The Loss of Bare Singular Noun Phrases in the History of English

Kenneth Hanson

May 6, 2014

Contents

Acknowledgments

I would like to start by thanking Alan Munn and Cristina Schmitt for their invaluable roles in the work that led to this thesis: Cristina for originally suggesting the idea of a historical corpus study of English bare noun phrases, Alan for advising me in the preparation of two conference presentations

and this thesis, and both for serving as collaborators throughout the project. In addition to their invaluable contributions to this work, I am also indebted to both Alan and Cristina for imparting their knowledge of syntactic theory and training me to think about linguistic problems, as well as answering my endless, probing questions with honesty and insight.

Next, I would like to thank some of the other individuals whose assistance made this work possible: Beatrice Santorini for her help with CorpusSearch and the Penn Corpora, Karthik Durvasula for teaching me R and basic statistics, Ashley Hesson for guidance in analyzing multivariate data, and Aaron Ecay for his helpful suggestions regarding historical corpus work and for sharing his text date coding query.

Finally, I would like to thank and blame Cristina for getting me into linguistics. While I may have ended up here regardless, the excitement I found in her introduction to linguistics class and her recommendation to take syntax the following semester made it a near certainty.

1 Introduction

One common cross-linguistic pattern in the syntax of noun phrases is that in many languages, certain nouns cannot appear alone, but must be accompanied by certain modifying affixes, words, or phrases. In English, these include the definite and indefinite articles, numerals, demonstratives, quantifiers, and genitive possessors.

(1) Mary saw $\left\{\begin{array}{c}\text{the}\\ \text{a}\\ \text{one}\\ \text{this}\\ \text{every}\\ \text{John's}\end{array}\right\}$ penny on the floor.

While singular count nouns such as "penny" usually must be accompanied by an article (or another premodifier) in English, mass nouns and plural count nouns have no such requirement.[1]

(2) a. * Mary saw ∅ penny on the floor.
b. Mary saw ∅ crumbs on the floor.
c. Mary saw ∅ water on the floor.

[1] In the remainder of this paper, "singular noun" or "singular" without qualification will always mean "singular count noun".

Noun phrases that appear without one of these morphemes are known as *bare noun phrases*. The distribution which noun phrases can appear bare is complex both cross-linguistically and within individual languages, and is of considerable interest within research on the syntax of noun phrases.

The syntax of bare noun phrases is also known to change throughout the history of a language. In the case of English, there was formerly a much wider range of bare noun phrases. Old English (c. 500–1066) originally had no separate definite or indefinite article; instead, the same morphemes that marked definiteness and indefiniteness also had other functions. Indefiniteness, in particular, was not always marked with an article, so many indefinite noun phrases were bare by default (Traugott 1992, pp. 171–174). The definite and indefinite article were both established by the early Middle English period (c. 1066–1500), but many bare singular noun phrases that would be impossible today can still be found (Fischer 1992, pp. 217–220). Consider the examples in (3), which were found in the Penn Corpora of Historical English.[2]

(3) a. And to þe kyng of Frauns he was **grete enmye**... (CMCAPCHR,131.3063)
And to the king of France he was great enemy...

b. ...as **keper of a town**, þat was famed to hym... (CMWYCSER,256.559)
...as keeper of a town, that was famed to him...

c. ...þere he was i-made **preost of Agelbertus**; (CMPOLYCH,VI,187.1349)
...there he was made priest of Agelbertus;

d. ...the riche king of Lidia, whom Cyrus had taken **prisoner**... (ELYOT-E1-H,154.162)
...the rich king of Lidia, whom Cyrus had taken prisoner...

All of these are examples of *role predicates*, which are a class of nouns that frequently appear bare in Middle English. Role predicates are so named because they denote sets of individuals that are socially constructed and dependent on context, including professions (doctor, lawyer, teacher), clergy members (monk, nun, pastor), as well as a variety of other nouns (master, servant, keeper), standing in contrast to nouns such as "man" and "dog", which denote inherent categories.[3] This observation is interesting in that roles comprise one of the classes of singular noun phrases that often appear bare in the modern Romance and Germanic languages, including French, Italian, Spanish, Brazilian Portuguese, and Dutch, which suggests the possibility that bare noun phrases in Middle English could be analyzed in a similar way to these languages.

[2]The citations in these examples are token IDs, which identify the source text, section, and token (clause, roughly) of the example.

[3]Kupferman (1991), via Munn and Schmitt (2005).

This brings up several questions regarding the status of bare nouns in historical stages of English. First, what is the complete distribution of bare singular noun phrases in Middle English? Specifically, what lexical classes of nouns can appear bare in different syntactic positions? The exact distribution may turn out to be somewhat wider or narrower than the exactly the role predicates, so if we are to analyze historical English in comparison with modern languages allowing bare singulars, then we need to have a clearer picture of the descriptive facts.[4] Second, what changed between Middle English and Modern English that made these bare singulars disappear? Was there a single change in the grammar of English, or were there several changes affecting different subclass of nouns at different times? In either case, the nature of the change may shed light on the nature of the synchronic variation in the modern Romance and Germanic languages.

The aim of this paper to provide the start of an answer to both of these questions by means of a corpus study tracing the distribution of bare singular noun phrases between Middle English and Early Modern English. Thanks to the recent construction of digital parsed historical corpora such as the Penn Corpora, which can be searched using programs like CorpusSearch (Randall 2010), it is now possible to automatically collect large enough amounts of data for statistical analysis. Because of this, the present study is also test of such methods for the purpose of tracking changes in the properties of classes of lexical items. The data will be used to test the predictions of Munn and Schmitt's (2005) analysis of the cross-linguistic variation in the distribution of bare singular noun phrases.

The remainder of this paper is organized as follows. In Section 2, I will present an overview of the cross-linguistic distribution of bare singular noun phrases. In Section 3, I will introduce an adaptation of Munn and Schmitt's theoretical analysis. In Section 4, I will briefly review some what is known about the history of the syntax of noun phrases in English and discuss how the theoretical analysis bears on the patterns we expect to find. In Section 5, I will present the results of a corpus study tracing the distribution of bare singular noun phrases between Middle English and Early Modern English.

[4] In Middle English, there is allegedly some variability in the use of articles according to the pragmatic meaning of the noun phrase; for example when the noun phrase denotes a function rather than an individual (Fischer 1992, p. 219). This kind of variation is likely a distinct phenomenon from the basic lexical and syntactic patterns, and will not be discussed further.

