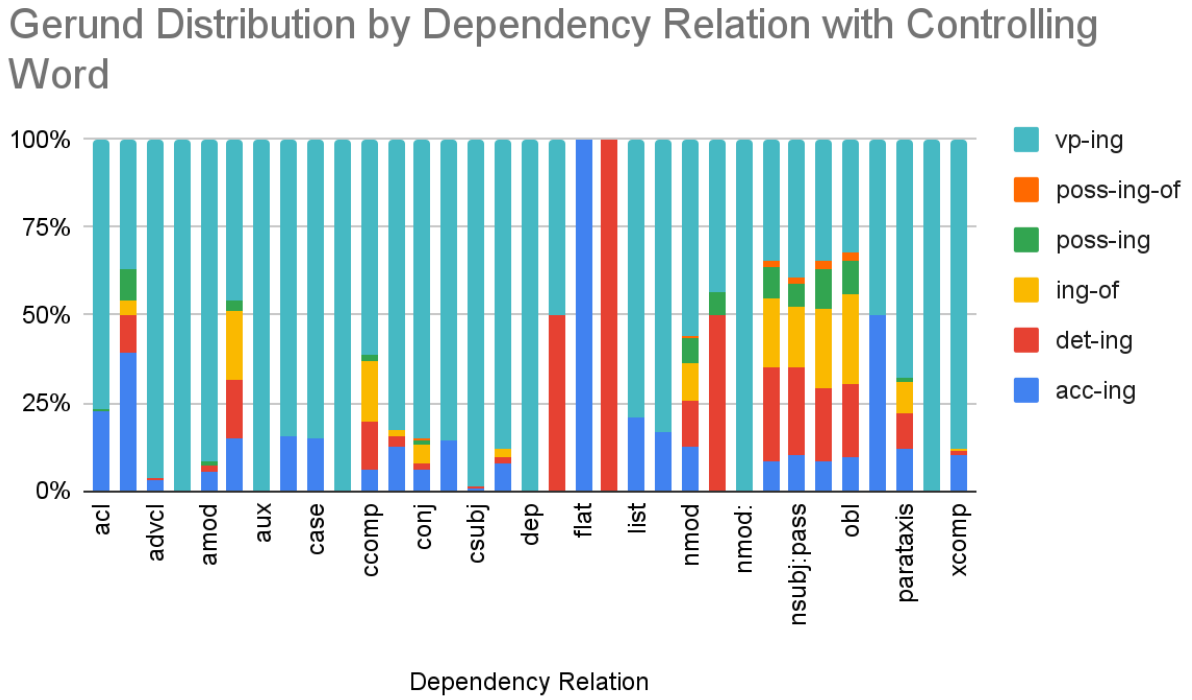


Figure 3. Normalized distribution of gerund type by sub-categories of dependency between gerund and controlling word.

Finally, we graphed the distribution of gerund types within each dependency category according to the categorizations of these dependency types by the Universal Dependencies website. As in Figure 3, the vast majority of the data in each category is VP-*ing* followed by ACC-*ing*, though many of them include most or all of the types. The raw data indicates that NCDCP and ND had by far the most tokens, followed by CDCP and COOR. CASE and SCD had under 5000 tokens each, making them significantly less common than the others, and the others have comparatively negligible data points.

