2015

# A group of subject-verb agreements: finding quantity in group and number

Samuel Drew Ascoli
*Iowa State University*

**A group of subject-verb agreements: Finding quantity in group and number**

by

**Samuel Drew Ascoli**

A thesis submitted to the graduate faculty
in partial fulfillment of the requirements for the degree of

MASTER OF ARTS

Major: Teaching English as a Second Language/Applied Linguistics

Program of Study Committee:
David Oakey, Major Professor
Bethany Gray
Gary Phye

Iowa State University

Ames, Iowa

2015

# DEDICATION

I'd like to dedicate this paper to the father who supported me as well as the students, staff, and faculty at Iowa State University who helped me both exceed my expectations and reign in my thesis work to a manageable scope.

# TABLE OF CONTENTS

Page

DEDICATION ..... ii

ABSTRACT ..... iv

CHAPTER 1: INTRODUCTION ..... 1

CHAPTER 2: LITERATURE REVIEW ..... 5

CHAPTER 3: METHODS ..... 12

CHAPTER 4: FINDINGS ..... 21

CHAPTER 5: DISCUSSION ..... 40

CHAPTER 6: CONCLUSIONS AND FUTURE RESEARCH ..... 46

REFERENCES ..... 50

# ABSTRACT

This study characterizes the subject-verb agreement that occurs with *group of NP* and *number of NP*. These two complex noun phrases can agree with a verb as a singular or plural noun. These two particular items were selected as *number of NP* has a relatively firm description of its quantification behavior described in existing literature while *group of NP* has not been shown to have describable rules governing its quantity. Using data collected from the Corpus of Contemporary American English (COCA), 1200 concordance lines centered on *group of* and *number of* which agree with a verb in a clause were extracted for study of several co-occurring features. Individual features such as determiners and modifiers are examined with respect to their distribution with singular or plural-agreeing verbs to identify patterns of agreement and potentially indicate trends, if not causal relationships. Some features, such as determiners preceding the first noun *number*, show trends with respect to the verb-demonstrated quantity of the noun phrase. Other features, such as premodifiers on either noun in *group of NP* do not appear to co-occur in demonstrable patterns. By creating a description of quantification in this way, this study lays the foundation for more targeted future studies of quantification in cognition, grammar, and semantics.

# CHAPTER I

# INTRODUCTION

Lexical items described as "quantifying collectives" in *The Longman Grammar of Spoken and Written English* (Biber et al., 1999) such as *group* or *number* have been problematic to teach in an EFL environment. The rules of quantification, that is to say the manner in which a noun phrase is determined to be singular or plural, can be unclear in complex noun phrases which contain multiple nouns. The major issue that arises with this grammatical group is the issue of subject-verb agreement (SVA). Differentiating the third person conjugation in English is an early and relatively simple series of lessons (examples invented for illustrative purposes)[1]:

1. *I* ***see*** *him.*
2. *She* ***sees*** *him.*

However, nouns of this "quantifying collectives" class will, in some circumstances, not agree with the verb in this way. Instead, the verb agrees with the second noun in the complex noun phrase (invented examples):

3. *A bunch of us* ***are*** *going to the movies.*
4. *A bunch of grapes* ***is sitting*** *on the counter.*
5. *In some states, a number of voters* ***seem*** *to have election fatigue.*
6. *The number of visitors* ***increases*** *every year.*

This is not a universal, though. Agreement can be varied, the rules of which are not totally complete (invented example):

7. *A group of fifth-graders* ***are*** *on a field trip.*
8. *A group of fifth-graders* ***is*** *on a field trip.*

[1] In all examples, the agreeing subject NP is underlined while the agreeing verb is in bold.

This variation potentially leads to confusion on the part of English language learners (ELLs), who may not have the implicit understanding of when or why the quantity of the whole noun phrase will change. It is the circumstances surrounding this issue that this paper seeks to describe in greater detail, specifically the types of modifiers and determiners that occur with the quantifying noun phrases and whether these features appear to have any trends of distribution with regards to SVA

In order to pursue this description, this paper will focus on two items that fit this pattern, *group of NP* and *number of NP*. According to *The Longman Grammar, group of NP* is one of the more common and more flexible "quantifying collectives" (Biber et al., 1999, p. 248). *Number of NP*, on the other hand, has much stricter rules governing agreement in SVA, given that there are different forms of *number of NP* with differing semantic roles (Biber et al., 1999, p. 185 and 248; Huddleston & Pullum, 2002, p. 339, 350, 502-503). For the purposes here, these nouns phrases will be referred to as quantifying noun phrases (QNP). The nouns that follow these QNPs are referred to as second nouns rather than subordinate nouns to avoid any presuppositions of pre-existing grammatical structure. An overview of this terminology and placement is demonstrated in Tables 1 and 2. Also dealt with are the differently-ordered sentence structures. The typical ordering of sentences follows a subject-verb-object ordering. However, there are instances in which the verb will precede the noun. This happens in, for example, instances of subject-auxiliary inversion and existential *there* forms.

Table 1: Quantifying Noun Phrases (QNP) in subject-verb-object ordering (SVO)

| **first noun** | **preposition** | **second noun** | **agreeing verb** | **object (optional)** |
|---|---|---|---|---|
| number | of | NP | X | X |
| group | of | NP | X | X |

Table 2: Quantifying Noun Phrases (QNP) in verb-initial ordering (VI)

| **agreeing verb** | **first noun** | **preposition** | **second noun** |
|---|---|---|---|
| X | number | of | NP |
| X | group | of | NP |

The pedagogical materials aimed at teaching general English grammar and SVA generally only address the basics of subject-verb agreement and higher-frequency "rules" that follow generally repeatable guidelines. For examples, descriptions exist of the relationship between determiners and the subject-verb agreements of *number of NP* (Huddleston & Pullum, 2002, p.350, 502-503; Biber et al., 1999, p. 185). Part of the issue is that certain forms, such as *group of NP*, have not been adequately described with regards to subject-verb agreement, let alone had pedagogical materials developed for them. This study aims to describe SVA with *group of NP* and *number of NP* and noun phrases similar to it as a way of narrowing the gap between grammatical description of the phenomenon and its teaching to nonnative speakers of English. Examining *group of NP* and contrasting it against the far more regular *number of NP* will not only allow for a better description and more solidified classification of *group of NP*, but also provide a precedent for more expanded studies of similar quantifying noun phrases. By describing *group of NP* forms and *number of NP* forms by their verb agreement traits, this study will hopefully lead to the development of methods and materials to aid non-native speakers in acquiring native-like comprehension of nouns of this class. By starting with a few important QNPs, such as *group of NP* and *number of NP*, this study serves as a first step towards the development of materials for later studies and ultimately the classroom.

This study first reviews SVA and collective nouns in the literature on grammar and on ESL/ EFL pedagogy, then describes the corpus based methodology used in this study to create a

description of these forms as seen in real-world language use. It then presents detailed quantitative findings, and discusses their implications in the light of the study's methodological limitations, and concludes by outlining avenues for further research.

# CHAPTER 2

# LITERATURE REVIEW

While there is a grounding for this research in longer grammar reference texts (Huddleston & Pullum, 2002; Biber et al., 1999; Radden & Dirven, 2007; Hunston et al., 1998; Quirk et al., 1985), and shorter research articles (Landman, 2011; Markman, 1985; Nicolas, 2008) which have addressed the complexity of quantifying noun phrases in English, none of them have looked at the issue from a subject-verb agreement perspective and will thus have only cursory introductions here. The shorter research articles look at smaller, more specific instances. Landman's (2011) article, for example, attempts to categorize noun phrases by their count/non-count features as well as the semantic roles that allow these features to shift between usages. Nicolas (2008), on the other hand, spends most of his article dealing with how context allows for plurality to emerge in the mind of the language user to justify the assessment of singular or plural qualities on ambiguous noun phrases.

Likewise, some of the other longer texts reviewed, such as *A Comprehensive Grammar of the English Language* (ComGEL) (Quirk et al., 1985) lacked the necessary depth of two later-written texts have with regards to group noun phrases such as those discussed here. Therefore, it was not used as a primary reference in this study. Two aforementioned recent grammar reference texts have greatly expanded on this area of description where ComGEL lacks. Another well-known grammar reference text that this study does not draw on is *Grammar Patterns 2: Nouns and Adjectives* (Hunston et al., 1998). This grammar reference book largely deals with pattern grammar rather than the semantic roles of the noun phrases. While useful as a means of perspective, this text provides descriptions of distribution and collocation; these targets are not features relevant to the specific subject-verb agreement observations that this study centers on.

The two grammar reference texts that provided the grammatical framework this study aims to build from are *The Longman Grammar of Spoken and Written English* (LGSWE) (Biber et al., 1999) and *The Cambridge Grammar of the English Language* (Huddleston & Pullum 2002). These two texts describe the grammatical nature of English: the former takes a corpus-driven approach and the latter a more theoretical and semantic interpretation. Interestingly, *The Longman Grammar of Spoken and Written English* shares a publisher, a corpus methodology, and a coauthor, Leech, with ComGEL. Comparison of the two texts shows some shared structure as well. However, its later publication and more advanced descriptive grammar techniques and corpus allowed for a more rounded description of many items, including group-noun phrases.

Huddleston and Pullum's text accounts for some of the distinctions between assessments of singular or plural of these noun phrases through verb agreement, such as in the following Examples 9 and 10 (Huddleston & Pullum, 2002, p. 502):

*9. A number of <u>spots</u>* ***have/*has*** *appeared.*

*10. Heaps of <u>money</u>* ***has/*have*** *been spent.*

Examples such as these do little to explain how or why the subject shifts from the first noun, *number* and *heaps* respectively, to the second noun (Huddleston and Pullum 2002, p. 501-503). Instead, they group them together with "number-transparent nouns," which have similar SVA irregularities, using the following example (Huddleston & Pullum, 2002, p. 501):

*11. The committee* ***has/have*** *not yet come to a decision.*

The analysis in Huddleston and Pullum's text deals with the focal shift of the language producer in Examples such as 9-11 above. The shift is a semantic one, examining whether the constituents or the collective are the intended target of agreement.

Of these common noun phrase forms, the most common is *a lot [of]*, which occurs so frequently in an unaltered sequence in both corpus data and elicited responses that it is frequently described as the lexical bundle *a lot of* (Biber et al., 1999, p. 276; Huddleston & Pullum, 2002, p. 349-350). Indeed, the LGSWE indexes *a lot of* under *A* rather than *L*, while *a number of* is categorized under *N* (Biber et al., 1999, p. 1,149 and 1,162). Several texts classify it as a quantifier (Biber et al., 1999; Berry, 1997; Hunston et al., 1998). It does not agree with verbs under these circumstances, while the second noun does. It appears as an adverb in other circumstances (invented examples):

*12.* ***A lot of*** *students are protesting the meal plan changes.*

*13. I like this movie* ***a lot****.*

Quirk et al. (1985) addresses this form in addition to other nouns like *deal*, *amount,* and *quantity* as determiners rather than head nouns of a noun phrase (p.264). This form has been apparent as an anomaly for some time. The form in Example 13 functions as an adverbial, modifying *like*. This all is to say that high frequency forms such as these have near-universal rules for their behavior surrounding subject-verb agreement which can be easily described. An unpublished study conducted last year (Ascoli, 2013) using data from Corpus of Contemporary American English (Davies, 2008) found 2,327 instances of *a lot of* that, as a noun, did not agree with the verb. This was contrasted against 155 examples using the same search criteria that apparently agreed with the verb. The anomalous 155 instances can be explained through conflicting patterns, which will be described in greater detail in the methods section. Ascoli (2013) also included the spoken subcorpus from COCA, which this present study excludes. *A lot of* is a singular example, whereas the whole class of noun phrases that fit into this category does not have the same transparent rules in usage. Other nouns within this community exhibit distinct

and interesting behavior with regards to subject-verb agreement when used by native speakers of American English. *Group of NP* is one such variable form.

Indeed, there is frequent variation within these categories. The Collins Online English Learners Dictionary, developed from data in COBUILD, has an odd and telling phrasing of the definition of *number*:

*14. If there **are** a number of things or people, there **are** several of them. If there **are** any number of things or people, there **is** a large quantity of them*

Example 14 demonstrates that *a number of* appears to behave as a quantity/quantifier, particularly in the second sentence. Conversely, data taken from the Corpus of Contemporary American English (Davies, 2008) indicate the inverse:

*15. Furthermore, as the number of pairs **increases**, the overhead scales very well.*

A pattern emerges with samples such as these taken from COCA: *a number of* and *the number of* have different relationships in subject-verb agreement. While *number* is described as a quantifier in texts such as *The Longman Grammar*, it is this kind of variation that is not addressed (Biber et al., 1999, 278). Huddleston & Pullum postulate that this is the result of the use of the indefinite article, which acts to "subdivide knowledge" (Huddleston & Pullum, 2002, p. 339). Articles, of course, are a constituent of the determiner class of lexical items. This is one example of the system they describe as "plural overrides," which does not describe a clear system by which speakers determine plurality. Instead, they are options (Huddleston & Pullum 2002, p. 501-503). However, we see that there is variation within subject-verb agreement with similar first nouns. If the indefinite articles *a* and *the* signal such a shift in meaning (as "cause" may be giving the articles too much agency) that subject verb agreement changes its target noun, one wonders why this would be, and what else could have this effect?