2 Crosslinguistic Variation

In this section, I will overview the basic distribution of bare noun phrases in the modern Romance and Germanic languages, followed by a theoretical analysis based on Munn and Schmitt (2005). The primary languages discussed will be English (Eng), Dutch (Dut), French (Fr), Italian (It), Spanish (Sp), and Brazilian Portuguese (BrP).[5]

There are three basic asymmetries in the availability of bare noun phrases that appear throughout the languages under discussion: (i) mass and plural nouns vs singular nouns, (ii) arguments vs predicates, and (iii) roles vs non-roles. We will look at each of these contrasts in turn, starting by comparing English to the Romance languages, and then introducing roles and the Dutch data.[6,7]

2.1 Bare Arguments and Predicates

In what follows, arguments and predicates are categorized based on their semantics. Noun phrases that denote entities, such as most verbal subjects and objects, will be considered arguments. Those that denote properties, or have open semantic arguments of any sort, will be considered predicates; these include post-copular noun phrases (be-predicates), small clause predicates, predicative objects of 'as' (as-predicates), and other secondary predicates.

In general, plural nouns pattern with mass nouns. Both are more likely to be bare than singular nouns, and are more likely to be bare as predicates than as arguments. Thus, we see that plural subjects and objects can be bare in English and Brazilian Portuguese, and objects can be bare in Spanish and Italian, whereas French does not allow subjects or objects to be bare. Because mass nouns are mostly beyond the scope of the current study, I will omit them from the examples in the remainder of this section.

(4) Plural Arguments

a. Mary saw crumbs on the floor. ✓Eng, *Fr, ✓It/Sp/BrP

b. Crumbs spilled off the table. ✓Eng, *Fr/It/Sp, ✓BrP

[5]The facts in Spanish are very similar to those in Italian. The differences are not relevant to the major points under discussion.

[6]The data in this section derived from Munn and Schmitt (2005) [Romance] and De Swart, Winter, and Zwarts (2007) [Dutch].

[7]There are also restrictions in the interpretation of bare noun phrases in certain positions (e.g. existential or generic), which vary by language. For the purposes of this paper, the crucial point is whether or not they are possible at all.

In contrast, bare singular arguments are rare in Spanish in Italian,[8] and non-existant in English and French. Only in Brazilian Portuguese are they freely available and relatively unrestricted in their interpretation.[9]

(5) Singular Arguments

a. Mary saw penny on the floor.	*Eng/Fr/It/Sp, ✓BrP
b. Penny fell off the table.	*Eng/Fr/It/Sp, ✓BrP

Bare singular arguments in Brazilian Portuguese are particularly interesting in that they are number neutral – this will be of considerable importance later.

Additionally, there is one special environment that conditions bare singular arguments cross-linguistically, namely conjunction structures, such as in the following example from Heycock and Zamparelli (2003).

(6)

a. A black cat and a brown dog were fighting in the street.	✓Eng/Fr/It/Sp/BrP
b. [Cat and dog] were equally filthy.	

Heycock and Zamparelli argue that these bare noun phrases are licensed by a kind of phrasal N-to-D movement that targets the conjoined noun phrases (which they analyze as NPs). While I do not examine this phenomenon is any further detail here, the fact that conjunction structures condition bare arguments in languages that otherwise do not allow them means that we will have to control for them in the corpus study.

Within predicates, the distribution is more complex, but shows a similar division. Here too, bare plural (and mass) noun phrases are widely available than bare singulars. English, Italian, and Brazilian Portuguese be-predicates allow bare plurals, while French again does not, and bare singular be-predicates are not possible in any language, including Portuguese.

(7) Be-Predicates

a. That is penny.	*Eng/Fr/It/Sp/BrP
b. Those are crumbs.	✓Eng, *Fr, ✓It/Sp/BrP

On the other hand, bare singulars *and* plurals are available in other (secondary) predicates in the Romance languages, but in English only bare plurals are possible.

[8] Some bare singulars can be found in the scope of negation in Italian (Longobardi 1994:613), and in a somewhat wider distribution in Spanish (Bosque 1996).

[9] Even in Brazilian Portuguese, bare singulars are not very common (Cristina Schmitt, p.c.). It is likely that various discourse factors condition against bare nouns much of the time (Alan Munn, p.c.).

(8) As-Predicates

a. We can use this box as table. *Eng, ✓Fr/It/Sp/BrP

b. We can use these boxes as chairs. ✓Eng/Fr/It/Sp/BrP

Just from this data, we can see that the option for noun phrases to appear bare is dependent on a variety of seeming independent factors. Bare plurals are more common than bare singulars, but there is significant variation in each category. Bare predicates are more widely available than bare arguments, but Brazilian Portuguese has fewer restrictions on bare arguments than on bare predicates. French has almost no bare noun phrases at all, but in the one place where it allows bare plurals, it also allows bare singulars.

2.2 Role Predicates

As mentioned in the Section 1, role predicates are often exceptionally bare in predicate position. For example, both singular and plural roles can be bare in be-predications in all of the Romance languages, including French, and (as is expected) they can be bare in other predicates as well.

(9) Be-Predicates (Roles)

a. John is doctor. *Eng, ✓Fr/It/Sp/BrP

b. John and Mary are doctors. ✓Eng/Fr/It/Sp/BrP

(10) As-Predicates (Roles)

a. John works as doctor. *Eng, ✓Fr/It/Sp/BrP

b. John and Mary work as doctors. ✓Eng/Fr/It/Sp/BrP

These bare roles also have a special interpretation, which is that the subject is an current participant in the activities associated with the role; I will call this the *active participant reading*. Thus, in the case of (9a) John cannot be retired. For bare singulars, this is the only interpretation, which is lost if the indefinite article is added. With an article, the interpretation is roughly the same as most other count nouns; I will call this the *set-membership reading*. For bare plurals, both the active participant and neutral set membership readings are possible.[10] This pattern is summarized in Table 1.

A more restricted class, the *unique role predicates*, which denote roles stereotypically filled by a single individual in any context (such as political offices and high ranking clergy members), can even appear bare in English.

[10]In in Romance languages, at least. Dutch behaves differently, as we will see later.

