# A MULTI-STEP ANALYSIS OF THE EVOLUTION OF ENGLISH DO-SUPPORT

Aaron Ecay

A DISSERTATION

in

Linguistics

Presented to the Faculties of the University of Pennsylvania in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy

2015

Supervisor of Dissertation

_____

Anthony Kroch, Professor of Linguistics

Graduate Group Chairperson

_____

Eugene Buckley, Associate Professor of Linguistics

Dissertation Committee:
Mark Liberman, Professor of Linguistics
Charles Yang, Associate Professor of Linguistics

For my parents

This dissertation is backed up by a small army of my colleagues who have contributed in ways both large and small to its empirical and technical underpinnings. It is only appropriate to acknowledge their contributions here – though of course any shortcomings in the dissertation should not be attributed to them. Pieces of this work were presented at PLC 34 and 38, and DiGS 14 and 16; audience members there provided valuable feedback. I am grateful to Amy Goodwin-Davies, Kajsa Djärv, and Heimir Freyr van der Feest-Viðarsson for providing judgments on their respective native languages. Heimir also provided illuminating of the early Icelandic *do*-support patterns. I am grateful to Hilary Prichard for sharing her geocoding of the PPCME2 texts, and thus infecting me with enthusiasm for exploring spatial patterns of variation in historical syntactic data. Don Ringe pointed out example (81) several years ago – and thus provided a crucial inspiration for the account of intermediate *do*-support that developed into the core of this dissertation. Ann Taylor and Anthony Warner created an electronic version of the coding strings in Ellegård's corpus, thus allowing me to check my results against this earlier source of data. Their hard work and generosity have been instrumental in allowing me to develop the quantitative analyses more fully. In culmination, the entirety of the Penn Parsed Corpora of Historical English (PPCHE) project deserves a distinguished mention. It would be impossible to overstate the debt that is owed, in the sense that writing this dissertation would have been impossible – indeed inconceivable – if the PPCHE did not exist. It was the elegant, thorough quantitative underpinning of English historical syntax that drew me to graduate school in the first place, making the PPCHE deeply important to me on a personal level as well. I hope to one day be able to create resources which unlocks as much potential for the field and for its future students as the PPCHE has done.

I have been especially lucky to have spent nine years at Penn with some of the best teachers in the world. I am grateful to all of them, but especially to the following. Lance Nathan first introduced me to the notion of formalism in linguistics. Julie Legate deepened that acquaintance and mapped out the landscape of modern theoretical syntax. I mean that literally – her careful blackboard taxonomies were a constant source of enlightenment in her classes, and one that I can only aspire to emulate in my own career. Beatrice Santorini has been a constant since the beginning of this dissertation, hatched in her LING-300 course in the fall of 2009. She has provided unerring guidance with her ability to cut right to the heart of an issue with an incisive question. She has also fielded many emails over the years reporting errors or inconsistencies in the PPCHE files. She has always been supremely helpful in fixing the errors, even the ones that are located between my ears. My committee members each contributed to my education, and this dissertation, in their own ways. Charles Yang through his teaching and research has impelled me to consider the "big picture"

and how it relates to any given specific problem. Mark Liberman exposed me to big data linguistics; the results of his influence can be seen most of all in chapter 5.

My fellow graduate student have made an equally important contribution to the last five years of my life. The voices and personalities of these colleagues remain in my own mental landscape, expressing some of the most crucial ideas to my scientific and personal success – ideas that they themselves helped inculcate. I will make the attempt to enumerate the most salient members of the chorus, cognizant of the fact that I will certainly fall short. Joel Wallenberg and Josef Fruehwald each served as role models in statistical methodology and corpus analysis, as well as a constant willingness to teach these skills to others. Jana Beck impressed on me the importance of the rigorous implementation and clear documentation of the technical aspects of the field. Chris Ahern was the best officemate I could have asked for, always excited to share his research with me, equally eager to hear about mine, and adept at finding conceptual links between these and any other areas of the field. Meredith Tamminga proved an excellent collaborator on the results in section 4.3.5. What's more, her advice on intellectual and professional issues is always spot on, and her passionate drive is a source of motivation. Caitlin Light has been like a cool older sibling – there's no other way to describe that mix of advice, inspiration, and camaraderie. Brittany McLaughlin and I made friends on my first day of graduate school, and she's been there for me ever since. Similarly, I met Betsy Sneller on *her* first day of graduate school. I could not ask for a better friend. Finally, I want to thank all the folks at 3810 Walnut for their companionship.

My parents have been my first and most important role models. They have always emphasized the value of education, and have sacrificed willingly of themselves so that I might pursue it. That's why I've dedicated this dissertation to them. I love you, mom and dad. I'm also grateful beyond words to Jamie, who has given me unconditional love and support especially in the last months of the writing process.

Last but by no means least, I would like to express my deepest gratitude to my advisor, Tony Kroch. I've been working with Tony for seven years now, and he has influenced my thinking, and my life, immensely. Almost everything that I know about being a linguist can be traced to him. His curiosity, versatility, and intelligence are a constant inspiration. I'll always be proud to call myself a student of Tony Kroch.

ABSTRACT

A MULTI-STEP ANALYSIS OF THE EVOLUTION OF ENGLISH DO-SUPPORT

Aaron Ecay

Anthony Kroch

This dissertation advances our understanding of the historical evolution and grammatical structure of English *do*-support through the application of novel historical data to this classical problem in historical syntax. *Do*-support is the phenomenon in English whereby a pleonastic auxiliary verb *do* is inserted in certain clause types. The phenomenon is characteristic of the modern language, and there is robust evidence that it emerged beginning in roughly the year 1500. The fine quantitative details of this emergence and the variation it engendered have been an object of study since Ellegård (1953). From the standpoint of generative grammar, Roberts (1985), Kroch (1989), and many others have treated the emergence of do-support as a closely-following consequence of the loss of V-to-T raising in the 15th and 16th centuries. Taking a cross-linguistic perspective, I show that though the totality of English *do*-support is uncommon in other languages, the phenomenon may be seen as the combination of several discrete building blocks, each of which is robustly attested. From this perspective, a question is raised about the genesis of English *do*-support: given that the present-day phenomenon is evidently composed of several separate subcases, why should its cause be attributed solely to the loss of V-to-T raising? I argue that the earliest emergence of *do*-support in English is in fact attributable to a different source: a usage of *do* as a marker of external arguments. This explanation addresses the following points, which under earlier accounts were unexplained:

- The different behavior of *do*-support across argument structure types.
- The appearance of *do*-support in affirmative declaratives at a peak rate of 10%, much more than can be attributed to emphatic assertions.
- The emergence of *do*-support from a Middle English causative.

This "intermediate *do*" spread through the language until roughly 1575, when the loss of verb raising triggered an abrupt reanalysis which transformed the argument-structure marking *do* into its modern form.

# Contents

# List of Tables

# List of Figures

# Chapter 1

# Introduction

This dissertation addresses a series of related theoretical and empirical questions in the historical syntax of EME, and specifically the *do*-support construction. *Do*-support is the phenomenon in English whereby a pleonastic auxiliary verb *do* is inserted in certain clause types (such as negatives, emphatic assertions, and subject-auxiliary inversion contexts including questions). The phenomenon is characteristic of the modern language, and there is robust textual evidence that it emerged during EME, beginning in roughly the year 1500. This change has been of long-standing interest to philologists and linguists, with many theories of its origin and spread advanced. The fine quantitative details of the emergence of *do*-support and the surface variation it engendered in texts has been the object of study for over 60 years, beginning with Ellegård (1953). From the standpoint of generative grammar, Kroch (1989) and Roberts (1985), and many others have treated the emergence of do-support as a closely-following consequence of the loss of V-to-T raising in the 15th and 16th centuries.

Taking a cross-linguistic perspective, this dissertation amasses evidence that, though the totality of English do-support is uncommon in other languages (if not indeed entirely unattested), the phenomenon may be seen as the combination of several discrete building blocks, each of which is robustly attested in languages, some of which are historically and typologically close to English and some of which are not. From this perspective, a question is raised about the genesis of English do-support: given that the present-day phenomenon is evidently composed of several separate subcases, why should its cause be attributed solely to the loss of V-to-T raising?

The core innovative proposal of this dissertation is a grammatical model of the evolution of *do*-support which explains the presence of these affirmative declarative *do*-support sentences. Using a database drawn

from the PPCHE which has never previously been applied to the study of do-support, this dissertation provides evidence for this proposal. This explanation addresses the following points, which under earlier accounts were unclear:

- The different behavior of do-support across argument structure types (latent in Ellegård's description of the phenomenon and elaborated more clearly in the more adaptable data from the PPCHE).
- The appearance of do-support in affirmative declaratives at a peak rate of 10%, much more than can be attributed to emphatic assertions (previously unexplained).
- The emergence of do-support from a Middle English causative, since there is a plausible semantic reanalysis from causative to agent-marker.

This "intermediate do" (between the ME causative and the present-day pleonastic support auxiliary) spread through the language until roughly 1575, when the loss of verb raising triggered an abrupt reanalysis which transformed the argument-structure marking "do" into its modern form.

In the process of establishing this result, several ancillary conclusions are reached. The PPCHE corpus of *do*-support constitutes a source of evidence about the diachrony of the construction which is independent of Ellegård's corpus, on which previous quantitative studies were based. It is thus possible to pursue replication of previous results on *do*-support. Many of these are broadly upheld, increasing our scientific confidence in the reliability of quantitative historical studies. However, certain revisions are necessitated by the new data, demonstrating the continued value of data collection.

Furthermore, this dissertation demonstrates that such a rich diachronic understanding is necessary from the standpoint of purely synchronic analysis. There are two families of analysis, both of which make identical predictions about the distribution of *do*-support in PDE. Broadly speaking, one analysis inserts *do* where it appears, and the other deletes it where it does not. However, I argue that the understanding of the history of *do*-support which is suggested by previous work and confirmed and deepened by this dissertation favors the insertion account over the deletion one.

Finally, in order to validate certain hypotheses about the behavior of individual lexical verbs, an entirely novel mechanically-annotated database of EME texts was produced, comprising one billion ($10^9$) POS-tagged words. From this corpus, it is possible to extract two orders of magnitude more data on *do*-support than were previously available. The result is an ability to test hypotheses about the detailed diachronic behavior of individual lexical items, as well as subject existing predictions to tests in minute detail. The insights thus

gained further advance the understanding of *do*-support, and the data and methods presented will continue to be of use in other historical syntactic investigations.

This dissertation is organized as follows. Chapter 2 gives an overview of the phenomenon of *do*-support in English, and provides comparisons to similar phenomena in other Germanic languages as well as more typologically diverse languages. Chapter 3 introduces a variety of statistical techniques which will be brought to bear in the investigation. Chapter 4 contains the discussion of many of the core results of the dissertation, including the replication and extension of previous accounts and the novel proposals about the structural analysis of *do* in early EME. Chapter 5 discusses the extension of these results in a large automatically-annotated corpus, with special emphasis on examining the behavior of individual lexical items and a search for broader patterns among them. Finally, chapter 6 concludes.

# Chapter 2

# Synchronic considerations

In this chapter, I will give an overview of synchronic data that inform the primary topic of this dissertation, i.e. the diachronic behavior of *do*-support in English. In section 2.1, I'll review the environments in which *do*-support appears in PDE. Section 2.2 discusses the parallels between English *do*-support and similar constructions in other languages. These two sections inform the construction of a structural description of the *do*-support phenomenon, a crucial step for articulating how it arose in English. Section 2.3 discusses the presence in other languages of constructions involving a *do*-like auxiliary verb which is not completely devoid of semantics (as in *do*-support), but rather is associated with certain semantic features of the lexical verb. Section 2.4 discusses in more detail variation within English and other Germanic languages in the use of *do* as such a non-vacuous auxiliary. These two sections relate to the proposed analysis (detailed in section 4.3) of the intermediate stages of English *do*-support.

## 2.1 *Do*-support in PDE

*Do*-support refers to the phenomenon whereby a semantically vacuous ("dummy") auxiliary verb surfaces in certain morphosyntactic contexts. The "support" nomenclature arises from an intuition that *do*-support sentences are somehow deficient without the presence of *do*, and more specifically that the function of the auxiliary is to allow bound morphemes to be spelled out which for morphosyntactic reasons otherwise could not be. The phenomenon is named after the dummy auxiliary in English, although as will be discussed below (Section 2.2), it has been observed in other languages as well. In standard PDE, *do*-support shows up in several contexts in the absence of another auxiliary verb (modals, *have* with a perfect participle, *be*, and

variably with so-called pseudo-auxiliaries such as *need*, *dare*, and *ought*; for information on the status and history of the class of auxiliaries in English see Warner 1993), and is obligatory where it appears. These contexts are discussed in the following subsections.

### 2.1.1 Negation

Sentences where *not* is a sentence negator show *do*-support:

(1) John didn't finish his chores.

In addition to negated indicative sentences (and negated questions, which also are an instance of subject auxiliary inversion discussed immediately below), *do* appears in negative imperatives:

(2) Don't lie to me, John.

Auxiliaries other than *do* are restricted in their ability to appear in imperatives. Modals are banned. Perfective *have* and progressive and passive *be* are both permitted in the appropriate context. When these are negated, they appear with *do* (unlike in negated indicatives):

(3) Don't have eaten everything before the guests arrive. Potsdam 1995 (18c)

(4) Don't be fooled by his shoddy argumentation. Potsdam 1995 (16a)

(5) Don't be ringing the doorbell incessantly. compare Potsdam 1995 (17b)

These facts indicate that there is a potential discontinuity between *do*-support in indicative sentences and in imperatives.

When *not* is not a sentence negator, but rather constituent negation attached to the main verb, *do*-support cannot salvage the sentence from ungrammaticality. On the other hand, if such a sentence has an auxiliary verb (including a *do* triggered by a sentence-negator *not*), there is no obstacle to its grammaticality. For a discussion of these facts, consult Embick and Noyer (2001, sec. 7.2).

### 2.1.2 Emphasis

Affirmative sentences that are emphatic (i.e. have verum focus) have *do*-support. Among indicatives, there are several varieties of emphatic sentences. The first kind places a pitch accent (represented by capital letters) on the auxiliary:

(6) John DID finish his chores.

A second kind has the auxiliary unstressed, and places the pitch accent instead on an adverb *so*, which follows the auxiliary:

(7) John did SO finish his chores.

For innovative younger speakers (including the author of this dissertation), it is possible to have this *so* without an accompanying auxiliary. That is to say, *do*-support is optional for these speakers in this context:

(8) John SO died on level 3 of Super Mario.

(A verb like *die* which cannot be graded is used to distinguish another, more archaizing use of *so* in similar contexts to mean "to a great degree," as in "For God so loved the world that he gave his one and only son…".)

Imperatives can also be made emphatic by the addition of *do*:

(9) A: Don't sit down B: No, DO sit down.

There is an alternative, unstressed, kind of *do*-support which is associated with obsequious politeness. This is perhaps most easily illustrable precisely in the case of imperatives (here *do* may have a – probably secondary – pitch accent of some sort, but it is crucially not the same as the pitch accent assigned to it in the previous example):

(10) Do sit down.

It has also been discussed in the literature as a stereotypical feature of flight attendants' speech (Banks 1994; Schütze 2013). The grammatical character of this unstressed *do* is not clear, but I shall tentatively hypothesize that it is related in its function and history to the verum focus version of the auxiliary, and leave further exploration of the details to one side.

### 2.1.3 Subject-Auxiliary inversion

Sentences where the verb inverts with an auxiliary display *do*-support. The most common such environment is *wh*-questions:

(11) What did John say?

When the *wh*-word is the subject, it does not invert with an auxiliary, and thus does not trigger *do*-support:

(12) Who said that?

(13) Who will say that?

Polarity questions also involve *do*-support:

(14) Did John finish his chores?

Other constructions can also produce Subject-Aux inversion. Examples include negative inversion:

(15) Never have I seen such a wondrous sight.

Equative clauses introduced by *so* and *as*:

(16) John finished his chores and so did Mary.

(17) John finished his chores, as did Mary.

Inversions triggered by *so* and *such*:

(18) So lazy did John feel that he didn't finish his chores.

(19) Such a heavy burden did John feel his chores to be that he didn't finish them.

Exclamatives:

(20) What pleasure did I feel on seeing that John had finished his chores!

Other types of fronting phenomena:

(21) Only later did we realize that John hadn't finished his chores

(22) Thus did we realize that John hadn't finished his chores

While this is a wide variety of environments for inversion with *do*-support, it is worth noting that all cases of surface inversion do not lead to the emergence of *do*-support. The above-listed inversion environments are all analyzed (or analyzable) as involving the movement of the verb to a high position in the clause (C, for example). There is at least one environment which involves auxiliary movement to C in PDE, but does not license *do* – Conditional Inversion (CI):

(23) Had John not finished his chores, he would have been in trouble.

(24) *Did John not finish his chores, he would have been in trouble.

However, CI was possible with *did* until the 19th century (and perhaps even the 20th in highly literate writing; see Visser 1963, §1473 for further examples):

(25) and surely I could move, did I but will it 1886, Visser (1963)

Furthermore, there is at least one other inversion construction which does not involve movement to a high position and which also does not license *do*-support, namely inversion with *come* and *go*. These two verbs invert *here* or *there* with non-pronominal subjects:

(26) Here comes John.

However, pronouns fail to invert:

(27) Here he comes.

Also worth noting in this context is the argument by Culicover and Winkler (2008) that the (optional) surface Subject-Aux inversion in comparative clauses headed by *than* (as in the following sentence) is in fact not generated by movement to C:

(28) John finished more chores than did Mary (/ than Mary did).

### 2.1.4 VP topicalization

In English, it is possible to topicalize the VP by fronting it to a sentence-initial position. In these cases, *do*-support is generated:

(29) John wanted to see the Great Pyramid, and see it he did.

### 2.1.5 VP ellipsis

English also has *do*-support in VP ellipsis contexts:

(30) John finished his homework and Mary did too.

In this context, *do* is in complementary distribution with auxiliaries:

(31) John has finished his homework and Mary has too.

(32) *John has finished his homework and Mary does have too.

Compare:

(33) John will have finished his homework and Mary will ??(have) too.

The precise conditions under which VP ellipsis is licensed in English are subtle, and the body of research on the question is large – see van Craenenbroeck (to appear) for an overview. Nonetheless, the generalization that English uses *do* in VP ellipsis contexts whenever they occur and otherwise lack an auxiliary is robust. There is one context in particular that is worthy of special mention. It combines subject-auxiliary inversion, VP ellipsis, and the insertion of *so*:

(34) John finished his homework, and so did Mary.

This construction involves bona fide *do*-support, as illustrated by the alternation of *do* here with other auxiliaries:

(35) John has finished his homework, and so has Mary.

(36) *John has finished his homework, and so does Mary have.

(37) *John has finished his homework, and so has Mary done.

This usage contrasts with a similar, yet non-inverted *do so* construction. The *do* in this non-inverted *do so* fails to alternate with modals:

(38) John finished his homework, and Mary did so too.

(39) John has finished his homework, and Mary has done so too.

(40) *John has finished his homework, and Mary has so too.

It is also restricted to agentive verbs:

(41) John knows that the Earth is round, and Mary does (*so) too.

For these reasons, it cannot be treated as a support phenomenon.

### 2.1.6 Non-*do*-support *do*

In addition to 'do so', there are other uses of 'do' in PDE that are not connected to the phenomenon of *do*-support. For discussion of a *do* which follows a modal or other auxiliary in VP ellipsis contexts in British English, see section 2.2.2.

### 2.1.7 Structural remarks

The property that the *do*-support environments discussed above share is a disruption of the local relationship between T and V. In the case of negation and emphasis, a head intervenes between these two projections – Σ in the terminology of Laka (1990). In the subject-auxiliary inversion and VP topicalization cases, movement disrupts the relationship. In the former case, head movement is the relevant phenomenon, whereas in the latter it is XP movement. In VP ellipsis, the V head is deleted by the ellipsis operation.

These facts suggest several architectural conclusions. The fact that head movement can disrupt adjacency suggests that *do*-support is created relatively close to the surface on the widespread (though by no means universal) view that head movement is a PF phenomenon (Chomsky 2001). The fact that a topicalized VP cannot fulfill the adjacency requirement in its base position (in a way analogous to reconstruction for interpretation at LF) is another piece of evidence in this direction.[1] In lieu of presenting an overview here of the different analyses of these facts, the reader is referred to section 4.4, where the proposals are spelled out and evaluated in their coverage of both synchronic and diachronic facts.

Having established the contours of *do*-support in (standard American) PDE, the next section moves on to consider the distribution of phenomena similar to *do*-support in other languages and dialects, another important consideration for grounding the historical syntactic inquiry of this dissertation.

## 2.2 *Do*-support in other languages

It has been noted that phenomena resembling *do*-support may appear in languages other than English. In this section, several such languages are reviewed, with the goal of placing the English *do*-support phenomenon

[1] Of course, the mere description of the phenomenon as relating to "adjacency" presupposes a linear – and thus PF-oriented – view of the *do*-support phenomenon. But adjacency is not a logically necessary component of the description; some other (structural) rule could instead have governed the distribution of *do*-support.

in a crosslinguistic context.

### 2.2.1 Northern Italian

The case of *do*-support in certain Northern Italian (specifically Lombard) dialects is described by Benincà and Poletto (2004, hereinafter BP). In the specific dialect analyzed by the authors, finite verbs raise to T in all sentences, as in French and unlike in EME or PDE (cf. Pollock 1989). This is demonstrated by the finite verb's appearance to the left of an adverb like *semper* 'always':

(42) *l tʃàkola semper*
he speaks always
'He always speaks' BP (7a)

Thus, in declarative sentences (affirmative or negative), *do*-support is not implicated:

(43) *l tʃàkola mìa*
he speaks not
'He doesn't speak' BP (8)

On the other hand, in questions (both polarity and *wh*) *do*-support is in evidence:

(44) *fa -l majà?*
does he eat
'Does he eat?' BP (18a)

(45) *ke fe -t majà?*
what do you eat
'What do you eat?' BP (21a)

Just as in English, this *do*-support applies only in cases where there is no auxiliary verb,[2] and not to subject *wh*-questions. There are, however, some differences with English *do*-support. In addition to the previously-noted lack of *do*-support in negative declaratives, it fails to manifest in emphatic sentences and in VP-ellipsis contexts. Each of these differences can be explained by an independent property of Northern Italian syntax. In the case of emphasis, the verb raises past the polarity projection Σ which hosts negation and emphatic affirmative features, rendering *do*-support unnecessary (because no bound morphemes are stranded by the syntax). As for ellipsis, it is generally banned in Northern Italian.

[2]For the Monnese dialect studied by BP, the class of auxiliaries includes *have* and *be*, certain modal verbs (but not all such verbs), and optionally the verbs *go* and *do*. The optionality of *do*-support with the latter two verbs is described as intra-speaker free variation.

More interesting is the case of subject *wh*-questions. In transitive *wh*-subject questions for example, the question construction merges an overt complementizer and is parallel to the syntax of embedded questions:

(46) *ki ke maja*
who C eats
'Who eats?' BP (25a)

(47) *el so mìa kü ke à majà*
it I.know not who C has eaten
'I don't know who has eaten.' BP (15a)

Unaccusative subjects have two options: they can either trigger the subject-*wh* pattern with a matrix interrogative complementizer, or the object-*wh* strategy with *do*-support:

(48) *ki ke l va a ka?*
who C 3sg.subj goes to home
'Who goes home?' BP (31a)

(49) *fa -l nda a ka ki?*
does 3sg.subj go to home who
'Who goes home?' BP (31b)

This, BP argue, is related to the fact that unaccusative subjects may optionally remain in a low position.

The Northern Italian situation, then, has striking parallels to the English one. There is (distributionally) a morphosyntactic requirement that C$^0$ be filled by a finite verb in matrix questions, except in sentences with a *wh* subject in a high position (as all English subjects must be). A lexically restricted class of verbs can fulfill this requirement by moving to C, but most verbs cannot. Those verbs rely on the insertion of a semantically vacuous auxiliary in C to satisfy the morphological requirements on that head. On the other hand, Northern Italian has verb movement to T in all circumstances (e.g. across negation), so *do*-support is never demanded by the requirements of T. These facts support an analysis of *do*-support as a last resort phenomenon in Northern Italian, and also increase the confidence that an analogous analysis of PDE is on the right track.

### 2.2.2 Scandinavian

In Mainland Scandinavian, *do*-support appears in VP topicalization, ellipsis, and pronominalization contexts only. The verb that appears is not the Scandinavian cognate of *do*, but rather a form cognate with ME *gar*

(attested primarily in the northern ME area as an alternant with *do*, *make*, and *let* in the system of causatives). The following example from Houser et al. (2006) illustrates the topicalization phenomenon in Danish:

(50) *Jasper lovede at vaske bilen og vaske bilen gjorde han (så sandelig)*
J. promised to wash the.car and wash the.car did he so truly
'Jasper promised to wash the car, and wash the car he did (indeed).' (6)

In Danish as well as Norwegian, there is variation between constructions where the topicalized V is finite as above, and those where it is an infinitive (51a). Swedish, on the other hand, allows only finite verbs to topicalize (51b):

(51) a. *og kørde/køre bilen gjorde han*
and drove/drive car.DEF did he
b. *och körde/*köra bilen gjorde han*
and drove/drive car.DEF did he
c. and *drove/drive the car he did Platzack 2008 (5)

PDE allows only the non-finite construction (51c), however ME "occasionally" (in Visser's words) displayed the finite construction:

(52) And touchede þe chest þo he dude with his honde c1450, Visser (1963, §1423)

Danish, Norwegian, and PDE – but not Swedish – allow VP-ellipsis with *do*-support. This is illustrated for Danish here:

(53) *Mona vaskede ikke bilen men Jasper gjorde*
M. washed not the.car but J. did
'Mona didn't wash the car but Jasper did.' Danish, Houser et al. 2006 (7)

Danish, Norwegian, and Swedish – but not English – allow insertion of a *do*-verb when the VP is replaced by a pronoun:

(54) *Mona vaskede ikke bilen men det gjorde Jasper*
M. washed not the.car but it did J.
'Mona didn't wash the car but Jasper did so.' Danish, Houser et al. 2006 (8)

It is important to distinguish the pronominalization here from *do so* substitution in English. English *do so* can only substitute for verbs of a certain aspectual class, being incompatible with statives as discussed previously. In Mainland Scandinavian, however, *do*-support+pronoun is compatible with all verb types including statives:

(55) *Maria gillar inte fisk men det gör Johan.*
M. likes not fish but it does J.
'Maria doesn't like fish but John does.' Swedish, Platzack 2008 (10d)

Thus, we may regard this phenomenon as a genuine instance of a pleonastic support phenomenon, unlike the superficially similar English construction.

Platzack analyzes the *do*-support in these constructions as being realized in *v*, not in T as in standard analyses of *do*-support in English. His evidence for this claim comes from the following crucial sentence, where the *gør* is inserted below negation, which in turn is positioned below T:

(56) *Maria liker melk mens Johan ikke gør det*
M. likes milk but J. not does it
'Maria likes milk but Johan doesn't.' (11)

A word-for-word identical sentence can also be constructed for Swedish.

Platzack approaches the questions raised by the above data from a Distributed Morphology perspective, where a category-neutral root (or $\sqrt{}$) combines in the syntax with a category-determining head (such as *v* to form verbs) in order to yield what appears on the surface as an open-class lexical word. The features of category-defining heads are relatively fixed – that is, such heads are drawn from a small set (possibly of cardinality one) in a given language. On the other hand, $\sqrt{}$s are subject to idiosyncratic variation in their content in terms of syntactic (as well as phonological and semantic) features.

*Table 2.1:* A table listing the surface restrictions imposed by different regimes of $\sqrt{}$-insertion. Languages such as Danish and Norwegian which allow tensed and untensed roots to be inserted are maximally permissive, being capable of deriving all available options.

| $\sqrt{}$ has Tense? | Language | VP ellipsis | VP topicalization | VP pronominalization |
|---|---|---|---|---|
| yes | Swedish | no | tensed | yes |
| no | English | yes | untensed | no |
| yes/no | Norwegian, Danish | yes | variable | yes |

Platzack argues that the microvariation in these facts is explained by a parameter controlling the insertion of $\sqrt{}$ with or without a Tense feature from the lexicon. The possibilities are laid out in Table 2.1. The presence or absence of a Tense feature on the root has the most straightforward effect on VP topicalization: if it is present, a topicalized VP will bear tense morphology; if not, it will surface as an infinitive. The ellipsis and

pronominalization cases are handled confusingly by Platzack's analysis. The core assumption is that in either case the root cannot move to *v*, since that position is occupied by *gøre* (etc.). The uTense feature on *v* is the beneficiary of checking by T, but in languages with tensed $\sqrt{}$ such as Swedish there is an extra uninterpretable tense feature on the root, present in the lexicon, which has no way to be checked. VP pronominalization is licit (and in fact required) after *göra* because it removes the root from the tree along with its extra tense feature, allowing the derivation to proceed. There is no satisfactory explanation, however, of why the lack of VP pronominalization in English follows from the lack of tense features on roots. Platzack merely avers that "it should be no surprise that comparing the structures [without tense on the root] and [with it], it is the more self-contained one, [the latter], that is the basis for pronominalization, whereas the more dependent one, [the former], is the basis for ellipsis." (18) Platzack's account thus falls short of successfully capturing in a coherent formalism all the properties of the data presented above. However, it does seem to be on the right track towards identifying the microvariation in the Mainland Scandinavian(+English) dialect continuum. I take Platzack's conclusions to mean that English shares with Scandinavian a kind of *do*-support which is unconnected in its synchronic syntax with the kind that is exclusive to English (i.e. that appearing in negative declaratives, questions, and emphatic sentences). The latter type of *do*-support is commonly analyzed as being associated with the position T; it is noteworthy that Platzack's analysis also is powered by tense features even though the action takes place in a different structural position. The proposal in section 4.3 about the diachrony of *do*-support in English precisely postulates that in early EME *do* is merged in *v*; this provides a diachronic link between the different kinds of *do*-support in English.[3]

**British English ellipsis *do***

There is a kind of *do*-insertion which occurs in British English and shares some superficial similarities with the phenomena discussed in the above section. It is illustrated here:

(57) The Eagles won't win the Super Bowl, but the Giants may do.

I will argue, however, that this phenomenon is not connected to the above paradigms. Rather, it is an alternative to absolute elimination for the realization of elided VPs. The evidence for this claim is the observation that what I will call the British ellipsis *do* is not in complementary distribution with auxiliaries, but in fact cooccurs with them. Sentence (57) above is one example of this fact with a modal auxiliary. Examples with *have* and auxiliary *do* follow:

[3]It also could be used to generate predictions about the spread of *do*-support through the language. The simplest of these proposes that *do* evolves from a causative to an external argument marker, and then bifurcates into T-*do*-support and *v*-*do*-support; due to lack of a solid dataset on the diachrony of *do*-support with VP ellipsis in English I won't explore this prediction further in the present dissertation.

(58) I didn't know the answer, but John may do. Pullum and Wilson 1977, (41b)

(59) John didn't go to the ball game, but Peter did do.

This contrasts with the *do* of VP topicalization which is in complementary distribution with auxiliaries in English – including British English, where the following judgments come from:[4]

(60) Listen though he may (*do), he won't be able to hear a thing.

Swedish minimally contrasts with British English in this environment:[5]

(61) *Maria lovade att köra bilen och köra bilen ska hon göra*
M. promised to drive the.car and drive the.car will she do
'Maria promised to drive the car, and drive the car she will do.'

### 2.2.3 Old Icelandic

In Old Icelandic, a construction similar to *do*-support appears briefly and never categorically.[6] This construction using the Icelandic synonym *gera* of the word *do* appears in very early (9th century) poetic texts, and in prose beginning in the 12th century. In early Skaldic poetry, *gera* appears freely, apparently in a VP topicalization construction. In the Eddic poetry (more vernacular in its syntax than Skaldic) and the earliest prose texts, *gera* always appears with negation:

(62) *ef hann görr eigi segia*
if he does not say
'If he doesn't say/tell.' Viðarsson 2009 (246a)

Later, it appears in contexts without negation as well, though apparently always in the presence of a contrastive interpretation of some sort (for this reason, Viðarsson (2009) analyzes the construction in terms of focus):

(63) *Klarus prestr gerdi þvi alldregi trua*
K. priest did it never believe
'(context: others believed but) Klarus the priest never believed it.' Viðarsson 2009 (261d)

This later usage also becomes less restricted in other ways: it can appear embedded under a modal (64) and with an inflected infinitive (65, i.e. with preceding *at* corresponding very roughly to *to* in English.)

[4] Thanks to Amy Goodwin Davies for this judgment, which she also confirmed with 2 other speakers of British English.

[5] Thanks to Kajsa Djärv for the judgment of this sentence, which she reported to be "grammatical but perhaps a bit awkward/clunky."

[6] The entirety of the discussion in this section is heavily based on information from Heimir van der Feest Viðarsson, for which I am very grateful. The section is partially based on Viðarsson (2009), and partially on personal communications. Any errors or mistakes are of course attributable only to me.

(64) *at hann munnde æi lægia gera*
C he would not make.laugh do
'(It was said about this man that there was no one) that he could not make laugh.'
Viðarsson 2009 (257)

(65) *þa gerdi hann eigi at ganga i tidagerd þeira*
then did he not to go in divine.service their
'Then he didn't go into their divine service.' Viðarsson 2009 (262)

At this time, Icelandic was undergoing a change in its negator, from a bound morpheme *-a* or *-at* which attached to a finite verb in C to a free (i.e. not bound) adverbial *ekki* (the latter system continues in present-day Icelandic). Viðarsson observes that suffixal negation survives longer with modals (and a certain subclass of lexical verbs) at a time when *eigi* (often but not always in combination with *gera*) is the negator for most lexical verbs. This suggests an account of the emergence of *gera*-support in Icelandic that assigns it initially the role of a last-resort element which supports lexical verbs when they cooccur with negation; modals are freer in their syntactic appearance as in English and do not demand a support auxiliary.[7] The system then quickly dissolves over the course of the following century, as *eigi* (and later *ekki*) become the default negators and any requirement for a support auxiliary is lost. In turn *gera* is reanalyzed as a marker of information structure before disappearing from the language entirely (in this role; it survives as a lexical verb and in VP pronominalization constructions). This development (insofar as we see it happen at all) is in a sense the inverse of what happens in English. Instead of a verb which takes verbal complements (i.e. *do*) gradually acquiring a more abstract grammatical character until it becomes fully semantically bleached and fully predictable in its distribution as an auxiliary, in Icelandic the most abstract and constrained distribution is the first observable, which subsequently erodes. (Unlike with *do* in English, *gera* appears to basically lack a causative use with an infinitival complement in Old Icelandic, outside of a few examples in direct translations from Latin. On the other hand, *gera* can freely take nominal, adjectival, and past participial complements.)

It is clear that the precise mechanics of the evolution of *gera* in Icelandic are not fully captured by this description. I mention the Icelandic case, albeit in this incomplete way, for two reasons: to point it out as the only Germanic parallel to the last resort *do* of PDE and to highlight the apparently quite pronounced

[7]The precise characterization of this last-resort type operation is somewhat mysterious, since lexical verbs in affirmatives are relatively unconstrained in their movement possibilities. With more analysis it may be possible to characterize the nature of this system, although the possibility is curtailed by the fact that the number of attested tokens is quite limited and that *gera*-support never reaches obligatoriness in the language.

differences in the diachronic trajectory of *do*-support through the grammar of the two languages. In English it spread from a causative to a support auxiliary, whereas early Icelandic shows exactly the opposite evolution.

As a final note, present-day Icelandic shares with its Scandinavian relatives the VP pronominalization construction in Section 2.2.2 – but allows neither VP topicalization nor ellipsis (Platzack 2008). The diachrony of these constructions in Icelandic has not been, to my knowledge, explored in the literature.

### 2.2.4 Korean

As reported by Hagstrom (1995), Korean possesses a kind of *do*-support which emerges in two contexts. The first of these is with negation. There are two possible types of negation in Korean: the so-called "short" and "long" negation. In short negation, a negative prefix attaches to the verb directly, which acquires tense and clause-typing suffixes and appears in the clause-final position:

(66) *Chelswu-ka* *chayk-ul* *an-ilk-ess-ta*
Chelswu-NOM book-ACC NEG-read-PAST-DECL
'Chelswu did not read the book.' Hagstrom (1995) (11a)

In long-form negation, on the other hand, the main verb appears with a participial ending,[8] followed by a free morpheme expressing negation and the verb *ha-* 'do' inflected for tense and clause type:

(67) *Chelswu-ka* *chayk-ul* *ilk-ci* *ani* *ha-ess-ta*
Chelswu-NOM book-ACC read-PRT NEG do-PAST-DECL
'Chelswu did not read the book.' (11b)

Thus, in Korean the verb cannot raise past *ani*, requiring the insertion of *ha*. This is exactly analogous to the situation in PDE with *not*.

The second environment for *do*-support in Korean involves a special information status for the verb:

(68) *Chelswu-ka* *chayk-ul* *ilk-ki-nun* *ha-ess-ta*
Chelswu-NOM book-ACC read-PRT-TOP do-PAST-DECL
'It's read the book that Chelswu does.' Hagstrom 1995 (35a)

The role of the verb in this construction is variously described as focus, topic, and contrastive topic; I will use the term "topic" in keeping with the most common nomenclature for the *nun* morpheme implicated, even among authors who refer to the whole construction as a focus construction. The topicalization can apply to any of several constituents: V, *v*, and T. (Aoyagi 2006) The cases of topicalization of V and *v* are

[8] Hagstrom (1995) remains non-committal about the nature of this morpheme, spelled *-ci*. The participial analysis comes from Yi (1994)

surface-identical, and yield structures like in (68) above. Topicalization of T yields two additional possibilities, doubling of the tense morphology on the topicalized verb, and displacement of the tense morphology from *ha* to the lexical verb:

(69) … *mek-ess-ki-nun* *ha-ess-ta*
eat-PAST-PRT-TOP do-PAST-DECL
'… ate.' Aoyagi 2006 (31d)

(70) … *mek-ess-ki-nun* *ha-ta*
eat-PAST-PRT-TOP do-DECL
'… ate.' Aoyagi 2006 (31f)

It is also possible to double the topic-marked verb rather than replacing it with *ha*:

(71) *Chelswu-ka* *chayk-ul* *ilk-ki-nun* *ilk-ess-ta*
Chelswu-NOM book-ACC read-PRT-TOPIC read-PAST-DECL
'It's read the book that Chelswu does.' Hagstrom 1995 (35b)

Korean, then, has the ability to insert *ha* at several different levels of the structure in order to provide a host for the clause-typing complementizer *-ta*, which cannot be a verb which is topic-marked by *-nun*.

### 2.2.5 Summary

*Table 2.2:* A summary of various languages with *do*-support phenomena, listing the context(s) where it occurs in each. For Old Icelandic, the availability of topicalization, ellipsis, and pronominalization is not precisely known; values in parentheses reflect the behavior of the modern language.

| **Language** | neg. decl. | ?s | VP topic | VP ellipsis | VP pron |
|---|---|---|---|---|---|
| English | + | + | + | + | - |
| Danish/Norwegian | - | - | + | + | + |
| Swedish | - | - | + | - | + |
| Monnese | - | + | - | - | - |
| Korean | + | - | + | - | - |
| Old Icelandic | + | - | (-) | (-) | (+) |

In this section, I have discussed the distribution of *do*-support-like phenomena in languages other than English. English exhibits *do*-support in a wide variety of morphosyntactic environments. Though no other

language is known to exhibit the same distribution of *do*-support phenomena as English, each of the specific contexts for *do*-support in English finds an echo in some other language. Table 2.2 summarizes the findings. Further discussion of these contexts can be found in Jäger (2006), who provides in his Tables 2 (p. 132) and 3 (p. 158) a list of up to 16 languages which show *do*-support in negative sentences, 21 which show it in questions (of some type or another), and 10 which have it in topic or focus constructions of some variety.

These findings demonstrate that *do*-support phenomena appear in a variety of languages and syntactic contexts. English happens to have *do*-support in most (but not all) of these contexts; however the existence of languages with *do*-support in fewer places confirms that the English phenomenon need not be treated as monolithic, but can rather be treated as the end result of several independent activations of *do*-support. (The intended sense of "independent" here refers to synchronic facts; of course any particular language's choices about where to use *do*-support is the result of contingent, and mutually dependent, facts about the language's history.) Having discussed the distribution of *do*-support in PDE and other languages, I'll now move on to a crosslinguistic discussion of non-*do*-support contexts in which *do* appears as a functional element.