Also emergent in data examined from Ascoli (2013) is the apparent lack of patterns with the quantification noun *group of NP*. The study revealed no distinct pattern of subject verb agreement: some verbs apparently agreed with *group,* others agreed with the second noun in the phrase. This study looked at overall usage, rather than the distinct variables that come into play (as with *number of NP*). Additionally, the scope of the study used only a handful over tag-searched queries into the corpus which limited the data output. However, the findings of this limited study contradict the claim in Biber et al. (1999) which describes *group of NP* as a collective noun, which is ostensibly singular (p.248), but essentially undefined in practice. The two approaches taken by these two texts, the corpus-driven *Longman Grammar* and *The Cambridge Grammar* both discuss and categorize these collective, quantification, and group nouns, but neither has been adequately able to explain their subject-verb agreement clearly.

Also important to this analysis are the distinctions between determiners. The description and breakdown of the determiners are based on several of the aforementioned texts as well as *English Guides 10: Determiners and Quantifiers* (Berry, 1997). The text very clearly defines the multiple types of determiners and the features that each constituent determiner or group of determiners carries. This text's description serves as a reference for the various determiners found within the data collected (Berry, 1997, p. 18-19).

This is an area of interest for ESL and EFL pedagogy. English language learners struggle with subject-verb agreement with regards to complex noun phrases. Dziemianko's (2008) study addresses the opacity of the rules and practices of similarly complex noun phrases in the context of Polish learners of English. Other prior research into native speakers' understanding and adjustment of quantification already exists in the research realm (Humphreys & Bock, 2005). The complexity of these arrangements makes creating a scheme to describe how quantification is

applied to such complex noun phrases useful for those non-native speakers who wish to achieve native-like proficiency.

Furthermore, developing a more rounded description of *group of NP* and *number of NP* and their quantification idiosyncrasies may be an element of understanding how native speakers understand quantification itself. Some literature indicates that there are links to other fields. Domahs et al. (2012) examines plural cognition in neuroscience while Nicolas (2008) and Cocchiarella (2009) take the semantic implications of plurality towards logic and philosophical implications. There are also immediate applications with software designed to check users' spelling and grammar. MacWhinney (2008, 2013), has written several articles on the Unified Model which attempts to explain how multiple competing rules apply which appears to be a factor here as well.

In order to assess the manner by which quantification is applied, there are several research questions which, when taken in order, develop an understanding of what grammatical features co-occur with differing realizations of quantification. The first research question aims to assess the extent and breadth of the issues regarding *group of NP* while also searching for outlying examples which don't have the expected SVA in the otherwise well-defined *number of* NP. Following those observations, the study will look at specific co-occurring features such as determiners and modifiers within the noun phrases to establish any trends. Finally, the study seeks to compare the observations made with existing literature. They are:

Research Questions

I. What do observations from a corpus of American English indicate about how *group of NP* and *number of NP* noun phrases determine quantity in subject-verb agreement? How does SVA vary for clauses containing subjects in the form of g*roup of NP* and

*number of NP?* How consistent is the target noun of subject-verb agreement with *group of NP* and *number of NP* in COCA?

II. How do modifiers and determiners within gr*oup of NP* and *number of NP* forms affect the semantic roles and language users' agreement of the noun phrases and verbs? How do modifiers and determiners in subjects in the form of gr*oup of NP* and *number of NP* affect the variation in Q1?

III. Do subjects in the form of gr*oup of NP* and *number of NP* function consistently as a determiner in the way that Quirk et al., Biber et al., and Huddleston & Pullum described other, similar noun phrases?

# CHAPTER 3

# METHODS

The aforementioned Ascoli (2013) study was very superficial in its approach to data collection, which is why a more intensive study of these complex noun phrases is to be undertaken here. Differences include using untagged corpus data, very broad search terms, and a much more thorough, qualitative assessment of the features in the collected concordance lines. The first issue is to select a corpus as the database from which to collect data. The next section discusses the corpus and the reasoning behind its selection.

## Corpus

The Corpus of Contemporary American English (Davies, 2008) was chosen for this study for a variety of reasons. The American context of the corpus allows for accessing a superordinate variety of English. It is comprised of 464 million words online and 440 million words in the downloadable version, the latter being the source of this study's data (Davies, 2010). This enables a large amount of data from a wide variety of sources to be examined. Furthermore, COCA attempts to present a balanced view of the language by including samples from several distinct registers collected into separate subcorpora of usage across a roughly twenty-year span, as can be seen in Figure 1. This corpus is designed to be a monitor corpus, one which continues to update to keep up with present language use, though it has not been updated since mid-2012. This allows for very recent language usage to be the source of observations. The texts sampled come from dozens of individual sources within each subcorpus, which in turn are comprised of

thousands of individual texts. The publications based on this corpus are numerous, with the website itself linking to more than five hundred publications citing it since 2008 (COCA 2008).

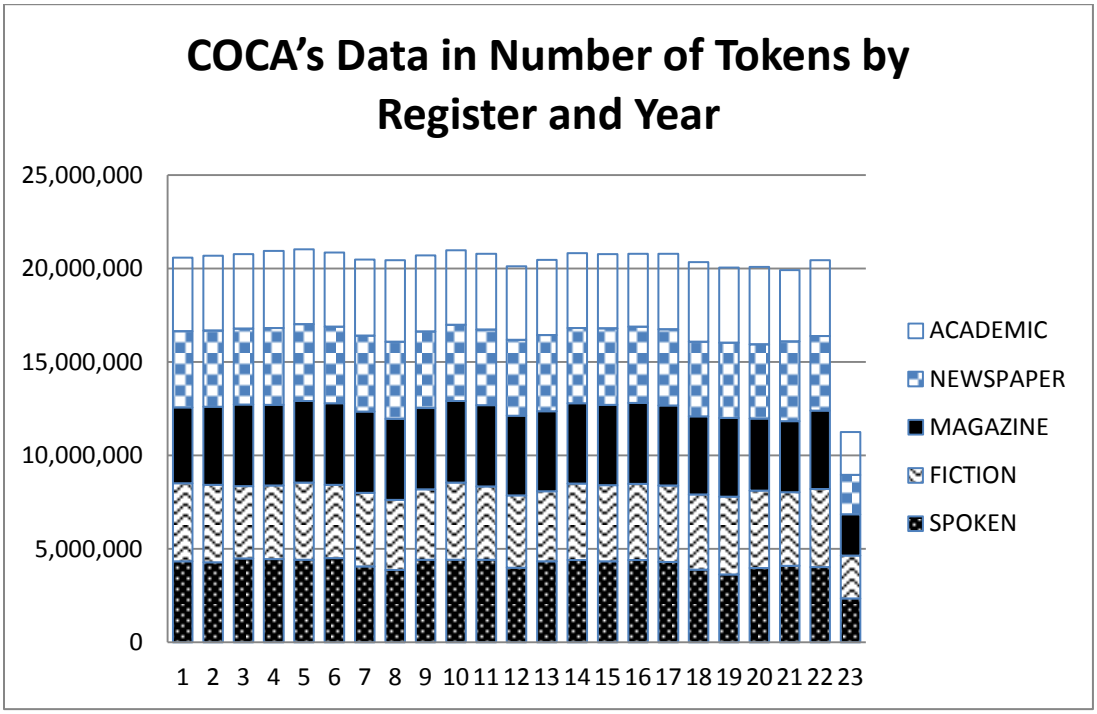



Figure 1: COCA's Data in Number of Tokens by Register and Year

COCA's balance amongst registers will allow for either further research for this study or another at a later date. COCA is broken up into several independently searchable subcorpora: Academic, Newspaper, Spoken, Fiction, and Magazine. As Figure 1 shows, the registers themselves are maintained at roughly four million words per subcorpus per year. The data for this study were collected separately for each of these register-subcorpora, excepting Spoken. This was done to allow for additional insights which could be made from the comparison of registers. Having the data sorted by register also allowed for additional analysis of the already collected data if register proved to be an important factor. Indeed, there were some register differences which are discussed in the findings section. However, the relatively small sample of concordance lines analyzed did not produce consistent enough samples from enough individual texts to complete an effective and thorough register analysis. In order to do a thorough register

comparison, a sampling that specifically targets the texts and journals--information which is available in COCA--with larger amounts of register-specific data would make for a more viable register study. The collected data could be used for or influence the development of a later study with a specific focus on register, which is discussed in more detail in the conclusion, while this study focusses on several potential trends.

Each of the five subcorpora of COCA were to be examined separately. Unfortunately, the data from the spoken register could not be relied upon[2]. In a future study, research could be conducted using spoken corpora that have been developed from data that were transcribed with linguistic study in mind, such as MICASE. Alternately, a purpose-specific corpus could be collected. The exclusion of spoken data also largely absolves the study of reconciling the speaker "errors" and the particular idiosyncrasies of the spoken subcorpus with the four written subcorpora. Speaking "errors," particularly in quotations occurring in the Newspaper and Fiction registers are identified and discussed, but the rate is lower than would be expected from a specifically spoken set of data.

The data in COCA are accessible both tagged and untagged. Tagged data allow for searches to be conducted for items of a lexical class, such as verbs or nouns. While more expedient, there are major drawbacks to the usage of tagged data. A tagged corpus may have far greater facility, as making generalizations across classes in this manner required sorting through tens of thousands of concordance lines to subsequently code twelve-hundred lines.

[2] While the website claims an accurate transcription method, the rationale on the COCA website is based on a very small control sampling (Davies, 2008). The transcriptions of the spoken words are opaque in origin, having been taken from television and newsreaders and transcribed via automated devices. As we cannot assume that the data were transcribed with linguistic features or linguistic awareness, the reliability of the transcriptions is questionable for studies which focus on minute details. As evidence, a sample transcription of an NPR interview (NPR being one of COCA's spoken register sources) was compared to the source audio. It did not match up with total accuracy (Davies, 2013). These transcripts also lack demographic information on speakers, though this is also true of the authors of COCA's other data.

Unfortunately, the tagging of data is not perfect—for instance, more than 800,000 instances of the word “no” are untagged in COCA, despite samplings overwhelmingly appearing to be determiners The tagger used in COCA, the CLAWS7 tagset, does not consistently tag items as either LGSWE ((Biber et al., 1999, p.68-71), Huddleston & Pullum (2002, p. 352-355), or Berry (1997, p. 18-19) define them. Indeed, LGSWE states that the determiner *no* is one of the two most highly frequent determiners used across all registers, but particularly in fiction (Biber et al., 1999, p.277). This is problematic in general usage of the corpus, let alone that this study is very much interested in determiners. Additionally, tagged text requires that the researcher have a search string in mind when searching. Given the vast amount of innovation and complexity that English allows, it would be unfeasible to predict accurately each and every formation in which *group of NP* and *number of NP* could possibly occur. Therefore, searching for the strings *group of* and *number of* in the untagged text and filtering the data to meet the inclusion (discussed below) criteria data is the logical, thorough, though not expedient approach.

## Collection and Coding

Once the remaining four subcorpora were accepted, data were collected from each via the offline corpus. The offline data were used primarily because it does not require retrieval through the online interface, which limits amounts of data displayed at a time and less flexibility in data displays. Using AntConc, a concordance program with a more flexible interface (Anthony 2014), the strings *group of* and *number of* were searched in the four subcorpora. The resulting body of data was large, as seen in Figure 2. The data collected were too large to be coded in its entirety. However, this collected data also needed to be sorted to meet a variety of conditions to be useful in the analysis of the feature: specifically, the *group of* and *number of* search terms first had to be determined to be the agreeing noun phrase.

These agreeing QNPs then had to be culled for verbs which did not have opaque agreement. This was more of a problem to the coding process, as verb conjugation and tense is frequently not transparent in the agreement. Either a verb or an auxiliary appearing as present tense was needed in order to determine singular or plural. Modals block any apparent singular or plural agreement in the verbs phrases in which they occur. Another problem occurred with regards to exceedingly long second-noun phrases. The concordance lines collected were twice AntConc's default length at one hundred characters. However, there were a number of second noun phrases which did not complete and show a verb before the concordance line ended. These were discarded rather than re-collecting data with longer concordance lines for several reasons, but primarily that the distance between the beginning of the noun phrase and the verb is also a variable.

Under these inclusion criteria, the study design took the first fifty usable concordance lines of each thousand then began sampling from the next thousand in examining *group of*. The usability rate varied quite a bit text to text, but the overall average was about one concordance line in eight was accepted. As a result, the bulk of the data come from roughly between the first ~2500 concordances lines, with a few lines taken from later in the data to replace initially miscoded lines. The lines are also in the chronological order of the corpus so much of the sampling is taken from the 1990's.