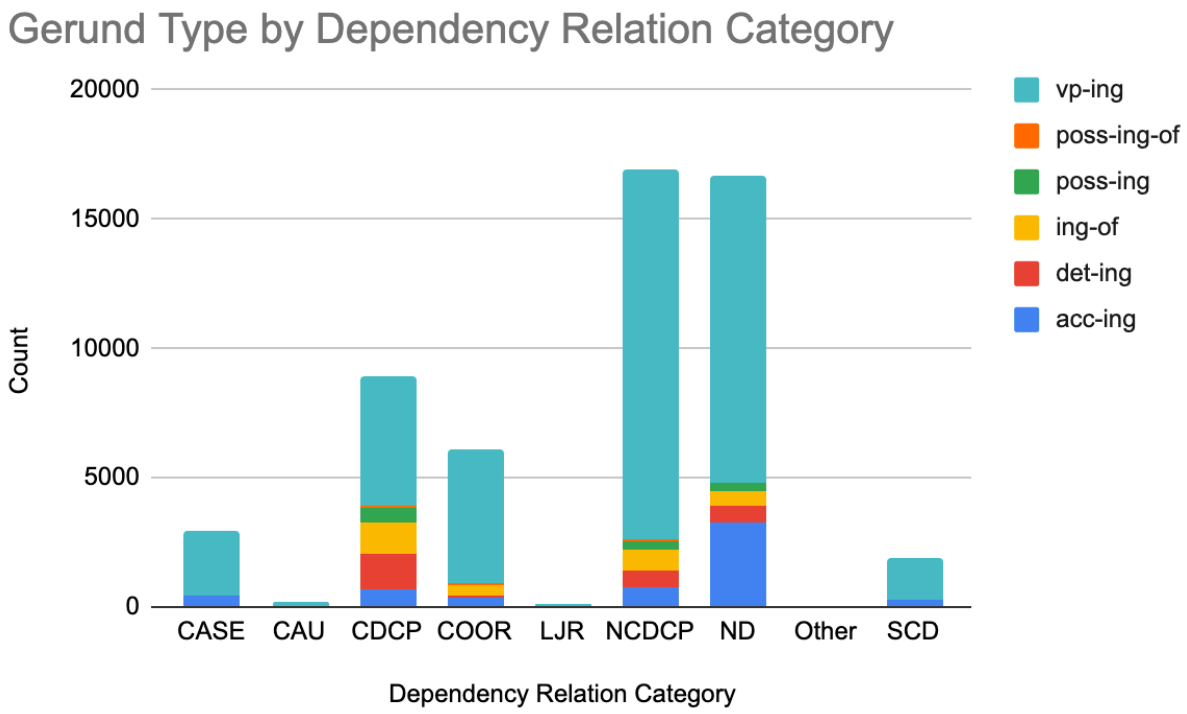


Figure 4. Distribution of gerund type by dependency between gerund and controlling word.

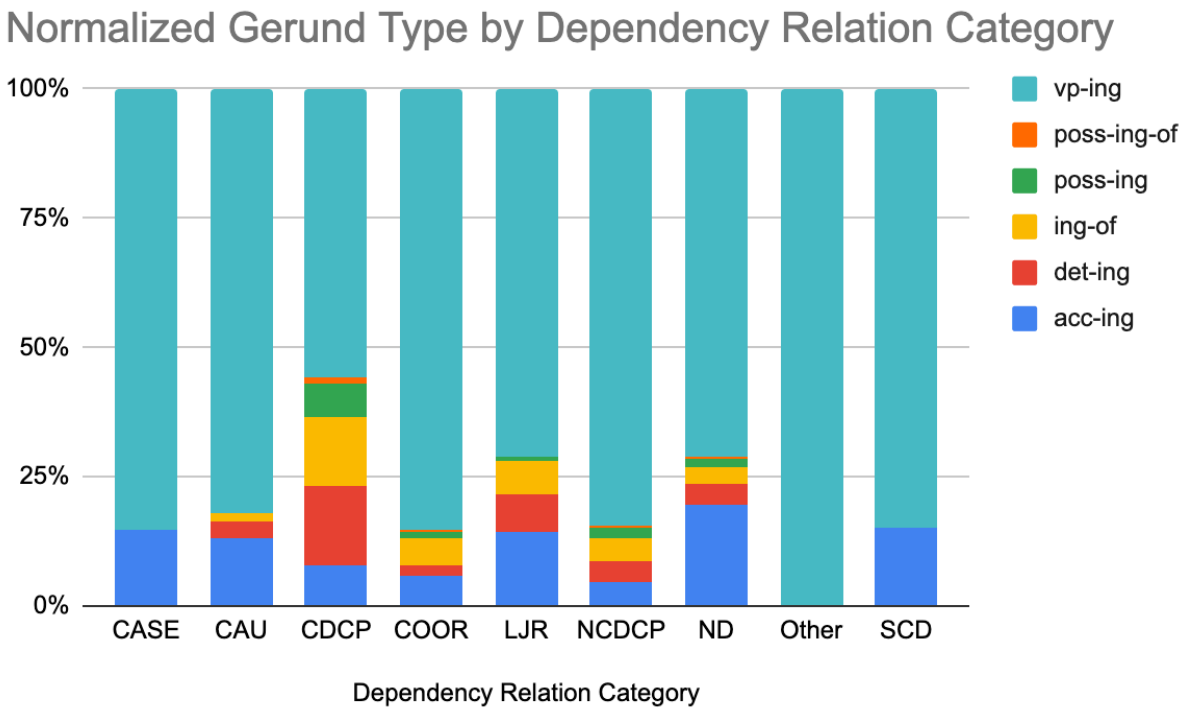


Figure 5. Normalized distribution of gerund type by dependency between gerund and controlling word.