## 2.3 *Do* as a non-vacuous light verb

A variety of human languages show phenomena where a synonym of *do* is used as a light or auxiliary verb, but which nonetheless are differentiated from *do*-support phenomena in that *do* in these constructions has some semantic value. The range and frequency of these constructions, especially those where *do*'s semantics are associated with causativity, agentivity, or other properties of the external argument, is important for understanding the stepwise development of *do* in the history of English, discussed in section 4.3. In this section, I provide an overview of some representatives of such constructions drawn from a variety of languages.

One common argument-structure related use for a *do* light verb involves the adaptation of loan words. Several languages show a pattern whereby borrowed verbs cannot participate in native inflectional processes, but rather are combined with a native light verb. The example below comes from Konkani (Indo-Aryan):

(72) *kazar -karuk*
marry make
'to marry someone'

(73) *kazar -zavunk*

marry be

'to get married' (Wohlgemuth 2009, p. 253)

Other languages which are claimed to have a similar system are Tamil (Dravidian) and Warlpiri (Pama-Nyungan) (Wohlgemuth 2009). Jäger (2006, p. 165) provides the additional example of Rama (Chibchan-Paezan) as a member of this class of languages. Jäger (2006, p. 171), following Alekseev (1994a,b), also points out languages which extend periphrasis to native vocabulary as well: Rutul and Budukh (Caucasian) use an auxiliary verb in every sentence obligatorily; for transitives the auxiliary is a synonym of *do* whereas intransitives take *be*.

Other languages have productive inflectional paradigms for lexical verbs, but use *do* periphrasis in a transitivizing role. Jäger (2006, sec. 5.3.3) provides an overview of many such languages. One example taken from that text are Apinajé (Ge-Kaingang), which has a productive transitivizer derived from *do* (*ɔ* in the below examples):

| intransitive meaning | intransitive form | transitive meaning | transitive form |
|---|---|---|---|
| drink | itkō | drink (something) | tɔitkō |
| go | tē | take | tɔtē |
| end | apeč | finish | tɔapeč |
| be full | dət | fill | ɔdət |
| be pretty | bɛči | beautify | ɔbɛči |

A second example comes from Lahu (Tibeto-Burman). In this language, *te* 'do' can have a causativizing function. However, it can also function as an optional transitivizer, as the following pair of synonymous sentences illustrates:[9]

(74) *yɔ̂ thà? dê chɛ̂ ve*

3SG ACC scold someone ?

'They're scolding him.'

(75) *yɔ̂ thà? dê te chɛ̂ ve*

3SG ACC scold do someone ?

'They're scolding him' / '*They're making him scold someone' Matisoff (1973, p. 246)

[9]The ? in the glosses does not correspond to a question marker; it is reported by Jäger (2006, p. 255) as an apparent indication of his uncertainty in interpreting the original unglossed examples.

Another similar case not treated by Jäger (2006) comes from the Austronesian languages of Northern New Caledonia. Several languages from this area (Nemi, Nêlêmwa, Nyelâyu) have grammaticized a verb meaning "take" to be a valency-increasing marker. In the former two languages, the marker does not introduce affected patients, but rather comitatives of a sort ("be lying down" > "sleep with"; "run" > "run away with"). However, in Nyelâyu this marker has been incorporated into a system of transitivity marking suffixes; specifically in causatives the system of allomorphic transitive markers is being replaced with a non-varying reflex of "take" (Ozanne-Rivierre 2004).

In conclusion, it is evident that in many languages *do* has a semantic and grammatical connection to notions of agentivity, transitivity, and causation. This is utterly unsurprising, given the core semantics of the verb *do*. However, it will come to play an important role in my analysis of EME. In the next section, I'll turn to uses of *do* in West Germanic. These are much less transparently relatable to the semantics of *do* generally, but their genetic proximity to PDE lends them an analytic importance nonetheless.

## 2.4 Synchronic variation in Germanic *do*

In various Germanic dialects – usually non-standard ones – it has been noticed that a cognate and/or synonym of *do* plays the role of a semantically-conditioned helping verb. For instance, in the English of Southwestern England, *do* is used freely (but not categorically) in non-emphatic affirmative declaratives. The precise semantic value of this *do* is widely commented on, but not entirely understood. Klemola (1998) surveys previous proposals that *do* is a habitual aspect marker, and concludes that whereas there is a tendency for *do* to be used in habitual (as opposed to punctual) contexts, this tendency is not categorical. There, evidence is also adduced from various corpus studies (the Survey of English Dialects and the Somerset Rural Life Museum recordings) which demonstrates that this *do* is used with predicates that resist *do* in EME (such as *come* and *look* in the sense of *seem*), indicating that the *do* of these dialects is unlikely to continue EME usage unchanged.[10] Cornips 1998 examines (nonstandard) *doen* usage in a regional Dutch dialect with L2 influences. The corpus of examples studied is small, but Cornips reports both an effect of agentivity (*doen* occurs only with agentive verbs, not unaccusatives or experiencer-subject ones) and of aspect (*doen* marks the habitual aspect). Usage of *do* (or cognates) as an auxiliary verb in nonempahtic affirmative contexts by child learners of English and various (child and adult) dialects of German has been argued not to show any categorical effects of argument structure or aspect on its distribution (Schütze 2004).

[10] Although we do see some few tokens of *do* with even the most resistant verbs in EME; a large-scale corpus study of the southwest English dialects would be needed to conclusively determine whether the modern-day usage patterns match those of earlier language stages.

There is little clarity in the literature on how the various types of Germanic *do* support may or may not reflect the constraints on *do* usage in EME. What is clear is that *do* is a very attractive target in West Germanic for generalization into a grammatical marker; an experiment is evidently run each time a language learner acquires one of these languages. The experiments often come to naught as in the many documented cases of learners who generalize an auxiliary-like *do* early in childhood but quickly lose it (reviewed by Schütze 2004). In other instances the innovation is successful enough to make an impact on a (geographically and/or socially) circumscribed dialect, as in the case of Southwest England English and Dutch. PDE represents the most large-scale success of this tendency to make *do* into a grammatical morpheme, in the sense that it has become an unremarkable feature of the standard language. (For another instance in which language acquisition can be treated continuing experimentation by individual learners with eventual results on the language of the community at large, see Johnson 2010.)

## 2.5 Conclusion

This section began by demonstrating the variety of contexts in which *do*-support may appear in PDE. These contexts of occurrence were then seen not to be the result of a single grammatical parameter (in the sense of a minimal point of crosslinguistic difference), but rather to simultaneously instantiate many discrete environments which are shared, in a piecemeal way, with a variety of other typologically distinct languages. In light of this evidence, it makes sense to disaggregate the various contexts for a diachronic study. Platzack (2008) provides a structural argument for considering VP topicalization, ellipsis, and focus cases separately from negatives and subject-auxiliary inversion. Traditionally, quantitative analysis of English *do*-support has taken the same tack (Ellegård 1953; Kroch 1989). For these reasons (and also because assembling a sufficient corpus of examples of rare VP topicalization constructions is difficult), I will not address the diachrony of VP topicalization or ellipsis in this dissertation, focusing instead in the first instance on negative imperatives and questions (both affirmative and negative; taken to be exemplars of the subject-verb inversion environment).

After disposing of that question, I then went on to consider the usage of *do* (and its cognates or synonyms in other languages) in non-*do*-support contexts. This provides two crucial insights for a diachronic analysis. The first is that *do* in a variety of languages can serve as a marker of argument structure, specifically of transitivity, agentivity, and/or causativity. Second, it is seen that a sort of *do* different from that found in PDE *do*-support recurrently impinges on a variety of West Germanic dialects. Both these observations play a key role in motivating my decision to undertake an analysis of the *do* which appears in EME affirmative

declarative sentences, seeing it as a crucial prerequisite to understanding the evolution of *do*-support as a whole. This theme is taken up in section 4.3 below.

Negative imperatives too show some synchronic oddities with respect to their relationship to the core system of *do*-support – such as the failure of full complementarity with some auxiliaries noted above. They are included in the investigation as another diachronic development potentially distinct from the other environments (and also because they have been traditionally analyzed alongside other type of *do*-support).

Having established these synchronic bases, I now turn to a discussion of the statistical methods employed in this dissertation.

# Chapter 3

# Statistical methods

The goal of this chapter is to give the reader an orientation to the statistical and computational issues which underlie the work in this dissertation.

There is an indubitable connection between statistical procedures and corpus methods as applied to questions of historical syntax (and indeed in general). Having extracted a dataset from a corpus (often, counts of construction (non)occurrences in a variety of different linguistic contexts), a researcher then desires a measurement of how meaningful these results are, and what, if any, conclusions may be drawn about the linguistic faculties of the speakers who generated the data. Statistical methodologies are a natural fit for this mode of inquiry. In this chapter I will describe the prevalent modes of statistical inquiry in historical syntactic research and their implementation in the present dissertation. Specifically, in section 3.1 I will discuss the CRH, the dominant mathematical model of the spread of syntactic changes through the population. In section 3.2 I will delve into further detail on logistic regression, the statistical procedure which underpins the CRH model. Section 3.3 addresses additional complications which arise when computing regression models, and describes the strategies this dissertation employs for addressing them. Finally, section 3.4 discusses a crucial issue in the interpretation of the statistical results in CRH analyses. The question of choosing a best model turns out to be crucial to the CRH.

## 3.1 The Constant Rate Hypothesis

Kroch (1989) formulated the Constant Rate Hypothesis (CRH; sometimes known as the Constant Rate Effect or CRE):

(76) **Constant Rate Hypothesis**: changes spread at the same rate in all contexts.

This hypothesis is important because "on the basis of [it] substantial progress can be made in understanding the relationship between the structural patterns uncovered by grammatical analysis and the frequency patterns revealed by sociolinguistic methods." That is, it is a useful tool for the study of language variation and change from a generative standpoint, since it links the domains of frequency and grammar, allowing corpus data to inform theoretical proposals and vice versa.

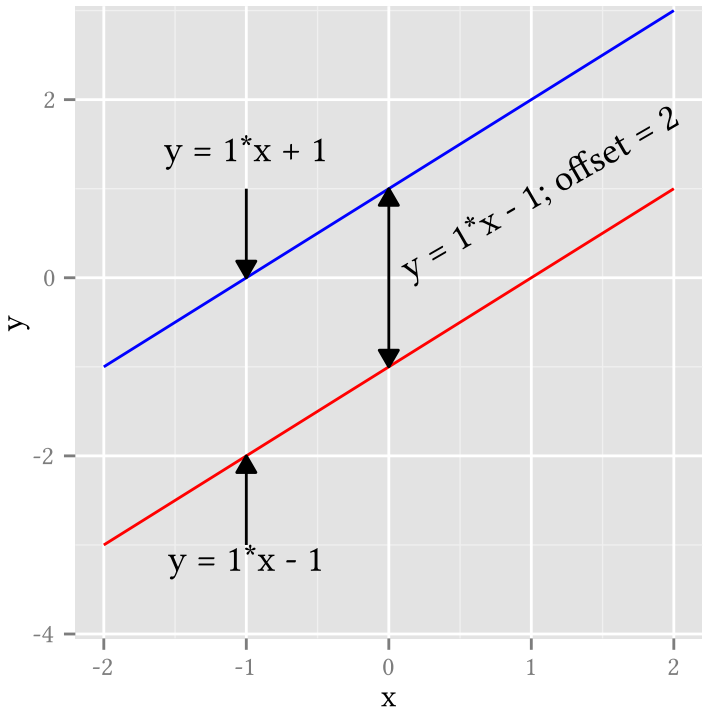



*Figure 3.1:* An illustration of two possible parameterizations of parallel lines. The left-hand system considers the lines as separate objects, assigning to each a slope and an intercept (for a total of 4 parameters). The right-hand system uses only three parameters: a slope and an intercept to describe one line, and an offset to measure the distance between the lines.

What motivates the posing of this hypothesis? It is fundamentally a parsimony argument. As illustrated in Figure 3.1, there are two possible mathematical descriptions of a system of two parallel lines. If we consider the two lines to be independent of each other, we must describe each fully, specifying its slope and intercept. On the other hand, if the two lines are taken to be part of a single system, it is necessary to specify the slope only once; the family of lines is then fully described by giving one intercept and the distance between the two lines. If we concretize by taking these lines to represent time courses of a change in two contexts, the first analysis is tantamount to proposing that there are two processes of change – one per context. These

two processes proceed at the same rate merely accidentally; since both slopes are specified in the model, there is no impediment to the lines being (or becoming) non-parallel. The second analysis amounts to a claim that there is only one thing (an abstract grammatical parameter) that is changing, corresponding to the single slope parameter. The intercept parameters are a measurement of context-specific effects which favor or disfavor the manifestation of the parameter change. The CRH counsels us to accept the former rather than the latter hypothesis because it uses fewer parameters to explain the phenomena.

Of course, the parsimony gain when there are just two lines is minimal. More convincing cases are adduced by Kroch (1989) – both the French V2 case compares 4 lines, and the English *do*-support case has 6. However, each of these scientific comparisons is not of equal importance. In each case, there is a large family of very surface-similar contexts (V2 with different subject types, or *do*-support with different clause types), which is compared with another, more distinct surface pattern (left-dislocation and verb movement to the left of *never*). While it is not a trivial discovery that V2 and *do*-support each evolve in parallel in various contexts, there is not very much at stake. Any theory of grammar which recognizes the identity of these syntactic constructions will be able to capture their diachronic unity. On the other hand, the out-group comparisons provide evidence bearing on questions of true grammatical abstraction – the effect of a prosodic change on syntax in the case of French and the existence of an abstract verb-raising parameter in English. Thus, we ought to concentrate most of our attention on these comparisons, and take the in-group comparisons to be less important. One extreme method of implementing such a scheme would be to in fact assume that the in-group contexts evolve in parallel, and calculate a pooled slope estimate from them for comparison with the out-group slope. This may or may not be satisfactory, especially in the context of analyses where (non)-parallelism is not immediately visually apparent (because of noise, data sparsity, or an abstract analysis which derives a slope estimate indirectly from observed data). A statistical analysis using the concept of *shrinkage*, where the in-group slopes can differ in the presence of especially compelling evidence but otherwise are constrained to be zero, may provide a sensible middle ground (with the disadvantage that it is not straightforward to implement; though Bayesian approaches which involve specifying a detailed implementation of the model should be able to cope at the cost of increased conceptual and implementational complexity).

Another factor which would help give the parsimony arguments more weight is the presence of more slope parameters which are collapsible across contexts, as would be the case if the diachronic trajectory is described not by a straight line, but rather by a higher-order polynomial or a spline function. Many syntactic changes unfold along an uninterrupted S-curve (which is equivalent to a straight line under the

transformations used in the statistical procedure underlying the CRH; see the following section for details). This is why the hypothesis bears the name *Constant* Rate Hypothesis, rather than *Equal* Rate Hypothesis (where the latter is a closer description of the hypothesis's actual claim). In fact, this circumstance presents a significant challenge to CRH analyses. The finding of a different slope conclusively demonstrates that two contexts are not related by a single underlying change, but the finding of equality does not guarantee that the changes are generated by the same change.[1] In practice, the slope values for syntactic changes which have been observed and described in the literature is constrained to a narrow range. On the short end, a change cannot take place more rapidly than one generation (and is likely to subsist in writing somewhat longer). Conversely, though some theories of change predict that very long-term syntactic changes are in principle possible, and putative evidence of such changes exists (Wallenberg 2013), it is difficult to distinguish such data from random drift (and indeed, fixation of former loci of variation by random drift over long time spans is predicted; see Kimura 1983).

In any case, the syntactic changes actually observed and described to date take place over roughly 100–300 years, a fact which necessarily restricts the observed slope values of these changes to those characteristic of S-curves with such lengths. The risk of two lines having statistically indistinguishable slopes by accident under these circumstances is thus heightened. The only antidote to this problem is to collect more data, a task which is addressed in chapter 5. Larger datasets will allow more precise measurements to be made of contextual slopes, and thus more precise comparisons to be made between them.[2]

## 3.2 Logistic regression

The specific statistical procedure used for testing CRH analyses is logistic regression. This procedure is appropriate for predicting binary observations (those what take on one of two discrete values, conventionally represented as 0 and 1). It involves the following mathematical model:

$$\ln \frac{p_i}{1 - p_i} = \beta_0 + \beta_1 x_{1,i} + \ldots + \beta_m x_{m,i}$$

Where $p_i$ is the estimated probability that the $i$th data point will have the value 1, and $x_{m,i}$ is one of $m$ numerical observations about the data point.

[1] This is a separate issue from the fact that standard statistical tests provide positive results of difference, whereas the case of identity is not distinguished in the test results from a lack of sufficient data. The problem described here would exist even if an oracle could tell us for any pair of curves that they were identical or differing in slope.

[2] However, as discussed below, the availability of larger datasets also lends greater power to detect extra-syntactic effects of various sorts. Thus the interplay of the various types of effect must be considered.

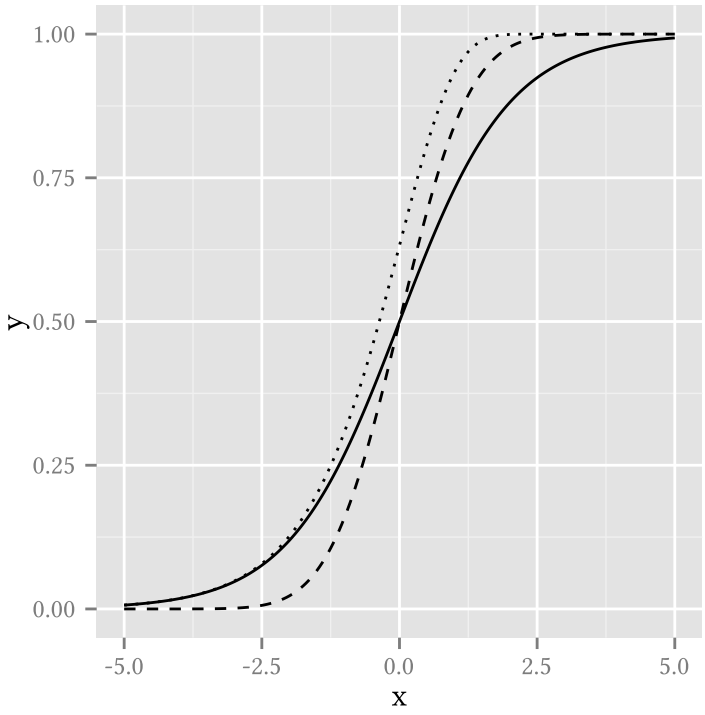



*Figure 3.2:* The logistic (solid) probit (dashed), and complementary log-log (dotted) link functions.

The $\beta$s are adjusted to maximize the correspondence between the model's output and the observed data. The function $\ln \frac{p_i}{1-p_i}$ is called the logit function. It sketches a particular S-shaped curve, which is illustrated by the solid line in figure 3.2. It is possible to see this curve as merely a convenient mathematical tool for mapping from real numbers to probabilities (which additionally are neither zero nor one). Under this interpretation, the logit alternates with other symmetric S-curve functions like the probit, which is based on the normal distribution. For many problems either works approximately as well as the other, especially if a higher-order polynomial is fit to the time variable (allowing the slope to vary at different points during the change, effectively liberating the functions from adhering to an S-shape at all). In many cases the logit model is preferred because its $\beta$s can be interpreted in terms of log-odds. In the case of syntactic change however, there is another reason to select the logit function as the link function, and to constrain the regression to

linear time functions.[3] The logit function is the solution to the following differential equation:

$$s' = s(1 - s)$$

This equation says that the proportion $s$ of some variant changes at a rate directly proportional to its prevalence in the population, and also directly proportional to the prevalence of the opposite variant. In other words, the spread is facilitated by interactions between individuals exhibiting opposite values of the varying trait. In population biology, this dynamic is familiar from cases of competition between two variants where one is strictly fitter than the other: the fitter variant replaces the less fit one. For the linguistic case, Yang (2000) specified a notion of fitness for grammars, and showed that a straightforward model of grammatical learning produces a logistic evolution when two differentially fit grammars compete.

There are also asymmetric S-like curves such as the complementary log-log function (also illustrated in figure 3.2), which rises gradually from 0 but approaches 1 abruptly. In some sense a vertically mirrored (around the line $x = 0.5$) version of this link function appears to be a better fit to written data, where apparently archaic examples may persist for some time. However, it is not clear that this property should be modeled as part of the grammatical change, or a social fact about language usage. The complementary log-log model has a solid interpretation in terms of hazard ratios, but the applicability of this interpretation to historical change is not obvious. (It may provide a plausible model of speakers adopting the change, especially later in life, to the extent that this is a one-time irreversible event similar to mechanical component failure or infection with a disease. It's far from clear that there is enough historical data from single speaker's lifespans to drive such a model however.) In any case, for the remainder of this dissertation I will assume a logit model without further discussion.

## 3.3 Further topics in regression

In this section, I'll discuss several issues which arise when applying logistic (or indeed any) regression technique to data on linguistic changes.

[3]The question of the degree of the time polynomial is different from (my interpretation of) the Constant Rate Hypothesis, though the two questions are conflated in the treatment by Kroch (1989). The CRH says that there is only one time polynomial underlying a syntactic change, whatever its degree. It is no contradiction to the CRH for there to be a change with underlying (for instance) cubic form, which meanders up and down probability space before going to completion – as long as this meandering ramifies equally in all contexts in which the change is observed. This is not predicted by the change-as-biased-learning model discussed immediately following this footnote, but this is a separate question. The combination of these two models means, in practice, that we also do not expect to observe parallel higher-order time functions: the change-as-learning model tells us that observed non-linearities are caused by secondary factors for which there is no general assumption of cross-context consistency akin to the CRH. The framework of this dissertation assumes both the CRH and the change-as-learning model, but either of these pieces could in principle be disentangled from the other.

### 3.3.1 Standardization of variables

Variables which are given on a continuous scale must be standardized before being used in regressions. The most common, and often only, such variable in a logistic regression will be the time variable (years); other options might crop up from time to time (such as (log) word frequency or the length of some constituent in a sentence). Because of the assumed normal distribution underlying many regression techniques, including logistic regression, the best kind of normalization is the $z$-score, which is computed as:

$$\text{zscore}(x_i) = \frac{x_i - \bar{x}}{\sigma_x}$$

Failing to perform this normalization can affect the output of a regression in (at least) two ways. If a variable is not centered, its associated terms in the regression model may become very large or small, and the computer implementation of the mathematical algorithm may not cope well with these extremes in the data, leading to errors. Secondly, the lack of normalization may affect the significance of certain individual terms in the model (though it should not, in general, effect the comparison of one model to another as discussed in section 3.3.3, assuming the model is fit without numerical problems).

*Table 3.1:* A comparison of two logistic regression models on *do*-support data from the PPCHE. One model standardizes the continuous time predictor, whereas the other does not.

| | Unstandardized | | Standardized | |
|---|---|---|---|---|
| | Estimate | $p$-value | Estimate | $p$-value |
| (Intercept) | −46.13 | 0.00 | −1.56 | $4.75 \cdot 10^{-12}$ |
| year | 0.03 | 0.00 | 1.05 | 0.00 |
| TypeNeg. Decl. | 5.20 | 0.72 | −0.88 | 0.00 |
| TypeNeg. Q. | −31.85 | 0.23 | 0.84 | 0.03 |
| year:TypeNeg. Decl. | 0.00 | 0.67 | −0.14 | 0.67 |
| year:TypeNeg. Q. | 0.02 | 0.21 | 0.77 | 0.21 |

Table 3.1 illustrates the output of two logistic regression models fit to data from the PPCHE. The data underlying the models is identical, but in the left-hand model time of authorship is represented by raw values (between 1410 and 1575), whereas the right-hand one uses the $z$-score of this variable. In the right-hand model the coefficient estimates are all of the same order of magnitude, which makes the fit more stable (a change of a given numerical quantity has roughly the same effect on any of the variables, meaning the space

which the optimizer has to explore is roughly an N-dimensional hypersphere rather than being stretched or squashed in any dimension). The $p$-values also differ across the two models, including some differences across the $\alpha = 0.05$ significance threshold.[4]

### 3.3.2 Coding of categorical variables

The formulas for logistic regression (and regression more broadly) require summing a vector of predictors weighted by regression coefficients. However, it is not possible to add together "negative declarative clause type" and "unaccusative verb." Thus, it is necessary to transform these predictors into numerical form. This is referred to as defining a contrast matrix for the categorical (i.e. non-numeric) predictor.

The conceptually simplest way to do this is, for a categorical variable with N categories $X_1, \ldots X_N$, to define N indicator variables $I_1 \ldots I_N$. $I_1$ takes the value 1 if an observation has the category $X_1$ and zero otherwise; the other indicator variables behave in a likewise fashion. However, including all N indicator variables in the model leads to problems with fitting the model. Regression models generally have an intercept term – a $\beta_0$ which is included for each observation, regardless of its properties. Since the N indicator variables completely partition the data, it is possible to balance arbitrary changes in the corresponding $\beta$s with equal and opposite changes to $\beta_0$. The model fitting algorithm will not converge, since there is an infinite family of models all of which fit the data equally well. One of the indicator variables must be left out. $\beta_0$ will include the effect of the category which does not have an indicator variable (among other effects); this is called the "reference category." The remaining $N - 1$ indicator variables (and their corresponding $\beta$s) will each yield an estimate of the difference of their category with the reference category.[5] This regime of contrasts is called "treatment contrasts" because it canonically corresponds to an experimental paradigm where there is one control group (the reference category) and the experimental hypotheses are whether any of several treatments (the other categories) provokes a significant difference from the control. The $\beta$s in such a regression and their associated confidence intervals and $p$-values provide an estimate of the effect of a certain treatment.

[4] Although these differences are for the main effects of clause type, which should not be scrutinized for their contribution to the model in the presence of an interaction term, which these models both have.

[5] For each N-category predictor, there is only enough information to add $N - 1$ terms to the model. An alternative strategy could be imagined: removing the $\beta_0$ term from the regression so as to compensate for the Nth indicator variable and its $\beta$. This works in the case of a single categorical variable, however it does not generalize. After adding a second categorical variable there is not another term that can be removed. $\beta_0$ must remain in the model to serve as the base case for all categorical predictors.

*Table 3.2:* A comparison of two logistic regression models on *do*-support data from the PPCHE. One has affirmative questions as the reference level, whereas the other has negative declaratives fulfilling this role.

| | Ref = Aff.~Q. | | Ref = Neg.~Decl. | |
|---|---|---|---|---|
| | Estimate | $p$-value | Estimate | $p$-value |
| (Intercept) | −1.38 | 0.00 | −1.76 | $1.27 \cdot 10^{-16}$ |
| year.std | 0.18 | 0.32 | 0.18 | 0.32 |
| Type[Treat: Neg. Decl.] | −0.38 | 0.40 | — | — |
| Type[Treat: Neg. Q.] | 0.85 | 0.20 | 1.23 | 0.03 |
| Type[Treat: Aff. Q.] | — | — | 0.38 | 0.40 |

Treatment contrasts are the R software's default, and (unless other arrangements are made) the first level of a factor (in alphabetical order) is treated as the reference level. However, there are drawbacks to this default. It is rare in syntactic analysis to be able to specify that one member of a set of contexts is the basic or control context, and all others are derived from this one. This means that the choice of the reference level is somewhat arbitrary. It is regrettably common practice to examine the significance values associated with individual coefficients. (Better approaches are discussed in section 3.3.3 immediately following.) Doing this without having made a considered choice of reference level does not yield sensible results. Whether the reference level has an intermediate or extreme estimated $\beta$ can affect whether the treatment effects appear significant. Table 3.2 illustrates this phenomenon using a subsample of data from the PPCHE.[6] In the left-hand model, with affirmative questions as the reference level, neither of the clause type effects has a significant $p$-value. Under some interpretations, it would be said that there is no effect of clause type in this data. However, on the right negative declaratives are treated as the reference level (this is the only difference between the two models). The effect for negative questions is significant at the $\alpha = 0.05$ level, indicating an effect of clause type despite there being no changes in the underlying data.

[6] It is necessary to use a subsample because the full dataset is large enough that a significant difference may be detected between any two clause types.

*Table 3.3:* A comparison of two logistic regression models on a subsample of *do*-support data from the PPCHE. Both models use sum contrasts.

| | Model 1 | | Model 2 | |
|---|---|---|---|---|
| | Estimate | $p$-value | Estimate | $p$-value |
| (Intercept) | $-1.22$ | $4.72 \cdot 10^{-8}$ | $-1.22$ | $4.72 \cdot 10^{-8}$ |
| year.std | 0.18 | 0.32 | 0.18 | 0.32 |
| Type[Sum: Aff. Q.] | $-0.16$ | 0.63 | $-0.16$ | 0.63 |
| Type[Sum: Neg. Decl.] | $-0.54$ | 0.03 | — | — |
| Type[Sum: Neg. Q.] | — | — | 0.70 | 0.07 |

An improvement on treatment contrasts is provided by R's built-in sum contrasts. These contrasts, instead of comparing the value of one category to another, compare the mean of a single category to the mean of all category means.[7] It is still the case that one of the levels is left out of the comparison. This is illustrated in table 3.3, which reuses the same subsample of the PPCHE data as table 3.2. The point estimate of its difference from the mean can be calculated by taking the negative of the summation of the other variables. With reference to the table, $0.70 = -(-0.54 + -0.16)$. As the regression output reflects, there is a significant difference in this data between negative declaratives and the mean of all clause types.[8] There is no significant difference from the mean for the other two clause types, however.[9] In this dissertation I'll use sum contrasts for models.

### 3.3.3 Model comparison

In the previous section, I explained (among other things) how certain modeling choices affect the interpretation of the $p$-values associated with individual coefficients. In the present section, I'll address some alternative methods of looking for meaningful effects in statistical models which recast the question in different ways, thus avoiding entirely the vagaries associated with the interpretation of these hypothesis tests.

[7] This figure is not equivalent to the mean of all the data points in the case when there are unequal numbers of data points in different categories.

[8] This result is related to the fact that there is more data for negative declaratives, thus their effect can be estimated with the greatest precision.

[9] Though negative questions come close – this time by having a large effect in spite of relatively little data.

### Likelihood ratio tests

A statistical model yields a set of probabilistic guesses about the dependent variable of a dataset, given the values of its independent variables. It is thus possible to calculate the likelihood of a certain model with respect to the dataset as the product of the probability of each of its guesses. For a given model, this figure provides a holistic picture of how well it fits the data. Given two models, one of which is nested inside the other, it is possible to compare the ratio of the two models' likelihoods to a $\chi^2$ distribution (with the number of degrees of freedom specified by the difference in number of free parameters between the models); this is called the likelihood ratio test (LRT). The null hypothesis is that the data is distributed according to the smaller of the two models, against the alternative that the richer description of the larger model makes the data more likely. This procedure gives the most principled version of a hypothesis test for whether there "is" an effect of a single covariate.[10] Specifically, one uses the LRT between the model including this term along with all other covariates and the model without the term. If the null hypothesis is rejected ($p$-value is small), then that variable can be said to have an effect.

### Information criteria

There are at least two unpalatable things about the LRT:

- It assumes that one of the two models in the candidate set is the true model, which is often not credible for linguistic studies. Specifically in historical syntax, we have hypotheses about the effects of phonology, information structure, style, and other factors on the data. Seldom is there sufficient data of sufficient richness to incorporate all of these hypotheses into our models.
- The LRT is confined to nested models; it cannot test the difference between models which are not nested.

There are two notional components to the LRT: a measurement of the correspondence between the model and the data, and a penalty for models with more free parameters (since a model with a larger number of parameters always fits the data better to some degree, even if minimal). Information criteria extend these notions to non-nested models, in the process dropping the assumptions that allow the confidence interval and $p$-value of a LRT to be computed. The Akaike Information Criterion (Akaike 1973) was developed first. It is grounded in the information theoretic concept of Kullback-Leibler (KL) divergence. KL divergence is

[10] The scare quotes stem from the fact that it is rare in a linguistic study to include a covariate whose effect we have a strong *a priori* belief to be zero. Additionally, there are limitations on the kinds of effects that can be sensibly dropped from a model to test their significance.

a measure of the difference between two probability distributions, giving the quantity of information (in bits) that are lost when using one distribution to approximate another. The AIC provides an estimate of the KL divergence between a model and the data-generating process. Because the actual data-generating process is not known (in the general case), it is not surprising that the AIC estimates this quantity only up to some unknown additive constant. However, crucially for two models fit to the same set of data the additive constant will be the same. Thus we can subtract the AIC of two models, and the ΔAIC gives an estimate of the difference in KL divergence between each of those two models and the data-generating process.[11] The model with the lower AIC is better, in the sense of being closer to the data. It has been shown that model selection using the AIC in this way is asymptotically equivalent to leave-one-out crossvalidation (Stone 1976).

Since there are no $p$-values associated with the AIC, a rule of thumb is necessary for interpreting differences in this statistic. One commonly used heuristic is that AIC differences smaller than two are to be disregarded, and the two models considered roughly equally well supported by the data.[12] Models with a difference of greater than two are taken to be less well supported by the evidence, and any with a ΔAIC of more than ten are not supported by the data at all.

An alternative information criterion called the Bayesian Information Criterion, or BIC, has been formulated (Schwarz 1978). This is very similar in structure and interpretation to the AIC, though it includes a larger penalty for the inclusion of extra parameters. It has certain theoretical properties that make it less desirable than the AIC in familiar cases in historical syntax. (For an overview, consult Burnham and Anderson 2004.) Most importantly, in line with the CRH we are often rooting for a more parsimonious model; thus using the BIC which is defined in a way that favors parsimony is in some way not playing with a full deck. Thus, in this dissertation I will rely on the AIC rather than the BIC.

**Multiple comparisons**

The information criterion approach gives a satisfactory answer to the question of which predictors overall have a meaningful effect on an outcome variable. It does not, however, answer a questions which are sometimes of scientific interest about which specific values of a predictor have effects which are large, small, positive, negative, or similar to the effects of other predictors. In order to address these questions,

[11] There is a small-sample correction that must be applied to the original AIC yielding the corrected AIC or AICc. (Hurvich and Tsai 1989)

[12] The threshold value of two corresponds to an evidence ratio of 2.7 – that is, between two models A with AIC=X and B with AIC=X+2, the data support model A to a 2.7 times greater extent than they do B. The threshold of ten corresponds to an evidence ratio of 148.8.

a model comparison methodology may be utilized from the `multcomp` package in R (Hothorn, Bretz, and Westfall 2008). This package fits the maximal model and calculates confidence intervals for the differences in the coefficients of interest directly. The result is a set of confidence intervals for the differences. If these confidence intervals exclude zero, then we can conclude that there is evidence that the two contexts in question differ. Although similar in function to the comparisons of various models considered above, this model comparison technique has several features which recommend it:

1. When equality of slopes is tested by combining various factor levels, the resultant slope parameter estimates are pooled. That is, all members of a class contribute to the determination of the estimate; when the distribution of tokens over classes is not even (as is nearly always the case in corpus-based studies) the pooled estimate will be biased towards the more frequent class. This methodology allows the calculation of the slope per group to be maximally faithful to the underlying data, while also allowing the hypotheses of interest to be tested.

2. The procedure controls the $\alpha$ level of the entire test to the specified level; there is no need for further corrections for multiple comparisons.[13]

## 3.4 Power analysis

In the Null-Hypothesis Significance Testing (NHST) paradigm, a procedure is desired that answers the question "is there an effect?" in the affirmative if and only if there is in fact an effect in the real world. As is suggested by the biconditional formulation, in practice the maintenance of this guarantee about statistical procedures is bifurcated into two parts: the false positive and false negative rates. The false positive rate describes how often a statistical procedure detects an effect when in fact there is not one in the real world. The significance threshold $\alpha$ which is traditionally set to 0.05 directly controls the false positive rate.[14] The false negative rate denoted by $\beta$, on the other hand, is not directly controlled in a hypothesis test. It must instead be calculated. Because it is not directly controlled, it does not exist independently of the specifics of the test. Rather, $\beta$, $\alpha$, the sample size, and the magnitude of the effect being investigated exist in relation to each other; fixing three of these parameters ($\alpha$, N, and magnitude of effect) allows one to solve for the fourth ($\beta$).

[13] Specifically, the family-wise error rate is controlled to the test's stated $\alpha$ level. For further details, consult the package's documentation.

[14] There is a large literature devoted to the topic of assuring that the nominal $\alpha$ of a test reflects its actual behavior. Sometimes tests are known to be (anti-)conservative; that is, to have a true false positive rate lower (higher) than $\alpha$. It is also necessary to correct when multiple tests are done in order to assure that the results of these tests, as a group, adhere to the promised $\alpha$ level. Nonetheless, the point remains that $\alpha$ is in principle a direct control over the false positive rate.

The *power* of a statistical test is defined as $1 - \beta$, and can be used interchangeably with $\beta$ in discussions of the false negative rate. If a test has a power of 0.8 (given a fixed $\alpha$, N, and effect size), then the test has a probability of 0.8 of detecting an effect if there really is one of that size in the data-generating process. There is a direct tradeoff between $\alpha$ and $\beta$: as one rises, the other falls. The tradition in the social sciences is to regard a false positive as 4 times more serious than a false negative, and thus to set $\beta$ to 0.20 ($= 4 * \alpha$). However, given that the CRH is a hypothesis about the absence of an effect, a false negative is actually more serious (since it leads to an unwarranted conclusion about grammatical structures). Thus it might seem reasonable to aim for a $\beta$ of 0.05. Such a high $\beta$ will necessitate a large sample size if $\alpha$ is to be maintained at 0.05 as well.

Whatever the choices made by the researcher in a particular study, it is vitally important to report a power analysis, since without this information it is impossible to evaluate any results against the deductive canons of NHST. This problem has been noticed in the social sciences before, and greater attention to power analysis is a common suggestion. (See for example Gill 1999 for political science and Cashen and Geiger 2004 for management research.)

Approximate closed-form solutions to questions about the statistical power of logistic regression are possible; see Alam, Rao, and Cheng (2010), Hsieh (1989), Hsieh, Bloch, and Larsen (1998), Schoenfeld and Borenstein (2005), and Væth and Skovlund (2004) for an overview of relevant literature. However, such solutions are difficult to construct in the best case, and can break down quickly as covariates are added to a model. Thus, simulation is often used in the power analysis literature (including as the standard of comparison for verifying the closed-form solutions proposed in some of the previously-cited articles), and will be used here. This introduces a variety of subtleties. First of all, the question of how to appropriately sample new datasets arises. Basically, this amounts to randomly creating a list of vectors of ⟨year, context1, …, contextN⟩ for tokens in an imaginary corpus. (In the case of *do*-support, the most interesting and often only context variable is the type of clause – negative declarative, affirmative question, negative question, etc.) One possible approach is to sample each of these variables independently from a uniform distribution. This is clearly inappropriate: tokens from a corpus are not uniformly distributed across years (tending to be scarcer in earlier time periods) nor linguistics contexts (some contexts are almost invariably more frequent than others). The effect of assuming a uniform distribution will be to overestimate the power of the model. A more plausible approach is to sample the existing tokens in the existing corpus (with replacement) – drawing ⟨year, context⟩ vectors from the rows of the existing dataset. This is also unsatisfactory in that it provides an underestimate of the model's power. It assumes that the distribution of particular contexts across years

does not vary, whereas in reality this is a contingent factor influenced by the particular texts selected for the analysis.

An ideal randomization procedure would replicate the steps undergone in an actual corpus experiment. Since this generally involves the selection of a series of texts, each of which has a characteristic author and date, this ought to be replicated in a power simulation as well. However, such simulations are complex, and not without their own difficult questions. (What is the ideal distribution of contexts across texts? Should this be nested in a notion of genre? How do we model multiple texts by the same author appearing in a corpus? …) Thus, I'll adopt the methodological principle that new data rows should be sampled from existing data, but with each covariate sampled separately. To give a concrete example, in the simulation of the results of Kroch (1989), I draw years from the list of years in the actual data (with replacement) and a *do*-support context from the list of contexts – but not a ⟨year, context⟩ pair from the attested list of such pairs. This procedure ought to give a relatively reliable – though not bulletproof – estimate of a study's power.