The data sets sampled here are quite small relative to the size of the initially collected data. This is an issue with the methodology employed. While roughly 120,000 concordance lines were collected, the majority of those coded were discarded as not meeting the inclusion criteria. To have an accurate sampling, say, 20% of the usable data, all ~120,000 concordance lines would have to be coded for the inclusion criteria, mooting much of the purpose of sampling.

Additionally, the *number of* dataset would be much larger than the *group of* data. While not wrong, there appears to be much more variation within the data of *group of* shown here. The samples used here, 600 lines of either set of data, is meant to provide a snapshot of the larger dataset without exceeding the limitations of a qualitative study of each sampled concordance line.

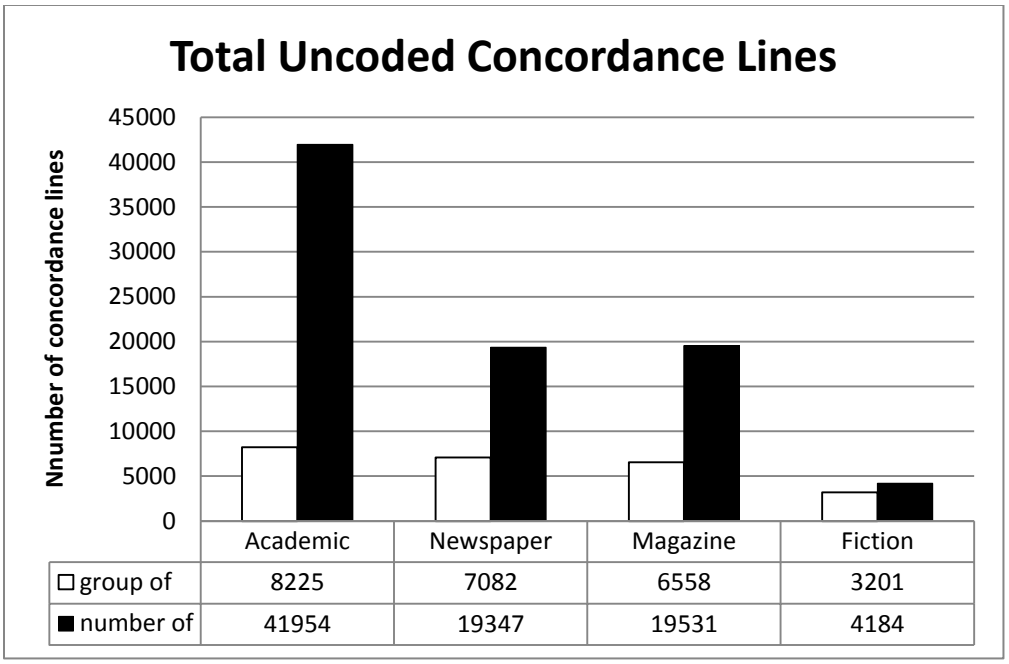



Figure 2: Total Uncoded Concordance Lines

Consequently, the approach to sampling *number of* was reconsidered and made to more accurately reflect distribution amongst years. The samples were taken from every hundred concordances lines rather than from every thousand (except academic). Because of the varied sizes of each of the subcorpora, a different number of samples were taken from every hundred lines. In fiction, 3-4 concordance lines were taken from each hundred, while in the academic register 3-4 were taken from each thousand lines. The logistical issue of having limited manpower for manually coding data represented an important factor in deciding how to proceed with sampling. Ideally, a larger sample would be taken for enhanced accuracy for statistical

analysis. Instead, the sampling provided enough natural language usage of the nouns phrases to provide preliminary observations about the roles of determiners and modifiers with these complex noun phrases.

In coding, there are several criteria. First and foremost, the quantity of the first noun, *group* and *number*, must differ from the second noun. Subjects containing coordinated noun phrases were discarded since coordinated phrases create ambiguity when assessing agreement:

*16. J.M. Burkholder and a group of her colleagues from North Carolina State University* ***have studied*** *toxic blooms[…]*

*17.* 29 *...United Parcel Service wants to expand a distribution center, but Ms. Messinger and a number of community groups* ***have*** *concerns that the area is already too congested*

*18. […]an individual or group of individuals* ***fails*** *to find sexual opportunities[…]*

Examples such as 16-18 were necessarily excluded, given the ambiguity of the agreement. As for the contrast in quantity, no assurances could be made of the second noun. However, only two instances of an uncountable noun occurring in the second position to *group* during coding:

*19. If a group of cannibalistic stickleback* ***approaches*** *a male's nest too closely, the male swims rapidly[…]*

*20. A diverse group of fish, salmon* ***are found*** *in a variety of settings*

This is appears similar to Example 11, which allows for these groups, *stickleback*, *fish*, and *salmon* to be perceived as either singular or plural. Additionally, the apposition of fish and salmon makes determining which of the two quantity-ambiguous nouns the verb agrees with ambiguous. Therefore, they must be excluded. *Number of NP*, on the other hand, had one type of singular second noun that regularly occurred:

*21. The number of the White House opinion line* ***is*** *( 202 ) 456-1111*

*22. And then he was in Memphis and found the number of the house. He opened the door and went through the silent foyer and mounted the stairs*

*23. When these images were catalogued, the number of the associated face or hand cast* ***was*** *often noted.*

Telephone numbers, addresses and numerical designations had to be excluded, though they were not in a noticeably high distribution. These are somewhat different from other instances of *the number of NP*s as they are not counts or quantities so much as they are numerals. Examples 21-23 are important though, given that each of them has the definite article *the* on each of the second nouns, a point which will be expanded upon in the findings and discussion sections. Another issue touched upon earlier illustrates a false instance of another issue: apposition. The above example could be misread to have the verb agree with the *group* noun phrase. However, unambiguous versions of this confusion did occur within the coded data:

*24. Suppose, further, that a group of potentially influential commentators, the retrenchment theorists,* ***are*** *deeply troubled by[…]*

*25. the dominant ethnic group of Yola , the Fulbe , have thus far been resisting full incorporation into this tradition[…]*

Examples such as these were also excluded. This is due to the addition of another variable: confused agents. In both of these instances, the agent has been stated and then restated. This leads to confusion in the quantity of the agent. In Example 24, the agreeing noun phrase could be the *group of potentially influential commentators,* or it could be *the retrenchment theorists.* Since the verb agrees as a plural, either could be the agreeing target, so the data are unusable. Example 25 takes this confusion a step further: *a group of Yola* itself cannot be clearly defined, as *Yola* is, as several previous examples are, both countable and uncountable in surface realizations. This is

also true of the tribe's name, *the Fulbe,* which also confuses the agency through apposition. While instances that contain a subject-auxiliary inversion or null subjects are included and addressed as a class, this ambiguous agreement represents an additional point of confusion that could not be reconciled with the data-inclusion criteria.

Finally, there were several instances found where *number of NP* had no actual determiner:

*26. ...disarticulated skeletons-mostly leg stiffeners. Even number of legs* ***was*** *no criterion, as both groups tended to lose legs to other uses.*

*27. The current yield on his debt will appreciate X months later than current appreciation date. Number of months* ***are determined*** *by the level of the infraction.*

The forms in Examples 26 and 27 are apparently different in configuration. In 26, *number of legs* appears to be a category. The context indicates that it is a category of criterion, similar to "how many legs the table has." As this was coded as a formation that did not affect the sought-after data, it was excluded and potentially any other concordance line similar to it, though no others were found. In Example 27, the expanded context seemed to indicate that this was taken from a legal document. The absence of the determiner was seemingly unnatural, unless it, too, was a category of criterion in the same vein as Example 26.

# CHAPTER 4

# FINDINGS

After the exclusion of examples described above 600 concordance lines of each form were analyzed. The data break down into two categories: *group of NP* and *number of NP*. As the latter set of data has been more thoroughly described in existing literature, it is analyzed first here so that similarities and differences can be identified and extrapolations can then be made with regards to the less well described *group of NP* data.

## Number of NP

To begin, the well-established *number of NP* forms were examined for the determiners occurring before *number*. Of the 600 lines coded, 326 instances of *number of* NP were preceded by the indefinite determiner *a/an*, 252 occurred after *the*, 18 by *any*, and the remaining four were grouped into the class titled *possessives*. The distribution can be seen in Figure 3. This study focuses on describing observed features which can in turn be used as the basis of a more statistically-minded focused study or studies in the future.

*12. Since a larger number of people **like** to play, it is easy to drum up a majority*

*13. The number of deaths at their hands **has risen** as they have increased their operations*

*14. But as any number of columnists and congressmen **have pointed out**, the United States imports only 12*

*15. ...our number of accounts **has grown** 29 percent*

As previously described (Biber et al., 1999, 184-185), most of the noun phrases determined by *the* saw the verb agree with the singular *number*, while most of those instances with the determiner *a/an* agreed with the plural second noun. Deviations from this

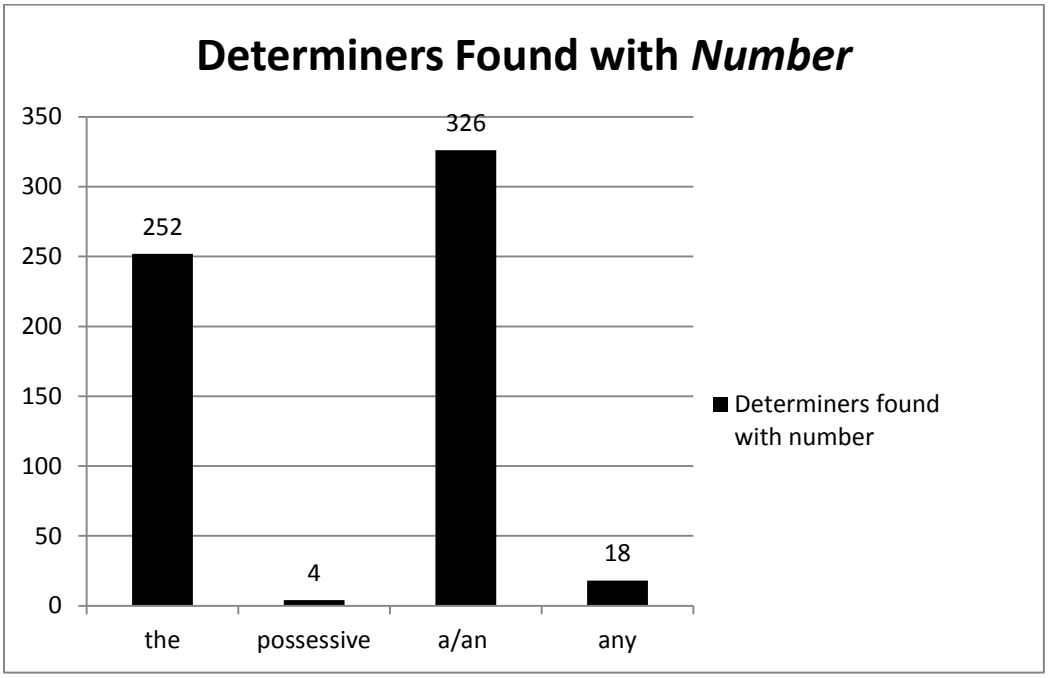



Figure 3: Determiners Found with Number

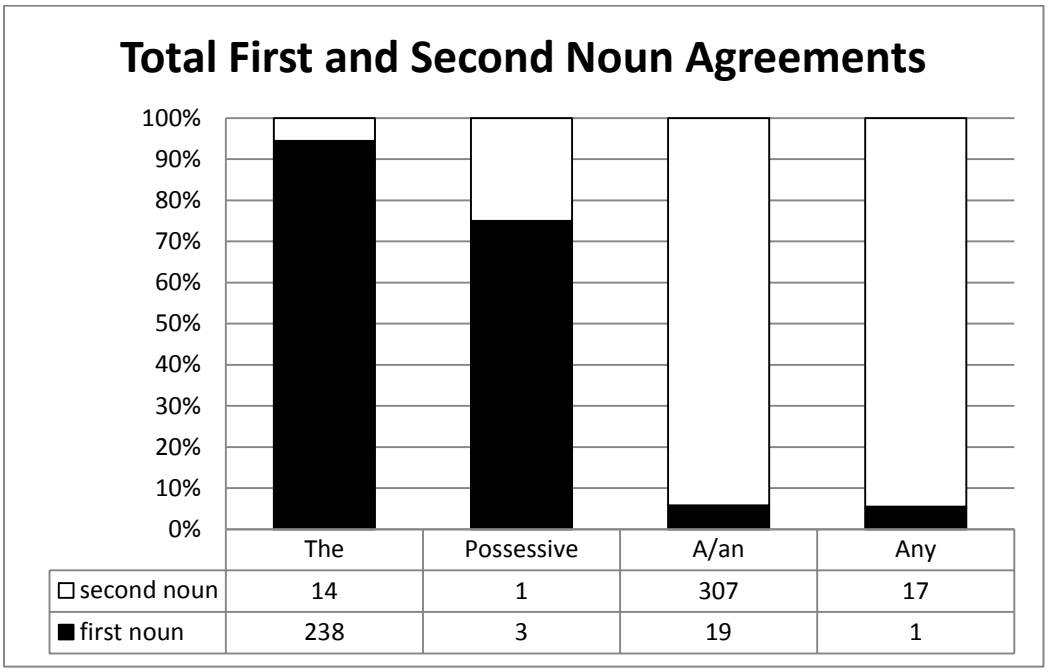



Figure 4: Total First and Second Noun Agreements

prescription are represented in Figure 4. The data only included phrases where *number* was determined by one of four determiner classes: *the*, *a/an, any*, and *possessive*. These latter two

categories, though having a relatively small representation in the data, seemed to largely follow along definite and indefinite lines in terms of agreement. *Any number of NP* primarily saw verb agreement with the second noun, similar to *a/an*. Of the four instances of *number* determined by possessives in the data, three saw verb agreement with the first noun, in a similar pattern as *the*. While only these four determiner categories were included, other examples occurred outside of the necessary inclusion criteria:

*16. No number of consultants or nurses, with their soothing and sympathetic hushed tones,* ***could help ease***

*17. This number of crushed lamps could release 2,250 to 9,600 mg of elemental mercury vapor*

While Examples 32 and 33 above lack transparent subject-verb agreement information as a result of the modal *could*, Example 27 posed a problem in coding which ultimately lead to its exclusion. English only allows singular nouns to be undetermined in isolated instances. In this case, the expanded context indicated that *number of months* is in a different semantic role than the two major categories of *number of NP* discussed herein, the definite and indefinite. It is possible that the phrase is referring anaphorically to a categorization. It also appears to be an instance of "legalese," a form of English which often appears awkward or unnatural to speakers outside of its discourse community (Trossberg 1997, p. 13-14). This type of register-based idiosyncrasy was also considered in the exclusion of Example 27. As such, it was excluded from the 600 lines of data for its idiosyncratic behavior. This example could potentially be a part of a more targeted study which will be discussed in the conclusion section.