Type	Example	Set-Membership Reading	Active Participant Reading
Marked Singular	John is a doctor.	✓	*
Bare Singular	John is doctor.	*	✓
Bare Plural	John and Mary are doctors.	✓	✓

Table 1: Interpretation of Role Predicates

(11) Unique Role Predicates

a. John is president. ✓Eng/Fr/Sp/It/BrP

b. John was elected president. ✓Eng/Fr/Sp/It/BrP

I will not have as much to say about bare unique roles, since they seem to be qualitatively different from other instances of bare nouns. In these cases, the bare form appears to be in complementary distribution with the definite article rather than the indefinite, and it is likely that whatever that licenses them in English is distinct from the factors that license the larger set of roles in other languages.

Dutch adds yet another variation to the themes we have seen so far. Like English, bare plural arguments are possible but bare singulars are not. Like the Romance languages, both singular and plural noun phrases can be bare in many predicate positions, but curiously, bare singulars are semantically number-neutral.

(12) Dutch Role Predicates[11]

a. Jan is leraar.
Jan is teacher
'Jan is a teacher.'

b. Jan is een leraar.
Jan is a teacher.
'Jan is a teacher.'

c. Jan en Sofie zijn leraar.
Jan and Sofie are teacher
'Jan and Sofie are teachers.'

d. Jan en Sofie zijn leraren.
Jan and Sofie are teachers
'Jan and Sofie are teachers.'

Similar to the Romance bare roles, there is an interpretive difference between the variants. In Dutch, however, the unmarked form only carries the active participant reading while both the

[11] From De Swart, Winter, and Zwarts (2007)

singular and plural marked forms carry only the set-membership reading. This is likely to be a separate phenomenon from the number-neutral bare singular arguments in Brazilian Portuguese. In any case, these facts also need to be taken into account in any analysis of bare nouns phrases.

The data in this section is summarized in Table 2. Of the languages discussed, French is the most restrictive: bare nouns are almost universally banned except for role predicates and some secondary predicates. English allows bare plurals quite freely, but allows no bare singulars except for unique roles. Spanish and Italian are like French with the addition of some arguments. Dutch is like English with the addition of role predicates. Finally, Brazilian Portuguese allows most arguments and predicates to be bare.

Noun Category	English	French	Sp/It	Dut	BrPort
Mass/PL Args	✓	*	Restricted	✓	✓
Mass/PL Be-Preds	✓	Roles	✓	✓	✓
SG Arguments	*	*	Severely Restricted	*	✓
SG Be-Predicates	Uniq Roles	Roles	Roles	Roles	Roles
SG As-Predicates	Uniq Roles	Roles + etc.	Roles + etc.	Roles + etc.	Roles + etc.

Table 2: Cross-linguistic Summary

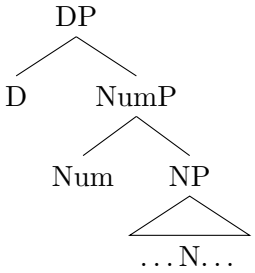
In the next section, I will present some theoretical assumptions that will allow us to uncover the deeper principles underlying the surface patterns.

3 Theoretical Background

3.1 Basic Assumptions

In order to make sense of the data in the preceding sections, we need a theory that can derive the distribution of bare noun phrases from a limited number of parameters. This can be done using a multi-layered analysis of the noun phrase, as is common in many recent Minimalist theories. A minimum of three layers is required: DP, NumP, and NP.

(13) Three-layer Noun Phrase Structure

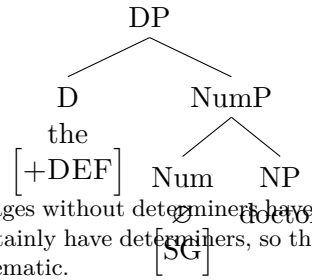


The typical assumption under such theories is that D is the locus of referentiality; hence arguments must have a D layer while predicates may lack it.[12] Num is the layer representing semantic number (at least for count nouns).[13] Again, the assumption is that most arguments have Num, since all quantificational determiners (including both the definite and indefinite) require a quantized noun, while predicates may be able to go without – whether a property is predicated of a singular or plural individual may not always matter. Finally, NP is the domain of thematic role assignment.

A further assumption is required: the definite article is in D, while the indefinite article is in Num.[14] This has several conceptual advantages. First, it captures the intuition that the indefinite article is fundamentally a number marker, as it appears in complementary distribution with plural marking. Second, it allows us to analyze both singular and plural indefinite (existential) arguments as having the same null determiner; the fact that indefinite article appears with both referential and non-referential noun phrases makes it odd to put it in the same category as the definite article.

In order to capture the fact that the indefinite article is in complementary distribution with all D-layer heads and phrases, we assume that singular Num is unpronounced in the presence of any overt D-layer material. Thus, definite and indefinite singular arguments will have the structures in (14a) and (14b), respectively.[15]

(14) a. Definite Singular Argument



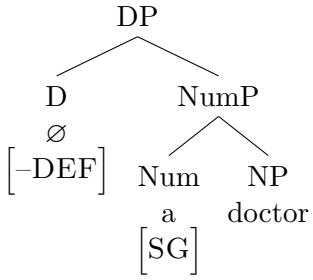
[12] While theories such as in Chierchia (1998) assume that languages without determiners have no D layer, all of the languages under discussion (except perhaps Old English) most certainly have determiners, so the assumption that all arguments in these languages have a D layer is relatively unproblematic.

[13] Numerals may be in SpecNumP, or another layer just above Num. The details of numerals are beyond the scope of this paper.

[14] Alternatively, the indefinite article may be in a lower projection in some languages. I will not pursue this possibility in this paper.

[15] Plural Num either values a feature on N, or causes N to move and adjoin to Num.

b. Indefinite Singular Argument



3.2 Two Parameters

Munn and Schmitt (2005) show that the variation within English and the Romance languages can be broadly accounted for using two parameters. The first, the strong/weak D parameter (Longobardi 1994), targets the D layer while the second, regarding optionality in semantic number realization, targets the Num layer.