Another question is which arises is how to set the other terms in the equation: sample size, effect size, and $\alpha$. I will in the first instance adopt the actual sample size of the corpus under consideration, and the conventional $\alpha$ level of 0.05. Using the effect size from an actual regression is not indicated, however. As discussed by Hoenig and Heisey (2001) under the name of "power approach paradox," a non-significant result with a comparatively larger effect size (and thus smaller $p$-value) will actually lead to assignment of higher power to the test by a *post hoc* analysis, and thus an inference that the test has given greater credence to the null hypothesis that the true value of the effect is 0. This is a paradoxical result – if anything a larger (but non-significant) effect ought to decrease confidence in the null.[15]

Instead of using the actual estimated effect size, it is necessary to construct a range of plausible effect sizes (partially informed either by actual results or by independent notions of the mapping between numerical effects and real-world importance). In the following subsection, I discuss how these may be interpreted in terms of years, further adding to the possibility of meaningfully anchoring their values and avoiding the paradox of a completely mechanistic approach to power calculation.

[15]Hoenig and Heisey (2001) actually advocate an approach to understanding the ability of tests to reject the null hypothesis which eschews traditional power analysis entirely, and instead inverts the usual NHST paradigm, making the null hypothesis that there is an effect larger than some $\Delta$ and the alternative that the effect is smaller than $\Delta$; the null will be rejected just in case there is good evidence that the effect is small (or zero). This procedure respects the familiar guarantees about false positives enshrined in $\alpha$. I do not adopt this approach (despite its merits) because it is not immediately obvious how to embed it in the framework of larger regression models which are used "off the shelf" in quantitative historical linguistics.

### 3.4.1 Power analysis in historical syntax

The CRH investigates slope effects within syntactic contexts. An overall slope effect (the main effect of a `year` variable in a logistic regression) measures the speed at which a change takes place. In the logistic model a change never actuates nor completes; it merely approaches 0 and 1 asymptotically. I will thus take the development of a syntactic change under the logistic model to be the progression from 0.05 to 0.95, recognizing that beyond these extreme probability values the evolution of syntactic changes is controlled by factors not expressed in the logistic model. There is a distance of 5.89 logit units between 0.05 and 0.95. We can thus divide 5.89 by the slope value given in regression output to yield the regression model's estimate of the duration of a change (the time, measured in years, it takes that change to progress from 0.05 to 0.95 probability of occurrence in the data). For example, Kroch (1989, Table 4) estimates a slope of 0.0374 logit units per year for the increase of *do*-support in negative declaratives before 1575. This is equivalent to a projection that the full change will take place in 157 years (= 5.89 / 0.0374). (Of course, in the case of *do*-support, other factors intervene to disrupt the change before it reaches completion.)

*Table 3.4:* A table of logistic regression interaction effects, interpreted in terms of their distances in years. The first two rows indicate the length of the change (main effect of year) in years and logit units per year. The left-hand column of numbers gives the magnitude of various interaction terms. In the body of the table are listed the differences in duration of the change (in units of years) that combinations of main effect and interaction implies. Thus, in the top-left cell of the table's content, the context in question has a slope of 0.118 - 0.1 = 0.018 logit units / year, and takes 50 + 281 = 331 years to take place.

| | Years | 50 | 100 | 150 | 200 | 250 | 300 | 350 | 400 | 450 |
|---|---|---|---|---|---|---|---|---|---|---|
| | Slope | 0.118 | 0.059 | 0.039 | 0.029 | 0.024 | 0.020 | 0.017 | 0.015 | 0.013 |
| Interaction | -0.1 | 281.253 | — | — | — | — | — | — | — | — |
| | -0.05 | 36.885 | — | — | — | — | — | — | — | — |
| | -0.02 | 10.227 | 51.429 | 155.770 | — | — | — | — | — | — |
| | -0.01 | 4.639 | 20.455 | 51.266 | 102.857 | 184.427 | — | — | — | — |
| | -0.005 | 2.217 | 9.278 | 21.892 | 40.909 | 67.365 | 102.532 | 147.987 | — | — |
| | -0.001 | 0.428 | 1.727 | 3.921 | 7.031 | 11.084 | 16.103 | 22.116 | 29.150 | 37.232 |
| | 0.001 | −0.421 | −1.670 | −3.726 | −6.569 | −10.181 | −14.542 | −19.635 | −25.442 | −31.946 |
| | 0.005 | −2.036 | −7.826 | −16.946 | −29.032 | −43.774 | −60.902 | −80.182 | −101.409 | −124.403 |
| | 0.01 | −3.913 | −14.516 | −30.451 | −50.704 | −74.503 | −101.250 | −130.474 | −161.798 | −194.920 |
| | 0.02 | −7.258 | −25.352 | −50.625 | −80.899 | −114.796 | −151.402 | −190.086 | −230.400 | −272.015 |
| | 0.05 | −14.901 | −45.918 | −84.025 | −125.874 | −169.940 | −215.426 | −261.877 | −309.013 | −356.654 |
| | 0.1 | −22.959 | −62.937 | −107.713 | −154.507 | −202.338 | −250.774 | −299.592 | −348.668 | −397.926 |

Logistic regression models testing the CRH can include not just a single slope term (effect of `year`), but interactions between the time variable and contextual variables. These can be interpreted as the model's

estimate (in years) of the difference between two contexts in the duration of a change. For example, consider table 3.4, which gives the tabulations of various main and interaction effects. Consider specifically the top left cell of the table. This indicates that, for a change which takes 50 years in context A, an interaction effect in context B of -0.1 predicts a difference in duration of roughly 280 years (and thus a total duration in context B of 330 years). This is clearly a significant difference. On the other hand, an interaction effect of -0.005 predicts only a difference of 2.21 years, for a total duration in context B of 52.21 years.[16] This is clearly not a meaningful difference, even if sufficient data is amassed to measure it precisely and reach statistical significance. The values in this table allow the abstract numbers generated by a logistic regression to be interpreted in conceptually intelligible terms. It is impossible to say whether an inter-contextual difference of (say) 0.005 logit units per year is meaningful. However, we have a very clear notion that a difference of two years, on the scale of syntactic change (and subject to the amount of noise we understand to be present in our data), is not meaningful. This notion of meaningful difference in turn is important for power analysis: in order to be fully confident in our analyses, they should have sufficient power to detect meaningful differences.

[16]Note that the values for -0.005 and 0.005 are not symmetric; this is due to the non-linear nature of the logit transform.

# Chapter 4

# The quantitative diachrony of *do*-support

In this chapter, I'll develop an account of the diachronic behavior of *do*-support in EME. Section 4.1 gives an overview of previous quantitative analyses. The following section 4.2 explores the possibility of using novel datasets to replicate and test the findings of these analyses. Some findings are successfully replicated, while others must be revised in the face of novel data. Section 4.3 combines some elements of previous analyses with new data and insights in order to propose a novel account of the grammatical history of *do* usage in EME. Section 4.4 takes up the question of how insights from our understanding of the diachrony of *do*-support can inform grammatical analyses, providing evidence that is different in kind from synchronic data. It will be argued that diachronic data can provide answers to questions of grammatical analysis that synchronic data cannot address conclusively. Finally, section 4.5 concludes.

## 4.1 Previous accounts

The first noteworthy quantitative study of the history of English *do*-support was undertaken by Ellegård (1953). He collected a corpus of approximately 21,000 tokens of potential *do*-support environments (sentences which either actually had *do*-support, or would have had *do*-support in modern English). He took his main research question to be the then-active debate about whether *do*-support originated from a Middle English causative, Celtic substrate influence, or some other source. This question is of secondary interest for the present inquiry (although some of the conclusions reached will bear on it). We are more interested instead

in the quantitative conclusions that Ellegård drew from his corpus and, by extension, the conclusions that others have drawn from the same data.

In presenting these conclusions, I have been aided by the electronic version of Ellegård's coding scheme prepared by Ann Taylor and further elaborated by Anthony Warner. This consists of ten columns:

1. auxiliary type
2. negation presence/absence
3. clause type (question/imperative/declarative)
4. question type (adverb, object, yes/no, …)
5. transitivity
6. special conditions (exclamative, emphatic, …)
7. various lexical classes
8. subject type, for questions
9. special conditions 2 (certain adverbials, tag questions, …)
10. position of *not* in negatives

### 4.1.1 Ellegård's results

One of Ellegård's most basic findings was the orderly pattern in which *do*-support develops. This finding is reproduced from his data in figure 4.1.

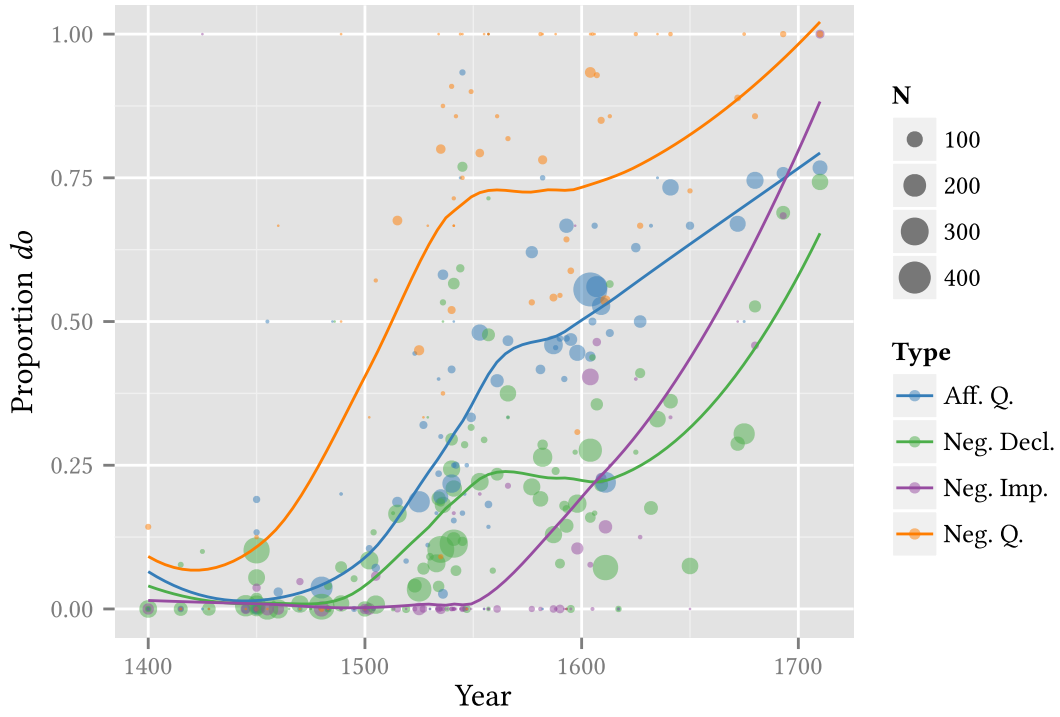



*Figure 4.1: Do*-support in Ellegård's corpus.

The data are broken down into several contexts, each of which is plotted separately as a series of points corresponding to the proportion of *do*-support sentences in the given environment in each text. The size of the points is proportional to the total number of tokens of the given context in the given year (that is, the denominator of the proportion). The trend line fit to the data is calculated by the LOESS nonparametric smoothing method with the $\alpha$ parameter set to 0.75.[1]

The LOESS method has that advantage that it produces a smooth fit to data without requiring the specification of a functional form. There are disadvantages associated with the method: it is not very powerful (in the statistical sense) – that is, strong trends require relatively more data to be detected well, as compared to a model which encodes the structure of those trends *a priori*. It also performs poorly near the edges of the domain (x-axis) of the data. (Ruppert and Wand 1994, remark 4) This can be observed in

[1] 0.75 is the default value for $\alpha$ parameter used by the `ggplot2` R package (Wickham 2009). I have thus adopted it for graphs in this dissertation; any deviations from this default will be noted. By default `ggplot` provides uncertainty intervals around the smoothing line; I have turned this feature off in my graphs because the width of the interval does not correspond to any specific easily interpretable property of the LOESS model when interpreted as a data-driven visualization strategy. At best the relative widths of the intervals (across different smooth lines on the same graph, or at different points along the same line) provide a qualitative estimate of the uncertainty inherent in the smooth.

figure 4.1 in the period beginning 1400, where the trend line has *do*-support decreasing from 5% in affirmative questions and negative declaratives; it would not be sensible to posit that *do*-support is used at a rate of even 5% in this corpus until 50–75 years later. (This is a relatively minor instantiation of this drawback of LOESS; it will reappear many times in the graphs in this dissertation however.) Despite these drawbacks, LOESS constitutes a solid methodological choice for the visualization of trends in datasets (like this one) where the trend is difficult to describe *a priori* in terms of a model, and where there is sufficient noise in the raw data that it cannot be merely read off the graph.

Ellegård did not use LOESS smoothing for visualizing his data – in fact it had not been invented at the time. He instead grouped his data into bins (of variable widths between 10 and 50 years). This practice has been continued by later researchers. It is not, however, an optimal visualization strategy.[2] Grouping of data into large bins obviates information about the variation contained within the bins. The traditional visualization technique does not represent the differing amount of data in each bin (though this could be rectified by allowing point sizes to vary, as in figure 4.1 above). More fundamentally, the traditional binning method does not show how much variation there is inside each bin. The data being measured is binomial, which means that the variance of the distribution of individual tokens is mathematically dependent on its mean (the proportion on the graph). However tokens come grouped into texts, and it is of interest whether the texts in a bin have all roughly the same proportion of *do*-support or conversely whether there are texts with differing sample proportions (whether because there are few tokens from the text available, or because of genuine inter-speaker variation in the propensity to produce *do*-support). For this reason, graphs in this dissertation present one point per text, without any binning.

Ellegård noticed about this graph that there is a decrease in *do* usage in the late 16th century, which he judged to be reflective of developments in the language (and not, for example, chance fluctuations due to sampling biases). This fundamental observation will prove to be quite important to later analyses of the phenomenon, including the present one. In the following subsections, I'll move on to consider more fine-grained aspects of the conclusions Ellegård reached about his data.

[2]Ellegård was aware that this is not an optimal strategy for uncovering the scientific truth about the behavior of *do*: "Strictly speaking we have no right to refer to the 'frequency of *do* at a certain period' without any further qualification. There was not one frequency at each period, since there seem always to have been both dialectal and stylistic differences with regard to the use of the *do*-form. The ideal procedure would be to single out a very narrowly defined dialect and study the development there. The same investigation should then be carried out for all other dialects, after which we should be in a position to observe and analyse the influences and cross-influences at work. But even a narrowly defined dialect exhibits individual variations, and the question would arise, for example, as to what should be regarded as the typical or representative frequency of *do*-forms in it. Still more important for us is that we are not able to get enough illustrative material from an narrowly defined dialect, quite apart from the fact that most texts cannot be placed dialectally even within very wide limits. It is obvious that the relative frequency figure for a period (and a dialect) must be expressed as an average." He is less concerned with data visualization strategies *per se*, but (so far) that is the question that has exercised our discussion here.

### Affirmative declaratives

Ellegård also studied the appearance of *do* in affirmative declaratives. As discussed in section 2.1.2, it is possible for *do*-support to appear in affirmative declarative sentences under conditions of verum focus. The prevalence of this environment is difficult to estimate precisely from written data, which lacks the prosodic signal which can identify verum focus. However, Ellegård attempted to judge whether a given example was, in context, unambiguously a token of verum focus (or "emphasis" in his term). He found a very low rate of such sentences: 78 out of 7065 affirmative declaratives were judged to be emphatic, or a rate of 1.1 percent. In any case, barring vertiginous shifts in genre, we do not expect the proportion of emphatic sentences to vary diachronically to a noticeable degree, given the constancy of the semantics and pragmatics of verum focus over the history of English.[3]

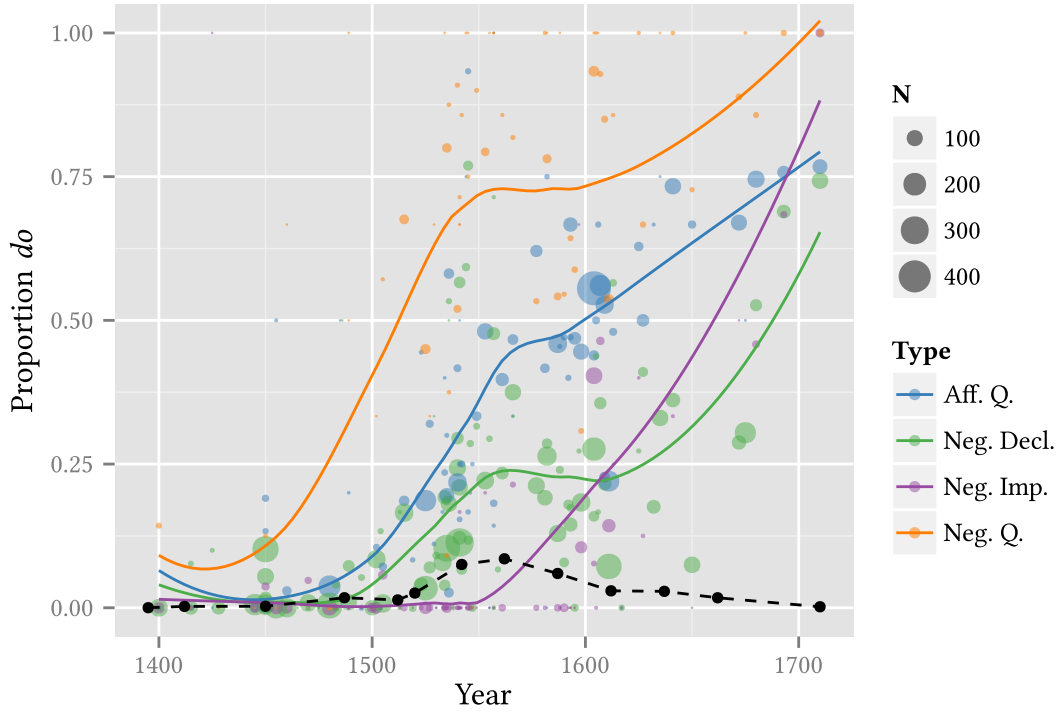



*Figure 4.2:* *Do*-support in Ellegård's corpus. Ellegård's estimate of the proportion of *do*-support in affirmative declaratives is represented by the black points (not scaled according to size), and the intervening dashed line gives a linear interpolation between the points.

[3]Indeed there are not genre shifts in Ellegård's corpus of the magnitude which would be needed to explain his results on affirmative declaratives, despite the lack of attention to genre balancing in the corpus's construction.

Ellegård did not collect a precise count of affirmative declaratives without *do*-support, so exact proportions cannot be calculated. However, he did estimate for each period the total number of affirmative declarative clauses (by subsampling small passages from each text in his corpus). Thus, approximate proportions can be derived. Figure 4.2 shows these approximations superimposed on the graph from figure 4.1. The figure demonstrates that the occurrence of *do*-support in affirmative declaratives varies by a large margin over the EME period, going from an occurrence frequency so small as to be unmeasurable at the beginning and end of the period to a frequency of roughly 10% in the middle century. Ellegård noted that the peak in the trajectory of *do*-support in affirmative declaratives is contemporaneous with the temporary decline of *do*-support in other contexts. The puzzle underlying these results is to explain why a construction which is not part of the PDE *do*-support phenomenon nonetheless appears in a non-negligible proportion of tokens for over a century, and why at first glance its inflection point should be simultaneous with respect to deflections in the trajectory of *do*-support in other contexts.

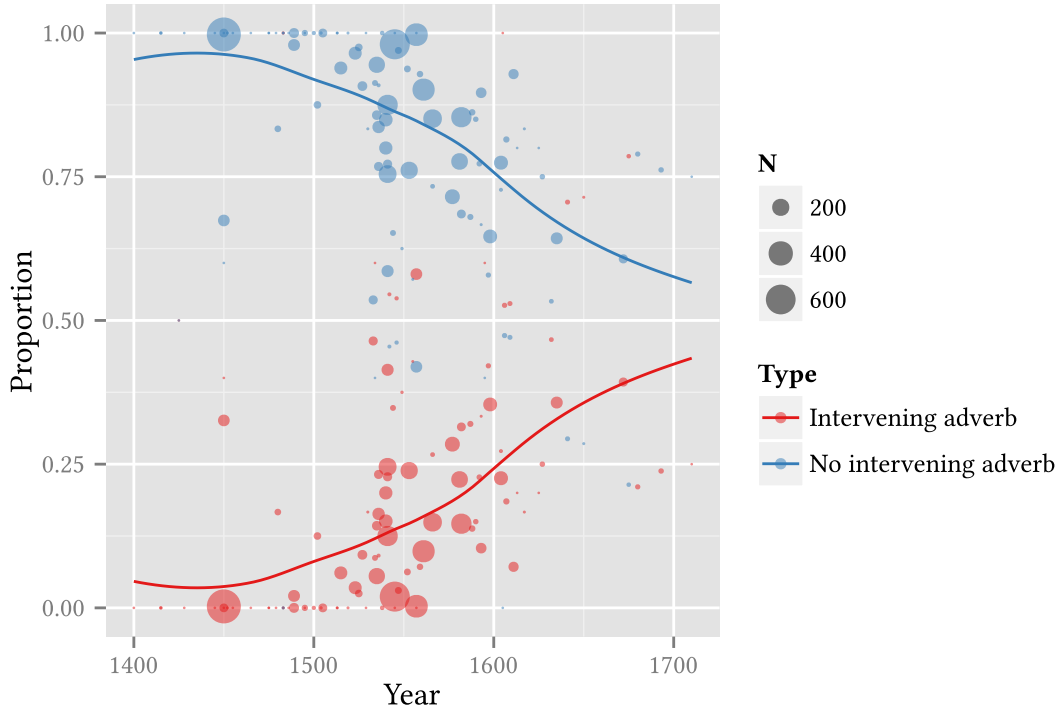



*Figure 4.3:* The behavior of affirmative declarative *do*-support in conjunction with adverbs in Ellegård's data.

Attending specifically to the affirmative declaratives which do have *do*-support, Ellegård made a further

observation that they are increasingly likely to occur with an adverb over time. This is illustrated in figure 4.3. The two sentence types considered are those affirmative declaratives where an adverb intervenes between *do* and the verb, and those where no adverb (or other element) does so. The graph illustrates that the frequency of the former sentence type increases drastically over time, at the expense of the latter.[4]

### Transitivity

Ellegård also noticed an effect of transitivity on *do*-support in negative questions and declaratives. Figure 4.1 is repeated, subdivided by transitivity, in Figure 4.4. As the graph makes clear, there are consistent transitivity differences in the above-named sentence types, whereas the effects are smaller in the others, if indeed they exist at all.

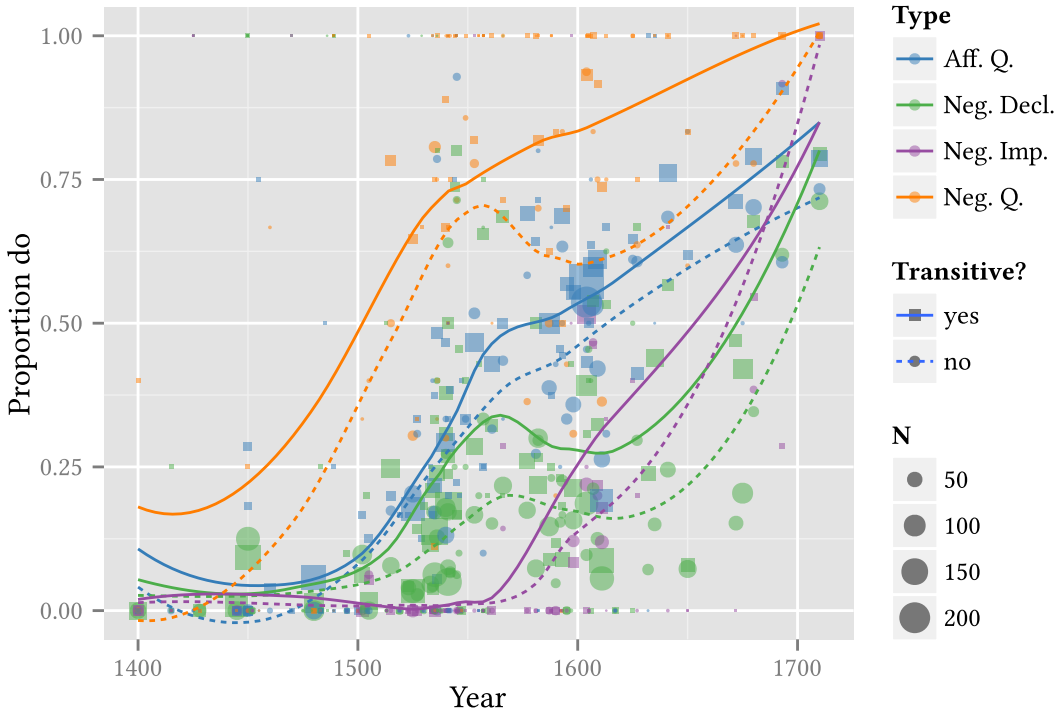



*Figure 4.4: Do*-support in Ellegård's corpus, partitioned by transitivity.

[4]This graph recapitulates Ellegård's Table 9 and the associated figure (p. 182). It removes his "a/o-inv" category – these correspond to sentences with subject-verb inversion, and their increase reflects the establishment of *do*-support in that environment generally rather than something about the behavior of affirmative declarative *do*-support in particular.

**Lexical class effects**

Ellegård also discussed a lexical class of verbs which tend to resist *do*-support. This class includes the following verbs:

- *boot*
- *care*
- *doubt*
- *fear*
- *know*
- *list*
- *mistake*
- *skill*
- *trow*

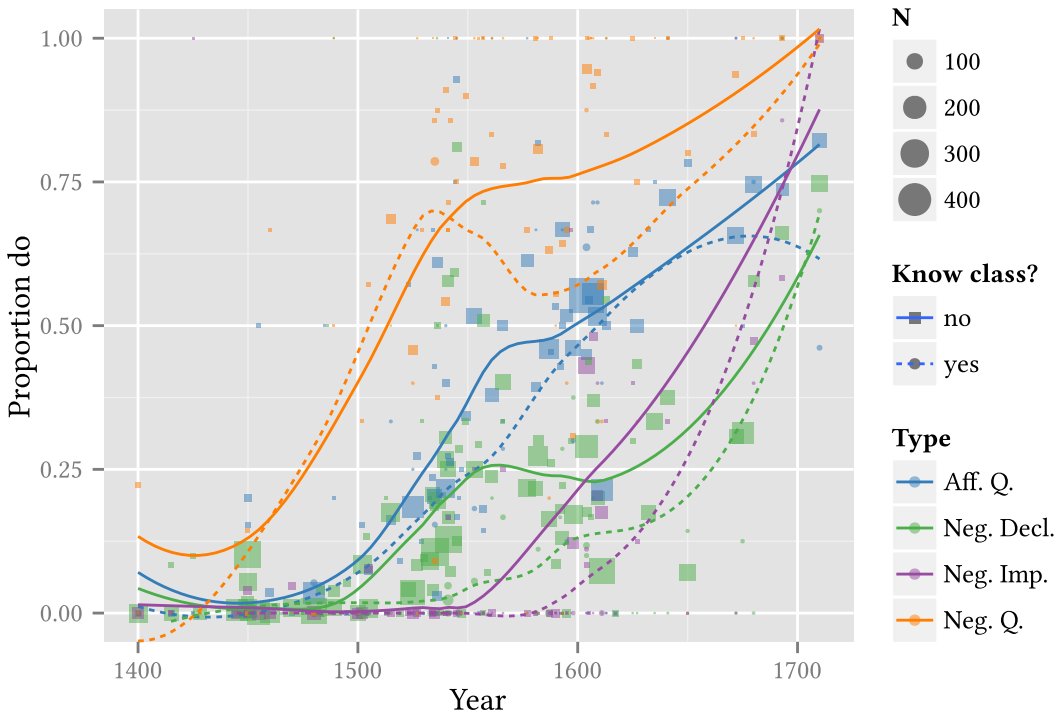



*Figure 4.5:* The behavior of *know*-class verbs compared to others in Ellegård's corpus.

The behavior of this class is depicted in the graph in Figure 4.5.

### Question types

Ellegård discovered differences in the distribution of *do* across different types of questions. This distribution is reproduced in the graph in Figure 4.6.

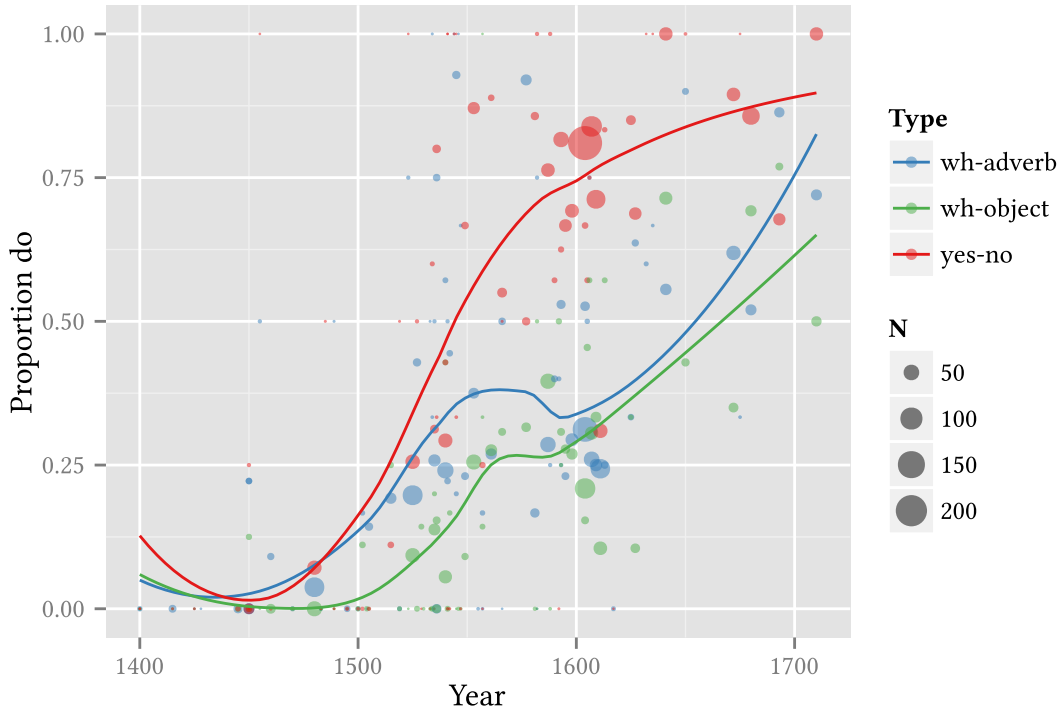



*Figure 4.6:* The trajectory of *do*-support in various types of questions in Ellegård's corpus.

### Summary

The incorporation of these results of Ellegård's into an explanatory generative framework is an important task for later investigations of the history of *do*-support, including the present one. In section 4.3, I'll introduce my own proposal, which addresses the results from section 4.1.1, 4.1.1, and (partially) 4.1.1. The results on adverbs with affirmative declaratives and question types remain (for the moment) mysterious. However, before moving on to presenting my own proposal, I'll review work on *do*-support subsequent to Ellegård, as well as the possibility of replicating results from Ellegård's results in other corpora.

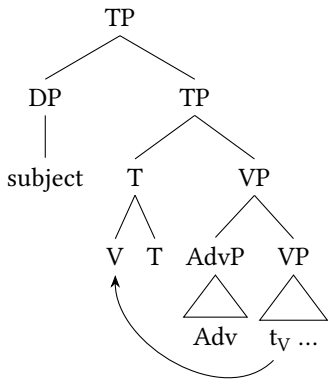
### 4.1.2 Later analyses of Ellegård's data

Various authors working since 1953 have used Ellegård's corpus to drive their own investigations. In this section, I'll review their contributions.

#### The Constant Rate Hypothesis

Kroch (1989) laid out the CRH framework presented in detail in section 3.1. One of the case studies which he used to justify his framework was the behavior of *do*-support and verb-raising in EME. Kroch made use of a grammatical analysis by Roberts (1985). According to this analysis, ME had the verbal syntax sketched in figure 4.7.[5] After the language lost verb raising, the analysis changed to the one in figure 4.8. Instead of raising to T in the syntax, the verb lowers in the morphology. (For a more complete description of several different analyses of the process, refer to section 4.4.)

*Figure 4.7:* A syntactic analysis of ME verbal inflection in the absence of an auxiliary. The verb raises to T, across any intervening adverbs. For reasons of simplicity, the movement chain headed by the subject is not shown, nor are a variety of intermediate projections. (cf. Roberts 1985)

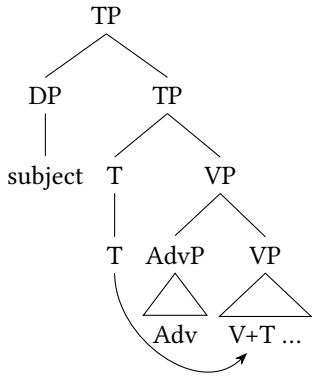



This reanalysis has (at least) two distinct effects on the strings that English generates. The first is directly implied by the trees in figures 4.7 and 4.8: the position of the verb relative to adverbs changes.[6] The second effect concerns the environments where the structural adjacency needed for morphological lowering is disrupted: the dummy auxiliary *do* is innovated. Kroch observed that since these superficially distinct patterns were tied to the same change in grammatical structure (the replacement of the tree in figure 4.7 by the one in 4.8), a CRH effect ought to hold between the two environments.

[5]The category labels have been updated from Roberts' original analysis, and the trees simplified somewhat.

[6]Since there are different classes of adverbs which appear probabilistically (not categorically) in different positions, the change in position must in general be measured quantitatively. However, there are some strings which are grammatical in ME but not later stages of the language, such as V Adv Obj (where the Obj is not eligible to participate in Heavy NP Shift or similar extraposition processes).

*Figure 4.8:* A syntactic analysis of late EME (and PDE) verbal inflection in the absence of an auxiliary (in affirmative declaratives). T lowers onto the verb by a morphological operation (Embick and Noyer 2001). The tree is simplified somewhat as in figure 4.7. (cf. Roberts 1985)




This is indeed what Kroch found: all the environments for *do*-support that Ellegård studied except affirmative declaratives have a parallel slope up to 1575. That the data exclude affirmative declaratives from this class is important, since this environment is not part of the PDE *do*-support paradigm. What's more, Kroch found that the slope describing the loss of verb raising over *never* is the same as the slope describing the rise of *do*-support (correcting for the opposite directionality of the two changes). In order to derive this result, Kroch had to introduce a term to correct for the phenomenon whereby *never* can appear between the subject and a modal, as in:

(77) He never will know the truth.

Kroch made an empirical estimate of the frequency of this phenomenon: roughly 16%. He used this estimate to deflect the slope of the curve for verb raising over *never*, thus changing slightly the analysis from a simple comparison of logistic regression coefficients. However, the exact value of this correction term does not appear to matter very much – Kroch found that the CRH effect obtains at values between 5 and 20%, inclusive.[7] (Kroch did not test a value of 0% for this parameter. 25% was the next largest value which he tested, and he found that the CRH did not hold there.)

The reason Kroch only measured the slopes of the various contexts until 1575 is that, in that year, there is a manifest discontinuity in the data. The rise of *do*-support is halted, and in many contexts reversed. A decline is never predicted by the population-biological model underlying the logistic model; any meaningful deviation from monotonic increase must be taken as a sign that unmodeled complexity is obscuring the data. The suspiciousness of 1575 is heightened by the fact that this is roughly when *do*-support in affirmative

[7] In fact, the true rate of this phenomenon's occurrence is closer to 5% – see section 4.2.2.

declaratives, which reaches a level of roughly 10% in the population (far higher than its rate in PDE, where it is confined to emphatic contexts) begins to decline. Finally, the CRH effect between the *do*-support contexts fails to hold after 1575. (It is not possible to test the CRH between *do*-support and verb raising, because the latter change has already gone to completion by 1575.) From all this evidence, Kroch concludes that there was a grammatical reanalysis in 1575 (roughly), reflecting the definitive loss of V-to-I raising in the language.

**Stylistic effects**

Warner (2005) investigates sociolinguistic conditions on the evolution of *do*-support, using Ellegård's corpus. His aim was not to demonstrate a grammatical shift in the analysis of *do*-support but rather a social one. His investigation yields three important findings:

1. Lexical complexity has an effect on the rate of *do*-support usage
2. The trajectory of the evolution of *do*-support usage is qualitatively different at different levels of lexical complexity
3. Age grading in the usage of *do*-support exists after (but not before) 1575

He analyzes these facts to be attributable to a novel stylistic constraint introduced after 1575, which militates against word orders that lead to *n't* contraction, including *do*-support. He arrives at this analysis because Ellegård's affirmative questions don't participate in the deflection of 1575 (he also investigates Aux-*not*-Pronoun vs. Aux-Pronoun-*not* word orders in questions to bolster this analysis).

Warner constructed an index of lexical complexity, which he calculated for each text in Ellegård's database. This index was based on the average word length and type:token ratio in a 600-word sample from the text. With respect to point 2 in the above list, Warner showed that in the lower 50% of texts on this style index, the rise of *do*-support is uninterruptedly monotonic, whereas the trajectory in the upper stylistic half of the texts shows a basically flat trajectory after 1575. With respect to 3, he discovered that there is a propensity for speakers of different ages to employ a variable construction at different rates in the portion of the data after 1575. Specifically, after 1575 older speakers use less *do*-support than younger speakers; there is no such effect in the pre-1575 data. From these two sources of evidence, Warner concluded that what happened in 1575 was that the speech community adopted a novel evaluative principle militating against the usage of *do*-support. He further used the fact that in Ellegård's data affirmative questions do not undergo a deviation from monotonic increase to propose that the evaluative principle was not against the usage of *do*-support as such, but rather against the contraction of *not* to *n't* (and against contraction processes generally).

The use of *do*-support in negatives could lead to contraction, and was thus disfavored by writers utilizing a high style.

## 4.2 Replication experiments

### 4.2.1 Replications of Ellegård's results

Using a parsed corpus composed of the combination of the PPCEME (Kroch, Santorini, and Delfs 2005) and the PCEEC (Taylor et al. 2006), it is possible to attempt a replication of Ellegård and following researchers' results.

*Table 4.1:* Comparison of the size of Ellegård's corpus with the parsed corpora. In the latter category, only potential *do* support sentences occurring before 1700 are counted.

| | Ellegård | Parsed Corpora |
|---|---|---|
| Aff. Decl. | 7065 | 145491 |
| Aff. Imp. | 77 | 11358 |
| Aff. Q. | 3772 | 1453 |
| Neg. Decl. | 7604 | 6251 |
| Neg. Imp. | 1467 | 652 |
| Neg. Q. | 753 | 269 |

The parsed corpora of EME (PPCEME + PCEEC) contain data which is largely distinct from Ellegårds's corpus.[8] They also include a roughly comparable amount of data (Ellegård's corpus contains roughly 2–3 times as many tokens in each category, except affirmative declaratives and imperatives which Ellegård did not systematically collect). The precise counts of tokens in both corpora are given in Table 4.1. Thus, the parsed corpora present the possibility of replicating Ellegård's results, thus strengthening our confidence in them.

[8]Specifically, 6% of the data in the parsed corpora is in Ellegård's corpus as well; 94% is distinct. See appendix Overlap between Ellegård's corpus and the PPCHE for more details.

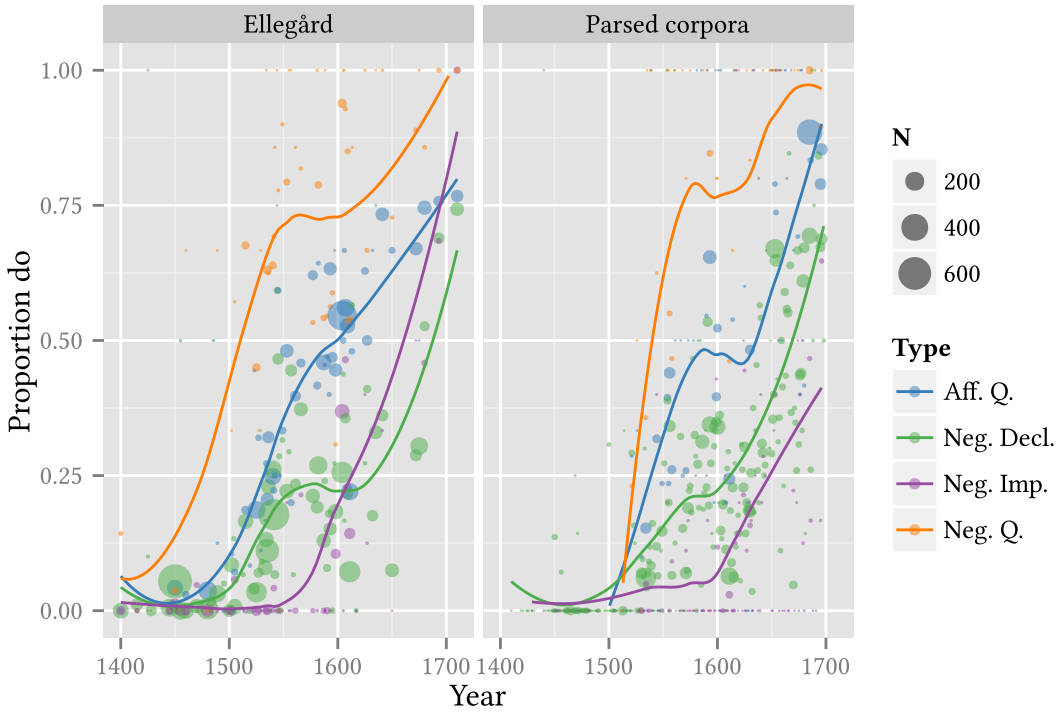



*Figure 4.9:* A comparison of *do*-support in the parsed corpora and Ellegård's data.