Colligation, the phenomenon in which words or strings of words occur more frequently with certain grammatical classes (Römer, 2005, p. 13; McEnery et al., 2006, p. 82), may also

play a role in SVA with regards to *number of NP*. In this case, the semantic class of the verbs should be considered in addition to the grammatical class. When *the number* occurs and typically agrees with the verb, the verb usually reflects something that a number would likely do in the context, but a second noun with agency would not. In Example 28, the verb phrase *like to play* agrees with the agent *people*; numbers, as a general guideline, don't play, though musicians can play a number. Additionally, numbers that *like* are generally anthropomorphized in some metaphorical context (e.g., "There are 60 minutes in an hour because the number 60 *likes* to be divided so many different ways.") Conversely, in Example 29, *has risen* agrees with *the number*, because numbers rise and fall. In this context, the second noun *death* does not logically rise or fall.

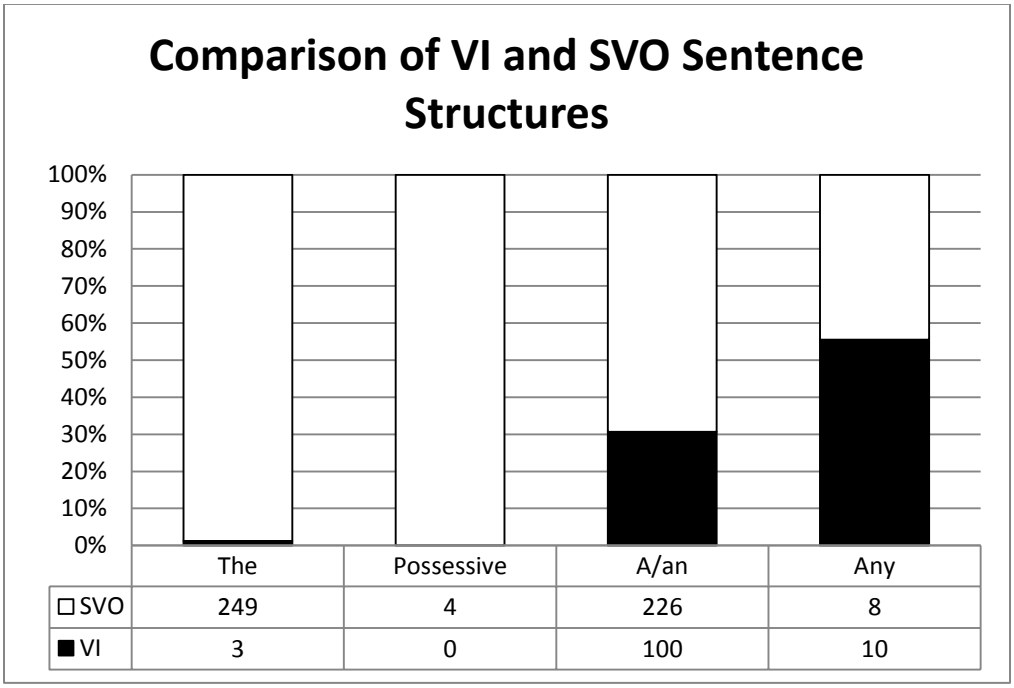



Figure 5: Comparison of VI and SVO Sentence Structures

The metric of Figures 5 and 6 is showing verb agreements in verb-initial (VI) ordered sentences. An initial observation is that *a/an* and *any* have relatively high frequency in VI-ordered this way, primarily in null/dummy subject configurations, as well as occasional other forms:

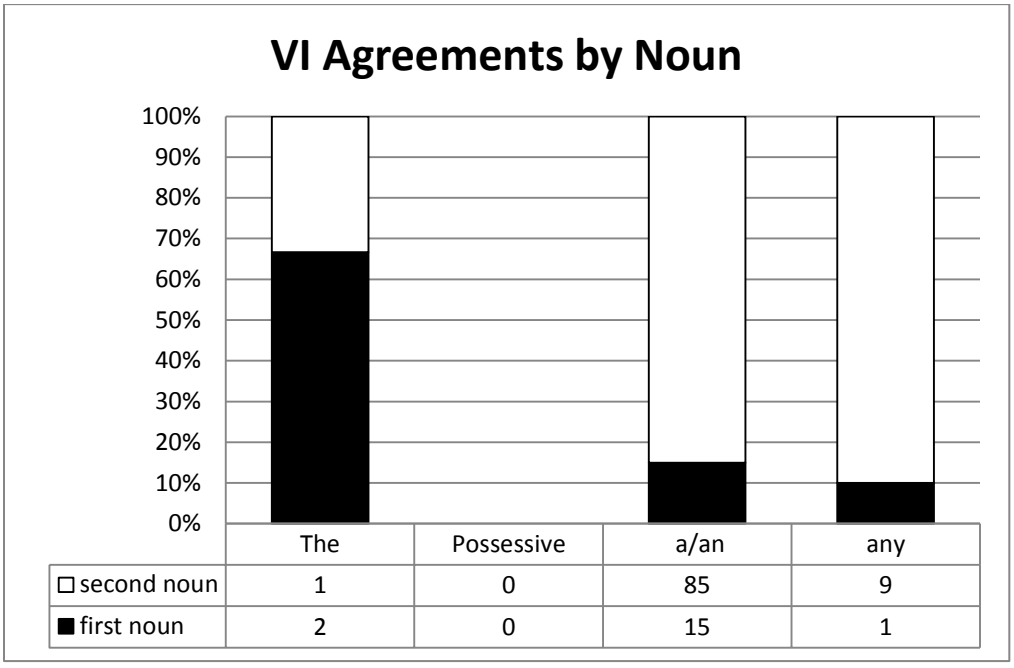
*18.* ***There are*** *an amazing number of these supplement products out there that are advertised and promoted*

*19.* ***There is*** *not a similar number of tribes waiting in line to take them on .*

*20.* ***There are*** *any number of ways for (the songwriter) to get taken.*

*21.* ***There was*** *any number of ways to search the list.*

*22. Complicating even the academic terrain* ***are*** *a number of initiatives reasserting the importance of individual actions,*




| | The | Possessive | a/an | any |
|---|---|---|---|---|
| second noun | 1 | 0 | 85 | 9 |
| first noun | 2 | 0 | 15 | 1 |

Figure 6: VI Agreements by Noun

The expectation is that the verbs will not agree with the second noun as frequently when the verb occurs before the full noun phrase. This appears to be the case with the three determiners that have observable data, though only the indefinite *a*/*an* has a relatively large--and therefore more accurate--set of occurrences. When contrasting this data against that presented in Figure 4, 15 of the 19 instances of the verb agreeing with the first noun occur when the verb precedes the noun. This means that the total number of first noun agreements of *a number of* is exactly 15% in

these VI circumstances while it is roughly 2.2% when the sentence is constructed in a more typical subject-verb-object (SVO) order. The single instance of *any number of NP* occurring with a first noun agreement is in a VI construction. This is demonstrative of an increase in atypical first noun verb agreements which appear to conflict with the trends observed but are, in actuality, superseded by another grammatical phenomenon. This fits in well with both the Competition Model and the more recently proposed Unified Model (MacWhinney 2013).

The remaining examples are so few that they can be qualitatively analyzed individually. While fifteen of the nineteen instances where *a number* is not the agreeing noun, there are four remaining instances. They are as follows:

*23. A typical number of hairs on a human head **is** about 100,000.*

*24. an odd number of ballots **is** cast*

*25. The hope is that if a great enough number of observations **is** directed at[...]*

*26. a higher number of lifetime partners **was** related to such risk factors as having casual partners*

As a group, the first noticeable factor appears to be that each of the first nouns has a modifier. the *numbers* are *typical, odd, great enough,* and *higher*. This may be the explanation itself. While these are not the only first nouns in the *number of NP* data to contain modifiers, the modifiers do seem to shift semantic focus to the numerical quality of the noun phrase. By increasing the salience and specificity of the *number* quality, there may be semantic shift towards the first noun as agreeing agent of the complex noun phrase. Additionally, the verbs are all link verbs, giving no indication to a colligational relationship with either the first or second noun. It is examples such as these that could be the basis of a later study.

The instances of *the number of NP* also have a few fairly clear differing agreements that result from different grammatical phenomena, several of which follow:

*27. The relatively small number of children who live with their fathers **seem** to be better off financially, at least.*

*28. The large number of people arrested **are crowding** Midwest jails.*

*29. The number of attacks this year **make** it the most violent of the decade.*

*30. But the number of distant dwarfs detected in the infrared survey **appear** about 10 times greater than the*

*31. The great number of rubs **indicate** that this is probably the buck 's core area.*

*32. The tiny number of things I do know well **are** subjects I 'll never try to use*

*33. Since February, 2001, though, the same number of positions **have been lost.***

*34. As the number of precise observations **increase**, something marvelous happens*

*35. But the number of academic failures **remain staggering***

*36. The large number of responses to the questionnaire **help** to elucidate the status of environmental education for*

*37. the number of practicing nurses **continue** to decline*

*38. As the number of master's degree programs **increase** in sport management*

Unlike the SVO first-noun verb agreements previously discussed, these examples do not seem to share a common thread. Some of the first nouns have modifiers, some do not. Two of the second nouns have relative clauses attached, other do not. Only one of the verbs is an ambiguous link verb while the remainder carry some colligational information. In Example 49, *the same number* is a specific number, referring either anaphorically or exophorically. As this is a specific, known numeral, *number* is realized as a quantifier, and the verb agrees with the second noun instead. On

the other hand, Example 50 which follows has no such explanation. *Increase* isn't a totally irrational action for *observations* to take. However, *Number* seems a more likely semantically-agreeing noun. The same can be said for Example 54, as *master's programs* can *increase* in quantity or size; the context seems to indicate the former, rather than individual programs growing in terms of students, instructors, classrooms, and offices.