Regarding the first parameter, languages with "strong" D, including French, Spanish, and Italian have strict restrictions on the availability of null determiners: they take mass or plural head nouns, must be governed by a lexical head, and receive an existential rather than a generic interpretation. In other arguments, D must be overt; N-to-D raising may be possible in the case of proper names, otherwise either a true or expletive determiner must be inserted. There may be additional language-specific restrictions. Thus, Spanish and Italian allow bare plural objects with existential readings (and some other restricted circumstances), whereas French requires an expletive determiner here as well. English and Brazilian Portuguese, on the other hand, have "weak" D, which is not subject to the restrictions that the other Romance languages have. This is what allows these languages to have bare plurals in both subject and object position and with both existential and generic interpretations.

The fact that only Brazilian Portuguese has unrestricted bare singulars, then, is explained by the second parameter, which dictates whether or not Num is obligatory in all noun phrases. The intuition is as follows: some languages require semantic number universally, while others allow it to be omitted when there is no true need for it – as I alluded in the previous section, this may be the case for some predicates as well as non-quantificational arguments.[16] Munn and Schmitt

[16]Munn and Schmitt encode this intuition by applying the Free Agr parameter of Bobaljik (2000) to the nominal domain: when Num is fused with Agr, it cannot be omitted, but when it remains separate, it can be. For the purpose of this paper, nothing critical hinges on the technical machinery employed.

propose that English is the first category, while the Romance languages are in the second. Since English requires Num, and the indefinite article is phonetically realized with null D (or no D), bare singulars in English are impossible. On the other hand, bare singulars can be found in Romance languages, including post-copular role predicates as well as most as-predicates.

Why, then, are there also bare plural predicates in French when the only places they are possible are those we claim Num to be missing? According to Munn and Schmitt, this is because the Romance languages still have *syntactic* number agreement even without semantic number. Roughly, there is an inflection number feature on all (count) nouns, and if it is not filled by the Num node then it will be filled by the next closest candidate, in this case the subject of predication. Finally, what is special about Brazilian Portuguese is that it has both weak D and optional Num. This allows for the possibility of a non-quantized null determiner that selects for an NP directly. Such a combination yields a bare singular argument, and this is exactly how Portuguese bare singulars are analyzed.

3.3 Dealing with Roles

In the above analysis, I left open the explanation for why some but not all predicates can be bare in Num-optional languages. For Munn and Schmitt, the reason has to do with theta-marking: a DP that merge within the NP domain is necessarily a thematic argument of the noun, and cannot escape the noun phrase due to whatever factor prevents structures like (15b).

(15) a. $[_{\text{DP}}\ [_{\text{DP}}$ The army's $]\ [_{\text{NumP}}\ [_{\text{NP}}\ t_i\ [_{\text{N}'}$ destruction of the city $]\ ]\ ]\ ]$

b. $*\,[_{\text{DP}}$ The army $]_i$ is $[_{\text{NumP}}$ Num $[_{\text{NP}}\ t_i\ [_{\text{N}'}$ destruction of the city $]\ ]\ ]$

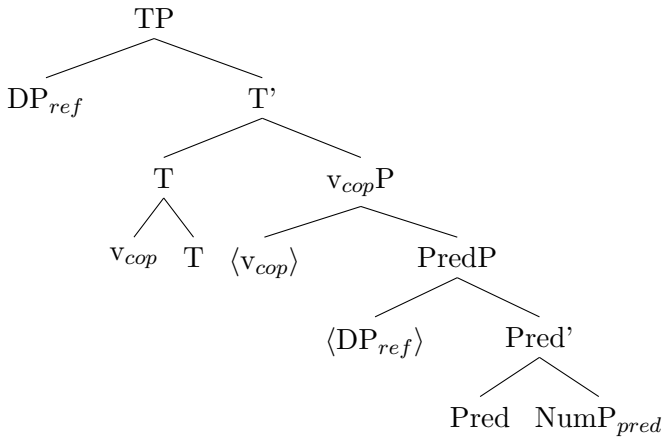
Under this account, in order for a typical noun phrase, with a referential argument to be bound, to act as a primary predicate, the subject of predication must be merged in another position. Munn and Schmitt suggest that the reason that Num is required in be-predicates is that the subject can be merged here, shielding it from becoming a theta-marked argument, but nothing critical hinges on this.[17] Under the analysis of be-predicates in Mikkelsen (2005), which is designed to capture the phenomenon of nominal predicate inversion,[18] the subject of predication is merged the specifier of a Pred node between the nominal predicate and the copula, and either the "subject" or "predicate"

[17] Thanks to Alan Munn for pointing this out to me.

[18] As in "John is the doctor" and "The doctor is John"

moves to become the subject of the clause. In this case, we can assume that NumP is selected by Pred for other reasons having to due with the semantics of primary predications.[19]

(16) Structure of a Predicational Copular Clause, Adapted from Mikkelsen (2005)



In any case, as-predicates are different; either *as* counts as an intervener or it lacks the semantic restrictions that Pred has, and the result is that singular nouns can generally be bare in as-predicates.

Along the same lines, Munn and Schmitt propose that what is special about roles is that they contain an event argument to be bound within the T domain, rather than a referential argument, and as a consequence they can take a thematic subject (merged within NP) but go without the Num and D projections. Instead, the copula merges with NP directly.[20] Whether roles are compatible with Num at all under this account is unclear, but it has the conceptual advantage of capturing the active participant reading associated with role predicates along with their enlarged distribution relative to other bare predicates.[21]

The Dutch data poses an additional wrinkle. In Dutch, singular role predicates are number neutral, and bare plural role predicates cannot have the active participant interpretation. Neither of these facts are not predicted simply by assuming that Dutch has weak D and optional Num, as we would under Munn and Schmitt's analysis. De Swart, Winter, and Zwarts (2007) analyze these noun phrases in a somewhat different way. Under their analysis, role predicates are derived from class of head nouns which can be converted using the semantic operators "CAP" and "REL" to two

[19] Mikkelsen assumes that all noun phrases are DPs, but this only is a technical detail in their analyses.

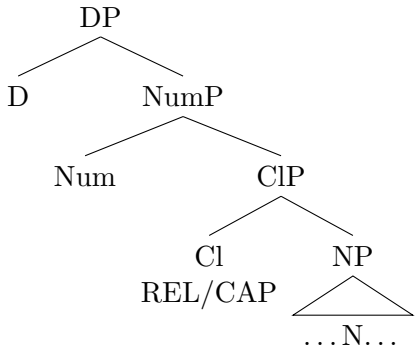
[20] Whether there might still be a Pred node in such cases is an open question.