Indeed, most of Ellegård's results can be replicated in the corpora, with a high degree of fidelity. Figure 4.9 shows the trajectories of *do*-support in both data sets. The extension of his results on transitivity and lexical classes is in section 4.3 below. In the remainder of this section, I will focus on aspects of Ellegård's data which were not replicated in the corpora.

### Timing of the dip

In Ellegård's corpus, the deflection that occurs around 1575 begins slightly earlier and ends slightly later. Figure 4.10 presents a direct comparison of the trajectory of negative declaratives in both corpora, which makes this difference apparent. Various explanations for this difference may be appealed to. Most basically, despite being seemingly a bigger corpus by token counts, Ellegård actually sampled a smaller number of speakers. There are 109 texts in his corpus (the vast majority, but not all, of which are single-author) whereas there are 903 different authors in the parsed corpora. Thus, Ellegård in some sense sampled much less of the variation in the population.

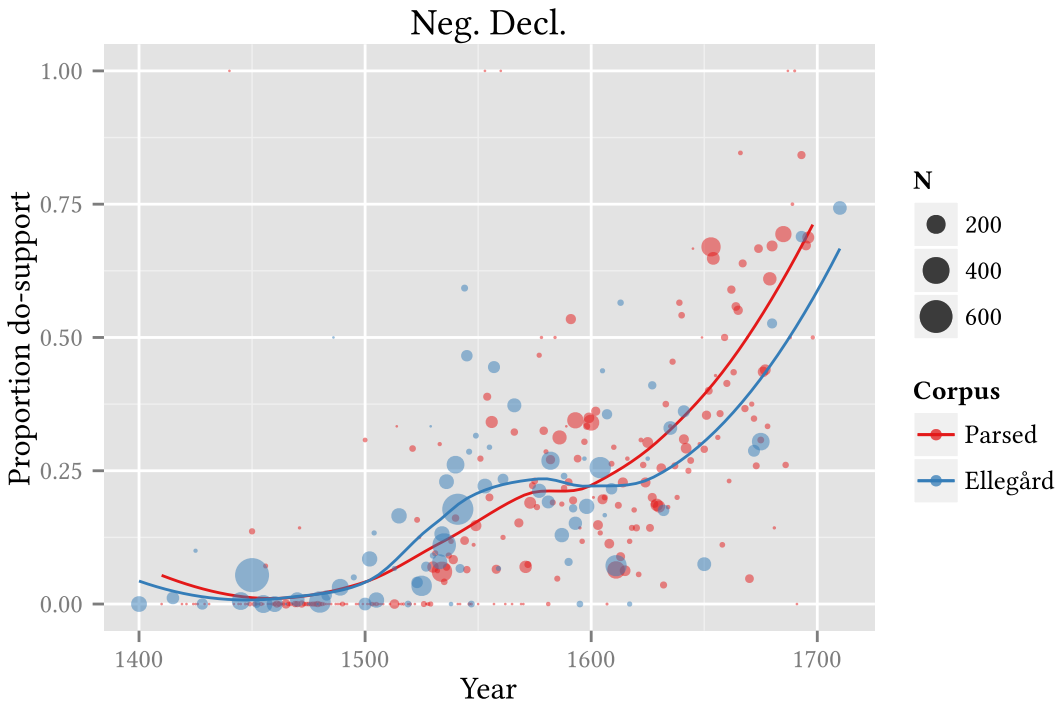



*Figure 4.10:* Negative declaratives in the parsed corpora and Ellegård's data.

Another possible explanation is that Ellegård collected his corpus specifically to study *do*-support, and with a detailed knowledge of the amount of *do*-support that various authors characteristically use. It is possible that he subconsciously biased his sampling towards the collection of "interesting" texts. One way of implementing such a bias would be to seek out texts which most sharply depart from the community pattern, thus sampling innovative texts at the beginning of the change and conservative ones later. Such a biased sampling pattern would produce exactly the lessening of slope that is seen in Ellegård's data. By examining the distribution of dots along the upper and lower limits of the y-axis in Figure 4.9, we can acquire some tentative support for such a view – there are fewer zeroes and ones in Ellegård's data, reflecting a possible dispreference for speakers whose usage is categorical. (An alternative explanation for this pattern however would be that Ellegård's larger samples of tokens per text/speaker make it less likely for him to draw 0 or 1 in a sample.)

#### Behavior of questions

In Ellegård's data, there is a clear and interpretable pattern to the behavior of questions (visible in Figure 4.6): yes-no questions, which do not have a *wh*-gap in the clause, behave differently than *wh*-questions, which do. Furthermore, these question types differ in whether they participate in the deflection of 1575; yes-no questions do not whereas *wh*-questions do. This finding is not replicated clearly in the parsed corpora, as can be seen in Figure 4.11; the reasons for this failure remain unknown. However, in section 4.2.3 we will see that the behavior of affirmative questions is important to other inquiries; thus understanding this difference is important to the investigation.

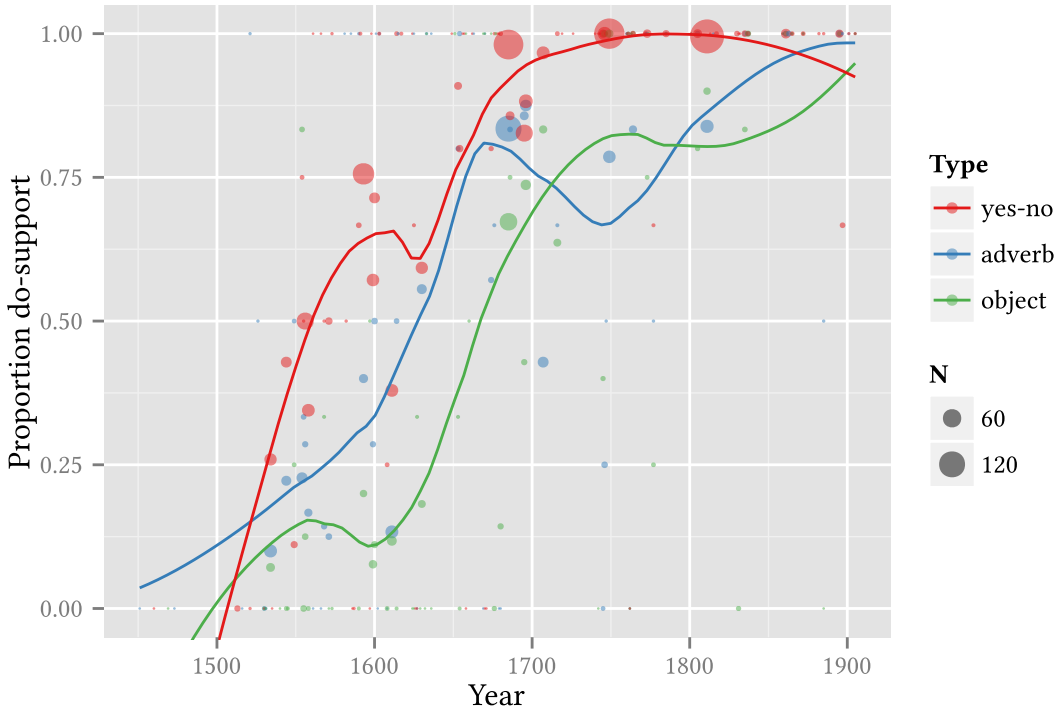



*Figure 4.11:* The behavior of various types of question in the parsed corpora. The $\alpha$ smoothing parameter of the LOESS lines has been set to 0.5.

### 4.2.2 Replication of the Constant Rate Hypothesis analysis

With the understanding of the CRH laid out in section 3.1, it is possible to attempt a replication of the findings of Kroch (1989) on *do*-support using the data from the parsed corpora.

### The adverbial correction term

The first issue which must be addressed is the correction to the slope of verb raising across *never* induced by pre-T adverbs, previously discussed in the neighborhood of example (77). A reanalysis of Kroch's data sources leads to the conclusion that he counted cases where the subject position is empty, either because it is a trace or because the verb is conjoined and shares a subject with a preceding clause. However, this condition triggers a higher rate of pre-T positioning of the adverb. Table 4.2 collects data on the difference between methods of counting the rate of pre-T adverbs with two different classes of verbs which move to T in the PPCHE corpus of EME. The two classes differ from each other in their baseline rate of pre-T positioning, but both obey the generalization that counting empty subjects leads to a higher apparent proportion of pre-T positioning. The overall rate, across all auxiliary types, is given in table 4.3. The proportion pre-T adverbs when counting empty subjects is estimated at 15%, very close to Kroch's estimate of 16%. However, I have elected to treat the estimate obtained by excluding empty subjects as closer to the underlying reality. This is because of the very high rates of pre-T positioning with empty subjects, coupled with the intuition that the lack of a subject may introduce pressures that induce pre-T adverbs to appear at a heightened rate.[9]

*Table 4.2:* Estimates of the proportion of pre-T adverb positioning in the PPCHE (EME portion) with two different classes of verbs which move to T.

| | Modals, auxiliary *have* and *be* | | Main verb *have* and *be* | |
|---|---|---|---|---|
| | w/ empty subjects | w/o empty subjects | w/ empty subjects | w/o empty subjects |
| Pre-Infl | 457 | 153 | 571 | 133 |
| Total | 4874 | 4570 | 1221 | 783 |
| Proportion | 0.09 | 0.03 | 0.32 | 0.15 |

[9]For example, it is plausible that English prosody is accustomed to dealing with XP-Aux sequences. In many cases XP will be the subject; when there is no subject present the placement of an adverb in the pre-T position may allow the default prosody to nonetheless be used. This is clearly a pretheoretic explanation, and indeed I have no insight to offer in this dissertation as to why this effect may obtain. Yet it suffices from examining the data to conclude that empty subject sentences do deviate from the default level of pre-T adverb positioning.

*Table 4.3:* Estimates of the proportion of pre-T adverb positioning in the entire PPCHE (EME portion).

| | Counting empty subjects | Excluding empty subjects |
|---|---|---|
| Pre-Infl | 1226 | 368 |
| Total | 6994 | 6136 |
| Proportion | 0.15 | 0.06 |

I have further chosen not to take this 6% correction factor into account in my calculations. It was easy for Kroch to accommodate the correction in his data, since he was working from Ellegård's data which was reported only over wide time bins. With the relatively exact year-by-year dating of texts available in the PPCHE and the more detailed consideration of other regression predictors in this dissertation, it becomes difficult to incorporate the correction term in the calculations. Furthermore, as Kroch demonstrated, a regression experiment of roughly the size of Ellegård's corpus (or the PPCHE) is not sensitive to variations in the value of the correction term of 5–10 percentage points. Thus, the calculated 6% is close enough to zero that it should be safe to disregard, on a provisional basis.[10]

**Overview of the data**

The plot in Figure 4.12 shows the underlying corpus data. Three modern *do*-support contexts are plotted, as well as the loss of verb raising past *never* (that is, 1 − the rate of overt raising). The dashed vertical line is placed at 1575, a point after which (following Kroch) we will discard the *do* data, as it departs from a monotonic upward trajectory assumed by the CRH model at that point (for reasons which are argued by Kroch (1989) and Warner (2005) to be extrinsic to the change). From visual inspection, it is not immediately implausible to think that the curves described by the data are parallel.

[10]The way to incorporate this correction in a regression with other covariates should involve either a simulation procedure which repeatedly recalculates the regression after dropping 6% of the relevant data, or a Monte Carlo-simulated Bayesian model. Both techniques are computationally intensive and require additional conceptual frameworks for the interpretation of their results. The conclusions of this dissertation must ultimately be compared to the output of such procedures, however it is not within the scope of the present work to do so.

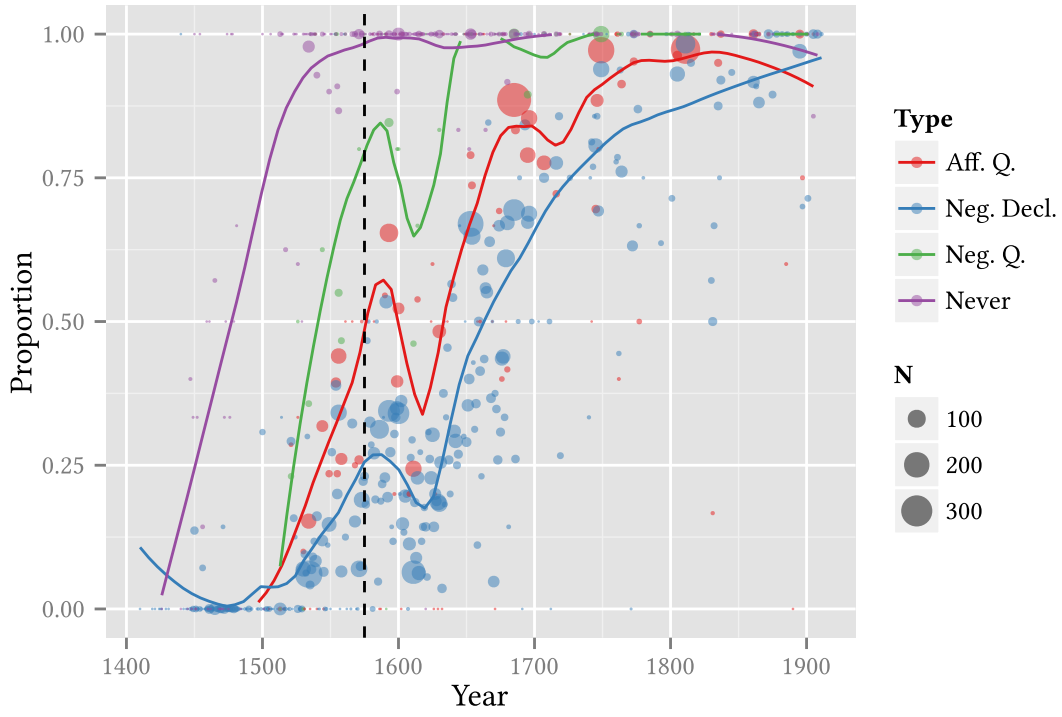



*Figure 4.12:* A plot, using data drwan from the PPCHE, of several *do* support environments alongside the incidence of failure of verb raising past *never*. The $\alpha$ parameter of the LOESS smoother is set to 0.3. A vertical dashed line is placed at 1575.

*Table 4.4:* An estimate of the duration of the emergence of *do*-support in years from a model fit to the PPCHE data from before 1575.

| | Aff. Q. | Neg. Q. | Neg. Decl. |
|---|---|---|---|
| Estimated duration (years) | 202 | 117 | 234 |
| Diff. with Aff. Q. | — | 85 | -32 |

Table 4.4 reports the slope estimate in years (as described in section 3.4.1) from a model fit to this data. The affirmative question and negative declarative contexts are estimated to be relatively close to each other (though not as close as was found by Kroch 1989). The negative question context's estimated slope clearly differs, though it is also poorly estimated (being fit from little data).

*Table 4.5:* Comparison between a model which assumes the CRH (left) and non-CRH-assuming models. The models are fit to data from the PPCHE. From left to right, these are two models with three-way varying slope: one that holds the slope constant between the types of questions (affirmative and negative) but allows variation in the other contexts; one that holds the slope constant across negatives (questions and declaratives). The rightmost model is one that allows full four-way slope variation across contexts. Non-parenthesized values are coefficient estimates; parenthesized values are standard errors. The stars report significance at the 0.05 (one star), 0.01 (two) and 0.001 (three) $\alpha$ levels.

| | Reduced | Declarative | Affirmative | Full |
|---|---|---|---|---|
| Intercept | 0.33 (0.08)$^{***}$ | 0.38 (0.10)$^{***}$ | 0.33 (0.09)$^{***}$ | 0.43 (0.11)$^{***}$ |
| Year | 2.33 (0.17)$^{***}$ | 2.45 (0.24)$^{***}$ | 2.45 (0.24)$^{***}$ | 2.77 (0.36)$^{***}$ |
| Sum: Neg. Decl. | −1.86 (0.09)$^{***}$ | −1.96 (0.12)$^{***}$ | −1.89 (0.10)$^{***}$ | −2.01 (0.13)$^{***}$ |
| Sum: Aff. Q. | −0.91 (0.11)$^{***}$ | −0.88 (0.12)$^{***}$ | −0.90 (0.14)$^{***}$ | −1.00 (0.15)$^{***}$ |
| Sum: Neg. Q. | 0.28 (0.16) | 0.35 (0.18) | 0.25 (0.17) | 0.57 (0.26)$^{*}$ |
| Slope: Decl. | | −0.37 (0.34) | | |
| Slope: Q. | | 0.45 (0.67) | | |
| Slope: Aff. | | | −0.05 (0.75) | |
| Slope: Neg. | | | −0.28 (0.34) | |
| Slope: Neg. Decl. | | | | −0.70 (0.40) |
| Slope: Aff. Q. | | | | −0.37 (0.62) |
| Slope: Neg. Q. | | | | 1.39 (0.92) |
| AIC | 2057.15 | 2059.04 | 2060.50 | 2059.42 |
| BIC | 2088.11 | 2102.39 | 2103.85 | 2108.96 |
| Num. obs. | 3614 | 3614 | 3614 | 3614 |

*Table 4.6:* Comparison between a model which assumes the CRH (left) and models that allow one context each to differ in slope from the other three.Non-parenthesized values are coefficient estimates; parenthesized values are standard errors. The stars report significance at the 0.05 (one star), 0.01 (two) and 0.001 (three) $\alpha$ levels.

| | Reduced | Neg. Decl. | Aff. Q. | Neg. Q. | Never |
|---|---|---|---|---|---|
| Intercept | 0.33 (0.08)$^{***}$ | 0.35 (0.09)$^{***}$ | 0.33 (0.09)$^{***}$ | 0.42 (0.10)$^{***}$ | 0.33 (0.09)$^{***}$ |
| Year | 2.33 (0.17)$^{***}$ | 2.51 (0.22)$^{***}$ | 2.32 (0.17)$^{***}$ | 2.29 (0.17)$^{***}$ | 2.20 (0.24)$^{***}$ |
| Sum: Neg. Decl. | −1.86 (0.09)$^{***}$ | −1.94 (0.11)$^{***}$ | −1.86 (0.09)$^{***}$ | −1.96 (0.11)$^{***}$ | −1.88 (0.10)$^{***}$ |
| Sum: Aff. Q. | −0.91 (0.11)$^{***}$ | −0.91 (0.11)$^{***}$ | −0.90 (0.14)$^{***}$ | −1.00 (0.13)$^{***}$ | −0.92 (0.11)$^{***}$ |
| Sum: Neg. Q. | 0.28 (0.16) | 0.29 (0.16) | 0.28 (0.16) | 0.58 (0.26)$^{*}$ | 0.26 (0.16) |
| Slope: Neg. Decl. | | −0.43 (0.34) | | | |
| Slope: Aff. Q. | | | 0.08 (0.74) | | |
| Slope: Neg. Q. | | | | 1.87 (1.21) | |
| Slope: Never | | | | | 0.25 (0.33) |
| AIC | 2057.15 | 2057.51 | 2059.14 | 2056.61 | 2058.59 |
| BIC | 2088.11 | 2094.66 | 2096.29 | 2093.77 | 2095.74 |
| Num. obs. | 3614 | 3614 | 3614 | 3614 | 3614 |

**Model comparisons**

We can create a model to test the CRH in this data set – that is, to test whether the slopes of the various contexts differ. In Kroch's formulation, there are two models implicated in this comparison, the leftmost and rightmost models of Table 4.5. The lack of significance of the slope coefficients is the traditional diagnostic, and on this criterion the CRH is upheld. The values of the AIC and BIC statistics are also included in the table (recall that a lower AIC or BIC indicates a closer correspondence between model and data).[11] The AIC advantage of the leftmost model is almost exactly 2, which is right on the threshold of a meaningful difference (as discussed in section 3.3.3). The BIC differences are larger – unsurprisingly, since it penalizes the addition of extra degrees of freedom more harshly than the AIC. Thus, the PPCHE data up to 1575 replicate the finding of Kroch (1989) that there is a CRH effect between the various *do*-support contexts and the loss of verb-raising past never.

There are two intermediate models included in the comparison as well. The full model adds three parameters to the model, which means it incurs a penalty of a certain size in the information criterion. If there is some difference between the slopes in various contexts, but not enough to justify the addition of three parameters, then the non-CRH model might be thought to have been unfairly rejected. We should test models that add one or two parameters to the base model, in addition to the full model which adds three parameters. The intermediate models in Table 4.5 seek to test such cases by constructing two linguistically

[11] The small-sample correction to the AIC is not used in the table, however its effect is on the order of one hundredth in this data.

plausible intermediate models that have three slope groups (two additional parameters):

1. negatives (questions and declaratives) / affirmatives (questions) / never
2. questions (affirmative and negative) / declaratives (negative) / never

Neither of these intermediate models achieves significant $p$-values nor favorable information criteria. Table 4.6 compares a different class of models to the same-slope model. These models each allow the slope of one context to differ with respect to the other three – they have two slope groups and thus one extra parameter as compared to the base model. None of them is clearly rejected by the information criteria (especially the AIC), nor is any of them accepted – they all appear to perform about as well as models of the corpus data (with the possible exception of the model which allows affirmative questions alone to differ from the aggregate of the other three contexts). But these models are *ad hoc*, in the sense that there is no linguistic reason to believe that (say) negative questions should differ in their slope from the set {affirmative questions, negative declaratives, verb raising over *never*}. (It is also possible that the uneven distribution of tokens across contexts is causing problems: negative questions are the least-well-sampled environment, but also the one that the model finds to be most different from the others.)

*Table 4.7:* Likelihood ratio test results between various alternative models and the CRH-assuming equal-slopes model.

| Model | LRT $p$-value |
|---|---|
| Declarative | 0.37 |
| Affirmative | 0.74 |
| Neg. Decl. | 0.22 |
| Aff. Q. | 0.92 |
| Neg. Q. | 0.11 |
| Full | 0.3 |

Another way to compare models is using a likelihood ratio test, which provides a frequentist $p$-value for the null hypothesis that the CRH-verifying model fits the data better, against the hypothesis that the alternative model is a better fit. The $p$-values for the various models considered above tested against the equal-slopes model which builds in the CRH is given in Table 4.7. None of the alternative models is an improvement over the CRH model on this diagnostic.

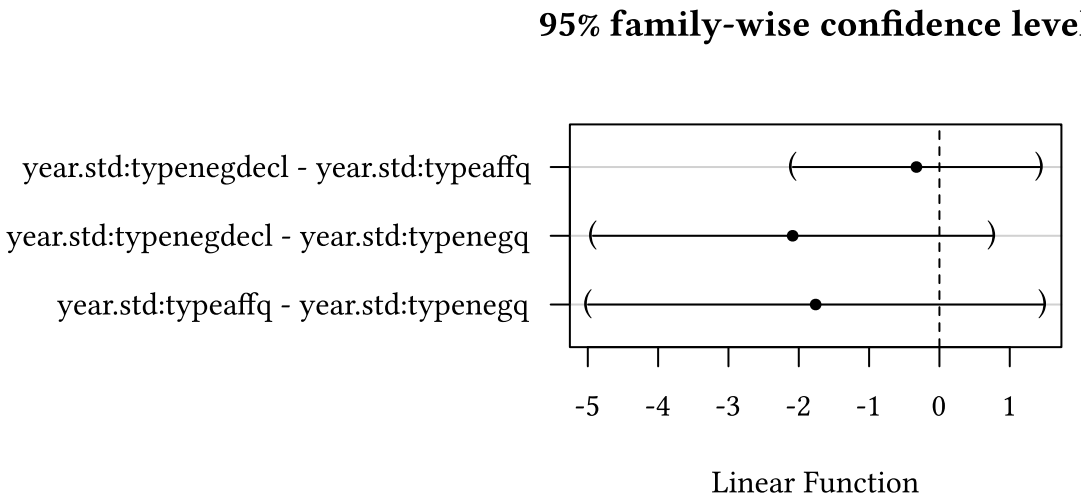



*Figure 4.13:* Confidence intervals for the differences between slope coefficients in the full model of the evolution of *do*-support and verb raising over never. (Calculated by the `multcomp` package in R.)

A final model comparison methodology uses the multiple comparisons approach outlined in section 3.3.3 to fit the maximal model (with four slope groups) and calculate confidence intervals for the differences in the slope coefficients directly. If the confidence intervals exclude zero, then we can conclude that there is evidence that the slopes of these two contexts differ. Figure 4.13 graphically presents the result of applying such a test to the full model. None of the resultant confidence intervals excludes zero; thus no evidence against the CRH is obtained.

**Power analysis**

In the below section the implications of power analysis for the *do*-support results of Kroch (1989) will be discussed, under a variety of assumptions. (For simplicity, the results pertaining to *never* will not be discussed).

*Table 4.8:* The sizes of the effects in Tables 4 and 9 of Kroch (1989), measured in logit units and years of change.

| | Slope | | Diff. with ND | |
|---|---|---|---|---|
| Context | logit/year | dur. of chg. | logit/year | dur. of chg. |
| Negative declarative | 0.04 | 157.46 | — | — |
| Negative questions | 0.03 | 170.69 | 0.00 | 13.24 |
| Aff. trans. qs | 0.04 | 162.68 | 0.00 | 5.22 |
| Aff. intrans. qs | 0.04 | 156.20 | 0.00 | −1.25 |
| Aff. *wh*-obj. qs | 0.04 | 146.85 | 0.00 | −10.60 |
| Aff. decl. | 0.03 | 208.83 | −0.01 | 51.37 |

To begin with, table 4.8 contains the effects for the CRH results on *do* from Kroch (1989). The contexts which were not detected to differ significantly in slope all have slope differences on the order of 10 years of change-duration. The affirmative declarative context was found to differ significantly, by ~50 years of change-duration. My intuition is that these results, in addition to falling on either side of the significance boundary, also differ in meaningfulness. A difference of 50 years (over a change that takes place in 150–200 years) is an important one, whereas a difference of about 10 is not. This analysis echoes an observation that Kroch makes: that all the slope estimates except for affirmative declaratives are within 15% of their common median. Though he does not elaborate on this observation, it seems that he had in mind a broadly similar notion of meaningfulness, which is in principle independent of statistical significance (though in his experiment the two notions happen to align).

*Table 4.9:* A power test of the CRH test for *do*-support in Kroch (1989). The proportion of likelihood ratio tests which reported a significant year×clause type interaction when the real difference between affirmative questions (reference level), negative declaratives (rows) and negative questions (columns) was as reported in the table. The other parameters (intercept, main effects of year and clause type) were as estimated from a regression on the original data. The year variable was $z$-centered, and thus the interaction values along the edges of the table are not denominated in meaningful units.

| | | Neg. Q. | | | | | |
|---|---|---|---|---|---|---|---|
| | | -0.5 | -0.4 | -0.3 | 0.3 | 0.4 | 0.5 |
| Neg. Decl. | -0.5 | 0.97 | 1.0 | 0.99 | 1.0 | 1.0 | 1.0 |
| | -0.4 | 0.86 | 0.89 | 0.85 | 0.99 | 0.98 | 1.0 |
| | -0.3 | 0.6 | 0.62 | 0.55 | 0.85 | 0.98 | 0.96 |
| | -0.2 | 0.54 | 0.36 | 0.36 | 0.67 | 0.86 | 0.86 |
| | -0.1 | 0.34 | 0.24 | 0.14 | 0.35 | 0.57 | 0.72 |
| | 0.1 | 0.52 | 0.35 | 0.32 | 0.22 | 0.35 | 0.4 |
| | 0.2 | 0.8 | 0.74 | 0.45 | 0.43 | 0.48 | 0.57 |
| | 0.3 | 0.85 | 0.82 | 0.8 | 0.51 | 0.64 | 0.66 |
| | 0.4 | 0.95 | 0.94 | 0.9 | 0.83 | 0.81 | 0.81 |
| | 0.5 | 0.99 | 1.0 | 0.97 | 0.91 | 0.91 | 0.89 |

After setting up these boundaries of meaningfulness, a simulation was performed to verify the power of the test of the CRH with respect to *do*-support from Kroch (1989). As discussed in Section 3.4 above, the simulated data were sampled independently from the columns of Ellegård's dataset corresponding to negative declaratives and affirmative and negative questions (excluding marginal clause types). The $\alpha$ and sample size were fixed as in Kroch's original experiment (0.05 and 6644 respectively). As in Kroch's original experiment, instead of using exact years, they were divided into the bins used in Ellegård's tables; the midpoint of each bin was used as the date for all texts in that bin. However, some deviations from the original procedure were implemented, namely centering of the year variable (as described in section 3.3.1). Treatment contrasts (section 3.3.2) were also used, with affirmative questions as the reference level. The simulation was run 100 times. For each simulated dataset, a logistic regression was fit, and the $p$-value was recorded of a likelihood ratio test comparing the model with a year×clause type interaction to one without. The proportion of such simulations where the $p$-value was below the chosen $\alpha$ is recorded in table 4.9. This indicates that the test has (roughly) power above the 0.95 threshold (corresponding to the traditional value for $\alpha$) to detect slope differences in negative declaratives of 0.5 logit units = 57.23 years of acceleration and 0.4 logit units = 94.16

years of deceleration, but not of 0.4 logit units = 48.56 years of acceleration nor 0.3 logit units = 63.21 years of deceleration. The performance of the model on other values is intermediate, and depends on the spread of values. For example, the model has greater power when there is greater difference between the slopes by context; that is, when one has a positive and one a negative interaction term. It has very no power to detect even large (~50 year) slope differences in negative questions in the absence of a large difference in negative declaratives. Indeed, since the simulation only recorded an overall LRT rather than per-context tests, the significant results in these cases are almost certainly driven by the negative declarative context.

Kroch's study has no power to detect differences in negative questions; indeed arguably this context should be excluded from evaluations of the predictions CRH on data of this size. The study was also moderately underpowered with respect to negative declaratives (the most abundant context), since it can barely detect differences in the ballpark of 50–75 years. Considering Kroch's results on affirmative declaratives indicate that true differences of roughly this magnitude should be treated as meaningful, and thus it is desirable to have the power to detect them. The differences among modern *do*-support contexts estimated by Kroch's study – and the present one – are smaller than this; thus there is reason to suppose that the study's lack of power did not adversely affect its results in practice. This conclusion is bolstered by the replication of the result in the PPCHE, though this is still somewhat equivocal (in the sense that the CRH model barely edged out alternative models, rather than presenting a noticeable improvement over them).

### 4.2.3 Replication of Warner's results

As the discussion in section 4.2.1 demonstrates, there are differences in the behavior of questions in the parsed corpora and those in Ellegård's corpus. Thus, reexamining Warner's analysis in light of the data from the parsed corpora will be seen to lead to different conclusions about the impact of social factors on the change.

#### Measuring style

Warner uses two measurements for style: average word length of a text, and type-token ratio. However, there are some problems that must be overcome when replicating Warner's experiements. Neither of these measurements is amenable to automatic calculation, as the parsed corpora are not lemmatized. However, it turns out that there is a nearly perfect linear relationship between these measurements as computed over lemmatized text and their cruder counterparts which treat each spelling variant as a novel lemma. A more serious worry is that the observed type-token ratio varies according to the length of the text sampled (Warner

overcame this problem by using uniformly 600-word long samples of texts). Warner reports that type-token ratio and word length correlate well with each other (he does not report the correlation in numerical terms). It's unclear whether this is true in the PPCHE data; the correlation is shown graphically in Figure 4.14. The correlation coefficient (Pearson's $R$) is 0.36; $R^2$ = 0.13.

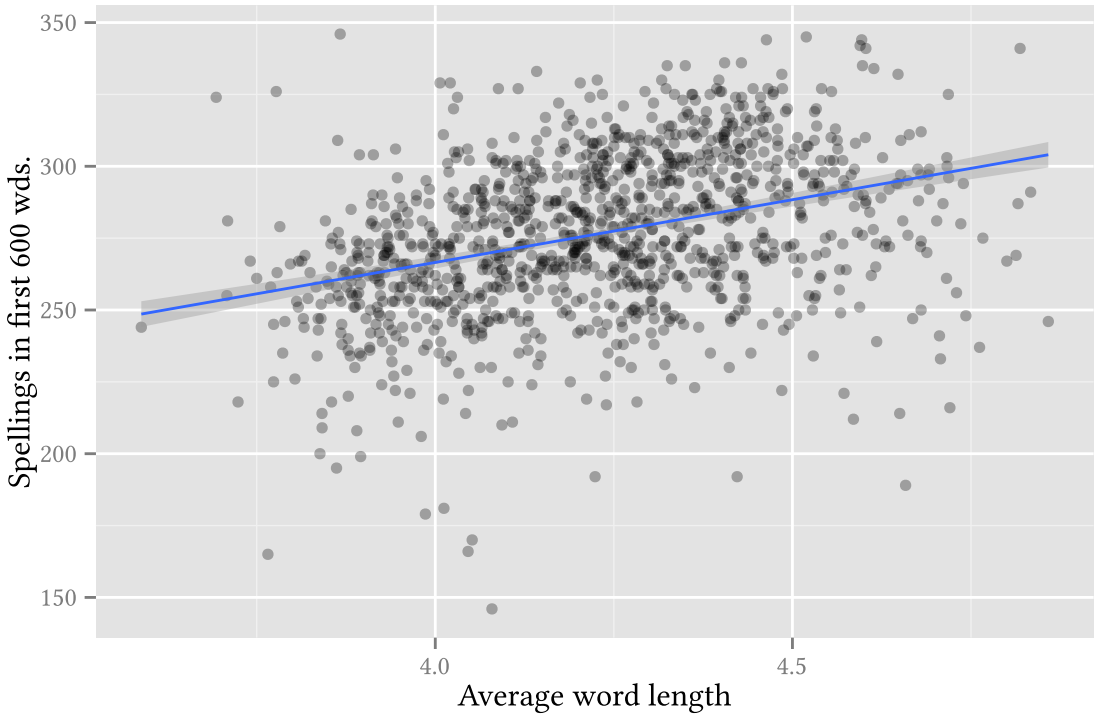



*Figure 4.14:* Relationship between two style measures proposed by Warner (2005) in data from the parsed corpora (only texts longer than 600 words).

*Table 4.10:* Information criterion model comparisons between models which include and exclude average word length as a predictor of *do*-support usage. A negative value means the model including the extra predictor has a lower *IC value. The model was fit using the `glm` function in R, with the formula `do support ~ clause type + year (+ word length)`; the year and word length variables were standardized to z-scores.

| | Pre-1575 | Post-1575 |
|---|---|---|
| ΔAIC | -26.44 | -32.99 |
| ΔBIC | -21.13 | -27.23 |

*Table 4.11:* Information criterion model comparisons between models which include and exclude type-token ratio as a predictor of *do*-support usage. The details of the table construction are identical to those of Table 4.10.

| | Pre-1575 | Post-1575 |
|---|---|---|
| ΔAIC | 1.81 | -90.7 |
| ΔBIC | 7.12 | -84.95 |

Word length is a good predictor of *do*-support usage both before and after 1575. Table 4.10 shows this fact by information criterion-based model comparison results. Type-token ratio, on the other hand, only predicts *do*-support usage well after 1575. Thus, I'll use only average word length in the discussion to follow.

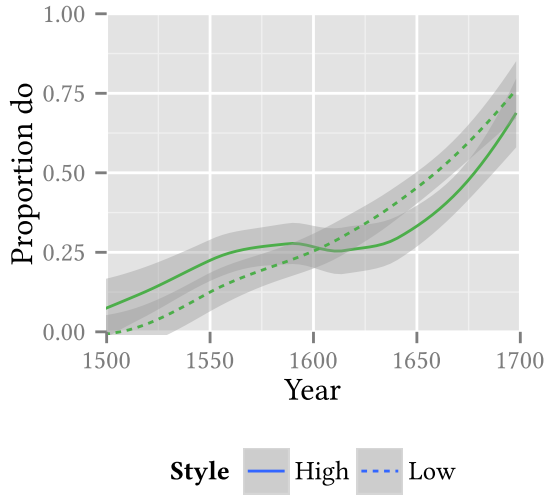



*Figure 4.15:* The behavior of negative declaratives in the high- and low-word-length halves of the parsed corpora.

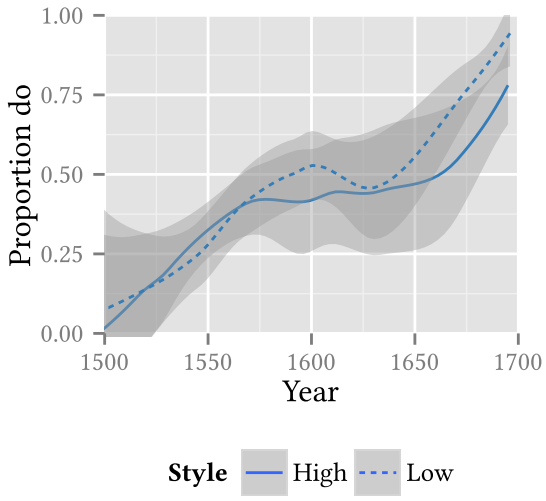



*Figure 4.16:* The behavior of affirmative questions in the high- and low-word-length halves of the parsed corpora.

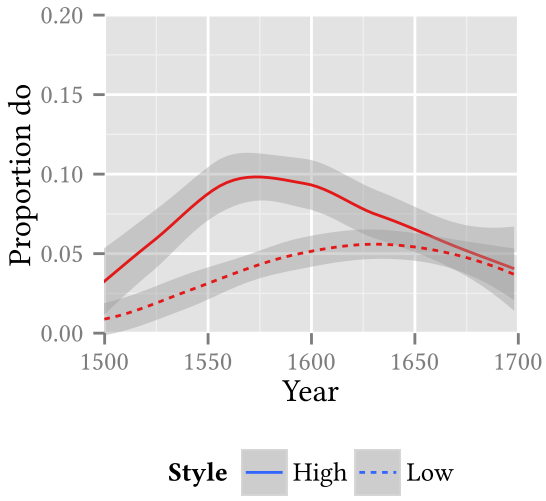



*Figure 4.17:* The behavior of affirmative declaratives in the high- and low-word-length halves of the parsed corpora.

I'll follow Warner in splitting the corpus into high- and low-style halves at the median value of average word length, and comparing the behavior of *do*-support in these two subsets. As Figure 4.15 shows, high-style negative declaratives participate in a deflection, whereas in the low-style condition the rise of *do*-support usage is monotonic and continuous. Figure 4.16 shows that the behavior of affirmative questions differs

from Warner's predictions: not only is a deflection evident, but it manifests itself in both stylistic contexts. Finally, Figure 4.17 shows that there is a style effect which favors *do* in affirmative declaratives early in the change, but it disappears after 1575.

The foregoing results test Warner's observations of the effect of style on *do*-support. He also tested for age-grading, another sociolinguistic phenomenon. His procedure was to fit two regression models to the data before and after 1575, and then examine whether age was a significant predictor in either model. We can imitate this procedure using corpus data.

*Table 4.12:* Models testing the predictions of Warner (2005) on the presence of age-grading in negative declaratives in the periods pre-1575 and 1575–1700.

| **Period** | Coefficient | $p$-value | ΔAICc | N tokens |
|---|---|---|---|---|
| **Pre** | 0.18 | 0.06 | −1.47 | 1332 |
| **Post** | −0.30 | $5.15 \cdot 10^{-15}$ | −61.31 | 3338 |

Table 4.12 shows the result of this procedure fit to data on negative declaratives. As the table indicates, the linear coefficient of age (standardized by $z$-score) makes a significant contribution to the performance of the model in the 1575–1700 period, as measured by $p$-value and AIC. In the period before 1575, however, this contribution is not evident. This is Warner's result.