Example 51 has a verb that could work with either the first or second noun. *Academic failures* can cause one to stagger, literally or metaphorically. The sheer *number* of them can also cause one to stagger. While not conclusive, this may be an instance of semantic confusion over agency. As for the two examples with relative clauses postmodifying the second noun, the simplest, but also inconclusive hypothetical explanation is simply distance causing the speaker to lose sight of the true agent. *Numbers* can *help,* but so can *responses*. Ultimately, some of these appear to simply have confused agency, while others may be a result of postmodification-related distance. A larger set of data would be needed in order to describe the phenomena with more

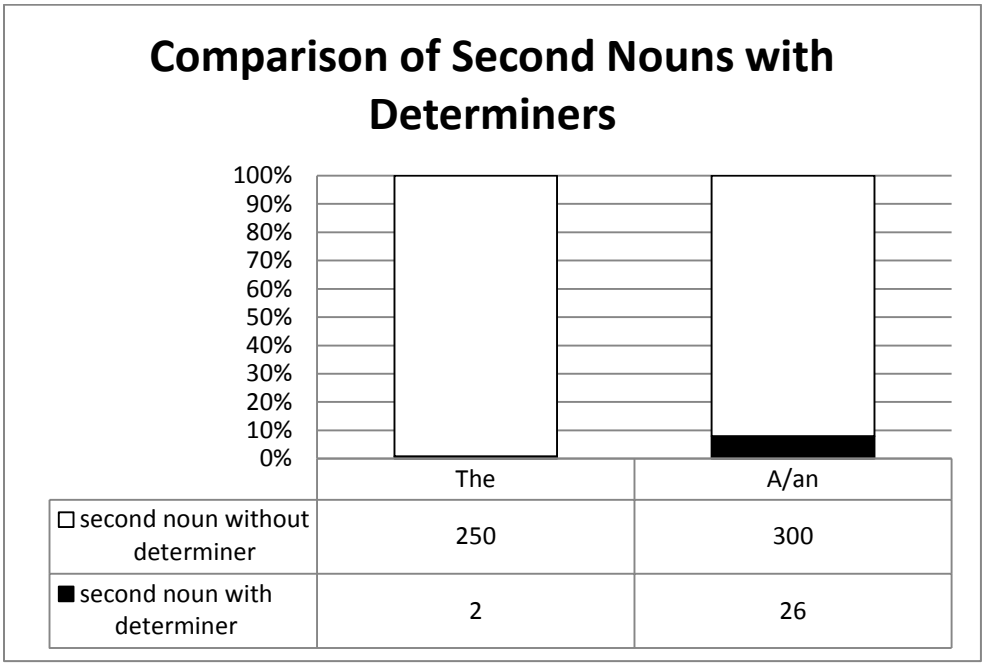



Figure 7: Comparison of Second Nouns with Determiners

reliability and accuracy, rather than an apparent attempt to grasp at straws.

Another observation that this data facilitate is that the indefinite *a/an number of NP* seems to permit, though certainly not require second nouns to have their own determiners, with approximately 8% of samples containing them. The second observation is that *the number of NP* seems to avoid such constructions, with only two instances occurring. Because of the small number of instances, both examples can be examined. The first instance appears complicated:

*39. The largest number of the victims in the sample -- 40 percent -- **were struck** in the legs and buttocks.*

This example is odd not only for being one of only two instances of a second noun having a determiner following *the number of NP* but also for the verb's agreement with the second noun in the same circumstances. Additionally, the sentence makes use of the superlative form, which requires a definite article on *number* for unrelated reasons. The multiple unique traits and rules that are being applied to this sentence make parsing the root causes of the verb agreement difficult. However, the multiple moving parts may be described as confusion resulting in ambiguous subject-verb agreement in these circumstances. It may also be a case of anaphoric reference, as the postmodifier *in the sample* indicates that there is a sample in the expanded context referred to earlier in the text. There is no absolute need for the determiner on the second noun, but its presence may be a device used to reiterate specificity to a group previously discussed. It may also be the very reason for the confusion here. As shown in the collected data, *the number of NP* doesn't commonly occur with a second noun with a determiner. The presence of the determiner may cause the focus of the speaker or writer to shift the second noun despite the usual first noun agreement in such constructions.

This sentence also seems to contain ambiguous agency, but flouts it. The verb agrees with a plural agent, so the two possible agents stated, *the largest number of victims in the sample* and *40 percent*, aren't ambiguous. The verb has to be in agreement with *victims*, as the other two would force the verb to agree as singular. This example is simply too complex and unique within the data to develop a concrete explanation of its seemingly unusual behavior. There is a further but related issue, that of the semantic meaning of *number*. In the methods section, a discussion of several unusable concordance lines described a differentiation between a count and a numeral meaning of *the number of NP*. While this instance is an apparent deviation from the standard usage for several reasons, it could also be that *the number of NP* is simply of a different semantic order, specifically, it is an actual number--in this case, *40 percent*. This could even be a case of ambiguity as a result of apposition again. If this is the case, it stands to reason that *the number* is acting as a modifier, despite the generally accepted notion that *the number of NP* will typically agree as a singular, rather than the plural in the above, with *number* serving as a quantifier. Finally, this instance has a verb that fails to be logical if it were to agree that *the number* is the head noun. Numbers, as a general rule, don't have legs or buttocks. Also, numbers are rarely *struck*, while a *victim* could be *struck* in their various body parts, legs and buttocks included.

*40. In contrast, the total number of <u>all</u> children adopted in 1992 was a substantial 127,441*

This second example is also confusing. Again, the presence of the determiner may have to do with the context and anaphoric reference. Additionally, it's worth noting that this example comes from a law journal, a genre which does not follow the same conventions as most natural language. Some of the features of legal writing are fossilized archaic expressions (Trossberg 1997, p 13-14), while a secondary objective is to contain as little ambiguity as possible so that readers and followers of law cannot find loopholes or claim exceptions. While a full account of

either of these samples cannot be given as a result of the data collected here, they raise the possibility of a future study which could examine such instances more closely.

*41. "I don't see this as precedent-setting," Hoffman says. "**There's** a number of refuges and parks on the Federal Register list. **There is** not a similar number of tribes waiting in line to take them on.*

Apparent norms in written grammar are not necessarily the norms in spoken language. The competing issue of differing agreement occasionally overrides the otherwise quite consistent verb agreement in written discourse as seen in the concordance lines analyzed here. In speech, the null subject form *there BE* is more frequently narrowed to *there is* regardless of agreement (Biber et al., 1999, p. 185-186).

For the remaining 22 instances not determined by *a/an* or *the*, the agreement appeared to follow definite and indefinite guidelines. Those *number of NP* instances determined by *any* displayed subject-verb agreement in the same distribution as *a/an*; in 17 of the 18 instances, the verb agrees with the second noun phrase. Of the 4 possessive determiners, 3 agree with *number* as singular, indicating a possible trend of possessives being akin to *the* in terms of definiteness and in subject-verb agreement, though the small sample size results in difficulty making a reliable claim. Having examined described some of the idiosyncratic issues that occur in outlying examples of *number of NP, group of NP* can now become the focus.

## Group of NP

In order to analyze the data with respect to *group of NP* agreement (singular conjugation) and second-noun agreement (plural conjugation), the 600 concordance lines were broken into two data sets. Once separated by differing agreement, the *group of* noun phrase data was shown to have 357 singular verb agreements and 243 plural agreements. This ostensibly means that the

verbs agreed with *group* as a singular noun 357 times and agreed with the second noun 243 times. Alternately, it means that approximately 59.5% of the data show verb agreement with the first noun *group* rather than the second noun.

To understand why this distribution occurs, the first level of coding was for the first-noun *group*'s determiners. Several of the categories merged multiple instances of determiners into larger grammatical classes as the data collected for individual words is too small. Possessives, for instance, include both the commonly named "possessive adjectives" (e.g., *my* and *your*) as well as possessive noun phrases. By coding the data separately, an image emerged into the distinction of singular and plural verb agreement. In this initial phase, the raw numbers of the indefinite *a(n)* were nearly the identical for both sets of data at 225 and 205 in singular and plural verb agreement respectively, and over two thirds of the determiners in the combined data. Also, every other determiner had a higher occurrence rate with first noun agreement in the data shown in Figure 9. The raw numbers shown below demonstrate the prevalence of the features throughout the 600 line data set. In Figure 11 a higher ratio of verb agreements with the second noun co-occur with indefinite determiners, specifically *a*/*an*, than with other determiners. For every other determiner, we see that they occur with greater frequency in singular, *group*-agreeing co-occur with indefinite determiners, specifically *a*/*an*, than with other determiners. For every other determiner, we see that they occur with greater frequency in singular, *group*-agreeing clauses. The one exception is *another*, which is also indefinite. Some of the determiners occurred only several times, so it is far from conclusive, but the data as a whole show a trend in distribution between indefinite articles and second noun agreements.

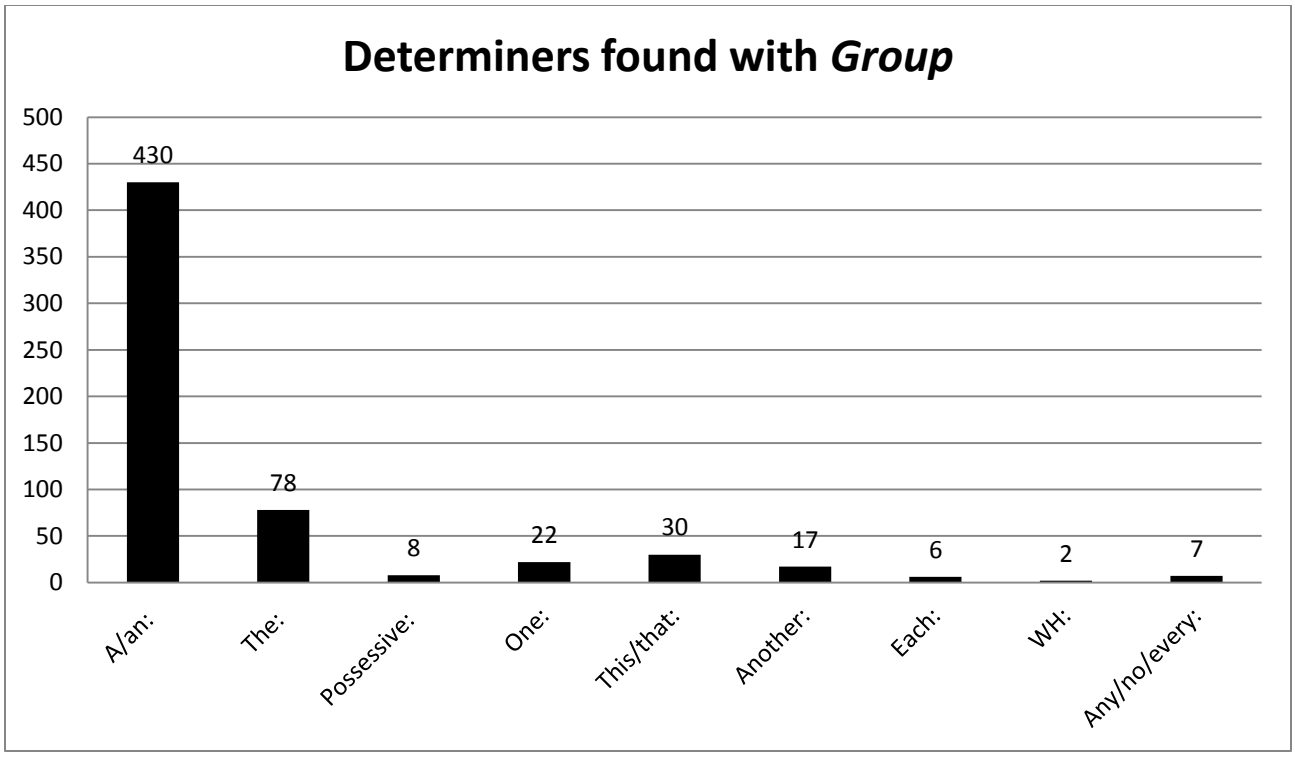



Figure 8: Determiners found with Group

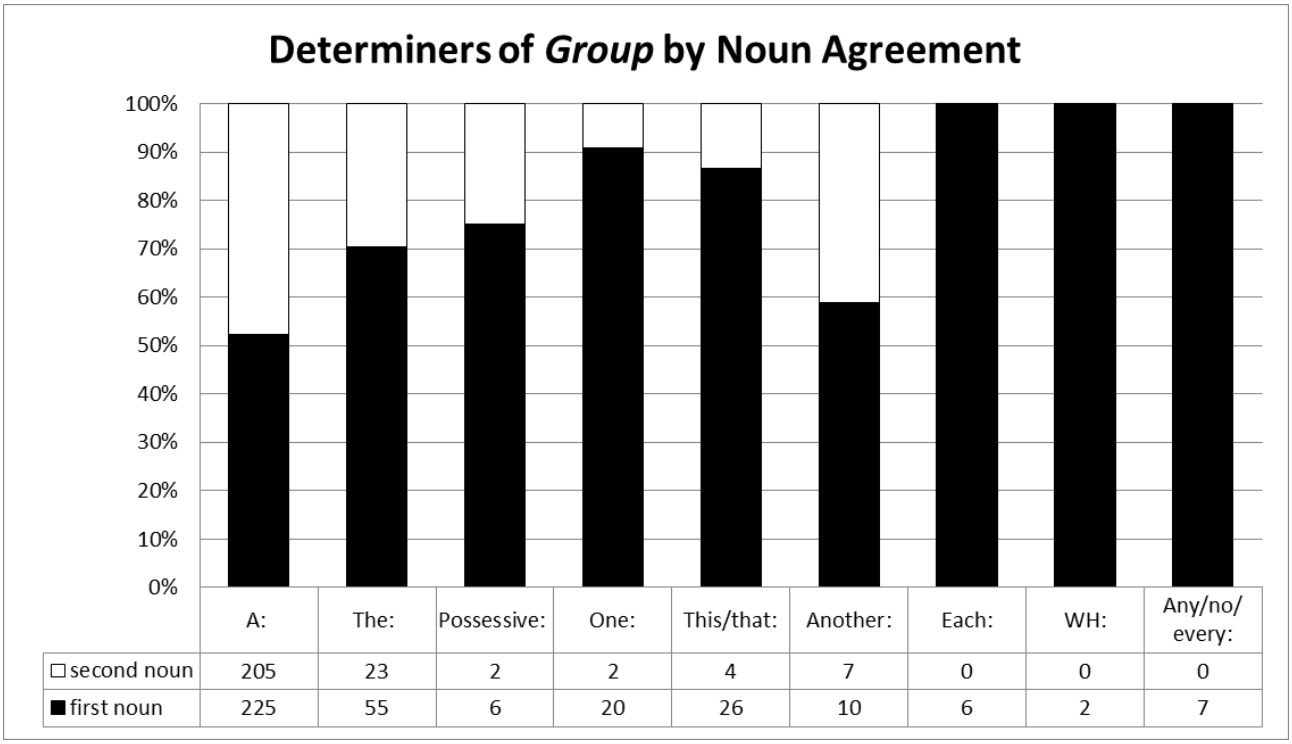



Figure 9: Determiner of Group by Noun Agreement

*42. A group of homes **are** unified and differentiated from other community sections.*

*43. The growing group of ranking Japanese officials who now work for foreign interests **runs** against conventional[…]*

One point that may be disputed is the inclusion of *one* as a determiner. This item can act as either a determiner or it can also be an adjective modifier as a part of a noun phrase that includes a different determiner. Given that *one* did not occur after another determiner in the included data, the conclusion that *one* acts as determiner in these instances was assumed. This is an important distinction once compared to the findings in the later stages of coding.