[21] De Swart, Winter, and Zwarts (2007) suggest that roles can be readily converted to non-roles via a semantic operator, as discussed below. Even if the active participant reading is not available, these nouns can certainly appear as quantized arguments.

kinds of predicates: one with the interpretation that the subject is actively performing the role, and another with the neutral set-membership reading. These readings correlate with number-neutrality and number marking, respectively.

De Swart et al suggest that these operators apply just above the NP layer, in which case it may be possible to analyze them as classifiers using a four-layer extended noun phrase in the spirit of Borer (2005), where count nouns take a classifier containing REL, and roles take another containing CAP.

(17) Four-layer Noun Phrase Structure



If number inflection is actually located at the Classifier layer rather than the head noun, then we have a way to account for the agreement differences between Dutch and the Romance languages: the role classifier has no number inflection in Dutch, while in Romance both classifiers inflect. This is not an unreasonable assumption if mass nouns take a third classifier, which does not inflect in any of these languages. The fact that the indefinite article is absent can then be explained if Num does not accompany the role classifier, just as it is absent with mass nouns.

The disadvantage to this solution is that it weakens the generalization that just two parameters of strong/weak D and optional Num account for the broad patterns. On the other hand, if we find other languages like Dutch that stray from the Romance pattern, then an additional parameter such as a difference in the classifier system will become inevitable.

4 Historical Background

4.1 The Syntax of Middle English Noun Phrases

There has been relatively little work on the historical syntax of noun phrases in English in comparison to work on clause structure (Fischer 1992, pp. 210–211). What we do know is that English originally had no definite or indefinite article (as with Proto-Germanic in general) and that both arose in the late Old English period. The definite article *the* was derived from the demonstrative *se/seo/þat* 'that', while the indefinite article *a(n)* split from the numeral *an* 'one' (Fischer 1992, pp. 217–218). Both of these additions appear to be established by the early Middle English period. While it is difficult to be certain whether the word *an* is a numeral or an article in Old English, the fact that the spellings were differentiated in early Middle English makes it possible to confirm that the change was indeed complete. This is demonstrated in Section 5.2 below. So while it is difficult to distinguish the emergence of the article system from any other changes affecting the availability of bare singulars in Old English, from Middle English forward we can be reasonably certain that any changes in the *use* of the indefinite article are separate from its initial appearance.

Aside from the article system, other aspects of noun phrase syntax that could be relevant to either the loss of bare singular noun phrases or a corpus study thereof are largely the same as in Modern English. Prenominal numerals, demonstratives, quantifiers, and possessors are similar, and the case and agreement system have already decayed (Fischer 1992, pp. 210–211). Even if there were changes in these areas, it is clear from the cross-linguistic data that the primary factor controlling whether a singular noun phrase can be bare at all is whether the indefinite article is required.

4.2 Using Parsed Corpora for Historical Research

A historical linguistic corpus is an organized collection of (digitized) texts. A part-of-speech (POS) tagged corpus provides morpho-syntactic labels for every word, and a parsed corpus provides simple bracketed trees in addition to POS tags. One of the primary advantages of a corpus study is the aggregation of data, such as frequencies of occurrence of certain structures over time; in this case, the frequency of bareness in noun phrases with different head nouns and in different syntactic positions. By running a logistic regression predicting bareness based on the date of the source text

combined with these other group variables, it is possible to estimate the rate change in frequency over time for each combination. This information, then, can be used to test for the Constant Rate Effect (Kroch 1989), which is the observation that for a grammatical change, the rate change in frequency of use over time across contexts (properly measured) will be constant. If the amount of data is sufficient, this tells us how many different grammatical changes occurred, an empirical observation that can be compared to existing theories.

There are a number of potential caveats to this method. First, because tagged and parsed corpora are created with the assistance of automatic parsers, there can be errors in the annotation. However, for studies that focus on relatively simple syntactic structures, the practical effect of such errors should be small. One possible confound in a corpus study comparing syntactic positions would be the difficulty in distinguishing subjects from preposed predicates in copula clauses (Tony Kroch, p.c.), but again if we believe that the number of bad tokens is small compared to the correctly annotated tokens then this may be a non-issue. By far the biggest problem for this kind of study is guaranteeing enough data. Although the syntactic structures under investigation are simple and presumably common, the number of such structures containing the head nouns we are interested in (role nouns) may turn out to be quite small. In this case, the conclusions that can be drawn from the data will be limited.

5 Corpus Study

This section presents the corpus study from Hanson, Schmitt, and Munn (2014) tracing the distribution of bare singular noun phrases between Middle English and Early Modern English. The study utilizes the CorpusSearch program (Randall 2010) in order to automatically search for and code the syntactic and semantic properties of singular noun phrases from the Penn-Helsinki Parsed Corpus of Middle English, Second Edition (PPCME2) (Kroch and Taylor 2000), the Penn-Helsinki Parsed Corpus of Early Modern English (PPCEME) (Kroch, Santorini, and Delfs 2004), and the York-Helsinki Parsed Corpus of Early English Correspondence (PCEEC) (Taylor et al. 2006). The resulting data will not only give us a more complete picture of bare noun phrases in history of English, but can potentially also be used to identify if surface changes in their distribution are reflexes of a one (or perhaps several) changes in the underlying grammar.

5.1 Predictions

Munn and Schmitt's analysis makes a number of predictions about the historical data we will find. First, it predicts that there will be a distinction between arguments and predicates. Assuming that Middle English had optional Num, it should be easier for Num to be missing in predicates than in arguments. Next, it predicts that within predicates there should be a division between roles and non-roles, where roles are bare more frequently and in a wider range of positions. On the other hand, we don't expect to see much of a division within the arguments. Finally, it predicts that if there are bare predicates, there may be at least a small number of bare arguments early on. Since English has had null D (and presumably weak D) since the Old English period, it would not at all be unexpected for these to remain marginally possible into the early Middle English period.

5.2 Preliminary Search

In order to ensure that the indefinite article 'an' is fully distinguished from the Old English numeral *an* in our data, we conducted a search of all possible spellings of 'an', 'one', and 'tone'[22] that co-occur with the definite article or a demonstrative in the Middle English. The result was that *an*-type spellings co-occurred with the determiners only in the earliest period, as shown in Table 3.