*Table 4.13:* Models testing the presence of age-grading of *do* usage in affirmative declaratives in the periods pre-1575 and 1575–1700, following the procedure given by Warner 2005.

| **Period** | Coefficient | $p$-value | ΔAICc | N tokens |
|---|---|---|---|---|
| **Pre** | 0.07 | 0.02 | −3.30 | 30734 |
| **Post** | 0.02 | 0.18 | 0.19 | 73598 |

Table 4.13 extends Warner's procedure to affirmative declaratives. There, the situation is precisely reversed. An age-grading effect exists before 1575, but does not carry over to the later period.

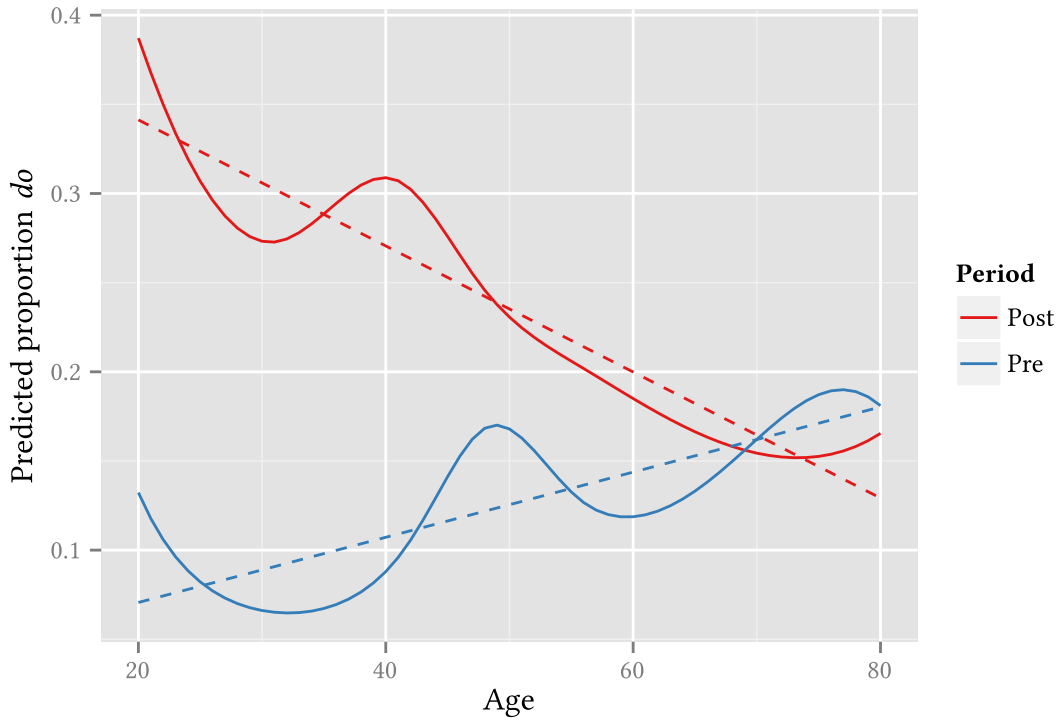



*Figure 4.18:* Predicted age effects of a model with non-linear effects of age and time in the years 1550 and 1600.

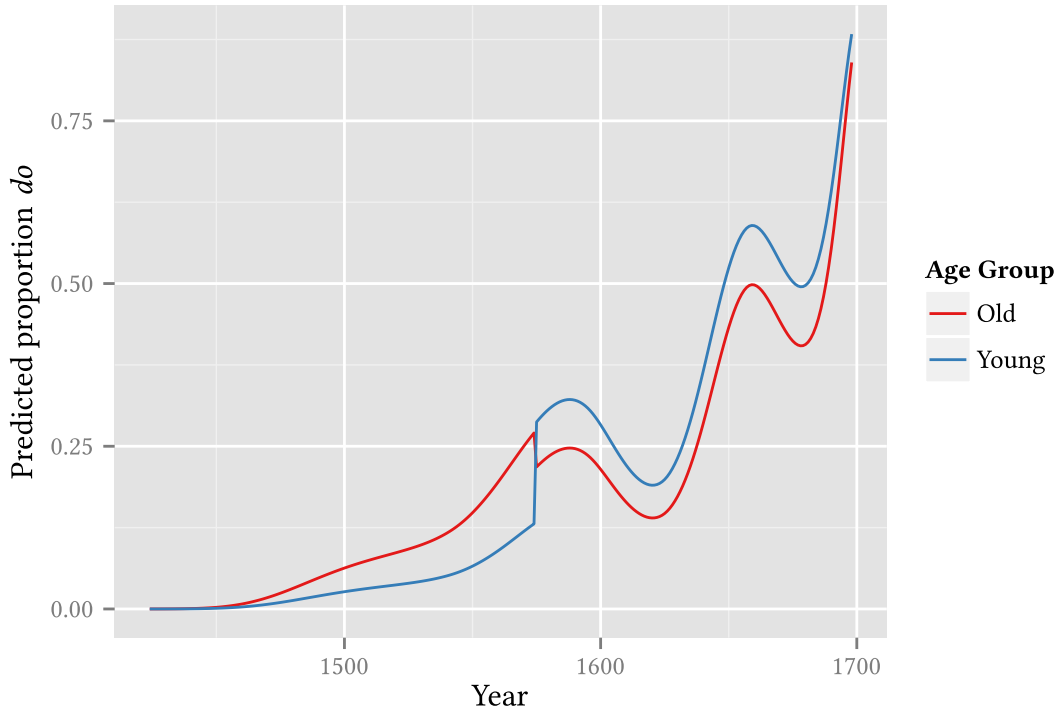



*Figure 4.19:* Predicted trajectory of *do*-support in a model with non-linear effects of age and time in the years 1550 and 1600. The two age groups are the 25th and 75th percentile of ages represented in the data.

It is also possible to illustrate this age-grading effect through a more complex regression procedure. I fit a logistic regression model to the PPCHE data on negative declaratives from the beginning of the dataset to 1700, with a non-linear effect (B-spline) with 7 degrees of freedom for both year and age. The age spline was allowed to interact with an indicator variable which was 0 before 1575 and 1 after.[12] Figure 4.18 shows the age effect that this model predicts in 1550 and 1600 (that is, 25 years on either side of the 1575 division). As is apparent, the age effect on 1550 is relatively shallow, and favors *do* usage among older authors. On the other hand, in 1600 the slope is steeper, and negative (meaning that older authors disprefer *do*, relatively speaking). Figure 4.19 gives another view of the model's predictions. It shows the behavior of authors at the 25th and 75th percentile of ages in the data (34 and 53 respectively). The young authors make a large adjustment in favor of *do*-support in 1575, whereas the older authors make only a slight downwards adjustment. Both of these findings are in agreement with Warner's description of the effect.

[12]The R formula used to fit this model was: `do.supp ~ bs(year.std, df = 7) + bs(age.std, df = 7) * post1575`.

**Discussion**

These results bear up most of Warner's core findings. The role of lexical complexity as a factor in determining *do*-support usage is replicated in the PPCHE. Furthermore, Warner's generalization that *do*-support is a feature characteristic of high lexical complexity texts before 1575 and of low complexity ones afterwards is upheld, under the condition that we consider style through the lens only of average word length. However, the behavior of affirmative sentences – especially questions – casts doubt on the conclusion that it is evaluation of *n't* contraction which drives the deflection of 1575. Rather, given that in the PPCHE dataset the lexical complexity effect affects affirmatives and negatives alike, the evidence points in the direction of a social evaluation of *do*-support directly.

## 4.3 Intermediate *do*

In this section, I will present a novel analysis of the emergence of *do*-support in EME.

This account improves on the status quo of the understanding of this construction's history in the following ways:

1. It explains the relative prevalence (compared to PDE) of EME affirmative declarative sentences with *do*-support. Previous accounts did not fail to notice the availability of such sentences, of course (by raw occurrence frequency, they are the most abundant type of auxiliary *do* usage for much of the EME period). Ellegård considers and discards a variety of previously proposed explanations based on phonological (meter or euphony) and processing (difficulty recalling the past tense of strong verbs) considerations. Ultimately he tentatively adopts a view that affirmative declarative *do* serves to smooth the transition from a +V-to-T grammar to a –V-to-T with respect to the linear position of adverbs. This explanation is difficult to cast in structural terms. It also fails to be totally satisfactory without a more rigid notion of what the difficulties surrounding adverb placement are (since speakers of PDE are able to make full use of adverbs without availing themselves of auxiliary *do*). For his part, Kroch relies on a stipulation that the ratio of sentences generated by affix hopping to those generated by *do* insertion be constant across the EME period to 1575. The waning of V-to-T then creates a steadily increasing number of sentences for which one of these options must be chosen. Kroch shows that under this assumption, the rate of loss of V-to-T raising implied by levels of affirmative declarative *do* obeys the CRH with respect to the other *do*-support contexts and overt raising across never. While this is a promising statistical result, it fails to explain why *do*-support should be used in affirmative

declaratives at all (unlike in the PDE *do*-support contexts, where affix hopping is ungrammatical). It also cannot escape the fundamental arbitrariness of the assumption that affix hopping and *do*-support should maintain a constant proportion until 1575. The present account proposes a coherent evolution of *do* from its roots in ME as a causative. It follows an upwards structural trajectory and undergoes gradual semantic bleaching, as described by work in Grammaticalization theory (Hopper and Traugott 1993).

2. The present account furthermore allows the observations about the differing behavior of certain lexical classes with respect to *do*-support to be explained. While the explanation is far from complete (see chapter 5 for much more detail), palpable progress can be made, again by reference to the function of *do* as a causative in ME.

In the following sections, I'll review five pieces of evidence that this account is on the right track, in addition to spelling out in structural detail the analysis that this evidence suggests.

### 4.3.1 Cooccurrence with other ME causatives

*Do* was originally a causative in Middle English. In that function, it has the synonym *make*, which each have different geographic distributions, as shown in the map in figure 4.20. The centroids of the distributions of each causative demonstrate that *make* is characteristic of western dialects whereas *do* predominates in the east. (However, as the individual points show, this tendency is not categorical, and there are ample attestations of each causative throughout the data.) Causative *let* is also attested in the PPCME2 at a high frequency in all locations, though it is most concentrated in the northwest quadrant of the data; this is not shown on the map for purposes of clarity. In northern dialects, the causative *gar* was also used; this is also not shown on the map, nor reflected in the PPCME2 data.

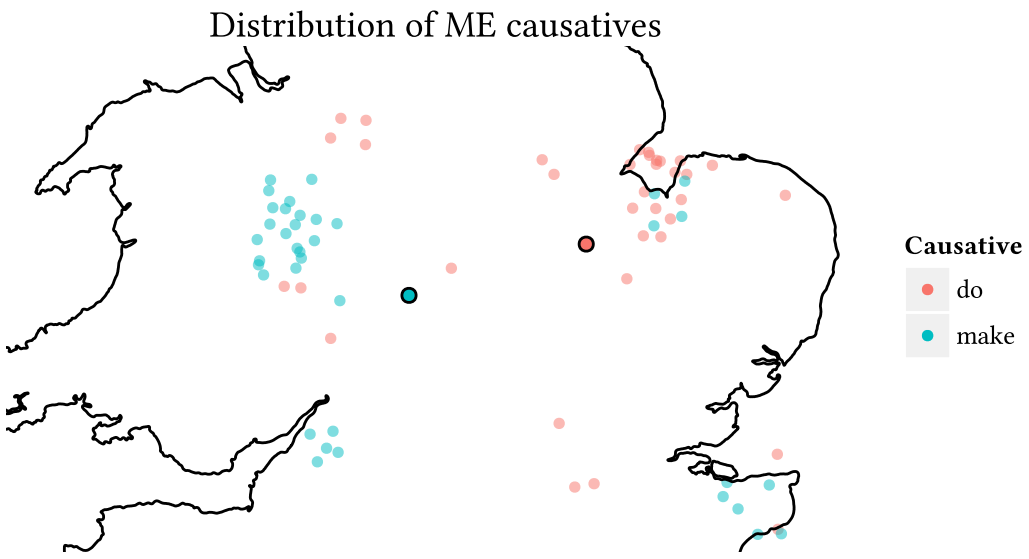



*Figure 4.20:* Causative sentences with a non-overt causee formed with *let* and *do* in the PPCME2 (Kroch and Taylor 2000). The solid points with a black border are the centroid of each cloud of points. (To avoid excessive overplotting, a small amount of random noise is added to each point, which is why some data from coastal regions is located in the nearby ocean. Thanks to Hilary Prichard for the geocoding of the PPCME2 texts on which this map is based.)

One theory about the origin of *do*-support involves language contact between the different ME dialects. Under this account, speakers of western ME heard *do* causatives produced by eastern speakers, but (knowing that, in their dialect causatives are formed by *make*) misinterpreted them as tokens of a pleonastic auxiliary verb *do*. This auxiliary was then pressed into service in *do*-support constructions in EME. The first attested tokens of *do*-support indeed occur in western texts, bolstering this account.

Putting aside the question of the origin of *do*-support, it is clear from the data that towards the end of the Middle English period, *do* begins to be bleached of its causative meaning. The first indication of this development comes from instances of *do* occurring with other causatives, together contributing only one causative meaning to the sentence. Some examples follow:

(78) He leet the feste of his nativitee
Don cryen thurghout Sarray his citee,
'He had the feast of his birthday cried throughout Surrey, his city.'
(Chaucer *Canterbury Tales* "The Squire's Tale" c. 1400)

(79) gret plentee of wyn þat the cristene men han don let make

'Great plenty of wine that the Christian men have (caused to be) made.'

(PPCME2, CMMANDEV,47.1161 a. 1425)

These examples on their own might be amenable to an analysis under which both *let* and *do* together constitute a single causative (each one contributing some portion of the overall semantics of the construction). However, it is fully possible for two instances of *do* to co-occur in the same sentence:

(80) And thus he dide don sleen hem alle three.

'And thus he had all three of them killed.' (Chaucer, *Canterbury Tales* "Summoner's Tale" c. 1400)

This sharpens the demonstration that the bleaching of *do*'s meaning is truly complete. There is no *let* present in this sentence; the most plausible analysis is that one of the two *do*s is causative whereas the other is pleonsatic.

Further data concerns the co-occurrence of bleached *do* with other auxiliaries in non-causative contexts. These examples begin to appear around 1500:

(81) He hes done petuously devour

the noble Chaucer of makaris flour

'[Death] has piteously devoured the noble Chaucer, flower of makars [=bards]'

(Wm. Dunbar "Lament for the Makars" c. 1505)

(82) consequently it wyll do make goode drynke

'Consequently [barley] will make good drink' (A. Boorde *Introduction of Knowledge* a. 1542)

Specifically, these examples show that *do* is merged lower than T (the base position of modals) and lower than Asp (the base position of aspectual *have*).

Finally, we can adduce an example of *do* occurring below the nominalizer *ing*:

(83) Fro the stok ryell rysing fresche and ying

But ony spot or macull doing spring

'From the royal stock rising fresh and young / without any spot or blemish springing'

(Wm. Dunbar *The Thrissill and the Rois* 1503, in Visser (1963, §1419))

This example contains a *do* which is merged very low in the functional structure indeed.

Taken together, these attestations demonstrate that by 1400, *do* has been bleached of its causative meaning, and can co-occur with other causatives. Further, by the early 1500s, this bleached *do* is found in environments other than causatives, indicating that it has become an independent, low-merged auxiliary verb.

### 4.3.2 Distribution relative to adverbs

The previous section gave a distributional argument that semantically bleached *do* appears in a low position in late ME and early EME. This section will give another distributional argument of a slightly different flavor – instead of relying on single tokens which demonstrate the possibility of a certain construction, we will examine the relative frequency of certain constructions in corpus data.

Modals (which are merged in T throughout EME and through PDE) and auxiliary verbs like perfect *have* (which categorically moves to T during the same period) almost always precede clause-medial adverbs (those which are merged to the right of the subject but to the left of the lexical verb).[13] As illustrated in figure 4.21, for both modals and perfect *have* there is a roughly 5% rate of preceding adverbs consistently throughout the EME period. If auxiliary *do* consistently has its modern distribution in EME, we should expect it to follow this same distribution. It is not surprising that there is no data on auxiliary *do* in the first roughly 100 years of the dataset, since *do*-support has not yet taken root. From the earliest attestations, however, it is clear that *do* has a different distribution. It occurs more often to the right of an adverb. If one assumes that the placement of clause-medial adverbs remains constant during this period (an assumption bolstered by the constancy of the behavior of modals and have), this indicates that *do* is occupying a lower position in these early attestations.[14] By the end of the EME period, *do*'s behavior in this regard has become more or less indistinguishable from the other auxiliary types. Note that the data displayed in this graph are not generated by the loss of verb raising. The rate of pre-verbal adverbs (in relevant structural positions, i.e. adjoined between Spec,TP and V) climbs to 100% over the 16th century, just the opposite of the behavior of *do*. That is, "often saw it" becomes the only available word order, whereas "often did see it" becomes noticeably rarer.

[13] For further discussion of this phenomenon, see section 4.2.2.

[14] There is a bit of a puzzle with this data, since auxiliary *do* ought to move to T, and therefore be indistinguishable in its positioning from modals (which are base-generated there). One possible solution to this puzzle might lie in the notion of scope. Though pre- and post-T positions are available for adverbs, it is not clear that their use is absolutely unconditioned. If preserving on the surface a scope relationship between certain kinds of adverbs and the auxiliary is important, and if functional heads are distributed in an order which is semantically "natural" such that items which demand a lower scope appear lower and vice versa, then we expect a *do* which takes narrower semantic scope than modals to be more permissive of a wider variety of adverbs appearing to the left of it.

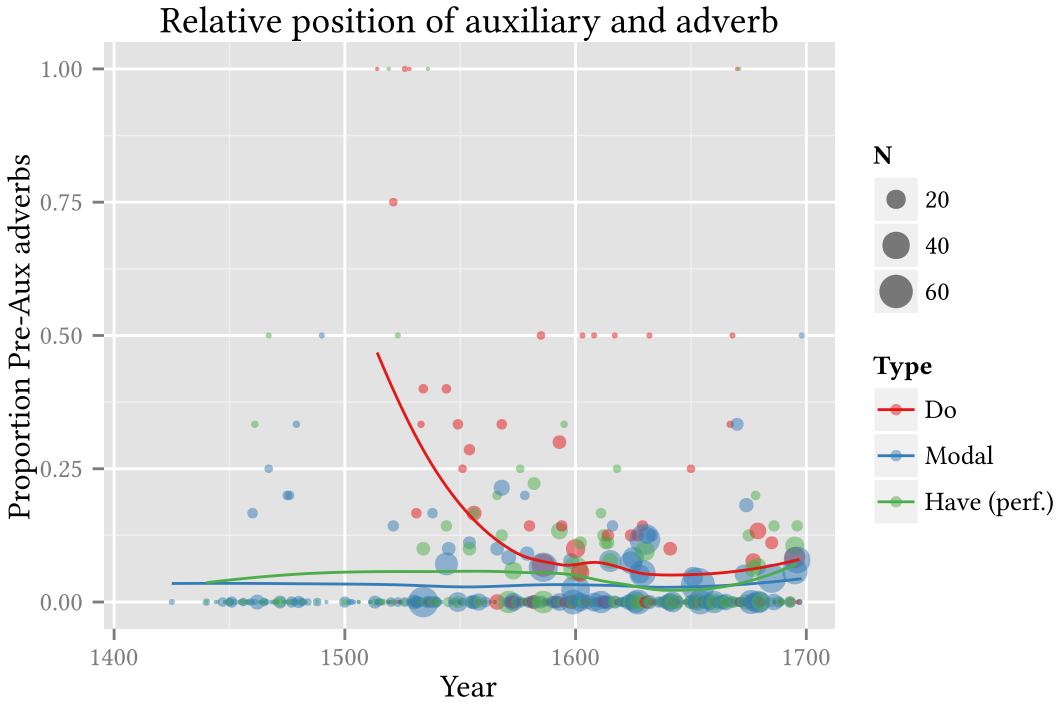



*Figure 4.21:* Data on the position of adverbs relative to certain types of finite auxiliary verbs in the PPCHE.

### 4.3.3 Argument structure effects

The previous two sections presented a pair of arguments that there is a semantically bleached, structurally low auxiliary *do* in the earlier stages of EME. This section will pivot to discussing the specific semantics of this auxiliary.

The argument structure of a sentence's main verb affects the incidence of *do*-support. In order to quantify this effect, it will be necessary to operationalize a definition of argument structure. In order to do this, I picked from the list of verb spellings in the PPCHE, arranged in decreasing order of frequency, the six most frequent prototypically unaccusative verbs, and the six most frequent experiencer-subject verbs.[15] I then looked through the list for all variant spellings of these verbs, in order to identify all usages of these verbs in the corpus. The frequency of these verbs in the combined EME corpus is given in Tables 4.14 and 4.15.

[15]Experiencer-subject here is used as a cover term for verbs whose subjects do not bear a canonical agent theta-role. It may turn out to be the case that it is possible to draw distinctions between the different members of this postulated lexical class. Its membership and definition were inspired by an effort to make a semantic generalization over the "*know* class" presented in Ellegård (1953).

*Table 4.14:* Unaccusative verbs in the combined corpora, pre-1700. "Total" indicates the #+caption: number of occurrences in all sentences, whereas the "With possible *do*-support" column indicates the number of occurrences in potential *do*-support sentences (whether or not *do*-support actually occurs in the sentence).

| Verb | Total | With possible *do*-support |
|---|---|---|
| *arise* | 181 | 138 |
| *come* | 11307 | 7228 |
| *die* | 949 | 557 |
| *go* | 7694 | 4511 |
| *rise* | 380 | 248 |
| *stand* | 1798 | 1184 |

*Table 4.15:* Experiencer-subject verbs in the combined corpora, pre-1700. "Total" indicates the number of occurrences in all sentences, whereas the "With possible *do*-support" column indicates the number of occurrences in potential *do*-support sentences (whether or not *do*-support actually occurs in the sentence).

| Verb | Total | With possible *do*-support |
|---|---|---|
| *care* | 193 | 143 |
| *doubt* | 1118 | 947 |
| *dread* | 32 | 19 |
| *fear* | 788 | 631 |
| *know* | 6806 | 4847 |
| *like* | 848 | 597 |

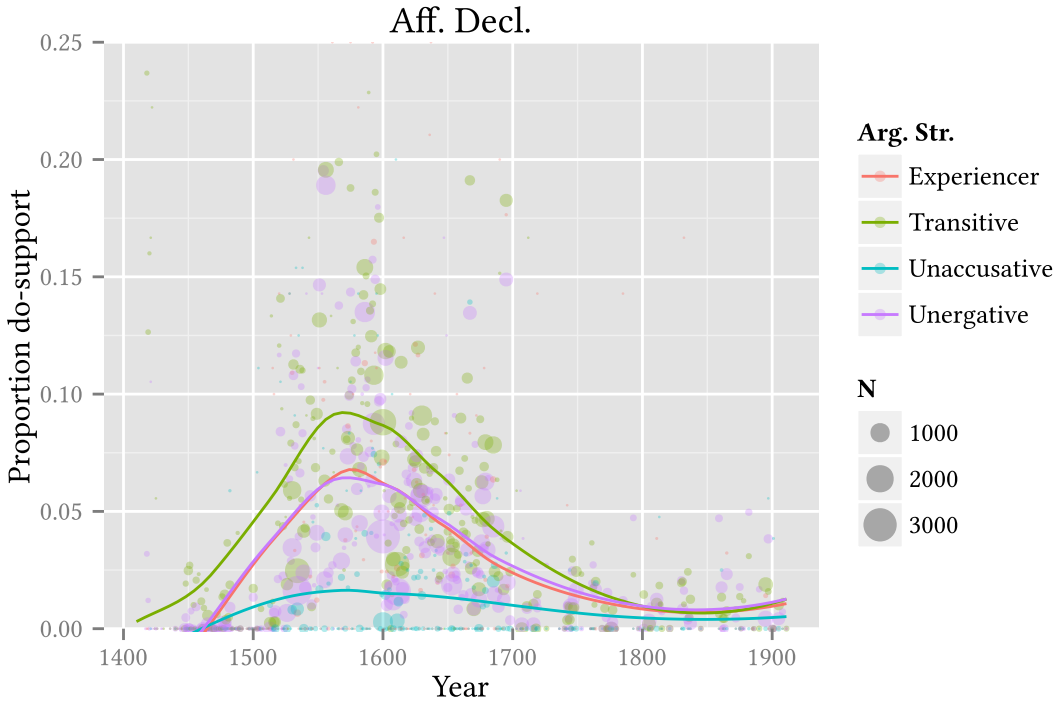



*Figure 4.22:* *Do*-support in affirmative declaratives, by argument structure type. (Some data points are off the top of the graph.)

Any verb lacking a direct object which was not a member of the unaccusative class was considered an unergative, and any verb with a direct object not on the list of experiencer-subject verbs was counted as a transitive. These verb classes behave differently in terms of their incidence of *do*-support. Figure 4.22 shows the rate of incidence of *do*-support in affirmative declaratives, stratified by argument structure types. As is evident, *do*-support is robust in all types except unaccusatives; in the latter type it peaks at only ~2% (by the loess smooth). I argue that this low frequency should be interpreted to mean that affirmative declarative *do*-support never happens with unaccusatives. The most robust generalization is that *do*-support in affirmative declaratives is generated by a grammar which uses *do* to mark the presence of an (agentive?) external argument. Since it is possible to coerce verbs into an agentive interpretation, the apparent tokens of unaccusatives with *do*-support may be in fact agentive.

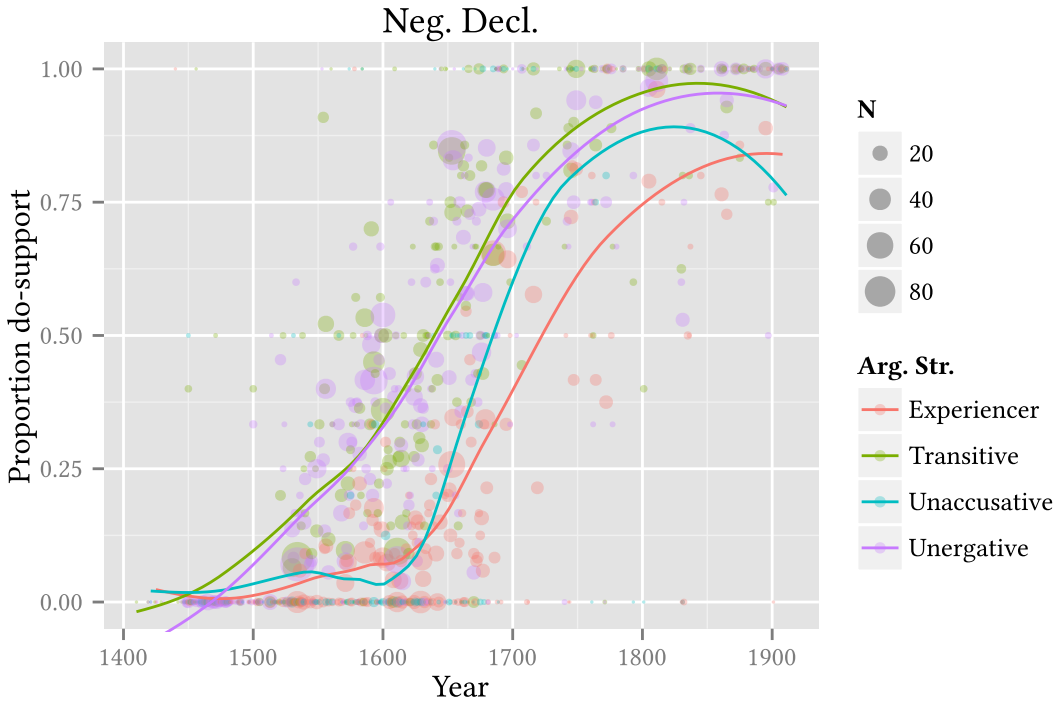



*Figure 4.23:* *Do*-support in negative declaratives, by argument structure type.

With negative declaratives, a similar pattern may be seen in Figure 4.23. Before 1600, there is little *do*-support in unaccusative or experiencer-subject negative declaratives, whereas in the agentive-subject types the rate of *do*-support rises steadily to roughly one third of all sentences. This split begins to wane in about 1625, when unaccusatives and experiencer-subject verbs begin to move towards 100% *do*-support (the latter more slowly than the former). This is consistent with the tokens in this corpus being mainly generated by a "*do* as argument structure marker" analysis before 1575. There is an event of reanalysis, or reorganization of the grammar of speakers, at around this date, as observed by Kroch (1989) and Warner (2005). From a grammatical point of view, this reanalysis consists of reassigning *do* to its modern role.[16] As a consequence, unaccusatives and experiencer-subject verbs are pulled up to 100% over the next centuries, joining their more-advanced agentive counterparts.

[16]Warner (2005) discusses other sociolinguistic events which occur concomitantly with this analysis; his thesis is that these account entirely for the data, and that appeal to grammatical structure or restructuring is not necessary. Given the new richness of the data presented here, that account is not tenable, as is discussed in section 4.2.3.

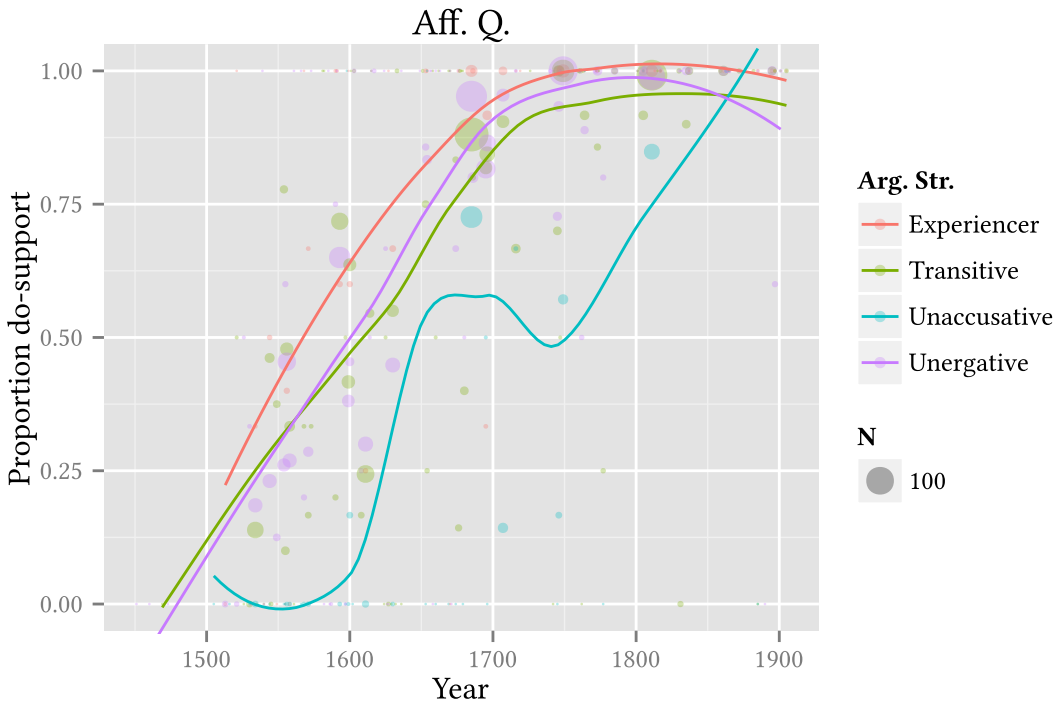



*Figure 4.24: Do*-support in affirmative questions, by argument structure type.

The data from affirmative questions are much less abundant than either of the foregoing types. They can be seen in Figure 4.24, and are largely consistent with the other types. Negative questions and imperatives do not provide enough data for analysis.

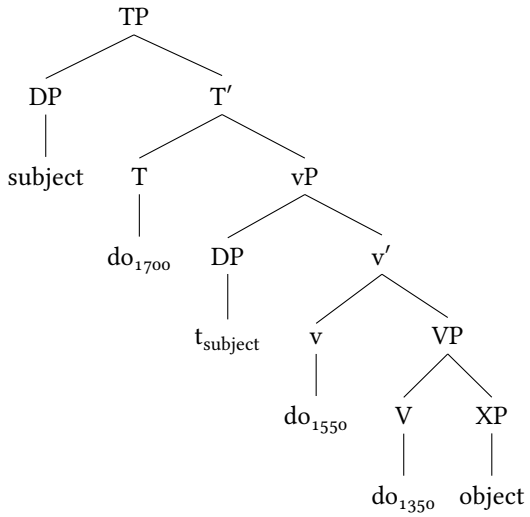
From the data presented so far, a picture emerges of the differing behavior of unaccusatives (verbs lacking an external argument): these verbs lag in their proportion of *do*-support everywhere. The behavior of verbs with experiencer subjects presents more of a puzzle. They lag in the negative declarative context only. A tentative generalization which emerges is that these verbs lag in negative environments but not affirmative ones. Without sufficient data to measure negative questions and imperatives this generalization cannot be evaluated (the trend in these environments does go in the predicted direction, however). In any event, this examination of the data contributes to our understanding of the behavior of auxiliary *do* in the early EME period (before 1600): it occurs mainly with external-argument bearing verbs, and barely at all with verbs which lack an external argument.

I will now move on to discussing my analysis of the structure underlying these surface generalizations.

The reader is also referred to chapter 5 for more discussion of the boundaries of the lexical classes implicated in this analysis.

### 4.3.4 Interlude: structural analysis

*Figure 4.25:* The positions of *do* at various points in the history of EME.




The structure of my analysis is reflected in the tree in Figure 4.25. In ME, *do* is a causative predicate, which is merged in V and takes a clausal object of some (small) size. In early EME, it is reanalyzed as the head of a projection low in the functional hierarchy. Kratzer (1996) introduced the notion of a separate syntactic head which introduces external arguments, which she named Voice. Later work has subsumed this function, along with others, under a head *v*; I adopt this usage in this dissertation. My proposal is that in EME *do* spells out the flavor(s) of *v* which introduce external arguments.[17] This option for spelling out *v* is never the only choice: at all times, speakers have available to them an alternative analysis under which *v* of all flavors is silent. Ultimately, this version of *do* is reanalyzed as *do*-support of the kind observable in PDE; this reanalysis enters the population of competing grammars in 1575 and has reached 100% adoption by the end of the EME period (1700).

The "intermediate" *do* to which this section has referred, it can be seen, is intermediate in two senses. It is temporally intermediate between the ME causative and PDE last-resort analyses of *do*. It is also structurally intermediate, occupying a head in the structure between V and T (which I have identified as *v*, given that

[17]For further exploration of – though not a conclusive answer to – the question of precisely which kinds of external arguments are implicated, seen through the lens of verb classes, see chapter 5.

head's association with external arguments/agents).

Now that a specific structural analysis has been spelled out, I will proceed to give two more pieces of evidence that the analysis of *do* differs before and after 1575.

### 4.3.5 Priming data

Priming is the phenomenon whereby a recently-used linguistic form is preferentially reused by a speaker. This effect has been observed in naturalistic speech (Sankoff and Laberge 1978), textual corpora (Estival 1985; Weiner and Labov 1983), and experimental settings (Bock 1986, in each case citations are to the earliest work on a topic). Work in this tradition has demonstrated that the relevant notion of repetition is structural, and not merely lexical. For instance, Bock and Loebell (1990) showed that usage of the preposition *to* in ditransitive sentences could prime further usage of *to* in that context; however, use of *to* as an infinitive marker does not prime subsequent usage in ditrnasitives. Thus, if there are two *do*s in EME (one associated with argument structure marking and one with *do*-support), we expect to be able to tell them apart in priming data. In this section, I will review this evidence and show how it bolsters my analysis which divides EME *do* into two classes.[18]

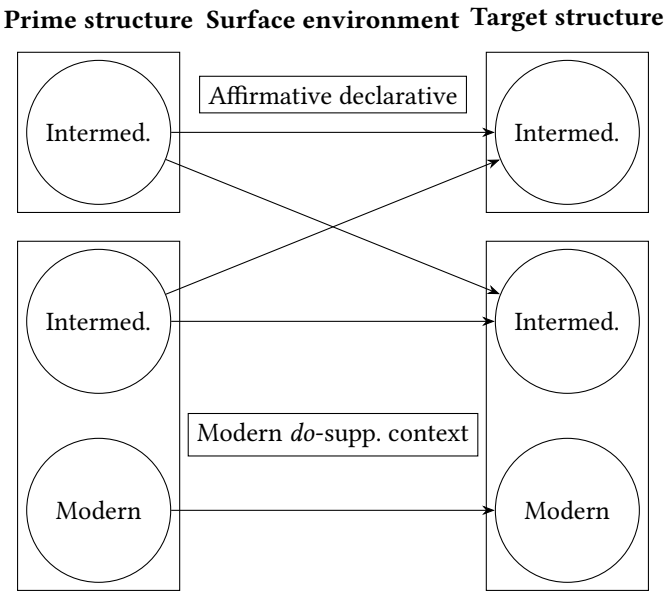



*Figure 4.26:* The predicted priming behavior of *do* in early EME.

[18]This is joint work with Meredith Tamminga, which has previously been presented as Tamminga and Ecay (2014).

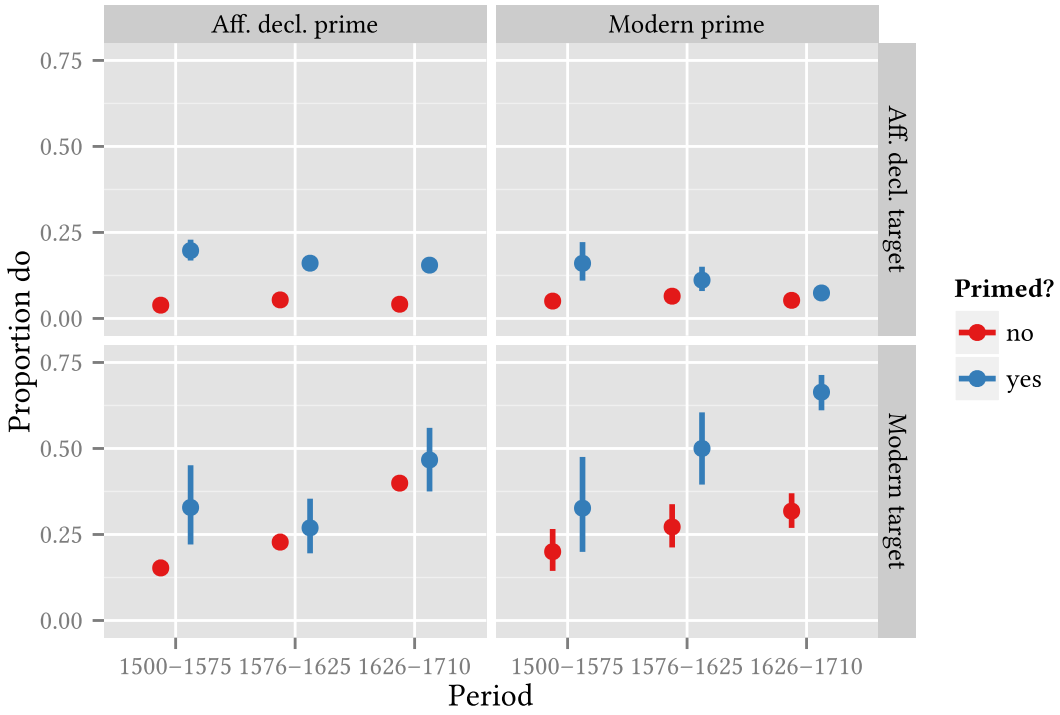



*Figure 4.27:* Priming data on *do*-support in EME. The two possible clause types are affirmative declaratives and all other clause types (those which would be expected to have *do*-support in PDE, referred to as "modern" *do*-support environments). The graph partitions all eligible prime-target pairs by which of these two classes the prime and target belong to. The error bars represent 95% confidence intervals based on the binomial distribution. Dots which seem to lack error bars have a 95% CI which is smaller than the diameter of the dot.

Figure 4.26 lays out the predictions of the priming account. The fact that intermediate structures can surface in apparently modern contexts gives rise to a prediction of some cross-priming.