Once the determiners for the first noun *group* were described, the second stage was to analyze the determiners for the second noun. The initial hypothesis was that there might be a trend in distribution between determiner choice on both noun phrases and the verb agreement. However, the data showed an entirely different sort of relationship that does not seem to involve subject-verb agreement at all. Out of the total 600 concordance line sample, only four of the second nouns had determiners, or 0.666%. One of these four second nouns does not agree with the verb, the other three do. This number of determiners is too small to extrapolate into a larger trend. However, it is an appropriate size for a qualitative analysis of individual contexts.

*44. […]where a group of **her** friends talks animatedly.*

*45. the largest group of **the** millions of Jews killed in German death camps were Polish citizens*

*46. In an annual report , the Paris-based group of **the** world 's leading industrial nations forecast a 12 percent gain in Mexico's gross[…]*

*47. Only a month into this season, a group of **his** boyhood friends from Virginia were living in the three-bedroom Philadelphia condo*

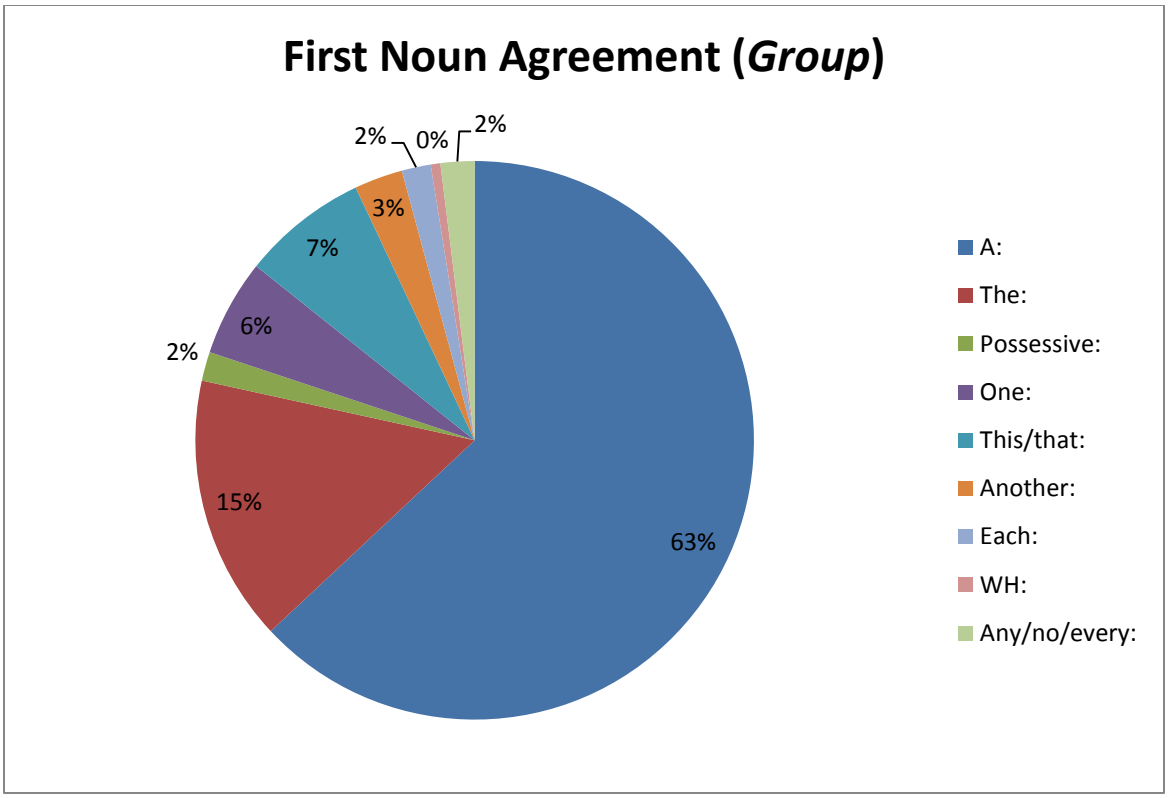



Figure 10: First Noun Agreement

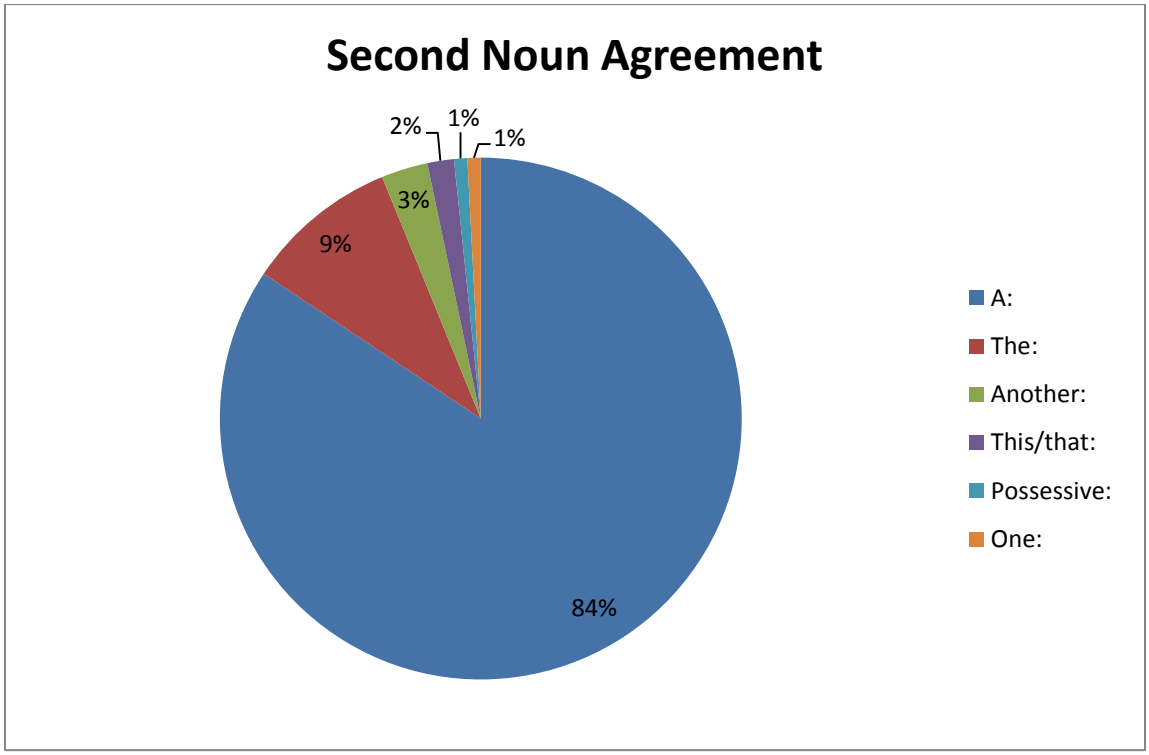



Figure 11: Second Noun Agreement

The initial observations are that two are the definite determiner *the,* and two are possessive determiners. These all count as definite articles (Huddleston & Pullum 2002 p. 271). Of the two *the* determiners, both seem to be special cases. They both appear as determiners on even more complex noun phrases, making their validity as actual determiners of the second noun suspect. In Example 61, the noun phrase "the millions" appears to be a quantification of *Jews killed*, rather than a determiner of *Jews*. This appears to be a part of the multilayered complexity of subject-verb agreement in complex noun phrases, of which assessing *group of NP*'s agreement is the first of the strata. Given that *millions* and *Jews* are both plural, determining which noun is the target of the verb's plural agreement is not feasible. The second *the*-determined noun phrase in Example 62 is also complicated by additional grammatical structures. It is, in fact, part of *the world's*, a possessive noun phrase. This sample is more along the lines of *group of NP's NP,* making this a unique circumstance in the data. The agreement of the verb is not opaque, as it agrees as a plural, of which only last noun of the three nouns in the example is.

This means that there are only three second nouns with determiners, each of which is possessive. While definite, possessives' primary function is to grant ownership or possession of an NP to an agent or entity. In Example 62, removing *his* from *boyhood friends* changes the meaning; these *boyhood friends* could be any male friends anywhere anytime, instead of those friends of the male referent of *his*. This feature cannot be separated from definiteness.

This then leads into the third stage of the coding, that of adjective pre-modifiers. There appears to be a slight trend relating to second-noun verb agreement when adjective pre-modifiers are present. Figures 12 and 13 show the distribution of modifiers in second nouns phrases within both the *group*-agreement and second-noun agreement data sets. This slight tendency towards

agreement with modified nouns is shown in data of the modified *group* nouns as well.

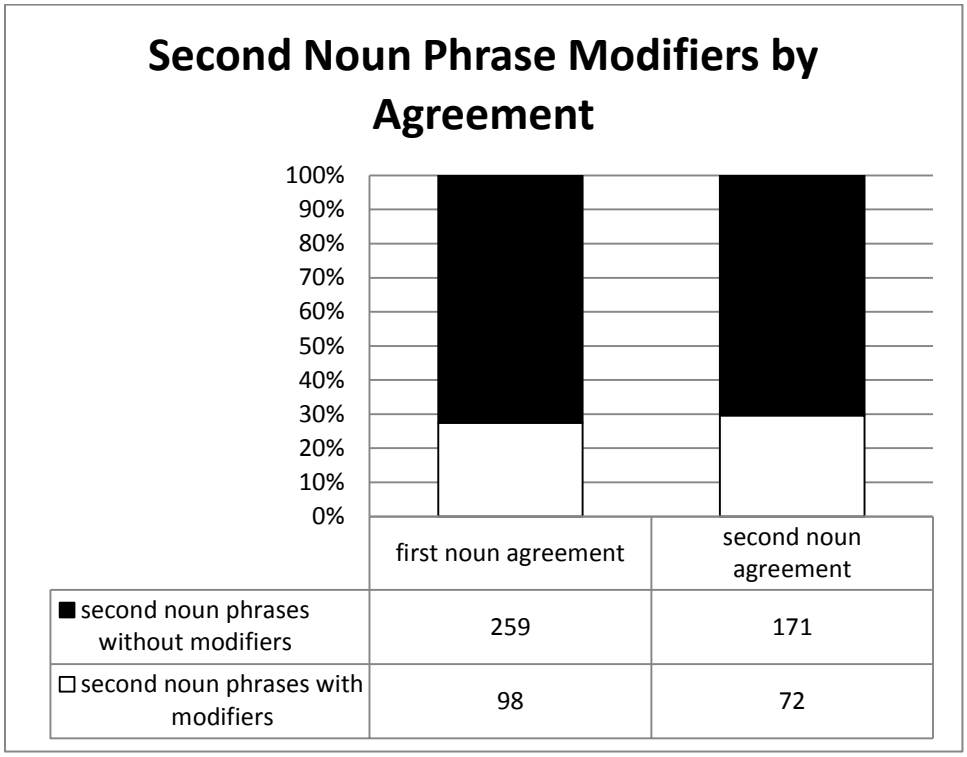



Figure 12: Second Noun Phrase Modifiers by Agreement

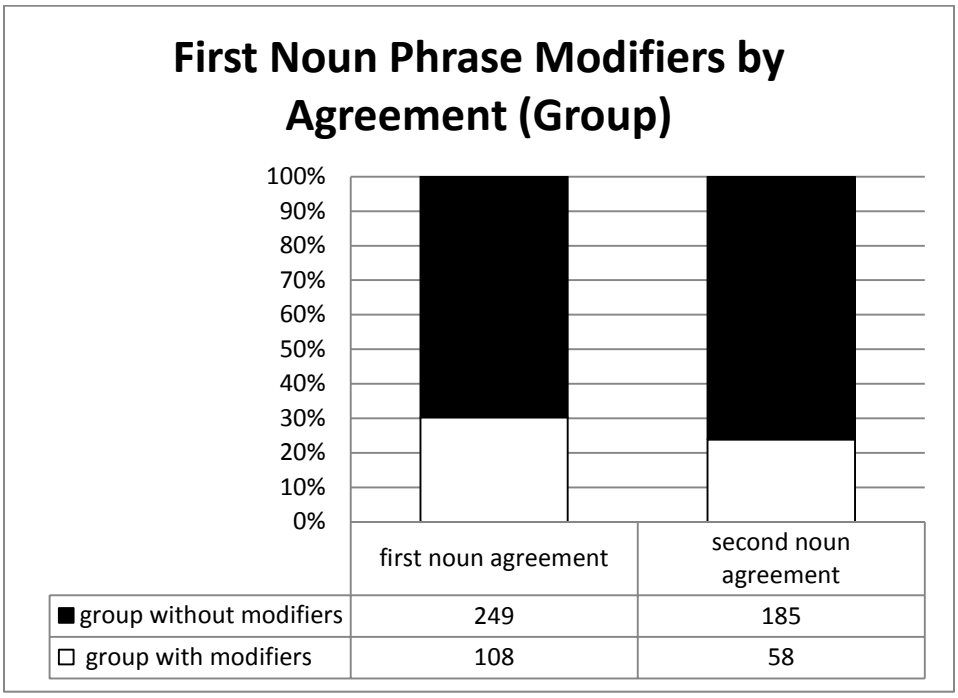



Figure 13: First Noun Phrase Modifiers by Agreement (Group)