Spelling Group	Time Period			
	1150–1250	1250–1350	1350–1420	1420–1500
a/an	8	0	0	0
one-type	1	0	37	23
tone-type	0	1	0	7

Table 3: Spellings of an, one, and tone co-occurring with the definite article or a demonstrative in the PPCME2

From this we can conclude that Middle English writers did indeed distinguish the indefinite article from the numeral, and any ambiguity is restricted to the earliest period.

5.3 Methods

The corpora were coded using a series of CorpusSearch revision, search, and coding queries, coordinated using shell scripts and Unix makefiles.

[22] A form of 'one' that often co-occurs with the definite article

First, a revision query was used to extended the set of node labels used for noun phrases to include additional syntactic functions needed for the study. Both the original and added labels relevant to this study are listed in Table 4.

Labels provided	NP–SBJ – verbal subjects
	NP–OB1 – direct objects
	NP–OB2 – indirect objects
	NP–SCP – small clause predicates
	NP–SPR – secondary predicates
	NP–POS – possessors[23]
Labels added [24]	NP–OB1–BE – object of 'be'
	NP–POBJ–AS – object of 'as'
	NP–CONJ – conjuncts of a larger NP [25]

Table 4: Labels for syntactic functions of noun phrases

Next, lists of singular count nouns representing the lexico-semantic classes relevant to the study (unique roles, non-unique roles, and non-roles) were created, which are shown in Table 5. We then searched the corpora for all noun phrases with head nouns on these lists, which were then coded accordingly. The corpora used are not lemmatized, so all unambiguous spelling variants that appear in the corpora were compiled manually.

Class	Lexical Items
Unique Roles	bishop, duchess, duke, emperor, empress, king, lord, pope, queen
Non-unique Roles	ally, doctor, friar, guard, keeper, knight, monk, nun, priest, prince, princess, prisoner, servant, sinner, slave, soldier, teacher
Non-roles (Animate)	beast, ghost, maiden, man, woman
Non-roles (Concrete)	body, book, face, house, mouth, voice, weapon
Non-roles (Abstract)	deed, end, heart, name, soul, tale, thing, word, world

Table 5: Classes and lexical items included in search

Lexical items were chosen based on their intuitive meanings and their distribution in languages allowing bare role predicates. As much as possible, we tried to include words that appear frequently throughout the time period under investigation, and do not have word senses in more than one of the semantic classes. Because what counts as a role and what counts as a unique role is context-

[24]The complete coding scheme includes additional categories that are not included in the analysis for the present study.

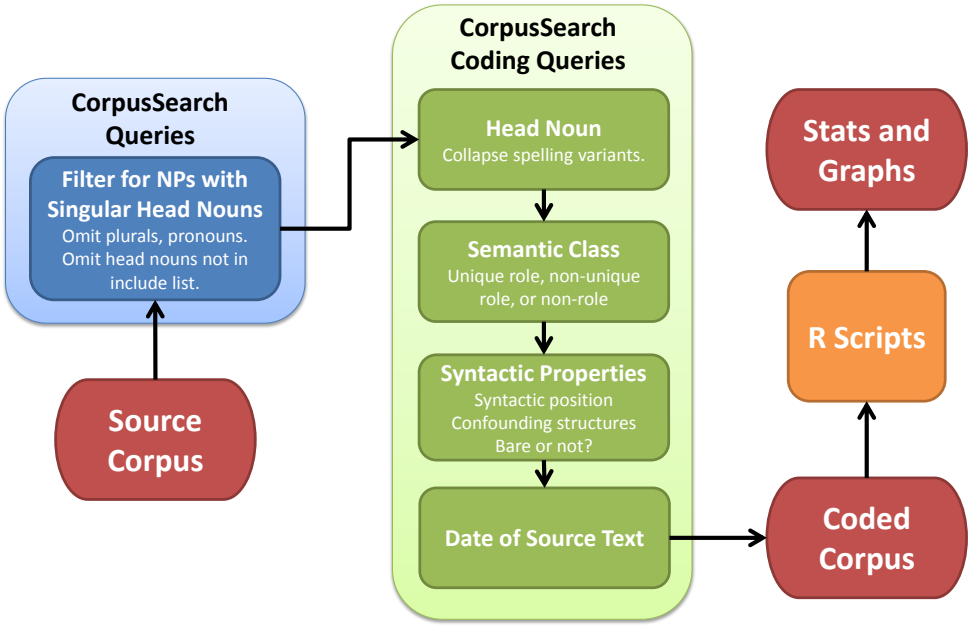
[25]NPs containing conjuncts need to be filtered out due to the independent licensing of bare singulars in conjoined noun phrases.

sensitive, we selected the class that we believe the word to fall into more readily. For example "bishop" without qualification is a non-unique role (as evidenced by the fact that "He is bishop" is odd in English), but is unique when it appears with a prepositional phrase specifying the domain (as in "He is bishop of Canterbury"), so we coded it as a unique role. Non-roles are broken into subclasses in this table in order to make it clear that we included a broad range of lexical items; all three subclasses were collapsed for analysis.

After coding for head noun and semantic class, we coded for the remaining factors: syntactic position (using the labels created using the revision query), bareness (as defined in Section 1), any confounding structures that condition for or against bareness across languages (including conjunction structures and adjective modification), and (approximate) date of source text.

The resulting coding strings for all data tokens were extracted, and processed using a series of R scripts. Confounding data tokens were filtered out, and coding values were collapsed to remove irrelevant contrasts or raise the number of data tokens in some categories to an acceptable level for graphing. Graphs of logistic regressions of frequency of bareness over time were created using *ggplot2*, partitioned by semantic class and syntactic position. The entire process is summarized in Figure 1.

Figure 1: Corpus Coding and Analysis Process



5.4 Results

Data token counts for all syntactic positions and semantic classes are presented by time period in Table 6. By merging all verbal arguments, and we were able to obtain enough data for most combinations of class and position, small clauses had merged with secondary predicates to obtain reasonable counts, and the counts for other combinations remains low. While the conclusions that we can draw from the data are limited in these cases—in particular, we must be careful about extrapolating our findings beyond the data time frame when the confidence intervals are large—we can see many broad patterns clearly.