These predictions were tested in a dataset drawn from the PPCHE. All clauses were taken from the corpus sequentially. Then, those clauses which are not potential *do*-support environments (such as non-finite clauses) were removed from the dataset. Clause pairs from each text were then formed in such a way that each clause (except the first and last clause) is a member of two pairs: once as prime and once as target. The clauses were coded for their type and presence of *do*-support.

The results of this analysis are shown in figure 4.27. As can be seen in the upper left quadrant, *do* in affirmative declaratives always primes itself – that is, there is always a significant difference between the red and the blue dots. This is the simplest case, since *do* in an affirmative declarative can only represent the

intermediate grammar.[19] I have argued that *do* usage in modern *do*-support contexts in the period before 1575 in fact represents tokens of intermediate *do*. If this is the case, then *do* in modern *do*-support contexts should prime *do* in affirmative declaratives, and vice versa. That this prediction is borne out may be seen in the lower left and upper right quadrants of the graph.

On the other hand, these effects disappear in the latter two time periods. Conversely, modern *do*-support environments do not begin to prime each other until after 1575, once *do* in these contexts is in fact generated by a modern *do*-support grammar.[20]

Before 1575, priming data demonstrate that the affirmative declarative context is entangled with the others in a way that suggests structural identity (or similarity). This apparent identity disappears after that date. Thus, these facts bolster the hypothesis that affirmative declarative *do* is different in kind from modern *do*-support, and that the early EME period is dominated by the intermediate *do*.

### 4.3.6 Shared constraints

Research in sociolinguistics has established the principle that variable usage can furnish information about the grammatical organization of a language.

This principle is extended by Tagliamonte (2013), who argues that the identities and differences in the relative strength of different effects on language usage, understood as variable rules (Cedergren and Sankoff 1974), can establish phylogenetic relationships among English varieties. In this section, I'll deploy a similar strategy for the analysis of *do* usage in EME.

Figures 4.28 and 4.29 depict the effect of subject type in various *do*-support environments. Considering the first figure, we observe that affirmative declaratives exhibit a robust effect throughout the EME period whereby non-pronominal subjects have a higher rate of *do* usage than either pronominal or wh-trace subjects; the latter two types have virtually identical rates of *do*. (Later, in the post-1700 Modern English period, *do* disappears entirely from wh-trace subjects but survives at a low rate with both pronominal and non-pronominal subjects.) The second figure illustrates that before 1575 this effect exists in the negative declarative and affirmative question clause types, which are environments for modern *do*-support. At 1575,

[19] The very rare incidence of emphatic *do* including (but not limited to) affirmative declaratives has been ignored for this experiment.

[20] More needs to be said about why priming does not hold in this context before 1575. If in the early period *do* in modern contexts is generated solely by an intermediate grammar, then it should be as capable of priming itself as it is in the other three prime/target configurations. However, the reanalysis in 1575 is not an instantaneous event: some minority of apparently-modern *do* tokens before this date are in fact generated by a modern grammar (and conversely afterwards). This means that the modern environment is in fact a (largely) surface-indistinguishable mixture of two types of *do*. This mixture, I argue, prevents the emergence of a priming effect since many token pairs will be assigned discordant structures and thus not be able to prime each other. This supposition might also be able to explain why the pre-1575 magnitude of the modern prime/affirmative declarative target priming effect (difference between the red and blue dots) is smaller than either of the affirmative declarative prime effects in this period: only a fraction of the modern *do*s are actually licit primes.

however, a noticeable shift in the pattern takes place, and the non-pronominal subjects' advantage disappears. (The advantage actually seems to be reversed in the case of affirmative questions, though there is little data for us to examine.) The pattern shift does not seem to take place in negative questions – but here, *do*-support with non-pronoun subjects reaches 100% in 1575 (or just a few years afterwards). Thus, it may not be subject to reversal.[21] These facts, in combination with the theoretical considerations discussed above, constitute further evidence that the grammar of *do* usage in affirmative declaratives and other contexts is related before 1575, but becomes distinct in and after that year.

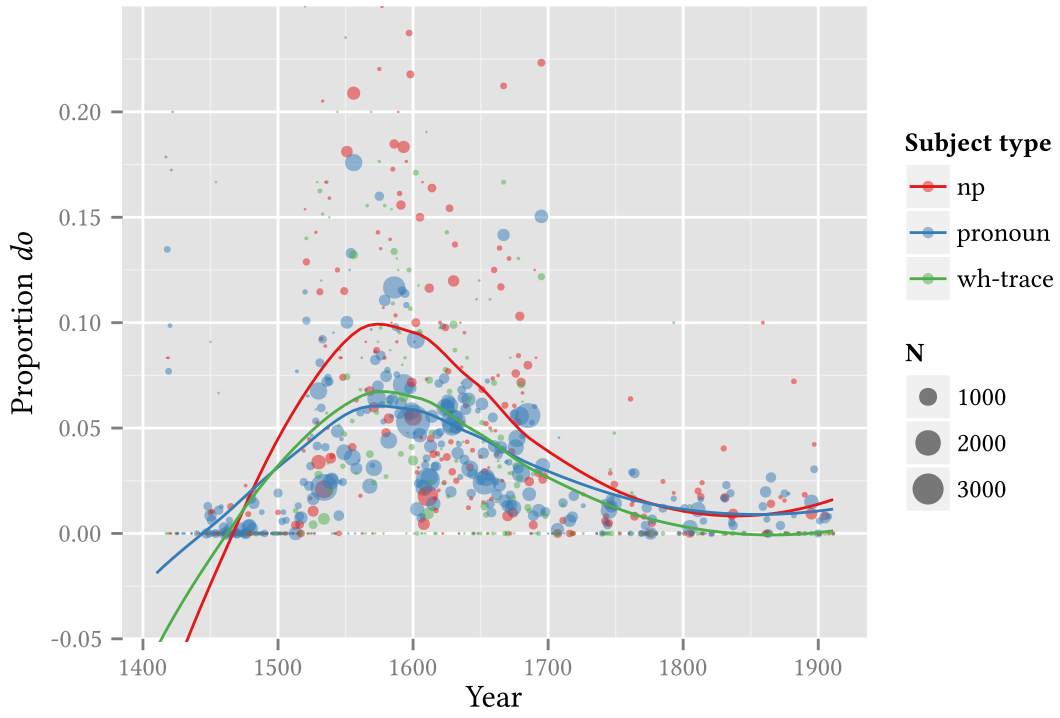



*Figure 4.28:* Subject type effects in affirmative declarative *do*-support in the PPCHE data.

[21] For more discussion on this topic, see the following section.

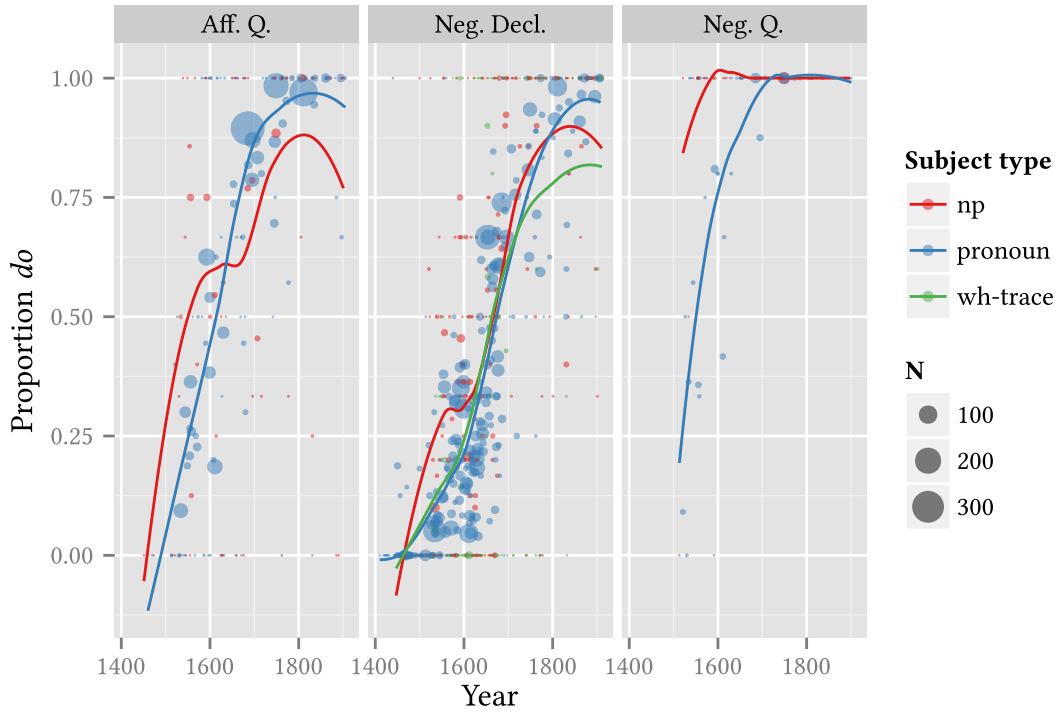



*Figure 4.29:* Subject type effects in various modern *do*-support environments in the PPCHE data. Note that wh trace subjects are only attested in declarative clauses, since wh subject questions are not a *do*-support environment.

### 4.3.7 An analysis of the events of 1575

The data presented so far in this dissertation all indicate significant disruptions in the patterning of *do*-support occurring during the EME period. This is not a novel observation, nor a particularly difficult one to arrive at after even a cursory examination of the data. Kroch (1989) claimed there was a grammatical reanalysis at this time to the effect that V to T raising was lost from the language. Warner (2005) on the other hand proposed that the disruptions in the behavior of *do* could be entirely explained by sociolinguistic principles. I have demonstrated in section 4.2.3 that Warner's specific proposal about the relevance of *not*-contraction to the sociolinguistic picture is not borne out in data from the PPCHE; however his observation that stylistic factors do play a role remains sound.

In this section, I will endeavor to give an account of the reanalysis of the syntax of verb movement in purely grammatical terms, leaving the social factors aside. Before discussing this point, however, it is necessary to note that I do not assume that the reanalysis is a sudden event. Since the object of study in

generative syntax (whether historical or synchronic) is the competence of an individual native speaker, a change from grammar G to grammar G′ at the population level lasts at least from the time that the first speaker of G′ is born to the time that the last speaker of G dies. When studying written data, we should substitute "begins/ceases to be reflected in the written record" for "is born / dies", in addition to taking account of other peculiar features of the written medium (such as uncertain dating of texts, the appearance of archaic structures through quotation or mimicry). The subdiscipline of sociolinguistics contains numerous examples of the ways in which this simple model requires even further refinement. In the present case of *do*-support, it was demonstrated by Warner (2005) (and further confirmed in section 4.1.2) that we cannot treat speakers as idealized sources of variation (analogous to the proverbial spherical cow of introductory physics textbooks), but rather must take account of their age and stylistic orientation.

Bearing all this in mind, it is nonetheless clear from the data presented so far that there were two grammatical systems underlying the surface *do*-support construction in EME. One of these predominated before 1575, and the other afterward. Thus, when speaking of "the reanalysis of 1575," I am not advancing the claim that something suddenly happened to English grammar in this year. Rather, I am sketching from a purely grammatical point of view the syntactic changes revolving around *do*-support which took place in EME, and which are seen to make an abrupt transition in textual sources of evidence in roughly the mentioned year.

In the period before 1575, V-movement is being gradually lost. This can be directly observed by the decline in verbs crossing *never*-adverbs. It can also be observed in the increasing tendency to insert *do* – at this time a marker of an agentive external argument – in agentive sentences (transitives and unergatives). In these contexts, agentive *do* is a direct replacement for verb raising: *do* patterns as a modal, so when it is inserted the question of raising the lexical verb to T does not arise. Thus, from the logic of the CRH discussed in section N, we expect to observe a parallel slope between verb raising past *never* and transitive and unergative *do*-support. By the logic of the CRH, we must accept that V-raising is also being lost with unaccusative main verbs at the same rate as in the other contexts. This conclusion is bolstered by data from raising of verbs over *never*, discussed in section 5.2.2 below. Though there is a small (and unexpected) delay in the timing of the loss of raising past *never* for unaccusative verbs of motion in particular, the rates of change are visibly indistinguishable, and all contexts have gone to completion by 1575. The question arises of why this loss is not reflected in the surface data. No *do*-support is possible with unaccusatives (because *do* is semantically incompatible). There is not a noticeable trend towards failing to raise across *not* in negatives as in the example given here:

(84) much breade I not desyre PLUMPTO,11.184

Though this construction is attested sporadically, it is extremely marginal.[22] One possibility would be to claim that this construction is in fact the competitor to V-raising in unaccusatives, just one which appears incredibly rarely. However, this would require two tenuous premises. The first is empirical: non-raising over *not* seemingly appears with all predicate types, and is not specific to unaccusatives. Thus it fails to be a satisfactory solution to the question of why non-raising fails to expand specifically in unaccusatives. The second is a broad theoretical concern: putting analytical load on such a rare construction requires us to posit that children can track infinitesimally small probabilities (on the order of $10^{-5}$).

Another solution would be to propose that unaccusatives are not affected by the loss of V-raising, in the same sense that auxiliary verbs in PDE are not (modals, *be*, and sometimes *have*). This proposal is lent plausibility by the fact that the class of auxiliaries has been subject to gradual erosion (the elimination of pseudo-modals like *need* and *ought*, as well as possessive *have* in American and later British English). It is also suggested by theories about the interaction of verb movement and $\theta$-assignment. Pollock (1989) and Roberts (1985) proposed that the class of auxiliary verbs is constituted by those which do not assign any $\theta$-roles. Their reasoning behind this proposal was highly theoretically specific, rigid, and ultimately not satisfactory (since there has been no change in the $\theta$-properties of possessive *have* that can motivate the differences in its distribution in American and British English). However, if the spirit of their proposal is on the right track, it stands to reason that, before arriving at a class of auxiliaries that assign no theta role at all, English could pass through a stage where it singles out a class of verbs that assign no *agentive* $\theta$-role. However, one crucial consideration rules out such an approach. If it were true, then unaccusatives should differ from other main verbs in their placement with respect to diagnostic adverbs – and pattern with auxiliaries. This prediction is falsified by figure 5.17.

Taking a step back from the strong assumptions of statistical independence underlying the CRH framework, we can go only so far as to say that, pre-1575, the grammar contains a provision which variably (but decreasingly) raises the verb to T. It also contains a resource for side-stepping the question in agentive sentences, namely the insertion of agentive *do*. Apparently, the fact that learners can analyze some unaccusative sentences as being derived by V-raising (using the rump of the V-raising rule) allows them to analyze all negative unaccusatives in that manner.[23] The two relevant tools in their grammatical toolbox are V-raising

[22] One might also wonder about the situation of V-raising in questions. However, it is not easy to distinguish questions in which V-to-T has applied from those with Scandinavian-style direct V-to-C (discussed below), Furthermore, Han and Kroch (2000) lay out a number of syntactic issues for the analysis of imperatives. For simplicity I will focus here on the issues raised by declaratives.

[23] The issues for theories of variation, change, and learning that are raised by this analysis are not insubstantial. The core issue is: what is the probability assigned to a grammatical phenomenon? Clearly speakers know that certain structures should be more or less frequent, as demonstrated by Labov (1989) and many follow-up studies of learners' probability-matching behavior. However, it may be

and agentive *do*. By some combination of these they analyze all relevant sentences.

This situation changes in 1575. The PPCHE data in figure 4.12 show that verb raising past *never* disappears from English in this year. Learners are faced with a crisis: they have not inherited any syntactic resources which can cope with negative declarative sentences with an unaccusative verb. This situation induces them to analyze agentive *do* as a support auxiliary which takes the place of V-raising – that is, they innovate the modern grammar of *do*-support. Because *do*-support does not immediately jump to 100%, learners must also develop an analysis for the residual tokens of non-*do*-support sentences which are attested in the data. There are two cases to be considered: negatives and questions. In the case of questions, Kroch (1989) points to the case of V2 in mainland Scandinavian languages. In these languages, a main verb moves to C (above the subject in Spec,TP) in main clauses:

(85) *I dag ville Lotte inte läsa boken*
today wanted L. not to.read book
'Today Lotte didn't want to read the book.'

However, in embedded clauses it remains in V, below negation:

(86) *att Lotte inte ville läsa boken i dag*
that L. not wanted to.read book today
'that Lotte didn't want to read the book today'

The same surface distribution of lexical verbs applies to late EME (restricted to questions): the verb moves to C in that context, but does not surface in T in any other context. A grammatical process similar to that applying in Scandinavian (which is yet not fully understood) could apply to the English cases. An explanation is more difficult to find in the case of negatives. Kroch (1989) suggests that *not* moves to the right of the verb by a cliticization process, analogous to the cliticization of *not* to auxiliaries. This is an unusual proposal, given that *not* does not participate in verb movement of a main verb to C (either before or after 1575), very much unlike how it does move to C with an auxiliary. Thus, the precise nature of the non-V-raising analysis

inappropriate to interpret summary statistics over corpus data (even the productions of an individual speaker) as the object of mental representation. I have suggested – backed up by the data in this dissertation – that the loss of V-raising is spreading through EME in an ordinary way before 1575, even as it is anchored to zero in the context of unaccusatives. This anchoring does not arise from an architectural cause, since the lack of V-raising is logically compatible with unaccusatives (as in later EME). It is also not a lexical accident or exception, as we might describe the behavior of auxiliaries with respect to *do*-support (especially those which eventually lose their auxiliary status). Rather, the cause has to do with the structure of other parts of the grammar. The change progresses where there are other alternatives, but is held back in the case of unaccusatives where there is not. How do learners become aware of these interactions between contexts by availing themselves only of innate knowledge and patterns in the data? Why do they entertain such scenarios at all, rather than immediately reanalyzing their input to either change(/vary) or not change(/vary) consistently in all contexts? These are deep questions, and merit further study and elucidation. It is in some sense unsurprising that they arise in the study of *do*-support, a change whose history has been well-studied and is evidently quite complex. Because of that very complexity, however, the data on *do*-support presented here cannot of themselves constitute a simple and convincing case that revisions to the theory are necessary. Thus, I must merely highlight these questions without being able to advance any firm answers.

*Table 4.16:* A test of the CRH between *do*-support and verb raising past *never* on PPCHE data coming only from transitives.

| | CRH | No CRH |
|---|---|---|
| Intercept | −0.86 (0.13)*** | −0.93 (0.14)*** |
| Slope (std.) | 2.20 (0.27)*** | 1.88 (0.39)*** |
| Neg. Decl. (sum) | 0.26 (0.28) | 0.15 (0.34) |
| Aff. Q. (sum) | 3.76 (0.27)*** | 4.00 (0.35)*** |
| Slope: Neg. Decl. | | −1.24 (1.54) |
| Slope: Aff. Q. | | 0.68 (0.55) |
| AIC | 713.34 | 714.78 |
| Num. obs. | 1131 | 1131 |

of negatives remains elusive. But in any case, these distorted vestiges of V to T movement begin to compete with *do*-support after 1575, and are eventually extinguished by it. The lack of grammatical unity among these constructions explains, in the CRH paradigm, why their trajectories do not share a common slope after 1575.

A question also arises of why Kroch (1989) found a CRH effect between *do*-support and V to T raising across *never*. Under the analysis I have proposed, such an effect is not predicted: the occurrence of pseudo-auxiliary *do* pre-1575 is not connected to verb raising at all, except insofar as it co-occurs with unaccusatives.[24] The true, and very low, rate of *do*-support before 1575 is measured by the occurrences of *do* with unaccusatives. On the other hand, the usage of *do* with transitives represents the intermediate *do* grammar.[25] There is not enough data on unaccusatives to directly test for the expected CRH. However, testing the transitives for a CRH is possible. The results are given in table 4.16. The results mildly favor the CRH even here; the $p$-value model comparison method favors the CRH, as does a likelihood ratio test ($p = 0.28$; no evidence that the more complex no-CRH model is a better fit) whereas the AIC is within the two-unit margin of uncertainty. This must be treated as an accidental fulfillment of the conditions of the CRH, and not a genuine effect.

Finally, it is worth noting that Han and Kroch (2000) take a different approach to explaining the disruptions of 1575, proposing two positions for negation and a two-stage loss of verb movement, with 1575 as the pivot between the two stages. They hypothesized that verb movement to T is lost up to 1575 (as suggested by the *never* data), whereas "short" verb movement (i.e. movement out of the verb's base position, but not as high as T) only begins to be lost after 1575. The empirical grounds of their account are diminished by the failure to

[24]Under an instantaneous reanalysis, there should be no occurrences of *do* with unaccusatives before 1575. Of course, an instantaneous reanalysis is implausible; even if on a certain date all new speakers of the language learn the reanalyzed grammar, the older grammar will survive until all older speakers die. Thus, the occurrences of *do* with unaccusatives before 1575 represent leakage of particularly innovative productions into the written record. (It is also possible, of course, that some of these tokens represent usage of *do* to coerce a non-agentive lexical verb into an agentive frame.)

[25]With a small additional contribution from innovative *do*-support, which I will ignore in the following quantitative analysis.

replicate in the PPCHE Ellegård's finding of absolute non-occurrence of *do*-support in negative imperatives before 1575 (for them, imperatives do not have a T position, thus only the loss of short verb movement after 1575 can induce *do*-support in imperatives). My own attempt to substantiate their short verb movement proposal focused on a search for verbs which move past adverbs like *completely*, which occur lower in the clause than *never*. A search in a billion-word corpus for strings of the form listed below uncovered only 4 examples – not enough to postulate the robust existence of a short verb movement phenomenon.[26]

(87) Aux (Adv) V *ly*-adverb Oblique-pronoun

### 4.3.8 Comparison to other analyses

This account is consistent with an explanation which sees the origin of *do*-support in the Middle English causative verb *do*. The debate about the origin of the *do*-support construction goes back very far; Ellegård defended a version of the causative-origin hypothesis against a variety of other theories including most prominently the suggestion that *do*-support is a syntactic borrowing from Welsh.

Denison (1985) extended this hypothesis, proposing four stages of evolution:

1. *do* is one among many causatives
2. *do* causatives spread at the expense of others
3. *do* becomes an auxiliary
4. *do* acquires its modern distribution

Denison's description raises the question of whether and how the intermediate *do* account has been proposed in previous literature. Denison proposes an account which comes close to the present one. For him, early (that is, ME) uses of *do* are "factitive," and vague between a causative interpretation and a non-causative (directly agentive) one. That is, *do* is not a causative in ME on Denison's account. This construction is then pressed into service as a marker of non-perfective Aktionsart (a development prompted by the decay of the OE system of Aktionsart marking through verbal prefixes), before being discarded in that function. Denison's account nonetheless differs from the presently proposed one in several ways. If Denison's account

[26] The examples are:

- "But you haue heard already how he and his brother haue deuised so with the Turke, that he **might oppresse sodainly vs** only and our fellowes." (1560)
- "yf they do thus they **shall so worshyp veryly me**" (1581)
- "And Escobar with others having no regard at all to this distinction, **will condemn absolutely them** both of mortal sins" (1670)
- "so that no Passion can rise or mutiny within, but it **must betray presently it** self without" (1700)

Notably, only in the second example is the object pronoun bereft of following modifiers, making possible an analysis of the other three examples which treats them as instances of extraposition of heavy objects (of a type, to be sure, which is degraded in PDE).

is accepted as an explanation of the behavior of *do* in the EME period, then we must accept that there is a 300-year time gap between the first attestations of the "vague *do*" (in the 1200s) and its rise in affirmative declaratives (beginning ca. 1500). On its own, this explanation is not satisfactory, and some other change in the language must be adduced to explain why the rise of affirmative declarative *do* occurs when it does, and not in any of the earlier centuries that seem available to it.[27] Secondly, the data from the corpora do not support the Aktionsart-marking hypothesis. Figure 4.30 shows that agentive non-perfectives such as "believe" and "desire" have high levels of *do*-support while non-agentive ones such as "please" and "seem" are much lower. Figure 4.31 shows the same graph for perfective verbs. Though (the commonest) perfective verbs are generally more agentive, the non-agentive *die* shows a low trajectory (as does *tell*, anomalously). Indeed, the association of perfectivity with agentivity gives the impression that Denison's proposal that *do* is associated with non-perfectives in the earliest period of the change is precisley backwards. However, the data support my arguments above that argument structure (specifically related to the presence and role of the external argument) is a more important determinant of *do* usage behavior than Aktionsart. It may be the case that Denison's account can shed light on the earliest stages of the emergence of *do*-support, but it cannot explain the behavior of *do* in EME. Denison himself admits this fact, identifying his final stage four with the period after 1400 ("mainly [the] fifteenth and sixteenth centuries", p. 55). Thus, the account I have sketched above does not directly compete with Denison's account, but rather addresses a different part of the trajectory of *do*-support through the language.

[27] The loss of V-to-T raising is a noticeable change that takes place at the right time to be a candidate to influence the development of *do*-support. However, as I have extensively argued above, *do* usage in affirmative declaratives is not related to the other *do*-support contexts, and thus a separate explanation is needed under Denison's account for its apparent 300 year lag.

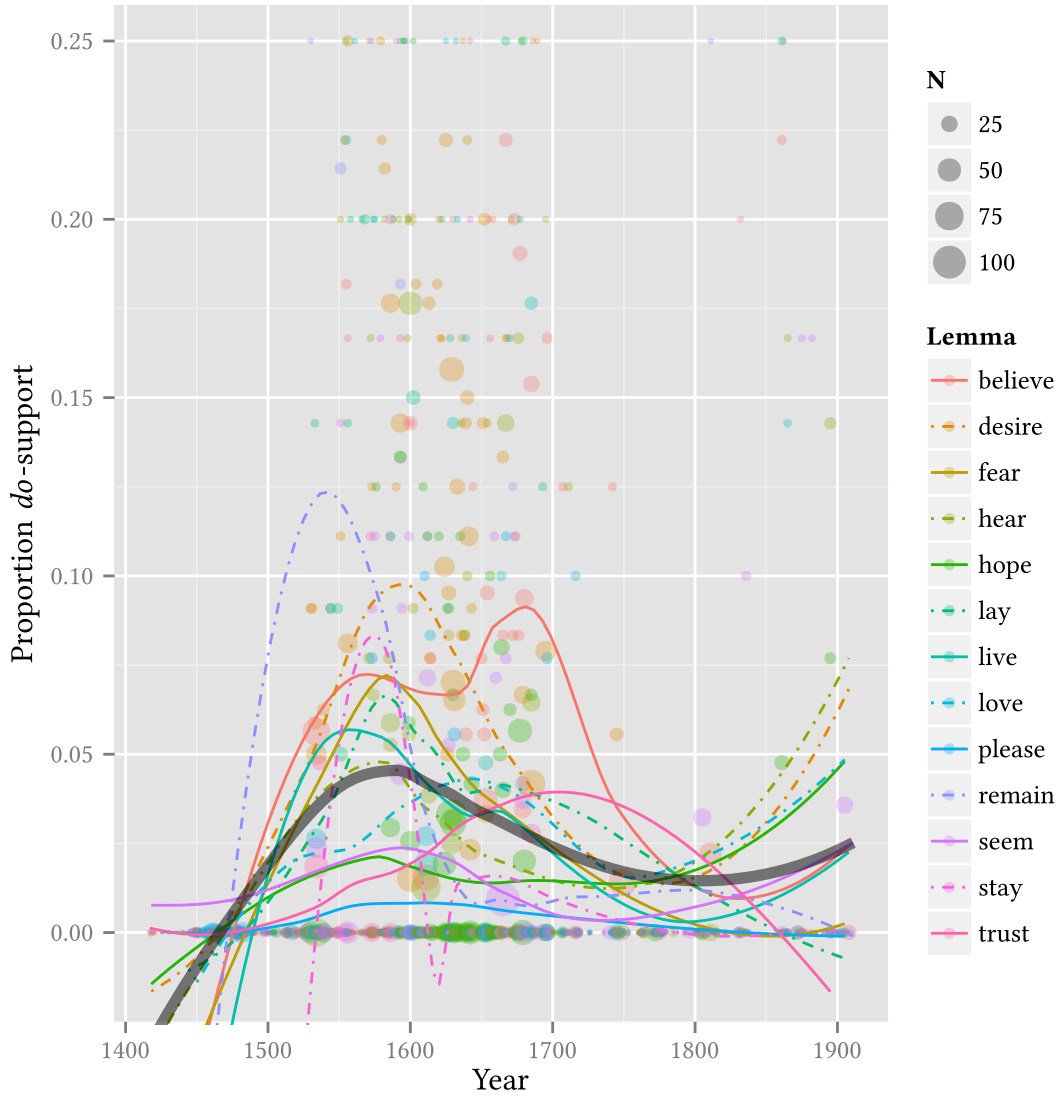



*Figure 4.30:* A graph of the evolution of affirmative declarative *do*-support with the most common non-perfective verbs in the parsed corpora, as diagnosed by the "for/in" test. The black line is the (weighted) average trajectory for the class.

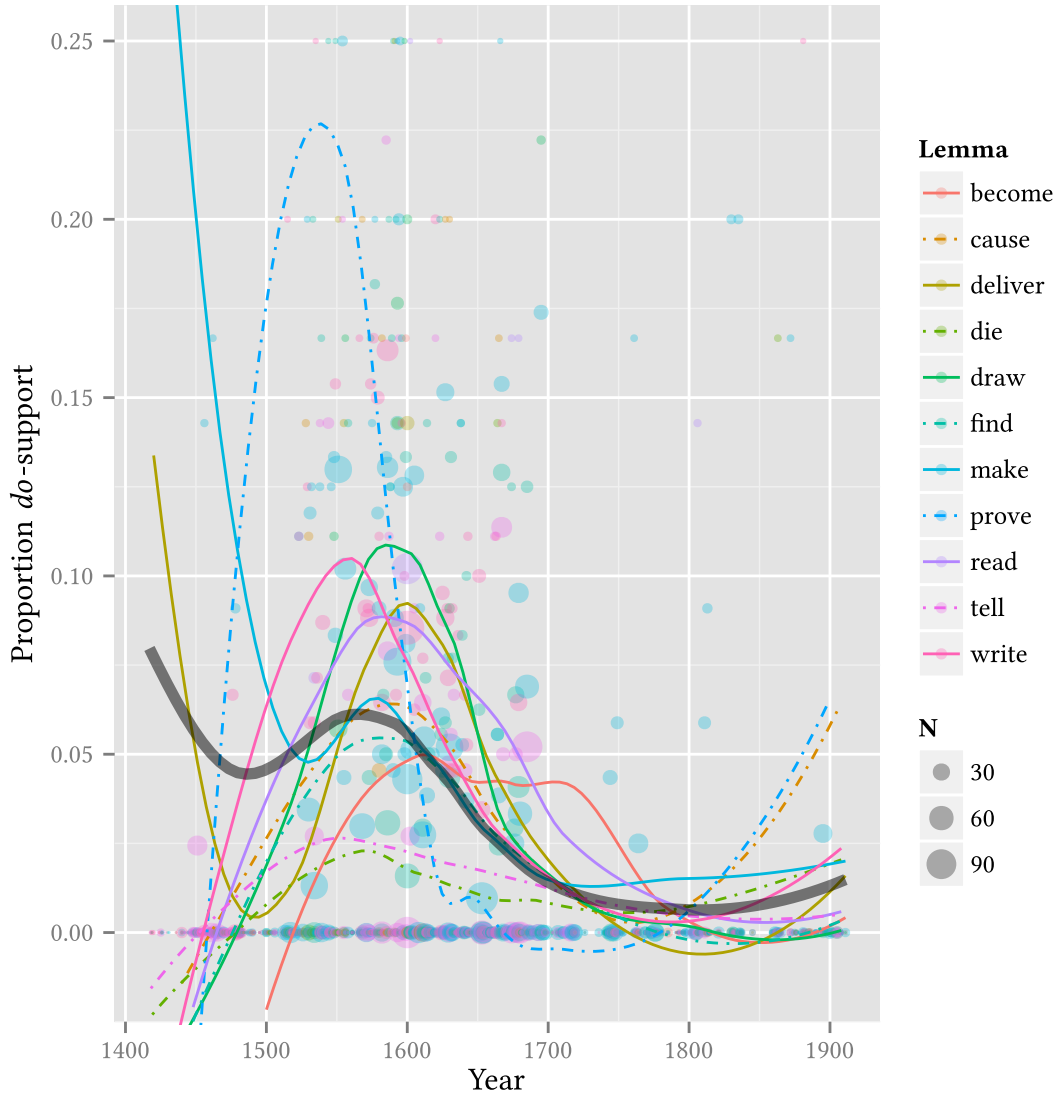



*Figure 4.31:* A graph of the evolution of affirmative declarative *do*-support with the most common perfective verbs in the parsed corpora, as diagnosed by the "for/in" test. The black line is the (weighted) average trajectory for the class.

Roberts (1993) also claims an analysis with an intermediate do: "Sixteenth century *do*, then (until after 1575), was intermediate between the ME main verb and the [ModE] 'supporting' auxiliary in terms of its distribution." (p. 296) However, this is not an intermediate analysis in the same terms as that proposed here. Roberts proposes a reanalysis of *do* from a main verb to a modal in one fell swoop, occurring in the early 1500s (he proposes 1530). He concludes that after this date, *do* is "semantically empty" – which is clearly not

the case, given the data on argument structure presented above. Roberts' intermediate stage just marks the time between the loss of the ME *do* and the emergence of *do*-support, and is not attributed any properties of its own. Though it shares the name "intermediate," this account is very different in spirit from the one I propose.

The account presented here accords with the causative-origin theory, advocated by Ellegård and (in a modified form) Denison and Roberts. It is natural to suppose that *do* might be reinterpreted as a marker of agentivity or external argument presence, both those features being present in causatives. The data presented do not, however, resolve the question of whether it is truly agentivity or the presence of a (not necessarily agentive) external argument which drives the insertion of *do* in the intermediate grammar. The data from affirmative sentences (declaratives and questions) contradict that from negative declaratives; in the former sentence types experiencer-subject verbs pattern with the agentive verb types, whereas in the latter they behave like unaccusatives. The root of this difference could lie in the interaction of the lexical semantics of these predicate types with negation, especially if *do* is agentive. On the other hand, an account within the framework of Grammaticalization (Hopper and Traugott 1993), where semantic bleaching is an important element of the notion of language change, would prefer to see the intermediate grammar as marking external arguments, since Agent is not a bleaching of Causer but rather a strengthening. That is, for $X$ to have the thematic role Causer with respect to an event $e$ means $X$ is the cause of $e$. For $X$ to be the agent of $e$, $X$ must both be the cause of $e$ and $X$ must act volitionally.

## 4.4 From diachrony to synchrony

Various structural explanations for *do*-support as a synchronic phenomenon in PDE have been proposed. These analyses fall into two broad categories. The first class of analyses generate *do* freely in all clauses not containing another modal, and delete it from those clauses in which it does not appear on the surface (insert-and-delete models). The second class only inserts *do* when it surfaces (last-resort models). I will briefly lay out in this section exemplars of both families of accounts. Then, I will go on to argue that they are fundamentally similar in specific ways. Because of their structural similarities, I will claim that synchronic evidence underdetermines the choice between the two families of account. Consideration of diachronic evidence, on the other hand, can differentiate between the two accounts, and shows the last-resort family of analyses to be superior to the insert-and-delete one.

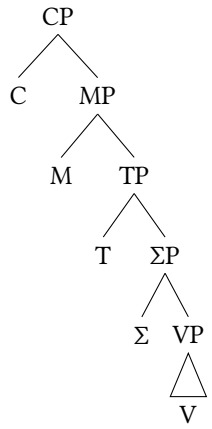
### 4.4.1 Insert-and-delete

#### Pullum and Wilson (1977)

Pullum and Wilson (1977, PW) gives an account of *do*-support in PDE which unifies the treatment of constructions exhibiting *do* with those having other auxiliary verbs (*have* and *be* with verbal complements, and modals). Under their treatment, auxiliaries have the same categorial status as main verbs and are thus are generated in V. Auxiliary verbs take a CP complement (S′ in PW's terminology), which in turn contains the lexical verb and its arguments. Furthermore, main verbs select a complementizer that is incompatible with being a root clause on its own; it must be embedded under a modal of some type (PW p. 776). Thus, in a clause without another type of auxiliary, *do* must be inserted at the top of the clause. After this, transformations apply. A rule applying late in the derivation (PW (63)) deletes *do* when it is string-adjacent to its complement CP. This rule thus applies in precisely the set of cases when the grammar of PDE requires *do*-support: when adjacency is disrupted by *not*, by subject-Aux inversion, or by topicalization of the VP (= embedded S′).[28]

#### Schütze (2004)

*Figure 4.32:* The structure of an English sentence for Schütze (2004). For simplicity, specifiers are omitted.




Schütze (2004) proposes another version of the insert-and-delete approach. Superficially, it is quite different from PW: the advent of the Split-INFL Hypothesis (Pollock 1989) changes the assumptions about clausal structure, and Minimalism (Chomsky 1995) changes the nature of the syntactic rules proposed. Nonetheless,

[28] PW bracket the issue of how *do* is retained in emphatic clauses, such as:

(i) (A: You didn't take the garbage out last night.)
B: But I *did* wash the dishes.

the core insight remains that *do* is generated in all English sentences lacking another modal, and is filtered out where it does not surface.

Schütze's theory of the clause structure is given in Figure 4.32. He assumes that main verbs are generated in V and move to T (if not blocked by an intervening head).[29] For him, T is not the target of verb movement in French (and other languages with "V-to-T" movement) as was argued by Pollock (1989); rather, French finite verbs target some higher projection, perhaps in a Split CP system (Rizzi 1997). Σ is the host of negation; modals are generated in M. Schütze argues that *do* is also generated in M. In clauses with negation, *not* blocks the raising of V to T on morphological grounds (it is not possible to spell out *not*+V). M must therefore be spelled out as *do* in order to provide a host for the inflectional features in T. Presumably a similar analysis applies to cases of emphasis (Schütze does not specify). Questions (which in PDE require *do*-support) pose a challenge to the analysis. Something (presumably M) must raise to C to support the +wh feature there (evidently, given the distribution of auxiliaries in PDE questions). It's necessary to stipulate that this C-supporting head have phonological content, otherwise a silent (empty) M could fulfill C's support requirement. With this stipulation in place, it follows that V cannot raise to T; M must spell out as *do* (in the absence of another modal), and T must raise to M and then to C. Presumably another stipulation is called for to prevent V from raising to T, M and C in that order, and spelling out in C (as it did in ME).

For Schütze, given that M can sometimes spell out as *do* (in a semantically vacuous way), there is no syntactic resource which can block it from always doing so. In order to perform such blocking, he argues, it would be necessary to engage in transderivational comparison, a conceptual impossibility under his assumptions about syntax. Indeed, sentences with unemphatic affirmative declarative *do* are grammatical for speakers of PDE. They are filtered out by an "extrasyntactic principle" (512) of economy – "use as few words as possible."

There are empirical and conceptual problems with Schütze's account. Turning first to an empirical argument, in PDE, there are three different possible positions of *never* in the following sentence:

(88) a. Sam never will have seen such a marvel.

b. Sam will never have seen such a marvel.

c. Sam will have never seen such a marvel.

Schütze argues (511) that in such sentences *have* must be below Σ (in order to rule out sentences like *"John will haven't been drinking"), and thus below T. The claim is that *have* heads its own VP, which takes the VP

[29] I ignore in this discussion the treatment of *have* and *be*. Schütze's account of these verbs is complicated, and not without instances of confusing equivocation. (For instance, he criticizes Roberts (1985) for appealing to the semantic vacuity of *have* and *be* in an account of their auxiliary-like behavior, but then precisely relies on such a hypothesis himself (511).)

headed by *see* as its complement. But this cannot be the case. It is possible to have a passive as a complement of *have*:

(89) Such a marvel will never have been seen.

Furthermore, the constituent [ *have seen X* ] cannot be VP-fronted: compare (90) to (91).

(90) *John wanted to see the Great Pyramid, and have seen it he would if not for a sandstorm.

(91) John wanted to see the Great Pyramid, and seen it he would have if not for a sandstorm.

Nor can it be targeted by VP-ellipsis:

(92) John wanted to see the Great Pyramid, and he probably would have ~~seen it~~ (if not for the sandstorm).

(93) *John wanted to see the Great Pyramid, and he probably would ~~have seen it~~ (if not for the sandstorm).

These facts indicate that the complement of *have* has more functional structure than a bare VP – compare these judgments with other bare-VP contexts, such as the complement of the restructuring verbs *try* and *go* (Cable 2004). Thus, *have* must in fact be in a functional head below T, and *never* must have an available position below T as well, in order to appear to the right of *have* in (88c). Yet, if main verbs (such as *saw* in the following example) raise to T, their failure to permute with *never* cannot be explained:

(94) a. Sam never saw such a marvel.

b. *Sam saw never such a marvel.