We were going to run a chi square test on these data but ultimately decided against it; the only information that it could have given us was whether the distribution of a given category deviated from the expected distribution, and it would not indicate how they differed. In this respect, the visualizations themselves give a more complete picture of not only the distributions themselves but also of their significance.

4. Discussion

While our data was largely very helpful in drawing conclusions, there were some categories that, while initially looking promising, wound up being counterproductive. We noted early on that the *conj* dependency type would not be helpful—these relations would link a gerund to a preceding noun phrase in a list, rather than the controlling word of that list, which would actually be determining its role in the sentence. There were also some dependency types which had a very small sample size, and looked to be mostly miscategorized. These included the *vocative*, *nmod:tmod*, *iobj*, *dep*, *cc*, *aux*, *advmod*, *flat*, *obl:tmod*, and *expl*, all of which had fewer than five data points. The *nmod:poss* category mostly consisted of bad data, where nouns such as "painting," "building," and "dwelling" were incorrectly classified as gerunds. Due to the infrequency of these occurrences, and the added issue of miscategorization, we avoided drawing conclusions based on these dependency types.

The categories of dependency types in de Marneffe et al. (2014) proved to be less helpful than we hoped. These categories revolved around mostly syntactic concerns rather than semantic ones—for example, whether certain phrases were arguments or adjuncts. They do offer us some

insight, though. Even though there were relatively few dependency relations that licensed all or most of the gerund types, the majority of the dependency categories also licensed all or most of them, indicating that the environments that do license multiple types cannot be easily grouped syntactically. Therefore, even though we did find some interesting trends from the dependency relations, it seems unlikely that syntax is the main factor contributing to the distributions.

While the larger categories were unhelpful, several of these had overlapping subcategories which did provide useful context. In particular, the "nominal dependency" subcategory of the Core Dependents of Clausal Predicates (CDCP), Non-Core Dependents of Clausal Predicates (NCDCP), and Noun Dependents (NP) categories showed incredibly similar distributions, as shown below in Figure 6.