Period	Class	As-Pred.	Be-Pred.	SC/SP	Verb Arg.
ME	Non-roles	44	226	28	5812
ME	Non-unique roles	8	75	42	323
ME	Unique roles	12	448	185	1722
EME	Non-roles	89	298	21	5688
EME	Non-unique roles	10	109	48	472
EME	Unique roles	4	42	27	1085

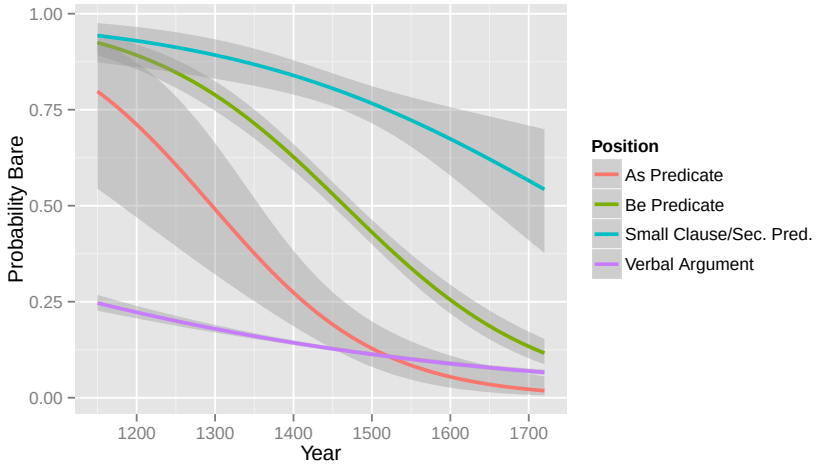
Table 6: Data token counts

I did not notice any strong patterns specific to individual lexical items, with the exception of the pair "man" and "ghost", which were unusually likely to appear bare in predicate positions. Considering that a large portion of the texts in the corpora are of religious subject matter, it is likely that these are behaving as roles when describing the forms that Christ takes in the Gospels. This being said, the effect was not enough to skew the overall trends, so these words were left with their original coding.

First, as shown in Figure 2, the overall frequency of bare singular noun phrases declined over the time period under investigation, with predicates initially overwhelmingly bare and arguments overwhelmingly non bare in early Middle English. Small clauses and secondary predicates are the most likely to be bare overall, followed by be-predicates and as-predicates, with arguments low throughout the time period. This is consistent with what we might expect to find based on Munn and Schmitt's analysis. If Middle English started with weak D and optional Num, then there should a sizable number of bare predicates: those that don't require number marking. The fact that bare arguments are also slightly more frequent in the earliest period suggests the English may have gone through a phase with bare singular arguments like in Brazilian Portuguese. Another possibility is

that this slight decline is due to the completion of the emergence of the articles. Whether these bare singulars are also number neutral will require further investigation, and could potentially lend support to one of these two explanations: if number neutral bare singulars are possible at during this time and disappear by late Middle English, then the emergence of the article alone cannot explain the pattern.

Figure 2: All nouns by syntactic position



The patterns become more interesting when we subset the data by semantic class. First, we observe in Figure 3 that bare non-roles are already in decline in all predicate positions by the beginning of the Middle English period. While the confidence intervals are now much wider, the overall trend is unmistakable.

In contrast, when we look at the non-unique roles in Figure 4, we see that they are much more likely to be bare in predicate positions well into Middle English and even into the Early Modern English period. This makes sense if, as in the modern Romance languages and Dutch, bare role predicates are licensed by factors partially separate from those that allow all nouns to be bare. While the as-predicates appear to be doing something different from the other predicates, there was very little data for the as-predicates, and as a consequence the confidence interval fills the graph.

Unique roles follow a third pattern, as shown in Figure 5. This time, bare small clauses and secondary predicates as well as be-predicates remain highly probable throughout the time period. This makes sense considering that they are still possible in Modern English. The observation that