There is a long-standing point that distance and noun phrase ‘weight’ causes focal shifts,

There is a long-standing point that distance and noun phrase ‘weight’ causes focal shifts,

which could in turn be the driving force of this distinction (Negro et al., 2005; Gillespie & Pearlmutter, 2011). This is mirrored by the number of null-subject and subject-auxiliary inversion samples which do the opposite; they are much more likely to see verbs agree with *group* or *number* as the VI ordering causes the verb and the first noun to become much closer (Haskell & MacDonald, 2003; Franck et al., 2006; Roberts, 1985) as is demonstrated in Figures 7 and 13. Again, while there are several realizations of VI ordering, the null- or dummy-subject formation is the most common in the data for *number of NP* while subject-auxiliary inversion (SAI) was the most common circumstance for *group of NP*. Four of the second noun verb agreements in Figure 14 are null-subject lines, and remaining instances involved subject-auxiliary inversion. While 59.5% of the concordance lines show verb agreement with the first noun, the percentage is only ~55.26% of SVO-ordered sentences while it is ~92.65% of VI sentences.

Unlike the determiners, there appears to be a minor co-occurrence between the nouns’ respective modifiers. The mere presence of modifiers in either of the noun phrases show a minor increase in the likelihood that the very will end up agreeing with the modified noun. However, there is at least one exception to this rule: quantifying or numerical modifiers. Some of the second nouns in *group of NP* have quantities:

*48. …where a group of <u>four</u> hoatzins **has** just begun to forage.*

*49. A group of <u>four</u> thick tentacles **were** affixed in a semicircular pattern.*

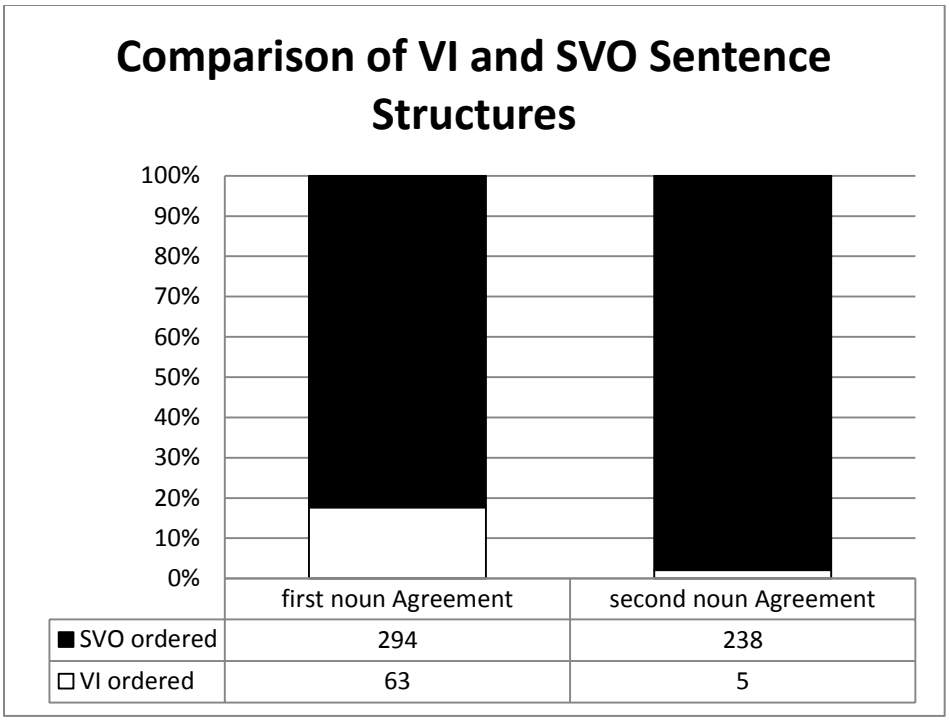



Figure 14: Comparison of VI and SVO Sentence Structures

In the data, a trend emerged. While there appears to be a trend of modifiers shifting focus towards the noun phrase it is in, the numerical modifiers have a much more frequent trend of appearing in second noun phrases that don't agree with the verb. The relationship in the data examined shows that, after normalizing, second-nouns with numerals or specific quantities were roughly two and a half times more likely not to agree with the verb than agree as shown in Figure 15. While it doesn't appear to be as common a trend of agreement with the first noun *group* in verb-initial sentences, the finding that the numerical modifiers follow an inverse pattern of distribution to verb agreement in this way is more surprising.

# CHAPTER 5

# DISCUSSION

## Discussion of Findings

This set of data presented show several trends with regards to subject-verb agreement between *group* and *number* with respect to the second nouns with which they occur. While the initial round of determiner coding supported the hypothesis that determiners on the first noun have some role in subject-verb agreement in this form, the results of coding the second noun determiners were a surprise. This second round of coding did not undermine the previous results, but it was a surprise to discover the dearth of second-noun determiners with respect to the *group of NP* sampling and the fairly clear distinction found in *number of NP* second nouns.

The trends shown here answer many of the research questions. The first, broadest question has a wide set of trends that illuminate, if not fully explain the subject-verb agreement trends of *group of NP.* In addition to confirming previous insights into the distinctions between *a* and *the number of NP*, some data collected show potential reasons for disagreement and exceptions to the established first-determiner rules. We see that the modifiers in both the *group* noun phrase and the second noun phrase have some marginal relationship with subject-verb agreement, with "heavier" modifier-laden nouns having some increased probability of agreement. We also see that the first noun, *group,* is more likely to agree with the verb generally and has a broader distribution of determiners than *number* in the sampling. The second noun has an increased probability of agreeing with the verb when the indefinite determiners occur with *group,* however. The surprising revelation that the second noun, when modified with a number or quantity has an apparent trend of inverse distribution to verb agreement with the second noun is the most interesting and ponderous of the observations. This seems to be a good point to retest in future

studies. However, the second part of Research Question 1 with respect to *group* doesn't have a firm answer. While there are trends in the data, there seem to be no absolute rules governing subject-verb agreement with this form. However, in both *group of NP* and *number of NP*, a trend related to the definiteness of the determiners rather than specific determiner seems to be motivating agreement, though not as consistently as with *number of NP*.

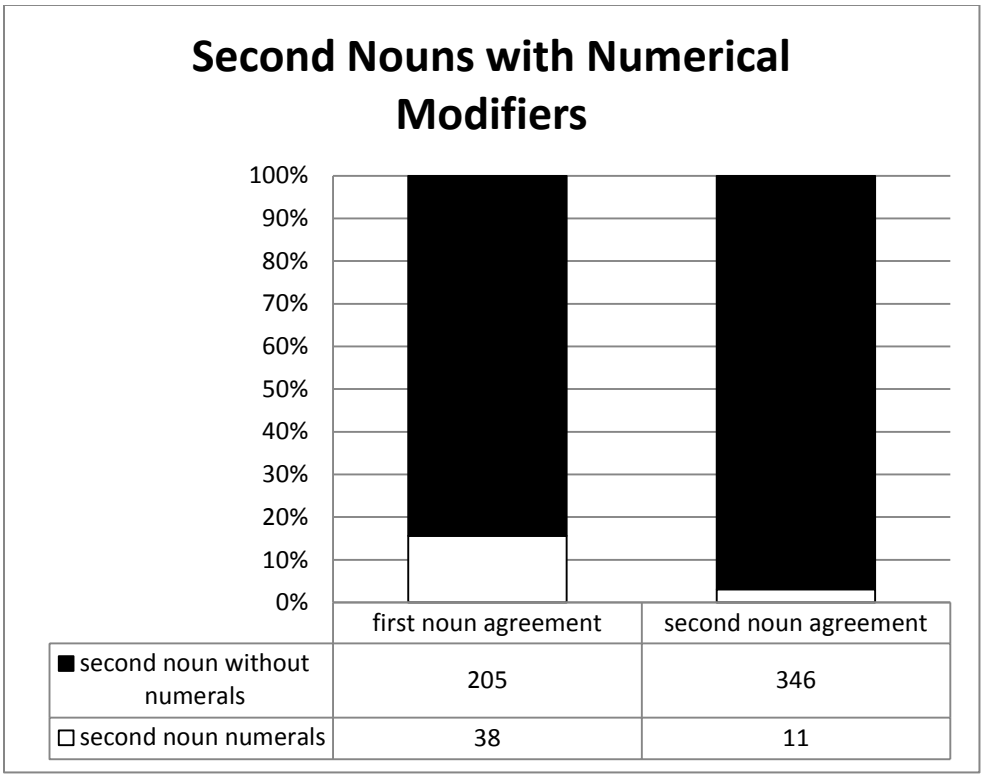



Figure 15: Second Nouns with Numerical Modifiers

Research Question 2 has some interesting revelations which tie into RQ 3. With *number of NP*, the determiners are the prevailing indicator of whether *number* behaves as a quantifier or a head NP that agrees with the verb. In *group of NP*, *group* seems to behave in much the same way as Huddleston and Pullum describe the class they call "number-transparent nouns" (2002, p. 501-502). The so-called transparency is, in effect, a distinction between the collective (*group)* or the constituents (the second noun phrase) being the focal target of the sentence. It appears that modifiers occur with some common distribution with the noun phrase which agrees with the

verb. This trend may be causation, may result from the focus itself, or may simply be coincidence, given the only slight difference observed. However, the determiners have more common co-occurrence, though they are not necessarily the cause of the focus but may just as likely be chosen as a result of the focus. The evidence taken from the analyzed data seems to show that the definiteness of determiners of *group* has a focusing effect on (or from) the author, increasing the likelihood that the verb will agree with *group*. Alternately, one might describe *indefiniteness* as having an un-focusing effect.

Potentially the oddest observation regarding RQs II and III is the distribution--or effectively absence--of determiners on the second nouns. With *group of NP*, *t*hose determiners that do occur as a part of the second noun arguably do not carry definite or indefinite qualities. The possessives, which arguably do carry definiteness, can be described as a workaround; since possession and definiteness are inextricable, the possession feature overrides. As determiners can carry several loads, it seems that these few examples are workarounds. This concept has greater credence when considering that seemingly backwards trend with numerical modifiers. Plural determiners denote nonspecific quantity; since there are no real determiners, the numerical modifiers appear to take their place. The possessives don't have a modifier equivalent, so they are irreplaceable. This leads me to the tentative conclusion that *group* is, in some way, carrying the second nouns' definiteness feature, at least in the case of these plural second nouns.

This position is lent some support by the distribution of second-noun determiners in the *number of NP* data. The near absence of determined second-nouns in *the number of NP* data seems indicative that the second nouns in these phrases have no need of determination. When *number* is the head noun, quantity is unnecessary, while possession and definiteness are not excluded, but are very low in terms of occurrence. The data did not yield any second nouns

determined by possessives, but this is not conclusive that they cannot be present. For instance (invented example):

*50. <u>The number of her students</u> who've finished the test **is** almost 100%.*

While artificial, the example above isn't necessarily incorrect. This is an issue that can be the basis of future studies that either do not rely on corpus data or do not require the exclusion of data that does not demonstrate subject-verb agreement in the way that this study's methodology required (invented example):

*51. I saw the number of her students who had enrolled in her new class and was amazed.*

By removing the subject-verb agreement criterion from the data collection methods, data could be collected and analyzed for a study focusing on the definiteness issues outlined here with greater expedience.