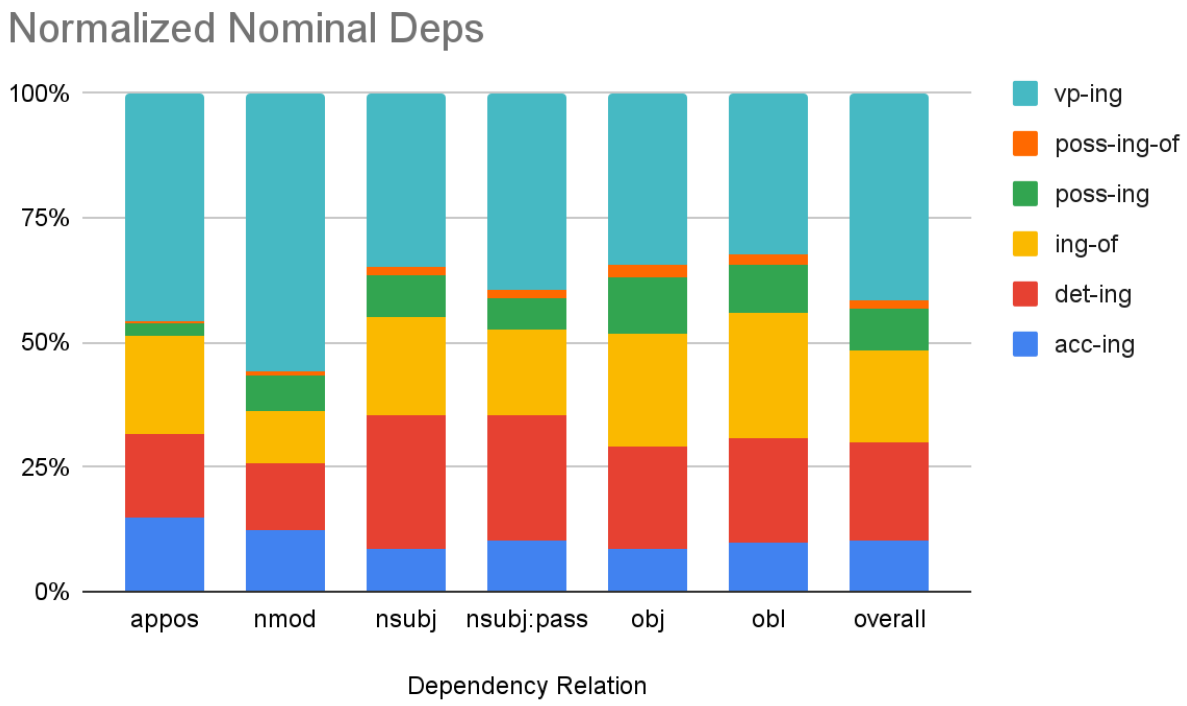


Figure 6. Normalized nominal dependency relations.

While there are some variations here (in particular, *nmod* has far more VP-*ing* gerunds than the other relations), these five dependency types have undeniably similar distributions. The *nsubj* and *obj* relations are used to show the subject and direct object of a sentence, respectively, while the *appos*, *nmod*, and *obl* relations are used to modify noun phrases in specific ways. These dependencies had six of the seven most diverse distributions of sufficient sample size. The fact that all gerund types appear to be licensed under these circumstances, while most of the other dependencies are limited to VP-*ing* and ACC-*ing* gerunds, is the main conclusion we were able to draw from our exploration of the data. That being said, the previous class noted that they had trouble with getting accurate categorizations of ACC-*ing* gerunds, so it is unclear if the size of the ACC-*ing* category is truly significant or if it is due to misclassifications.

There were three other dependency relations that had sufficient sample size and relatively diverse distributions of gerund type: *acl:relcl*, *ccomp*, and *parataxis*. The *acl:relcl* tag is a subcategory of the broader *acl* tag, which describes an embedded clause acting adjectivally, while *acl:relcl* is reserved for relative clauses acting adjectivally. Notably, the standard *acl* relation had a fairly typical distribution consisting of mostly VP-*ing* gerunds together with some ACC-*ing* gerunds. The *acl:relcl* distribution was significantly more diverse, however—only 37% of gerunds governed by an *acl:relcl* dependency were vp-ing gerunds, compared with 77% of the general *acl* dependencies.