The conceptual problem stems from the nature of the extrasyntactic economy principle that is invoked to rule out affirmative declarative *do* sentences. It is formulated as a general cognitive principle, but is allowed to vary in dialect specific ways (to capture the dialects that do in fact allow this construction). It also fails to explain why affirmative declarative *do*-support in EME has clear syntactic properties: there is no reason why a cognitive economy principle should weigh less on transitive sentences than on unaccusatives.[30] The account also assigns a different source to *do* in questions as opposed to negative declaratives, failing to capture the observed diachronic relationship between them – that is to say, identity of their causal source.

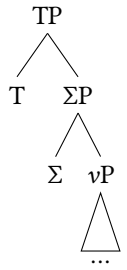
### 4.4.2 Last-resort models

A distinct class of syntactic theories of PDE *do*-support insert *do* only when it surfaces. I refer to these as the "last resort" models owing to the intuition that the insertion of *do* takes place to salvage what would otherwise be an ungrammatical structure.

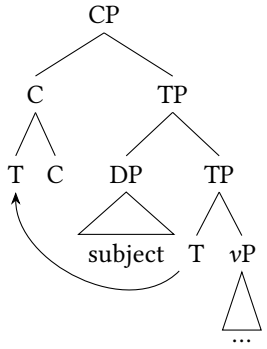
[30] And a similar argument could be constructed for other dialects with affirmative declarative *do*, insofar as any grammatical influences on *do*'s occurrence frequency are understood.

**Embick and Noyer (2001)**

Embick and Noyer (2001) frame their account of the behavior of English *do* in terms of a general Distributed Morphology (DM) theory of post-syntactic operations, which are postulated to obey strict structural-locality conditions. For them, a *v* head is adjoined to T by the syntax whenever T finds itself not in a sufficiently local relationship with *v* – that is, without a *v*P complement or *v* adjoined sister. This head is morphologically realized (Spelled Out) as *do*. This theory handles the case of negation and emphatic affirmation, since in such structures the complement of T is ΣP, not *v*P:[31]

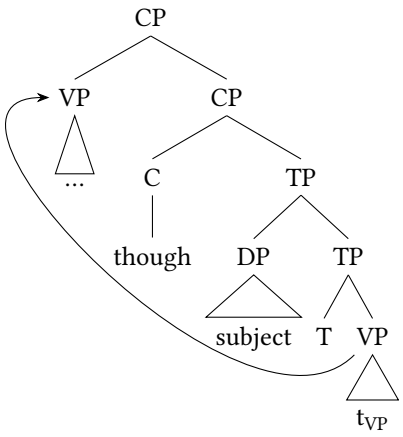



It also deals with cases of subject-verb inversion, since when T moves to adjoin to a higher head it (specifically its higher copy) no longer has *v*P as a compelment:




In cases of VP topicalization the account works, assuming the lower copy of VP does not count as a "complement" for the purposes of the *v*-insertion rule:

[31] Embick and Noyer (2001) also successfully capture a larger set of negation-related facts related to the interaction of constituent negation of the verb and sentential negation.




Embick & Noyer place their rule for *v* insertion in the syntax for conceptual reasons relating to the distinction between syntactic and phonological features: morphology, operating on the PF branch of computation, cannot access syntactic features, and vice versa for the narrow syntax. However, as the treatments of subject-auxiliary inversion and VP topicalization have shown, the insertion rule is heavily attuned to surface facts. Indeed, Embick & Noyer's treatment of *do*-support bears a striking similarity to that of Pullum and Wilson (1977), with a modern overhaul of the framework and a promotion of the notion of adjacency from strictly linear structure-sensitive (but only very local structure: head-complement relations).[32]

### 4.4.3 Difficulties

One difficulty that the insert-and-delete models face but last-resort ones do not is fundamentally diachronic. The results discussed by Kroch (1989) postulate that, given the observed identity in slope between the loss of verb raising across *never* and *do*-support, there must be a grammatical connection between these two phenomena. Insert-and-delete models cannot model such a connection, however, since their account of *do*-support is composed of two separate atomic grammatical operations: an insertion and a deletion. Taking first the case of PW's analysis, they might trace the availability of *do* as an auxiliary-like element to the *do* of VP ellipsis. This use of *do* dates back to ME at least:

(95) *ich him luuie & wulle do*
I him love and will do
'I love him and will.' St. Juliana, c1200, Visser (1963, §1751)

[32]The similarity of the two accounts is of course mirror-image with respect to the difference of whether it is a local condition which triggers insertion of *do* or its deletion.

(96) *ich þonke ȝou as ich wel aȝte do*
I thank you as I well ought do
'I thank you, as indeed I ought to.' Rob. Glouc., 1297, *ibid*

However, its modern syntactic properties which it shares with the modals cannot have been acquired earlier than the emergence of the latter category in the 1400s (Roberts 1985). In any case, for PW two independent changes must happen to the grammar simultaneously: *do* must begin to be inserted in all auxiliary-less sentences, and it must begin to be deleted from the affirmative declaratives. PW's account would fare better taking into account the proposal in section 4.3 that an auxiliary *do* is inserted productively in EME clauses including affirmative declaratives, since they could point to this as evidence that their insertion rule in fact does observably enter the language before the deletion rule does. They cannot, however, explain the argument structure effects on the insertion of *do* which are observed.

The picture is murkier still for the account in Schütze (2004). There, the analysis relies on several independent stipulations about the morphological properties of V: two of these are that it cannot spell out in a complex head with *not* and that it cannot spell out in a complex head with C. These conditions encode the necessity of *do*-support in PDE negatives and questions, respectively. Yet there is no explanation of why these conditions enter the grammar with the same slope, i.e. with the same underlying cause.

The question of whether this identity in slope posited by Kroch (1989) in fact holds in an expanded dataset is thus of critical importance for the question of the analysis of *do*-support in PDE – if Kroch's observations are borne out, then the shape of analyses of *do*-support in PDE, whatever theory of grammar they are couched in, are constrained to connect with the syntax of verb raising. It is possible that speakers of PDE have a different analysis of *do*-support. This sounds somewhat far-fetched at first blush, but it is not inconceivable. Recent developments in the English auxiliary system – such as the loss of V1 conditionals with certain auxiliaries and various developments in *be*+participle constructions (emergence of present progressive, which seems to occur separately – or at least with considerable time lag – for actives and passives) – may have pushed learners to a different analysis. However, to espouse such an analysis requires evidence, and the default position is of continuity in analysis across generations from the end of the *do*-support change to the present day.

## 4.5 Conclusion

This chapter has laid out many of the important theoretical and empirical contributions of this dissertation. I have demonstrated that *do*-support enters English not in a single step from a ME passive, but rather by passing through a discernable intermediate stage as an agentivity marker. I have explained how this picture builds on and refines previous accounts of the development of *do*-support, and how it can inform purely synchronic accounts of the phenomenon in PDE. In the following chapter, I will put these conclusions to the test in an unprecedentedly large corpus, discovering that they continue to hold in the face of new data, and in fact can be considerably enriched thereby.

# Chapter 5

# Data from a much larger corpus

## 5.1 Data gathering

The Early English Books Online Corpus (EEBO) results from a project of the Text Creation Partnership to make digitized scans and machine-readable text versions of English books (and other printed material, such as pamphlets) published between 1473 and 1700. This corpus is supplemented by a similar effort applied to the Eighteenth Century Collections Online (ECCO) corpus. (For simplicity, I will use EEBO to refer to both corpora.) The full text of the books which have been digitized is available to institutions which are members of the project. I have created a corpus from these texts in order to study *do*-support. Below, I will detail the steps used to create this corpus, including source code in R and Python.

### 5.1.1 Training a part-of-speech tagger

The EEBO texts are not annotated with any linguistic information. Part-of-speech (POS) tagging can be done automatically and highly accurately. State-of-the-art methods achieve per-word accuracy of around 97% on present-day written English (measured by the Wall Street Journal benchmark). (*POS Tagging (State of the art)* 2014) Given the preexistence of a large manually-annotated sample of linguistically similar material, namely the PPCEME, it is possible to apply an automatic POS tagger to the EEBO data. There are many POS taggers which perform basically at ceiling. For this work, I chose to use an averaged perceptron tagger (Collins 2002) implemented in Python (Honnibal 2013). Specifically, I used version `0.2.0` of the `textblob-aptagger` library (`https://pypi.python.org/pypi/textblob-aptagger`), released October 21, 2013.

The Python code used to train the tagger removes words which are tagged `CODE`, `LB`, and `ID` from the

PPCEME text, along with empty categories (*pro*, *, 0, etc.) and traces (*ICH*, *T*, etc.). These are artifacts of the corpus annotation, and are not present in EEBO texts. There are other possible improvements to this process which are not implemented here. These include:

- lowercasing all text (in the training data as well as the input to the tagger from EEBO)
- stripping dollar signs from the training text, which indicate textual emendations made by the PPCEME
- removing obvious artifacts of the OCR process (non-ASCII letter-like characters inside words)
- translating the PPCEME convention for contracted "n" (as in *jubilatiõ* with a tilde over the 'o') to the EEBO one (*jubilatiō* with a macron)
- making uniform the procedures used by the two corpora to account for and represent contraction processes orthographically

This training code runs single-threaded, and took several hours to run on an Intel i7 CPU.[1]

### 5.1.2 Tagging EEBO text

The next step is to use this tagger to analyze the EEBO text. The code extracts a list of sentences from an EEBO XML-format file, using the LXML library (Behnel and Faassen 2014). The function first extracts the content of the BODY tag from the file. It then removes any NOTE tags (corresponding to marginalia or footnotes, which are intercalated in the text where they appear textually without regard for linguistic structure) and L tags (corresponding to lines of metrical text, excluded from analysis for parallelism with the PPCHE). It then reformats any GAP tags, which represent unreadable spans of text. Specifically, the GAP tag is removed and its DISP attribute (containing a character to represent the discontinuity, often a bullet in the case of single-letter gaps and a pilcrow in the case of longer ones) is inserted in the surrounding text. A crude notion of sentence boundaries is created by splitting on all occurrences of a period or question mark. This algorithm misperforms on a small number of texts where sentence divisions are indicated with semicolons instead of periods, but in general it works well enough.

The code then simply runs the previously-trained tagger on each sentence in a given file. This code runs in 8 threads (corresponding to the 8 virtual cores of the machine it is run on), and takes slightly more than

[1]Specifically, the CPU used in these tests is a 4-core (8-thread) Intel i7. The specific model (as reported by Linux's /proc/cpuinfo) is: Intel(R) Core(TM) i7-2670QM CPU @ 2.20GHz . The computer has 16GB of RAM, and the corpus was stored on a 7200 rpm hard drive. (Disk I/O – whether reading the corpus or swapping excess RAM pages to disk – is almost certainly not a bottleneck in any of the workflows described.)

one day to complete (for a set of 44,422 texts from the EEBO portion of the corpus; the ECCO has 2,473 texts and completes tagging in a proportionally quick time). The code writes each file's word-tag pairs to disk, to allow them to be reused in later analyses without needing to re-run the tagger.

### 5.1.3 Recognizing negative declaratives

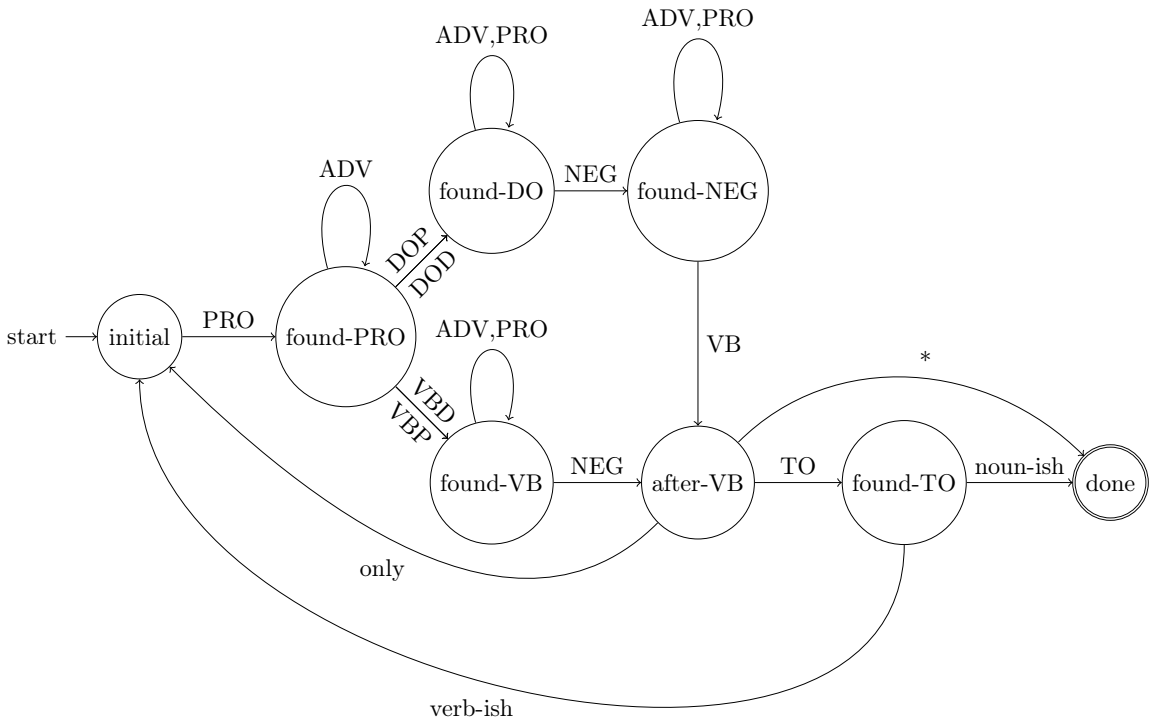



*Figure 5.1:* A visualization of the algorithm used to recognize negative declaratives from POS-tagged EEBO text.

This code to recognize negative declaratives in the tagged EEBO corpus implements the automaton in Figure 5.1. It looks for sentences which have a pronoun subject, followed by either a *do*-less verb or *do* + a verb, followed by negation in the appropriate spot. Adverbs are allowed to intervene between the pronoun and the verb, as in "he often does not see" or "he often sees not". Both pronouns and adverbs can intervene between (with *do*-support) *do* and negation, and negation and the verb; and (without *do*-support) the verb and negation. This encompasses examples such as "he does often not see," "he does not often see," and "he sees it not." (For each of these, there is a less-plausible alternative generable by swapping "often" for an object

pronoun or vice versa, but the acceptance of these alternatives is not believed to affect the performance of the detector in a serious way.)

When there is no *do*-support, negation follows the verb, and can sometimes be detected to be associated with some post-verbal constituent, rather than with the verb itself. In other cases, it is ambiguous which of these two positions *not* should occupy. Thus, the detector excludes two classes of sentence. The first consists of tokens where the negation is part of the string "not only." The second consists of tokens where the negation is followed by "to" and then a verb-like category (including verbs and adverbs, among others). With *do*-support on the other hand, the negation is bracketed between *do* and the verb, thus making it impossible (up to downright errors in the POS tagging) for it to be misconstrued with the verb instead of with a following constituent. However, in order to treat the *do* and non-*do* conditions fairly, these sentence types are excluded from both conditions alike. Thus, "he does not see only the good side of the situation" is excluded because "he sees not only the good side…" is; and similarly for "he does not know to take off his shoes in the house" and "he knows not to take off…" (the latter pair is disambiguated by world knowledge, but this is irrelevant to the automatic classifier of course).

There is one other obvious exclusion that this code fails to make: tokens with "not X but Y" constructions. There is a parameter affecting such an exclusion, namely how far along in the string to look for the *but*. Too short a window will miss too many necessary exclusions, whereas too long will spuriously exclude too many sentences. There is no length which is *a priori* optimal, but it would be possible to experiment with the rates of false in-/exclusion under various settings of this parameter (after first assembling a subsample of the data with suspicious "not…but…" constructions and hand-coding them). This investigation is not considered a priority.

For each negative declarative token, the following information is recorded:

- the text of the token
- the text of the recognized verb
- the text of the pronoun subject
- the file the token comes from
- whether *do*-support was detected
- how many words into the token the verb is

*Table 5.1:* A comparison of the number of tokens available in various corpora of *do*-support.

| | Neg. Decl. | Aff. Decl. |
|---|---|---|
| EEBO | 590361 | 6730949 |
| PPCHE | 6293 | 145902 |
| Ellegård | 7604 | — |

This procedure yields a dataset with 590361 negative declaratives. A similar procedure adapted to affirmative declarative sentences yields 6730949 tokens. These numbers are compared to the totals in the PPCHE and Ellegård in table 5.1.

### 5.1.4 Extracting dates from text info files

The EEBO is distributed with header files for each of the texts it contains, which include the details of original publication. The publication date information is not consistenly represented, however using some Python code it is possible to automatically extract this information from the header files. The script returns date information for 42,546 texts (95.8% of the number of texts in the corpus), with an additional 46 (0.1%) of the dates being greater than 1700, and thus invalid (these texts are discarded).

### 5.1.5 Lemmatization

For examining the behavior of different lexical classes in the EEBO data, it is necessary to map the individual spellings of verbs to a lemma. With a dataset of this size, it is infeasible to examine each sentence in context in order to assign it a lemma. It is even laborious to scrutinize all the unique orthographies for verbs in the dataset (of which there are 18323). Instead, a heuristic procedure was used. Spelling variants of a verb stem are combined with a variety of affixes. For regular verbs, the set of affixes is generated by combining, in order:

1. a linking vowel from the set {e, i, y, apostrophe, empty string}
2. a suffix from the list:
    - s, th, þ, t, tt: third person singular
    - st: second person singular (*thou*)
    - d: past tense

3. a final vowel: either ‘e’ or the empty string

To this list are added several ‘Vn’ endings corresponding to the infinitive and plural suffixes of ME and early EME, as well as a single final ‘e’ or apostrophe. For strong verbs, a modified procedure is used whereby the present/infinitive and past stems are specified; the past stems can be inflected with only a small subset of the possible endings listed above (corresponding to the 2sg ‘st’ and pl ‘Vn’).

This procedure is seeded with a list of 269 verbs. This suffices to lemmatize 2753 spellings in the negative declarative dataset and 4291 spellings in the affirmative declaratives. This translates to 80 percent of the negative tokens and 62 percent of the affirmatives being lemmatized.

## 5.2 Results

### 5.2.1 Validation of corpus

Using the methods detailed in the previous section, it is possible to assemble a corpus of *do*-support tokens. The picture that emerges is given in Figure 5.2. It is desirable to compare the results obtained from this corpus to previously available datasets, in order to assess the new results’ validity; this comparison can be seen in Figure 5.3. Strikingly differently from the data from Ellegård’s corpus and the PPCHE, the EEBO data does not show any deviation from upwards monotonicity (but rather only a lessening of the slope). A small downwards movement can be induced by setting the $\alpha$ parameter of the LOESS smoothing algorithm to a relatively low value (e.g. 0.3); however it is much smaller in magnitude than that of the other two corpora. The EEBO data also matches Ellegård’s corpus rather than the PPCHE in the steepness of its later trajectory.[2]

[2]When including data from the PPCMBE in the graph, the steepness of the PPCHE data decreases, bringing its trajectory closer in line with the other two corpora. However, smooth curves fit over different time ranges are very difficult to interpret relative to each other, given the properties of the LOESS smoother discussed in section 4.1.1.

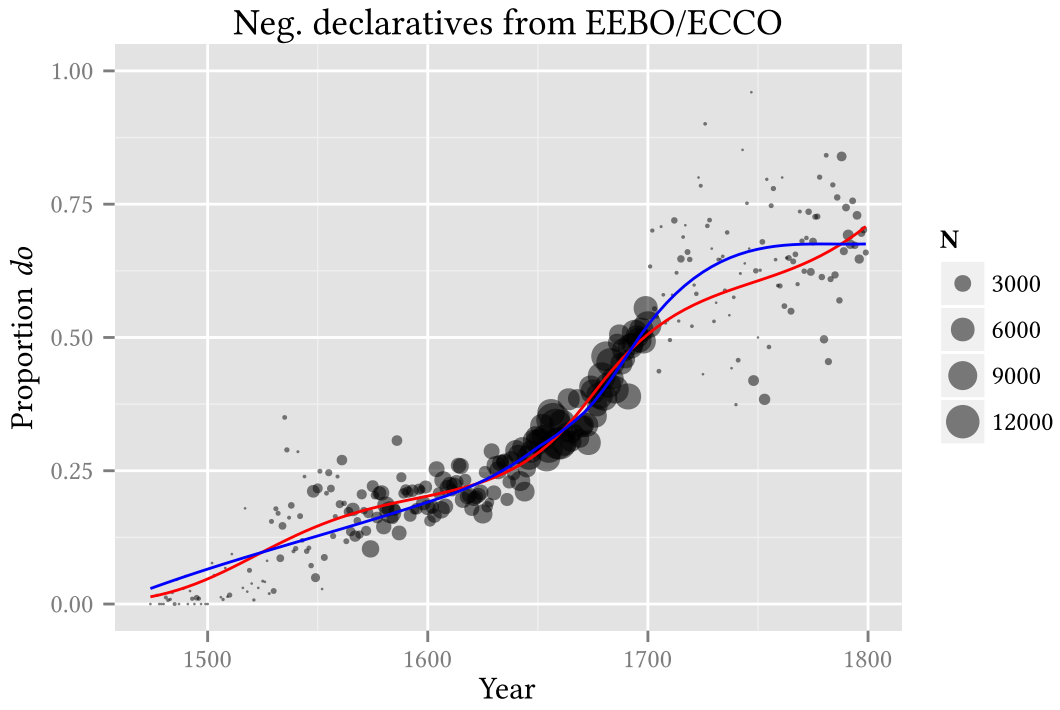



*Figure 5.2:* The trajectory of *do*-support in negative declaratives in the EEBO corpus. The blue line is a LOESS smooth with $\alpha = 0.7$. The red line is a smooth fit using a logistic regression over a cubic B-spline basis with three evenly-spaced knots at the 0.25, 0.5, and 0.75 quantiles of the data.

It is desirable to understand where this difference in the trajectories stems from. It is possible that the smaller sample sizes of the two previous corpora give a biased picture of an underlying reality which is more closely represented by the EEBO data. It is equally possible, however, that the uncertainties associated with the EEBO data obscure a subtle pattern which is more sharply reflected by the other two corpora. In order to test these hypotheses, we will conduct a series of resampling experiments.

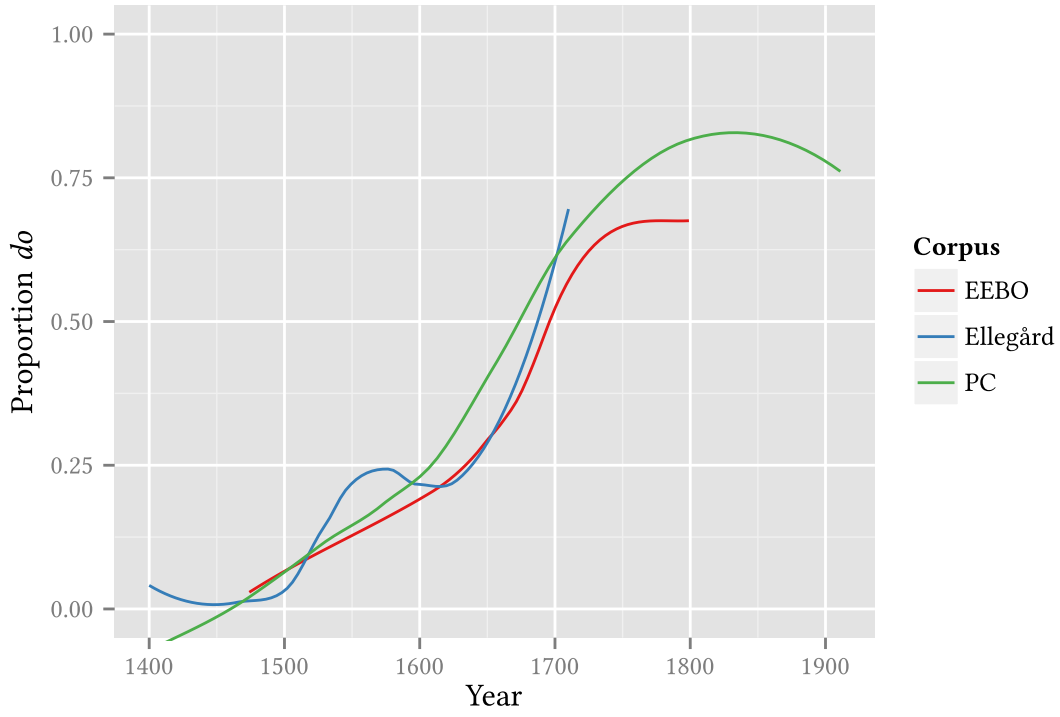



*Figure 5.3:* The trajectory of negative declarative *do*-support in various corpora.

In the first series of experiments, we will resample texts from the EEBO data. The population of texts available for resampling is restricted to those with more than 10 tokens of negative declarative *do*-support. There are 5 sampling regimes:

1. Draw a sample of text sizes from the empirical distribution of text sizes in the PPCEME, with replacement. There are 278 texts in the PPCEME, and thus 187 texts are used from EEBO. The EEBO texts' tokens are resampled (with replacement) to match the size distribution from the PPCEME.
2. The same as 1, using the size distribution and text count (5247) from the PPCEME+PCEEC.[3]
3. The same as 1, using the size distribution and text count (109) from Ellegård's corpus.
4. The same as 1, using 10 tokens per text and a text count of 175 (roughly halfway between the text count of the PPCEME and Ellegård).
5. The same as 1, but including all of the tokens from each EEBO text and using a text count of 175.

[3]This procedure overestimates the informativeness of the PCEEC, since it consists of many short texts by the same authors. Nonetheless, the additional complexity of a model which takes into account author-text relationships does not seem justified in light of the results obtained here, which clearly show that the non-motonicities seen in the PPCHE and Ellegård are plausible outcomes only of the reduction in sample size.

Repeating the sampling process 100 times for each condition, we compute a variety of measures of the trajectory of the evolution of *do*-support within each sample. These measures include:

1. The minimum year-to-year difference in the LOESS prediction ($\alpha = 0.7$) in the interval from 1550–1625
2. The number of years in which the LOESS-estimated slope is negative over that interval
3. The slope of the line between the minimum and maximum LOESS-estimated points in that interval

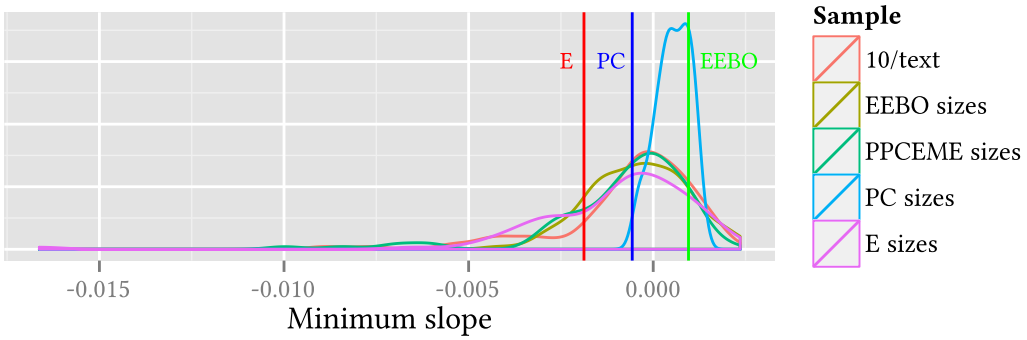



*Figure 5.4:* Density of the minimum yearly slope from 1550–1625 of a LOESS model fit to EEBO data resampled under various techniques. The actual values in Ellegård's corpus (E), the PPCEME+PCEEC (PC) and EEBO are indicated with vertical lines.

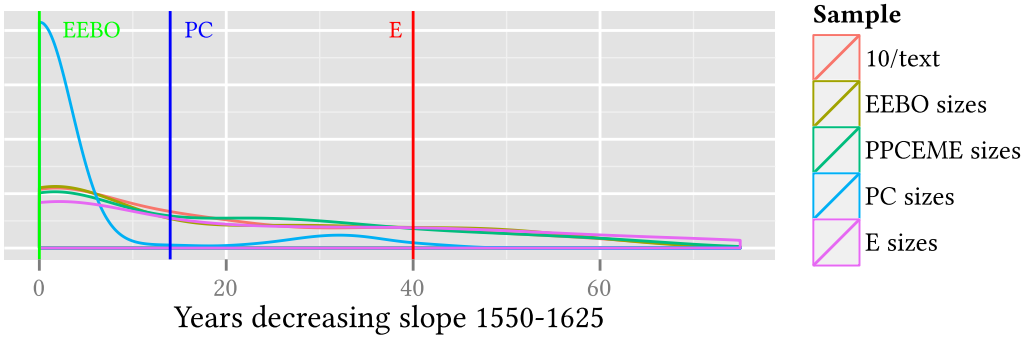



*Figure 5.5:* Density of the number of years between 1550–1625 that the slope of a LOESS model fit to EEBO data resampled under various techniques decreases. The actual values in Ellegård's corpus (E), the PPCEME+PCEEC (PC) and EEBO are indicated with vertical lines.

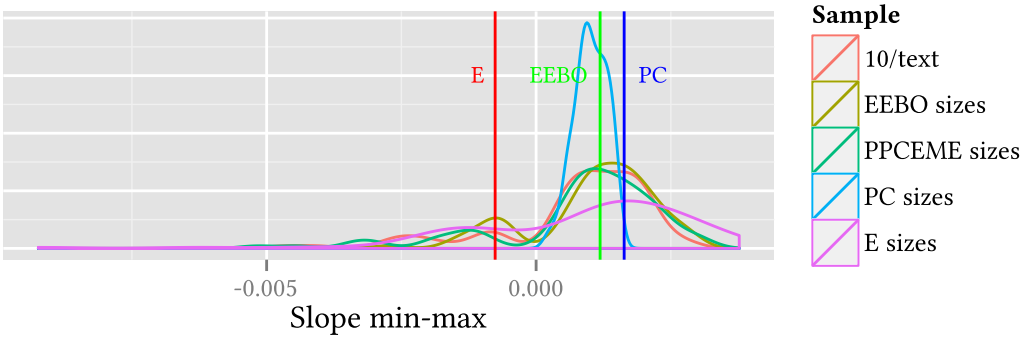



*Figure 5.6:* Density of the slope between the LOESS-fit minimum and maximum points between 1550–1625, using EEBO data resampled under various techniques. The actual values in Ellegård's corpus (E), the PPCEME+PCEEC (PC) and EEBO are indicated with vertical lines.

*Table 5.2:* The quantile of resampled data into which the actual minimum yearly slope in Ellegård's corpus falls.

| Sampling method | Min. yearly slope | Years decreasing | Slope min-max |
|---|---|---|---|
| 10/text | 0.09 | 0.84 | 0.14 |
| EEBO sizes | 0.01 | 0.82 | 0.11 |
| PPCEME sizes | 0.08 | 0.85 | 0.18 |
| PC sizes | 0 | 0.99 | 0 |
| E sizes | 0.08 | 0.73 | 0.23 |

The results of this experiment are shown in Figures 5.4, 5.5, and 5.6. Taking the graphs in order, Figure 5.4 shows that the minimum slope in the PPCHE and EEBO are both approximately equidistant from zero on opposite sides, and both fall within a region of high probability density. The minimum slope in Ellegård's corpus is somewhat farther to the left. For Figures 5.5 and 5.6, the situation is similar in that the EEBO and PPCHE values are close to the mode of the distributions while Ellegård's corpus is somewhat farther away. Table 5.2 gives the quantiles of the observed measurements from Ellegård's corpus under various sampling regimes; these are somewhat extreme but not generally beyond the 0.025 ritualized significance threshold (= 0.5 divided by 2, for a two-tailed test) except in the case of the overly informative sampling regime based on treating the PCEEC texts as independent.

### 5.2.2 Lexical classes

Using the data from the EEBO corpus, it is possible to investigate further the composition of the lexical classes broadly identified in section 4.3 above as being associated with argument structure.

#### Unaccusatives

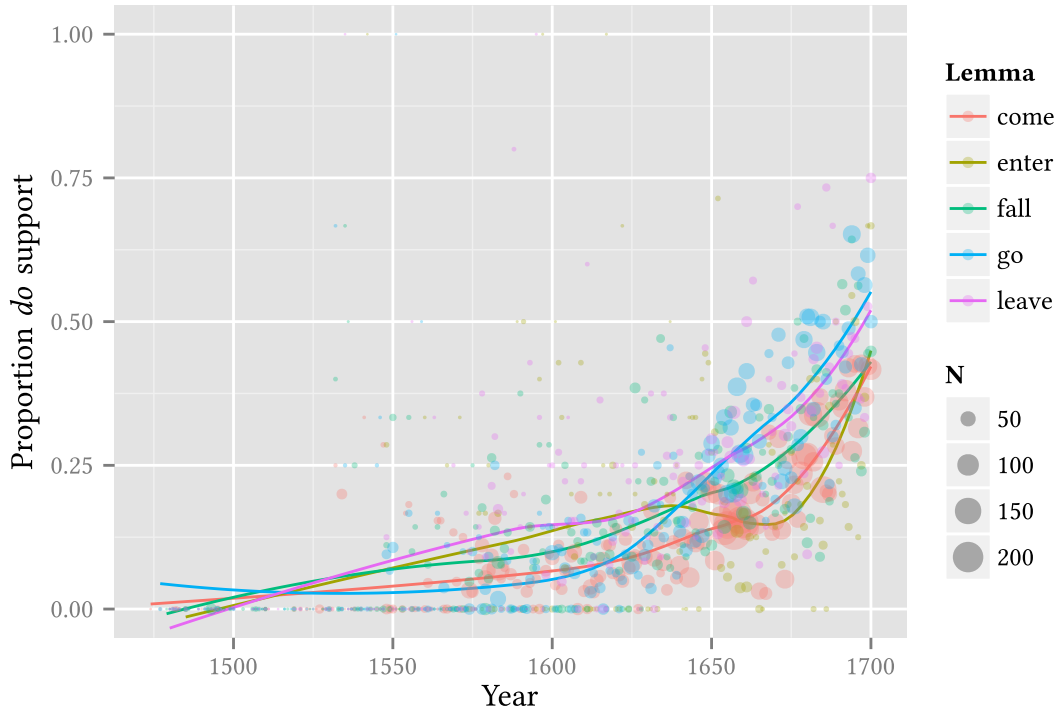



*Figure 5.7:* The behavior of several verbs of inherently directed motion in negative declaratives in the EEBO dataset.

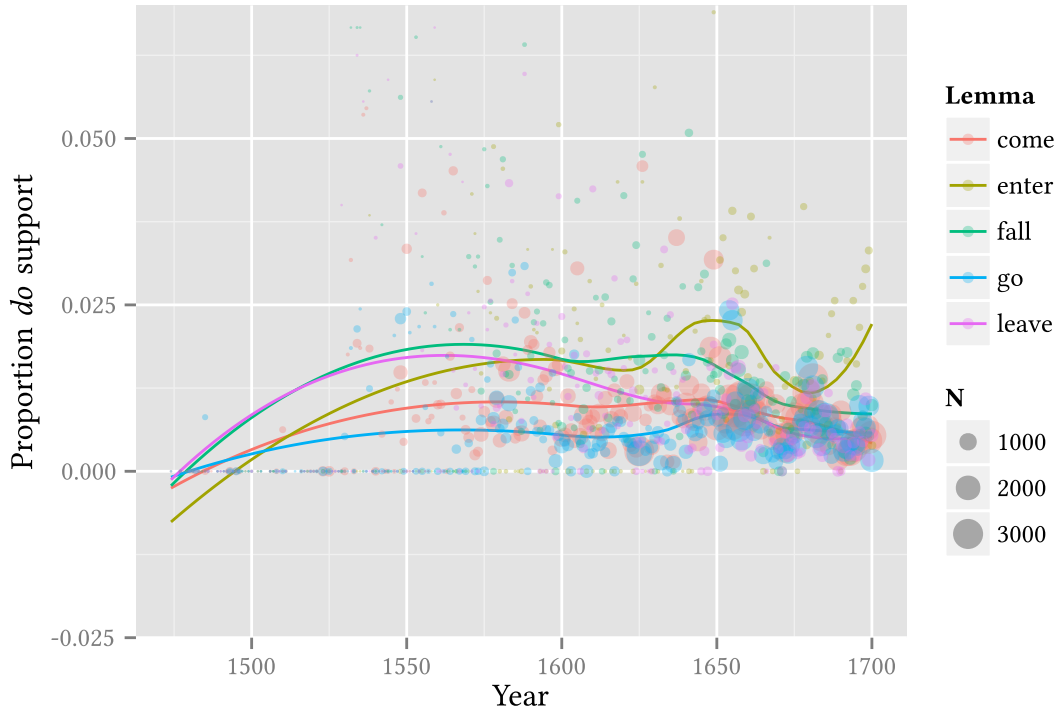



*Figure 5.8:* The behavior of several verbs of inherently directed motion in affirmative declaratives in the EEBO dataset.

Figures 5.7 and 5.8 present the behavior of several verbs of inherently directed motion in negative and affirmative declaratives, respectively. The list was taken from "class 2" of Levin (1993), and further filtered to those verbs occurring more than 1,000 times in the negative declarative data.

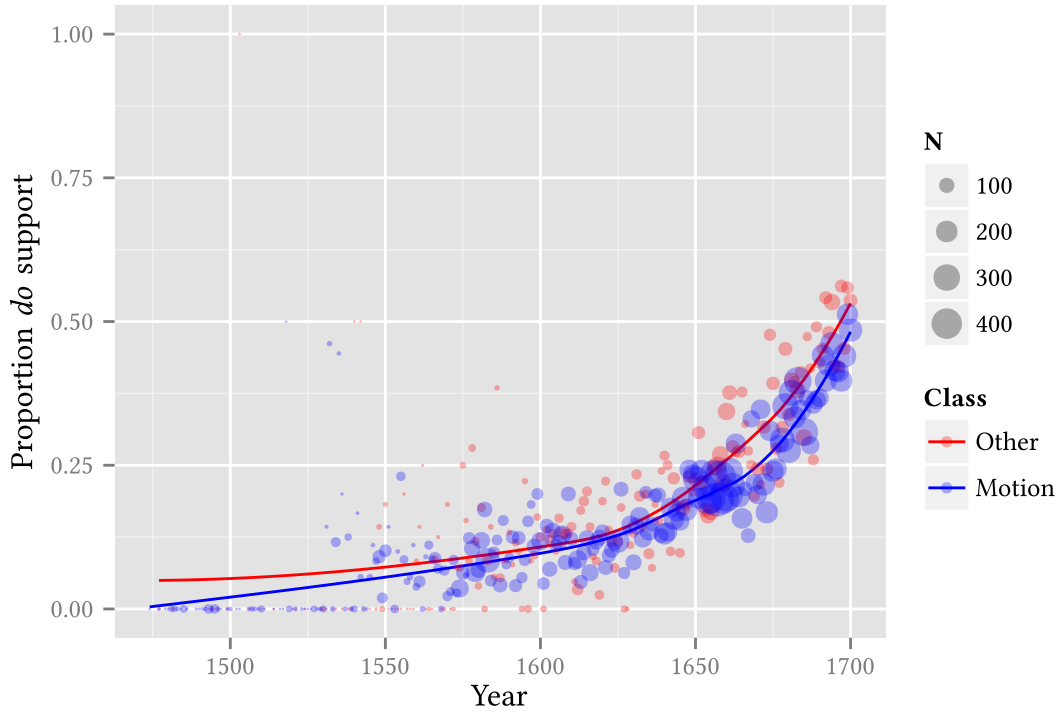



*Figure 5.9:* The behavior verbs of inherently directed motion and other unaccusatives in negative declaratives in the EEBO dataset.