## Limitations

### Corpus limitations

The corpus used is meant to be reflective of contemporary American English in use. However, both its sampling methods and the sampling methods used to extract data from COCA are imperfect. COCA uses many mainstream sources for its data sampling, but even these do not cover the full spectrum of any of its subcorpora. As a result, not every potential source was examined. Additionally, the sampling methods used within this study were imperfect. Because of the amount of data to be excluded and the high variability in the amounts of usable concordance lines gathered from each sample, the approach outlined in the corpus section of this paper was taken. For more accurate sampling in an expanded subsequent study, samples should be included from a broader array of individual texts, if not each of the year mini-corpora for each subcorpus.

This could also be recommended for any register-based study in the future. For instance, one concern that comes from the sampling of COCA's data concerns the relatively high frequency of scene-setting concordance lines acquired from the fiction register with regards to *group of.*

*52. Gunfire shatters the serenity of the night . CUT TO : 26 EXT . FUJI MARU DECK – NIGHT A group of tony British passengers* ***is shooting*** *skeet. Being night, the CLAY PIGEONS*

The narrative quality of fiction as a genre meant that there were far more examples of narrative scene-setting examples as well.

*53.* […] *Ernie jumping to the side to get out of its way. A group of boys* ***runs*** *after the dog, shouting, waving their arms in the air, up above them a* […]

While these samples included both singular and plural verb agreement, the sampling was of a much smaller data set, with roughly 1:5 concordance lines being included from the samples, rather than the 1:8 in the other three subcorpora. In future studies a potential issue to account for in the data is this kind of clustering. Sampling from a broader set of texts, specifically different types of fiction, would improve the breadth and representativeness of the collected data.

There is register variation amongst all of the data collected. While these data were included, discussion of the source materials is included in the discussion section as a rationale for a potential follow-up study of register variation.

**Coding limitations**

In coding the data, some items were necessarily excluded. Issues of ambiguity were a major issue, with verbs that did not transparently agree with one noun over the other. Some postmodifiers were ambiguous in which of the two nouns they modified. While several of the

relative clauses were headed by *who*, many were headed by other terms such as *that* or *which*. Given the possibility that *who* may refer to group, and given the majority of relative clauses were excluded, all postmodifying clauses were excluded from the modifier count. Similarly, second-noun phrases that exceeded the length of the concordance lines were excluded.

**Methodological limitations**

Much of the data in the various texts that informed this research came from existing language in corpora; this is problematic, as corpora examine language in use, rather than language knowledge. As a result, the more formulaic (i.e., commonly used and repetitive) forms were easy to find, discuss, and analyze, whereas novel uses were not. In order to describe these groups as a class—and consequently allow for a logical, systematic teaching method to be developed—a more systematic understanding should be had. This requires non-formulaic instances and lower-frequency forms that either didn't appear in the data or did not meet the coding criteria to be examined in addition to the more common instances. It is both plausible and possible to take a more targeted look at lower-frequency forms to develop a more holistic understanding of the phenomena at work.

# CHAPTER 6

# FUTURE RESEARCH AND CONCLUSIONS

Fortunately, this study has found a number of general trends and idiosyncratic issues that merit further study. The observations indicate several trends, though it stops short of defining new rules of grammar. The trends found indicate several possible factors in an overall system of subject-verb agreement. With that in mind, future studies could examine additional factors. There are far too few features to suggest a multidimensional analysis at this time. In essence, this study has successfully completed the task of providing enough observations to support conducting a broader study of the *noun of noun* verbs that have subject-verb agreement irregularities.

One of the frequently co-distributed features observed thus far has been in the contrast of SVO and VI-ordered sentences. While the VI-ordered sentences, particularly in *group of NP*'s case, are overwhelming, those results confirmed a pattern that has been repeatedly described (Negro et al., 2005; Gillespie & Pearlmutter, 2011; Haskell & MacDonald, 2003; Franck et al., 2006; Roberts, 1985 MacWhinney 2013). The contrast that they have against *number of NP* seems to indicate that the issue is not so much one of confusion on the part of the speaker as to the verb agreement, but one of cognition. The issue here is again that there is no assumed right or wrong way to apply quantity to a given noun phrase, only how the agreement is occurring in use. The language user supersedes the "correct" convention by agreeing with the nearer of the two nouns in the complex NP in *number of NP*. In the case of *group of NP*, the ambiguity of the rules of agreement simply allow for the language speaker to latch onto the nearer noun. Given that the reverse does not happen--that verbs do not overwhelmingly agree with the second noun in SVO-

ordered sentences, simple proximity cannot be said to be the definitive motivator of subject-verb agreement in this case.

The observations about definite determiners may extend to other *noun of noun* forms, such as those in *The Longman Grammar*'s quantifying collective class (Biber et al., 1999, p. 248-249). The initial hypothesis about determiner influence derived from similar *noun of noun* forms, such as *a*/*the number of NP*. Extending the *a*/*the* observation to include other determiners would be a start, as well as other noun forms. This hypothetical future study could also observe trends surrounding the second noun's determiners or lack thereof. If there is an overall trend with some of these *noun of noun* segments, a new classification could be developed around definiteness and determination.

While the adjective pre-modifiers were mostly uninteresting, the numerical modifier data show a distinctly surprising trend. In fact, the numerical modifiers are so different that the results should be confirmed with another study. Collecting more and/or different data to retest this observation seems prudent. If the results are truly indicative of a trend, the study could expand to examine similar noun clusters for the same feature. Again, if there is an overall trend, a classification may be developed.

Approaches to future studies could also potentially include data elicitation, rather than strictly corpus-based work. Grounding the observations in corpus data prior to data elicitation will allow researchers to find gaps in the language in use prior to testing language knowledge in a subsequent elicitation study. Cognitive linguistics may also play a factor in determining how definiteness is assessed by language users, given that definiteness is a feature that the data in this study seemed to avoid with regards to the second nouns. In sum, more research with a broader scope is merited.

## Corpus Studies

Low frequency determiners are simply under-described by the scope of this research. They do exist, and some of the data analyzed show that there may be a trend in their distribution with regards to their definiteness. A more targeted study of the determiners that precede *group of NP* and *number of NP* could help to develop and codify the systematic relationship between determiners as a class and the QNPs with which they co-occur. This would likely require continued observation of the subject-verb agreement in these noun phrases. By studying low-frequency determiners more closely, researchers could solidify the relationship that these QNPs' verb agreements have with determiners. In so doing, researchers could more accurately codify the function of indefiniteness and beyond.

A comparative study of number-transparent nouns and complex NPs could also shed light on the ways in which *group of NP* and *a number of NP* are similar to and different from number-transparent nouns. Given that *committee,* as in Example 11, does not explicitly show its constituents, does that affect the way in which it is quantified? Do determiners have a relationship with singular or plural verb agreement of *committee* or other such implicitly constituent-laden nouns? Do modifiers, such as *large* or *congressional*, affect the quantity? If so, why is the distinction shown in *group of NP* so low? Since there can be no confusion or dispute over a second noun in the noun phrase, directly studying the quantification on number-transparent nouns could elicit observations that the complex QNPs cannot. By examining these two different classes of noun, researchers could develop a better understanding of not only descriptive grammar, but potentially insights into the cognitive processing that determines linguistic quantification.

## Other Studies

The corpus methods have value as descriptive tools. This study has made minor attempts to resolve underlying causes, which is not possible without going beyond language descriptions and usages and moving into studying language knowledge. This means a study of live participants and eliciting data from them. This approach broadens the description that the corpus-based and corpus-driven studies provide. More importantly, this type of study allows for researchers to access observations about the cognitive systems at work that determine how a speaker or writer views plurality or quantity in complex noun phrases of the order that have been observed thus far.

Numerical modifiers on the second noun are also interesting. While there is room for a more targeted study of numerical modifiers in corpus data, a study of human subjects and data elicitation could be useful in controlling other variables such as the first nouns' determiners. This could also help expand on the definition of numerical modifiers put forth here, as other quantifying modifiers not observed may also play a role in assessing the quantity of the QNP.

# REFERENCES

Anthony, L. (2010). AntConc (Version 3.2.1) [Computer Software]. Tokyo, Japan: Waseda University. Available from http://www.antlab.sci.waseda.ac.jp/

Aremo, W. B. (2005). On some uses of singular collective nouns. *English Today*, *1*, 52-55

Ascoli, S. (2013) Quantifying noun phrases in American English. Unpublished course work.

Berry, R. (1997). English Guides 10: Determiners & Quantifiers. London: Harper-Collins Publishing.

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *The Longman Grammar of Spoken and Written English.* New York, NY: Longman Publishing.

Chiarelli, V., Yagoubi, R., Mondini, S., Bisiacchi, P., Semenza, C. (2011). The Syntactic and Semantic Processing of Mass and Count Nouns: an ERP Study. *PLoS ONE 6*(10), 1-15.

Chierchia, G. (2010). Mass nouns, vagueness and semantic variation. *Synthese 174* (1), 99-149.

Cocchiarella, N.(2009). Mass nouns in a logic of classes as many. *Journal of Philosophical Logic*, *38*(3), 343-361.

Davies, D. (2013, January 30) William H. Macy Is 'Shameless' On Showtime. Retrieved from http://www.npr.org/2013/01/30/170463939/william-h-macy-is-shameless-on-showtime.

Davies, M. (2008) The Corpus of Contemporary American English: 450 million words, 1990-present. Retrieved from http://corpus.byu.edu/coca/. Retrieved 11/11/2014.

Davies, M (2010). The Corpus of Contemporary American English as the first reliable monitor corpus of English. *Lit Linguist Computing* (2010) 25 (4): 447-464.

Domahs, F., Nagels, A., Domahs, U., Whitney, C., Wiese, R., Kircher, T. (2012). Where the mass counts: Common cortical activation for different kinds of nonsingularity. *Journal of Cognitive Neuroscience 24*(4), 915-932.

Dziemianko, A. (2008). Subject-verb concord with collective nouns or the count-mass distinction: Which is more difficult for Polish learners of English? *Poznan Studies in Contemporary Linguistics, 44*(4), 449-468.

Franck, J., Lassi, G., Frauenfelder, U., and Rizzi, L. (2006). Agreement and movement: A syntactic analysis of attraction. *Cognition, 101*(1), 173–216.

Haskell, T. and MacDonald, M. (2003). Conflicting cues and competition in subject–verb agreement. *Journal of Memory and Language, 48*(4), 760–778.

Hoey, M. (2005). *Lexical Priming: A New Theory of Words and Language*. New York, NY: Cambridge University Press

Huddleston, R., & Pullum, G. (2002). *The Cambridge Grammar of the English Language*. New York, NY: Cambridge University Press

Humphreys, K. and Bock, K. (2005). Notional number agreement in English. *Psychonomic Bulletin & Review*, *12* (4), 685-695.

Landman, F. (2011). Count nouns-neat nouns mass nouns-mess nouns. *The Baltic International Yearbook of Cognition, Logic and Communication 6*, 1-67.

MacWhinney, B. (2008). A unified model. In P. Robinson & N. Ellis (Eds.), *Handbook of Cognitive Linguistics and Second Language Acquisition*. Mahwah, NJ: Lawrence Erlbaum Associates.

MacWhinney, B. (2013). The logic of the unified model. In S. Gass & A. Mackey (Eds.), *Handbook of Second Language Acquisition*. New York, NY: Routledge.

Markman, E. (1985). Why superordinate category terms can be mass nouns*. *Cognition 19*, 31-53.

McEnery, T., Xiao, R., & Tono, Y. (2006). Corpus-based language studies: An advanced resource book. New York, NY: Routledge.

Negro, I., Chanquoy, L., Fayol, M., & Louis-Sidney, M. (2005). Subject-verb agreement in children and adults: Serial or hierarchical processing? *Journal of Psycholinguistic Research*, *34*(3), 233-258.

Nenonen, M. and Niemi, J. (2010). Mismatches between grammatical number and conceptual numerosity: A number-decision experiment on collective nouns, number neutralization, pluralia tantum, and idiomatic plurals. *Folia Linguistica, 44*(1), 103-125

Nicolas, D. (2008). Mass nouns and plural logic. *Linguistics and Philosophy*, *31*(2), 211-244.

Number (n.d.) In *The Cobuild American English for Learners Dictionary.* Retrieved from http://www.collinsdictionary.com/dictionary/english-cobuild-learners. Retreived 10/30/2014

Radden, G., & Dirven, R. (2007). *Cognitive English Grammar* (Vol. 2). Philadelphia, PA: John Benjamins Publishing.

Roberts, I. (1985). Agreement parameters and the development of English modal auxiliaries. *Natural Language & Linguistic Theory 3(1)* 21-58.

Römer, U. (2005). *Progressives, Patterns, Pedagogy: A corpus-driven approach to English progressive forms, functions, contexts and didactics*. Philadelphia PA: John Benjamins Publishing.

Rozumko, A. (2002). Countable, uncountable and collective nouns in the early eighteenth century English. *Studia Anglica Posnaniensia, 37,* 131-159.

Sinclair, J. (1998). Collins COBUILD Grammar Patterns 2: Nouns and Adjectives. London: Harper Collins Publishing.

Trosborg, A. (1997). *Rhetorical Strategies in Legal Language: Discourse Analysis of Statutes and Contracts* (Vol. 424). Tubingen: Narr