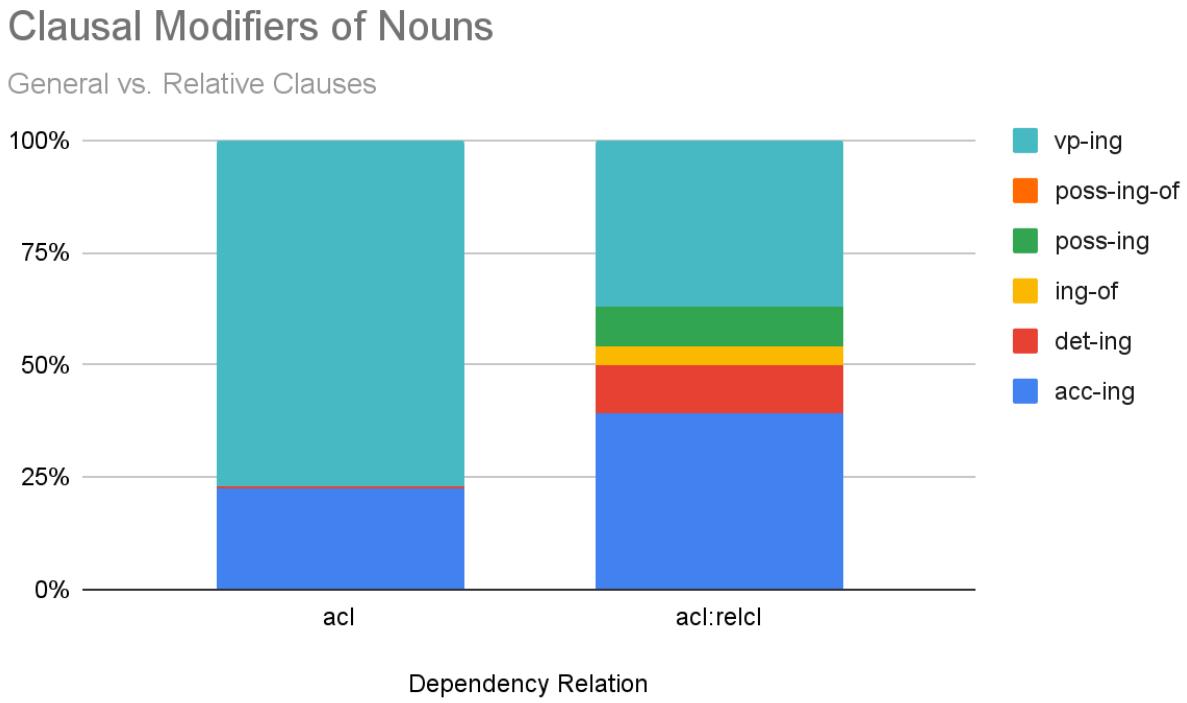


Figure 7. General vs relative clausal modifiers of nouns.

This suggests that relative clauses may license more gerund types than general clauses. It should be noted that the sample size for relative clauses (n=46) was significantly smaller than for general clauses (n>11,000), so this may be a relic of poor labeling or simply anomalous data. It seems that these relative clauses are somewhat rare, so collecting more data specifically about clausal modifiers of nouns may lead to future interesting results.

The *ccomp* relation was also relatively rare, with only 113 occurrences in all of the data. This relation describes a clause which functions as the object of the verb or adjective controlling it. There were significantly more DET-*ing* and -ing_{of} gerunds with this dependency relation than expected, indicating that these types of gerunds are licensed in this case. Notably, however, there were still fewer POSS-*ing* and POSS-ing_{of} gerunds than would normally be expected, indicating that not all gerund types are licensed equally in this environment.

The *parataxis* relation is also rare, with only 90 occurrences. Unlike the *ccomp* relation, though, many of the sentences involving the gerund being governed by a parataxis relationship

were malformed—either incomplete, or somehow involving a colon used outside of its normal grammatical context (e.g., introducing speech)—which suggests that this data was likely miscategorized. It is likely that, without these data points, the *parataxis* data would look very different. Unfortunately, without combing through the data by hand, it is hard to say whether *parataxis* is actually an anomalous relationship.

The only remaining dependency relationship with a particularly interesting distribution is *csubj* and its cousin *csubj:pass*. These relations indicate that the gerund is a standalone clause, and that clause is acting as the subject of the sentence in which it is embedded, *csubj* for active sentences and *csubj:pass* for passive sentences. The *csubj* relation is the only dependency relation of sufficient sample size with over 98% VP-*ing* gerunds, while *csubj:pass* has nearly 8% ACC-*ing* and 2% each DET-*ing* and -ing_{of} gerunds. This indicates that passive subject clauses license additional types of gerunds when compared to their active counterparts.

Apart from the relations we have mentioned above, all of the relations of sufficient sample size consist almost entirely (>95%) of VP-*ing* and ACC-*ing* gerunds. This seems to indicate that those gerund types are licensed in most, if not all environments, while the other gerund types (-ing_{of}, POSS-*ing*, POSS-ing_{of}, and DET-*ing*) require specific environments in order to be licensed.

5. References

De Marneffe, M. C., Dozat, T., Silveira, N., Haverinen, K., Ginter, F., Nivre, J., & Manning, C. D. (2014, May). Universal Stanford dependencies: A cross-linguistic typology. In *LREC* (Vol. 14, pp. 4585-4592).

Department of Computer Science & Department of Linguistics (2022). A Large-Scale Corpus of Gerund Nominalizations. University of Rochester.

Grimm, S., & McNally, L. (2015). The -ing dynasty: Rebuilding the semantics of nominalizations. *Semantics and Linguistic Theory*, *25*, 82.

Seiss, M. (2008). The English -ING form. *Proceedings of the LFG08 Conference,* Miriam Butt and Tracy Holloway King (Eds)*, 2008 CSLI Publications.*

Malouf, R. (1996). A constructional approach to English verbal gerunds. In *Proceedings of the Twenty-second Annual Meeting of the Berkeley Linguistics Society,* 255-266.

Universal Dependencies Contributors. "English Dependency Relations." Universal Dependencies, https://universaldependencies.org/en/dep/index.html.