Figure 3: Non-roles by position

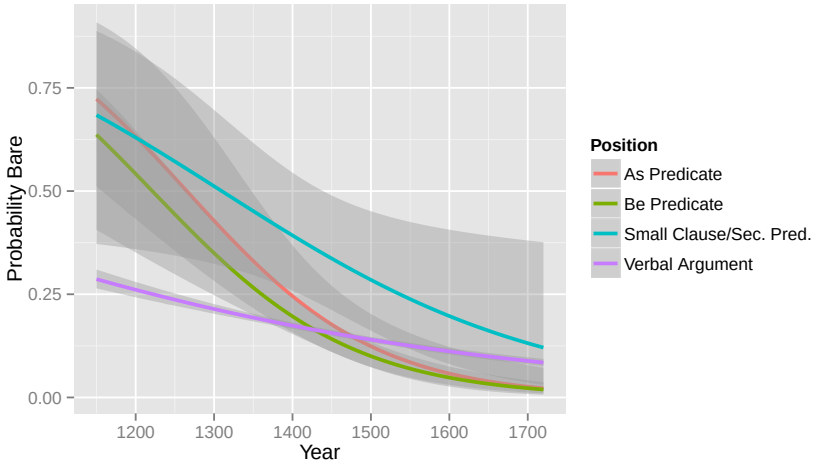


Figure 4: Non-unique roles by position

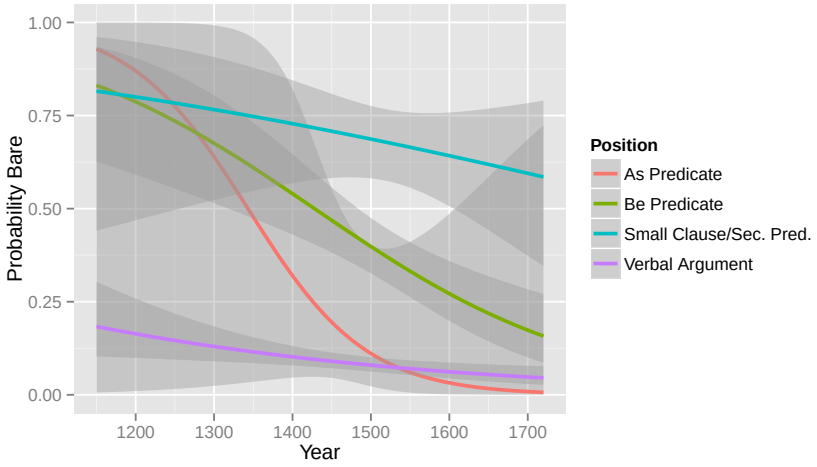
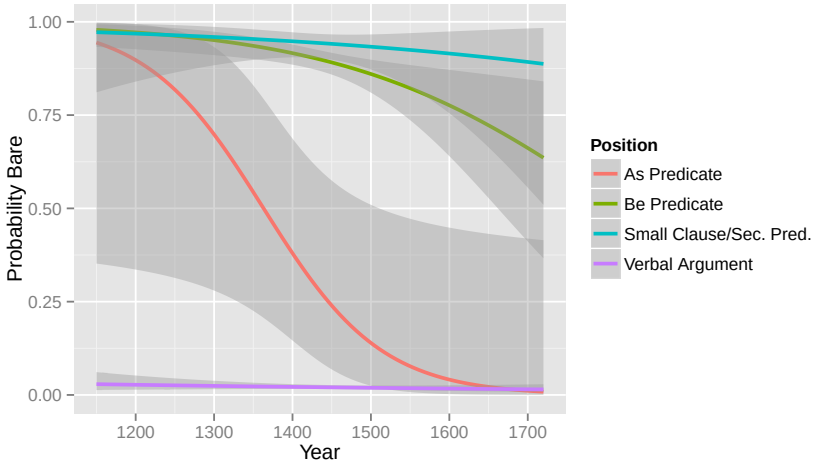


Figure 5: Unique roles by position



the number of bare arguments is lower than that of the other classes is also explainable by the fact that for unique roles, the bare form competes with the definite article rather than the indefinite; if definiteness was obligatorily marked with the definite article in arguments but not predicates, then there should not be any bare unique role arguments. The as-predicates, again, had too little data to be sure whether or not they pattern with the other predicates.

To summarize, the trends are as follows:

- Bare nouns decreased in all positions, with Unique roles > Non-unique roles > Non-roles and SC/SP > Be-predicates > As-predicates > Arguments
- Arguments and predicates followed different trajectories. Bare arguments were already low at the start of the time period and decline to marginal levels, while bare predicates were high and decline rapidly over the period.
- Roles, non-unique roles, and non-roles followed different trajectories. Bare non-roles predicates were already declining at the beginning of Middle English, soon to be followed by the non-unique roles, while unique-roles remained bare into Modern English.

Due to limited data obtained from the first iteration of the study, the constant rate effect could not be tested using numerical statistics at the time of writing. At this point, we can only hypothesize which curves represent the same change in progress using the graphical curves. Nonetheless, we were able to confirm the broad predictions of Munn and Schmitt's analysis: that arguments should be distinguished from predicates, that role predicates should be distinguished from non-role predicates, and that there should be at least some bare arguments.

6 Conclusion

While much work remains to be done on the syntax of bare noun phrases, the English historical data from the corpus study presented in this paper suggests that analyses along the lines of Munn and Schmitt (2005) are on the right track. There may be additional factors at work, but we can already see clearly that the argument/predicate and role/non-role distinctions, which are claimed to be related to the determiner and number systems, were significant during the loss of bare singular noun phrases in the history of English. By increasing the depth of such studies, the hope is that we

will soon be able to integrate historical English and the other Germanic languages into a principled account of the cross-linguistic distribution of bare noun phrases.

References

Bobaljik, Jonathan (2000). "The rich agreement hypothesis in review". In: *Ms., McGill University*.

Borer, Hagit (2005). *Structuring sense: Volume 1: In name only*. Vol. 1. Oxford University Press.

Chierchia, Gennaro (1998). "Reference to kinds across language". In: *Natural language semantics* 6.4, pp. 339–405.

De Swart, Henriette, Yoad Winter, and Joost Zwarts (2007). "Bare nominals and reference to capacities". In: *Natural Language & Linguistic Theory* 25.1, pp. 195–222.

Fischer, Olga (1992). "Syntax". In: *The Cambridge history of the English language, volume 2: 1066–1476*. Ed. by Norman Blake. Vol. 2. Cambridge: Cambridge University Press. Chap. 4, pp. 207–408.

Hanson, Kenneth, Cristina Schmitt, and Alan Munn (2014). "The loss of bare singular noun phrases in the history of English". In: Great Lakes Expo for Experimental and Formal Undergraduate Linguistics (GLEEFUL). q Undergraduate Society for Linguistics at Michigan State (qUALMS).

Heycock, Caroline and Roberto Zamparelli (2003). "Coordinated bare definites". In: *Linguistic Inquiry* 34.3, pp. 443–469.

Kroch, Anthony, Beatrice Santorini, and Lauren Delfs (2004). *Penn-Helsinki Parsed Corpus of Early Modern English (PPCEME)*. Department of Linguistics, University of Pennsylvania. URL: `http://www.ling.upenn.edu/hist-corpora/`.

Kroch, Anthony and Ann Taylor (2000). *The Penn-Helsinki Parsed Corpus of Middle English, second edition (PPCME2)*. Department of Linguistics, University of Pennsylvania. URL: `http://www.ling.upenn.edu/hist-corpora/`.

Kroch, Anthony S (1989). "Reflexes of grammar in patterns of language change". In: *Language variation and change* 1.03, pp. 199–244.

Kupferman, Lucien (1991). "Structure événementielle de l'alternance un/Ø devant les noms humains attributs". In: *Langages* 25.102, pp. 52–75.

Longobardi, Giuseppe (1994). "Reference and proper names: a theory of N-movement in syntax and logical form". In: *Linguistic inquiry*, pp. 609–665.

Mikkelsen, Line (2005). *Copular clauses: Specification, predication and equation*. Vol. 85. Linguistics Today. John Benjamins.

Munn, Alan and Cristina Schmitt (2005). "Number and indefinites". In: *Lingua* 115.6, pp. 821–855.

Randall, Beth (2010). *CorpusSearch*. URL: `http://corpussearch.sourceforge.net/index.html`.

Taylor, Ann et al. (2006). *York-Helsinki Parsed Corpus of Early English Correspondence (PCEEC)*. Department of Linguistics, University of York. URL: `http://www-users.york.ac.uk/~lang22/PCEEC-manual/index.htm`.

Traugott, Elizabeth (1992). "Syntax". In: *The Cambridge History of the English Language, volume 1: the beginnings to 1066*. Ed. by Richard Hogg. Vol. 1. Cambridge: Cambridge University Press. Chap. 4, pp. 168–289.