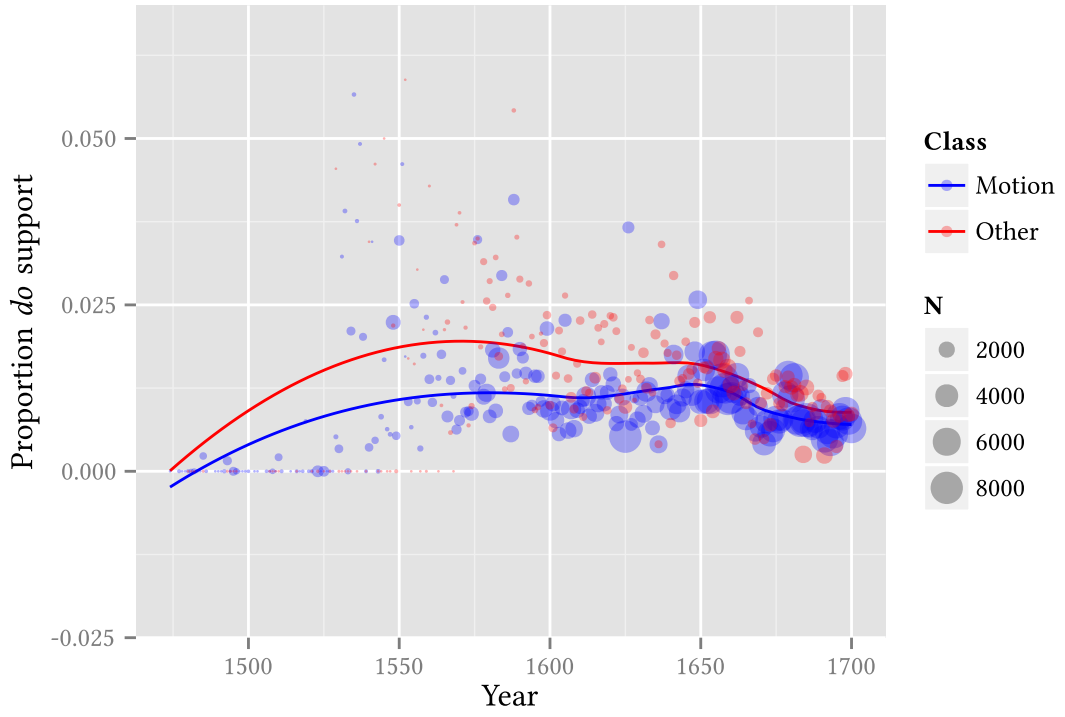



*Figure 5.10:* The behavior verbs of inherently directed motion and other unaccusatives in affirmative declaratives in the EEBO dataset.

Furthermore, figures 5.9 and 5.10 demonstrate that the behavior of the inherently directed motion class is very similar to that of other unaccusatives (represented by a heterogeneous group comprising the verbs *live*, *die*, *stay*, *perish*, *prevail*, *depend*, *extend*, and *remain*). In the case of the affirmative declarative graph, note that the visual difference between the two trajectories on the graph is accentuated by the very small range of the y-axis. Thus, the results from section 4.3 are validated over a broader class of unaccusatives, though the PPCHE can provide ample data on only two verbs both of the inherently directed motion class.

**Transitives**

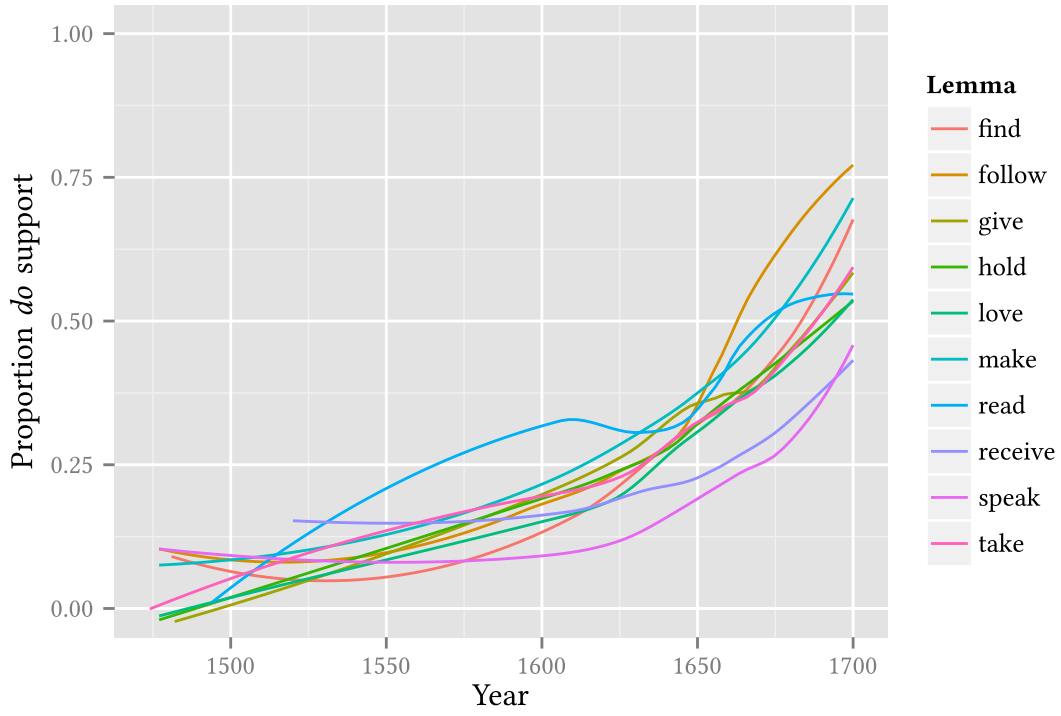



*Figure 5.11:* The behavior of various high-frequency transitive verbs in negative declaratives in the EEBO corpus.

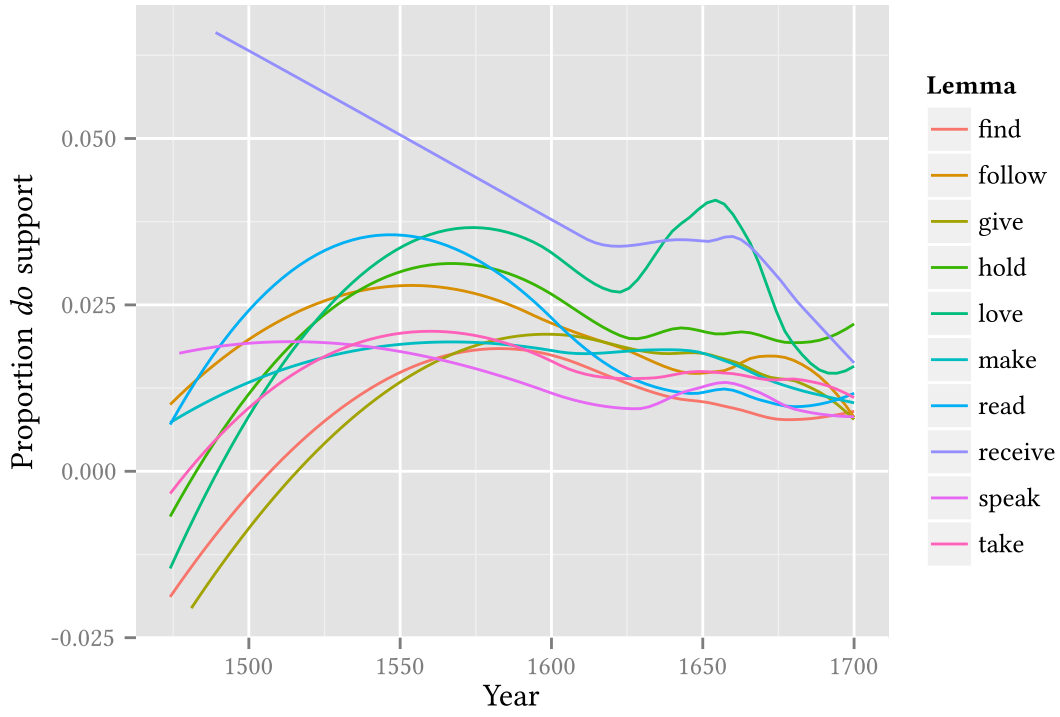



*Figure 5.12:* The behavior of various high-frequency transitive verbs in affirmative declaratives in the EEBO corpus.

Figures 5.11 and 5.12 show the behavior of the ten most frequent (in negative declaratives) transitive verbs which I have judged unlikely to have *that*-clauses as their arguments.

### Clausal object transitives

Figures 5.13 and 5.14 show the trajectories of some common verbs which take clausal objects. In negative declaratives, these divide into three groups. The highest group is composed of the verbs *remember*, *believe*, and *observe*, which have consistently higher rates of *do*-support than nominal-object transitives. I'll call this the high clausal object class. ("High" is a purely descriptive reference to rates of *do*-support, not a reference to high syntactic positions.) The second class is composed of the verbs *see*, *hear*, and *feel*. These have rates of *do*-support roughly comparable with NP-object transitives. I'll call this the sensing class. (This is of course related to the fact that these verbs, especially *see* and *hear*, are frequently used as NP-object transitives.) Finally, the group with the lowest rate of *do*-support is comprised of the verbs *know*, *regard*, and *doubt*. This class I will refer to as the low clausal object class. The verb *deny* seems to change its class affiliation,

transitioning from the high clausal class to the sensing class.

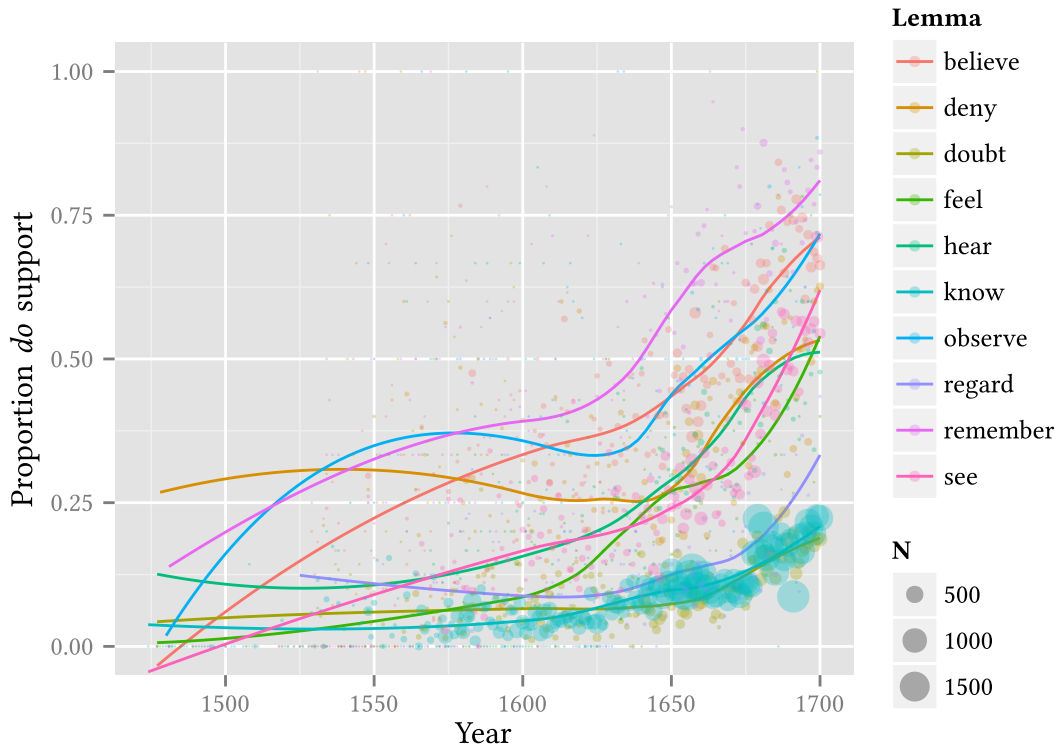



*Figure 5.13:* The behavior of various clausal-object verbs in negative declaratives in the EEBO corpus.

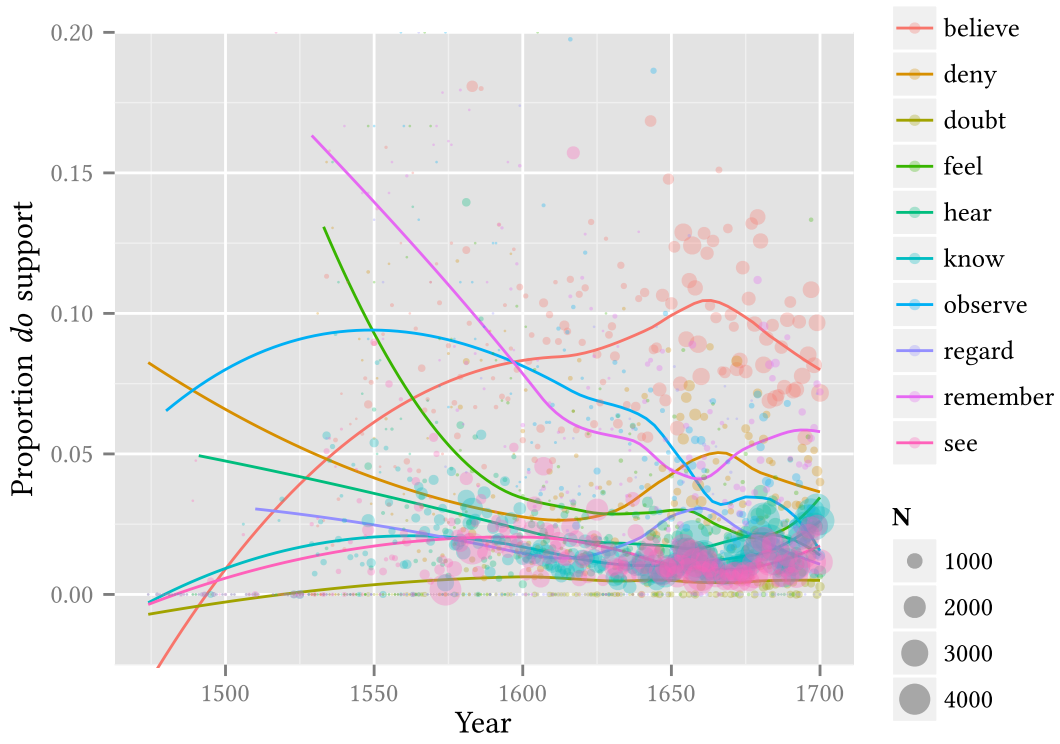



*Figure 5.14:* The behavior of various clausal-object verbs in affirmative declaratives in the EEBO corpus.

**Summary**

Figures 5.15 and 5.16 aggregate the *do*-support behavior of verbs in the EEBO data according to the lexical classes as shown above. Beginning with the negative declaratives, there are striking differences between the low and high clausal object groups and all others. The difference between (on the one hand) unaccusative verbs of motion and other unaccusatives and (on the other) transitive verbs and verbs of sensing is evident in the century between 1550 and 1650, though obscure before and after. In the affirmative declaratives, there is a large difference between the high clausal object verbs and all other classes. There is also a smaller difference between the motion unaccusatives and the other classes, which is clearly visible until the last 50 years of the dataset. On the other hand, unlike in the negative declaratives, the other unaccusatives do not pattern with the motion verbs, but rather identically to the other classes, including the transitives. The low clausal object class does not have distinctive behavior in this context.

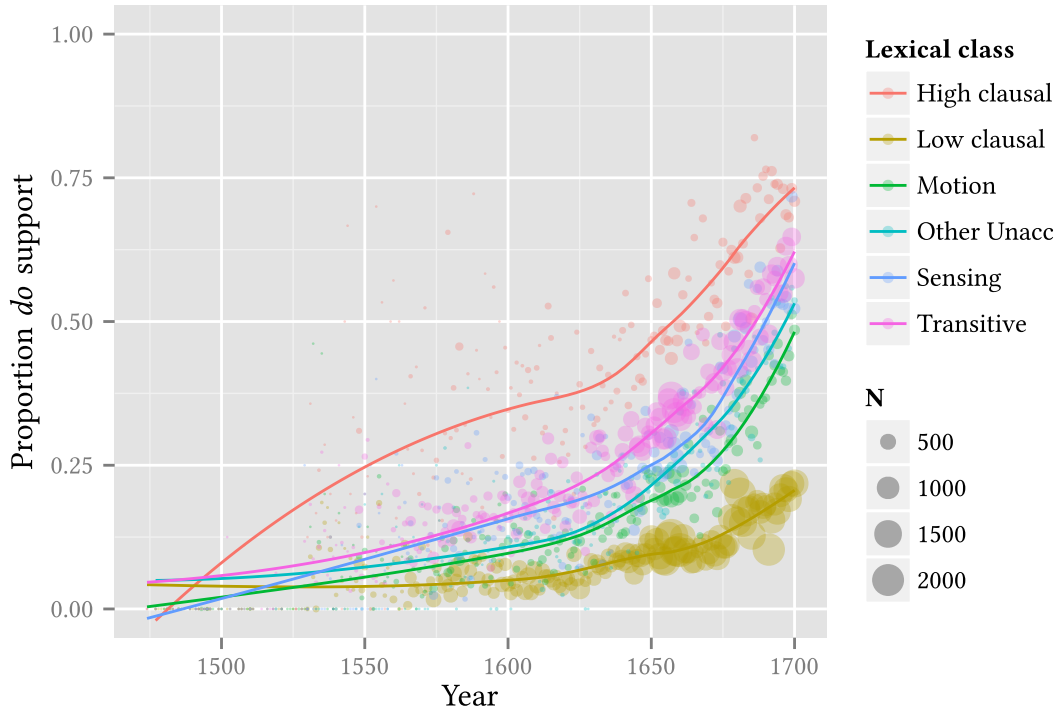



*Figure 5.15:* The behavior of various lexical classes (as defined above) in negative declaratives in the EEBO corpus.

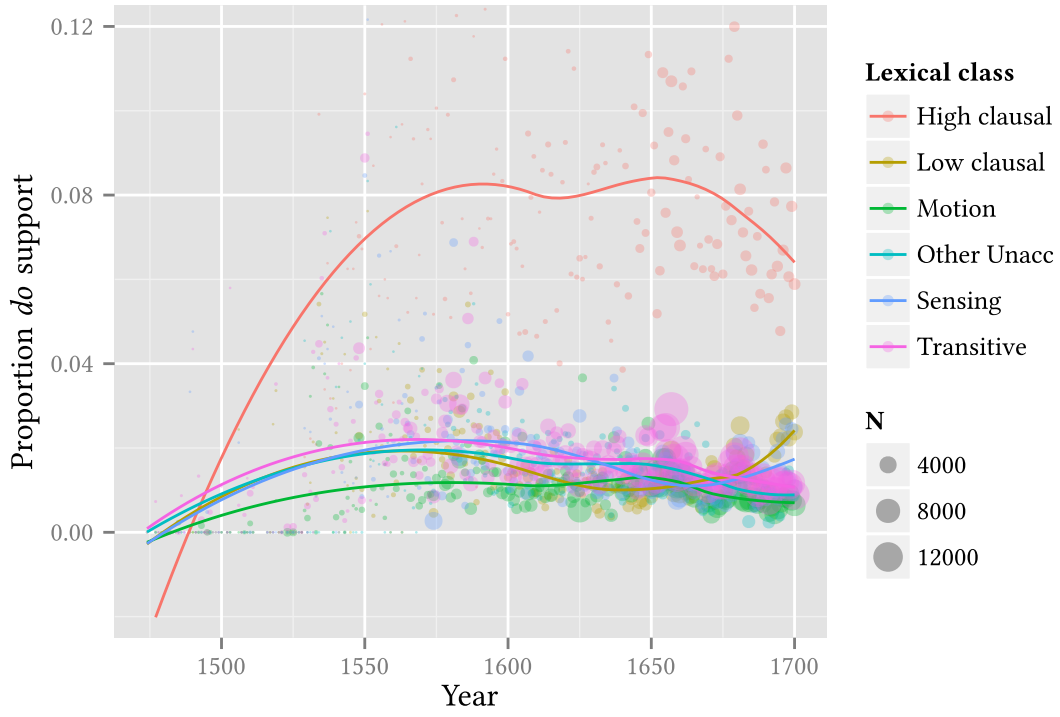



*Figure 5.16:* The behavior of various lexical classes (as defined above) in affirmative declaratives in the EEBO corpus.

This data extends the results on intermediate *do* from section 4.3. The previously-observed argument structure effect is borne out, in the main. However, the largest effects observable in the data lack a compelling underlying semantic generalization. They might be more fruitfully treated as some combination of inherently lexical effects and the covariance between lexical items and textual style. The latter explanation seems to be on the right track for the verb *regard*, which is in the low clausal object class. Examining a small random sample of occurrences of this verb from the corpus, it appears to be associated with texts of a religious nature to a degree which other verbs are not. Religious texts constitute a conservative style, and thus may be responsible for dragging down the surface frequency of *do*-support with this verb.

On the other hand, *know* seems a good candidate for an inherently lexical effect. Its usage does not seem to be particularly strongly associated with any particular style.[4] Rather, it is a frequent verb with a fundamental meaning, and thus can occur across a range of genres. At the same time, there does not seem to be a good reason to distinguish its syntactic behavior from other verbs with similar distributions and

[4]Though this impression has not yet been confirmed in a rigorous way.

meanings. The noticeable peculiarities of its behavior thus are good candidates for arbitrary properties associated with this verb's lexical entry (in the same way that the phonology of a verb is – barring cases of onomatopoeia and similar phenomena – an arbitrary accident).

The picture that emerges is that:

1. there is a single underlying trajectory to the change, which all contexts follow (upwards for negatives, up then down for affirmatives)
2. (certain) unaccusatives, as a group, are delayed in their progress along this trajectory
3. other specific lexical items may also show oddities

The situation is reminiscent of the organization of categorical (as opposed to variable) parts of grammars. Taking English plural noun formation as an example:

1. plurals are formed in general by adding the *-s* suffix
2. there is a basically semantically coherent class of large game animals which form plurals with a zero morpheme (as well as several other classes which are not conditioned by semantics)
3. there are a few lexical items which are genuine exceptions to any generalization, such as *person/people* and *child/children*

This parallelism indicates that the processes which lead to the observed diachronic stratification in the *do*-support data may be amenable to study using the same models of learning as have been developed and applied in the literature on the learning of morphological patterns.

**Lexical classes and *never***

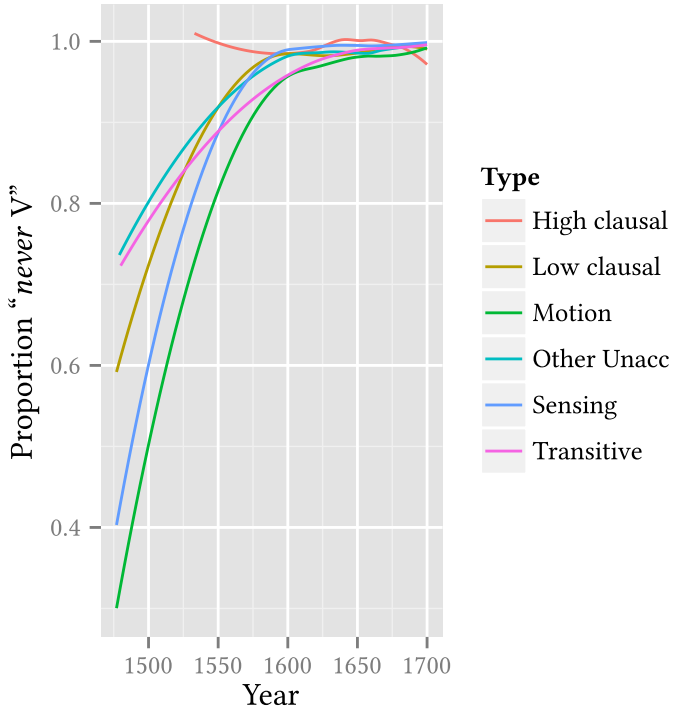



*Figure 5.17:* The behavior of lexical classes with V-to-T raising across *never* in the EEBO corpus.

It is also possible to extract a corpus of tokens of potential V-to-T raising past *never* from the EEBO corpus. That data is plotted in figure 5.17.[5] This data indicates clearly that the high clausal class has very unusual behavior, and appears to never allow raising to T past *never*.[6] The motion verb class has clearly distinct behavior from the other classes of verbs, and conversely does not pattern with other unaccusatives in this data. This is not unexpected, given the hypothesis that the special behavior of unaccusatives emerges from a reinterpretation of *do*'s causative function in ME, which is not implicated in the grammar of V-to-T.

### 5.2.3 CRH regression

Examining a dataset of this size presents challenges to the usual statistical implementation of the CRH, which proceeds by comparing a logistic regression model with and without a (set of) slope parameters in order to determine whether the CRH is obeyed, generally finding that it is. However, in some cases such

[5]The verb *believe* was removed from the high clausal class, because "I believe" is often inserted as a parenthetical before *never*. The low level of structural analysis available in the automatic annotation of the EEBO data means that these usages cannot be distinguished from instances where *never* and *believe* are members of the same clause, and the latter has raised over the former.

[6]Though note that there is a lack of data on these verbs in the beginning 50 years of the change.

findings are reached only after applying various corrections for superficially CRH-violating patterns in the data, which must be explained in terms of factors external to the syntactic change which interfere with its manifestation in the data. My contention is that this methodological practice should be interpreted as a sort of "Constant Rate Horizon:" the limit on the visibility of non-CRH phenomena when investigated through the lens of the CRH. Datasets where we expect to observe the CRH yet do not have yielded more information about factors which interfere with changes than they have about the truth or falsity of the CRH. This is certainly the case with *do*-support, where massive violations of the CRH (and indeed more broadly the population-based S-curve model of changes) have not called into question the truth of the models, but rather turned up evidence of social and grammatical reanalyses which punctuate the quantitative development observed in corpora. None of this is to say that the CRH is in principle unfalsifiable or useless as a scientific tool. However, as discussed above in the section on power analysis, previous CRH studies (at least as exemplified by Kroch (1989) on *do*-support) have been moderately underpowered. Thus, in principle we don't expect to find enough information in these studies to falsify the CRH.

In larger datasets, however, there is enough evidence to find a CRH-violating difference between virtually any two contexts. In other words, as datasets are enriched, the Constant Rate Horizon will grow more distant, and the number of discoveries made in its pursuit will grow. At some point, it may turn out to be the case that CRH-violating slope differences will be found in corpus data which defy reduction to other phenomena. When such a point has been reached, the usefulness of the CRH interpreted as a horizon will be at an end. However, in the case of *do*-support, given the large number of grammatical and social factors discussed by many authors, and summarized and extended above, it would be premature to conclude that we are close to the CRH horizon.

Now I'll move on to discussing some concrete examples of apparent CRH violations uncovered in the exploration of the EEBO dataset. To begin with, it is possible to compare the behavior of different subject pronouns in a dataset comprised of transitive verbs before 1575. (The inquiry was restricted to non-2nd person pronouns, to avoid complications introduced by the loss of *ye* during this period. One might also be suspicious about *thou*, though it remains robustly present until the 17th century; to be completely conservative this was also excluded.) This yields a finding that the model which allows the slope to vary across different subjects is preferred to a model which constrains the slope to be identical ($\Delta$AICc = 6.16, LRT $p$-value = 0.006). On its own, this might be interpreted as an effect of style or genre, since a different ratio of first:third person pronouns is expected to characterize texts of differing styles.

It is also possible to test whether individual verbs differ in their slopes, or share a common underlying

slope. The necessary model cannot be fit by a classical logistic regression (the number of parameters overdetermines the data); however it can be fit by a hierarchical model which constrains the lexical item effects to be drawn from a normal distribution, rather than allowing them each to vary independently of the others. This again yields a result that the CRH-violating model with individual slopes per verb is a better fit to the data than the model with a single common slope and per-verb random intercepts ($\Delta$AICc = 3.27, LRT $p$-value = 0.026). It is possible that this result is driven by genre differences appearing through the differential association of different verbs with genres. It is also possible that different verbs have different slopes for intrinsic reasons (as it was argued above that *know* demonstrates in a particularly dramatic way).

In either case, these results point to three important observations about the CRH:

1. As previously discussed, discovering the lack of a CRH effect often signals not that the CRH is somehow invalid, but that there is more investigation to be done of non-syntactic factors. In the present case, we have Warner (2005) to thank for a preliminary model of stylistic effects on *do*-support. This model should be extended using the data from the EEBO corpus in an attempt to cover the phenomena sketched above. Only in the event that this effort is unsuccessful should attention return to the validity of the CRH itself in this dataset. Put another way, the CRH is a model of the regulation of syntactic change; it is silent on extra-grammatical sources of influence on trajectories. A complete model of syntactic change must account for both grammar-internal and -external factors.

2. It is necessary to divorce two aspects of the CRH as originally proposed by Kroch (1989). One is the insight that syntactic changes tend to share a single underlying trajectory. This impression is confirmed by the data presented in this section (showing, for example, the unity of behavior among specific members of the broad lexical classes transitive and unaccusative). The second aspect of Kroch's proposal was that logistic regression be used to test mathematically for the presence of such behavioral similarities. In much larger datasets such as this one, that testing procedure cannot be used unmodified; rather some account must be taken of the evidence that inheres in the data about extra-grammatical facts. In the limit, we see that Krcoh's classical single-level logistic regression cannot even cope mathematically with a model that specifies a modest number of per-lexical-item terms. However, the augmentation of the classical model with more sophisticated procedures need not entail the discarding of the first aspect of the CRH.

3. The CRH is fractal, in the sense that the failure of the CRH across one set of contexts does not adversely impact the ability to test and measure CRH effects across other dimensions. What gives rise to this

state of affairs is the partitioning of variation that regression procedures implement. When fitting a regression model, an attempt is made to find components of the variation that correspond solely to a single predictor. Any variation not accounted for by a predictor is subsumed in an error term. Even in the presence of extra-syntactic influences, it is possible to recover and quantify the contribution of syntactic factors.

In order for regression models to fully realize the promise expressed above, it is necessary that the syntactic predictors of interest not be correlated with other influential factors which are not controlled in the regression model. In practice, however, various aspects of linguistic expression are intrinsically highly correlated. To name just one example which has already been touched upon in this dissertation, certain syntactic constructions and lexical items are more frequently found in certain styles. However, the precise numerical estimates of syntactic contextual effects are rarely of interest. Some attention to correlations between predictors of different types is needed to ensure that inferences about the direction of influences (positive or negative sign of regression coefficients) is not invalid. In the CRH paradigm, though, the slope coefficients of the regression are often of the most interest (and specifically whether they meaningfully contribute information to the model). Because of that, it is important to ensure that any extra-syntactic factors which might be correlated with the contexts under investigation are held constant across the data. Ellegård's dataset is not exemplary in this regard, being somewhat irregularly spotted with texts of a vernacular character (such as plays, Warner p.c.) The PPCHE are somewhat better in this regard, since in the selection of texts for the corpus an effort was made to balance genres across time. The EEBO corpus as produced by the TCP has not been sampled with an eye towards diachronic consistency over genres. Nor have efforts to classify the 36883 texts in the corpus according to genre yet yielded results. Thus, the investigation is launched with the understanding that the EEBO corpus is probably not optimal for testing CRH effects.

## 5.3 Conclusion

In this section, preliminary results from the examination of a much larger corpus of *do*-support data. The results of this analysis confirm previous results on *do*-support (Kroch's finding of a CRH), as well as the novel results on argument structure conditioning presented in this dissertation. As has been previously discussed, incorporating these results into a single coherent model of the history of English *do*-support is not without its conceptual challenges. However, that the empirical facts underlying these investigations are

confirmed by yet another independent dataset of considerable size means that the facts, and the puzzles they present, should be taken seriously.

This chapter further uncovered a significant amount of by-lexeme variation in the frequency of *do*-support. This is a dimension of historical change which has received little attention, as compared for example with the amount of attention paid to lexical effects in phonological and phonetic change. One framework which is advanced for the understanding of lexical effects in historical change is that of Construction Grammar. A very preliminary investigation has yielded the result that, while there are effects in the data which can usefully be regarded as multi-word linguistic units which defy prevailing structural generalizations, these are limited in scope and cannot account for the full range of lexically-specific effects in the data.

It is hoped that the EEBO corpus underlying the discussion in this chapter will provide novel insights on the history of *do*-support. However, such work will have to be left for the future.

# Chapter 6

# Conclusion

The results presented in this dissertation advance our knowledge of the diachrony of English *do*-support to a considerable extent. I have provided:

1. Cross-linguistic parallels to English *do*-support which demonstrate that, though it is quite unique among the world's languages, it is possible to identify subcomponents of the phenomenon – both in PDE and earlier stages of the language – which are attested robustly in other languages.
2. A confirmation (broadly speaking) of the body of literature on the quantitative diachrony of *do*-support, calculated over a previously unexamined dataset from the PPCHE.
3. A novel analysis of the evolution of English *do*-support in terms of intermediate *do* which explains previously mysterious facts about the distribution of the construction in earlier stages of the language.
4. A contribution to debates about the synchronic status of *do*-support in PDE which is crucially informed by diachronic evidence.
5. An extension of the analysis to a new, and very large, dataset.

At the same time as they stand as contributions in their own right, these results open the way to further research questions. The description that I provide of the evolution of *do*-support characterizes it as a rich ground for exploration of questions of cascades vs. catastrophes, the connection between learning and change, and the importance of lexical variation to grammatical processes. These questions animate large portions of the literature on historical syntax, and indeed on variation more broadly. There is also a connection to literature on phonological change from an Exemplar Theoretic framework, which postulates a large role for frequency effects in the regulation of change. Fortunately, the gigaword corpus of EME provides ample opportunity for such further study.

# Appendix A

# Bibliography


Akaike, Hirotogu (1973). "Information theory and an extension of the maximum likelihood principle". In *2nd International Symposium on Information Theory, Tsahkadsor, Armenia, USSR, September 2-8, 1971*. Ed. by B. N. Petrov and F. Csáki. Budapest: Akadémiai Kiadó, pp.267–281.

Alam, M. Khorshed, M. Bhaskara Rao, and Fu-Chih Cheng (2010). Sample size determination in logistic regression. *Sankhyā: The Indian Journal of Statistics* **72-B**:1, 58–75.

Alekseev, Mikhail (1994a). "Budukh". In *The indigenous languages of the Caucasus: North East Caucasian languages*. Ed. by Rieks Smeets. Vol. 4. Delmar: Caravan Books, 1994, pp.261ff.

Alekseev, Mikhail (1994b). "Rutul". In *The indigenous languages of the Caucasus: North East Caucasian languages*. Ed. by Rieks Smeets. Vol. 4. Delmar: Caravan Books, 1994, pp.215ff.

Aoyagi, Hiroshi (2006). "On the Predicate Focus Construction in Korean and Japanese". In *Harvard Studies in Korean Linguistics*. Ed. by Susumu Kuno et al. 11. Hanshin Publishing, pp.359–373. URL.

Banks, Stephen (1994). Performing public announcements: The case of flight attendants' work discourse. *Text and Performance Quarterly* **14**, 253–267.

Behnel, Stefan and Martijn Faassen (2014). *LXML toolkit*. Python software package. version 3.3.5. URL.

Benincà, P. and C. Poletto (2004). A case of *do*-support in Romance. *Natural Language & Linguistic Theory* **22**:1, 51–94. DOI: 10.1023/B:NALA.0000005565.12630.c1.

Bock, Kathryn (1986). Syntactic persistence in language production. *Cognitive Psychology* **18**, 355–387.

Bock, Kathryn and Helga Loebell (1990). Framing sentences. *Cognition* **35**, 1–39.

Burnham, Kenneth and David Anderson (2004). Multimodel Inference: Understanding AIC and BIC in Model Selection. *Sociological Methods & Research* **33**:2 (Nov. 2004), 261–304. DOI: 10.1177/0049124104268644.

Cable, Seth (2004). "Restructuring in English". Ms. MIT. URL.

Cashen, Luke H. and Scott W. Geiger (2004). Statistical Power and the Testing of Null Hypotheses: A Review of Contemporary Management Research and Recommendations for Future Studies. *Organizational Research Methods* **7**:2 (Apr. 2004), 151–167. DOI: 10.1177/1094428104263676.

Cedergren, Henrietta and David Sankoff (1974). Variable Rules: Performance as a Statistical Reflection of Competence. *Language* **50**:2 (June 1974), 333–355.

Chomsky, Noam (1995). *The Minimalist Program.* Cambridge, MA: MIT Press.

Chomsky, Noam (2001). "Derivation by Phase". In *Ken Hale: A life in language.* Ed. by Michael Kenstowicz. Cambridge, MA: MIT Press. Chap. 1, pp.1–52.

Collins, Michael (2002). "Discriminative Training Methods for Hidden Markov Models: Theory and Experiments with Perceptron Algorithms". In *Empirical Methods in Natural Language Processing.* URL.

Cornips, Leonie (1998). "Habitual *doen* in Heerlen Dutch". In Do *in English, Dutch, and German: History and present-day variation.* Ed. by Ingrid Tieken-Boon van Ostade, Marijke van der Wal, and Arjan van Leuvensteijn. Stichting Neerlandistiek, pp.83–101.

Culicover, Peter and Susanne Winkler (2008). English focus inversion. *Journal of Linguistics* **44**, 625–658. DOI: 10.1017/S0022226708005343.

Denison, David (1985). "The origins of periphrastic *do*: Ellegård and Visser reconsidered". In *Papers from the 4th International Conference on English Historical Linguistics.* Ed. by Roger Eaton, Olga Fischer, William Koopman, and Frederike van der Leek. Vol. 41. Current Issues in Linguistic Theory. John Benjamins, pp.45–60.

Ellegård, Alvar (1953). *The auxiliary do: The establishment and regulation of its use in English.* Engelska språket.

Embick, David and Rolf Noyer (2001). Movement operations after syntax. *Linguistic Inquiry* **32**, 555–595.

Estival, Dominique (1985). Syntactic priming of the passive in English. *Text* **5**:1/2, 7–21.

Gill, Jeff (1999). The Insignificance of Null Hypothesis Significance Testing. *Political Research Quarterly* **52**:3 (Sept. 1999), 647–674. DOI: 10.1177/106591299905200309.

Hagstrom, Paul (1995). *Negation, focus, and* do*-support in Korean.* ms. URL.

Han, Chung-hye and Anthony Kroch (2000). "The rise of *do*-support in English: implications for clause structure". In *Proceedings of the NELS.* 30, pp.311–326.

Hoenig, John and Dennis Heisey (2001). The abuse of power: The pervasive fallacy of power calculations for data analysis. *The American Statistician* **55**:1 (Feb. 2001), 1–6.

Honnibal, Matthew (2013). *A good POS tagger in about 200 lines of Python.* Blog post. Accessed August 8, 2014. Sept. 2013. URL.

Hopper, Paul and Elizabeth Closs Traugott (1993). *Grammaticalization.* Cambridge University Press.

Hothorn, Torsten, Frank Bretz, and Peter Westfall (2008). Simultaneous Inference in General Parametric Models. *Biometrical Journal* **50**:3, 346–363.

Houser, Michael, Line Mikkelsen, Ange Strom-Weber, and Maziar Toosarvandani (2006). "Gøre-Support in Danish". In *Proceedings of the Twenty-First Comparative Germanic Syntax Workshop.*

Hsieh, F. Y. (1989). Sample size tables for logistic regression. *Statistics in Medicine* **8**, 795–802.

Hsieh, F. Y., Daniel Bloch, and Michael Larsen (1998). A simple method of sample size calculation for linear and logistic regression. *Statistics in Medicine* **17**, 1623–1634.

Hurvich, Clifford M. and Chih-ling Tsai (1989). Regression and time series model selection in small samples. *Biometrika* **76**:2, 297–307. DOI: 10.1093/biomet/76.2.297.

Jäger, Andreas (2006). *Typology of periphrastic* do *constructions.* Vol. 12. Diversitas Linguarum. Bochum: Universitätsverlag Dr. N. Brockmeyer.

Johnson, Daniel Ezra (2010). *Stability and change along a dialect boundary: The low vowels of Southeastern New England.* Publications of the American Dialect Society 95.

Kimura, Motoo (1983). *The neutral theory of molecular evolution.* Cambridge University Press.

Klemola, Juhani (1998). "Semantics of *do* in southwestern dialects of English English". In Do *in English, Dutch, and German: History and present-day variation.* Ed. by Ingrid Tieken-Boon van Ostade, Marijke van der Wal, and Arjan van Leuvensteijn. Stichting Neerlandistiek, pp.25–52.

Kratzer, Angelika (1996). "Severing the External Argument from its Verb". In *Phrase Structure and the Lexicon.* Springer, pp.109–137. DOI: 10.1007/978-94-015-8617-7_5.

Kroch, Anthony (1989). Reflexes of grammar in patterns of language change. *Language Variation and Change* **1**:3, 199–244.

Kroch, Anthony, Beatrice Santorini, and Lauren Delfs (2005). *Penn-Helsinki Parsed Corpus of Early Modern English.* University of Pennsylvania. URL.

Kroch, Anthony and Ann Taylor (2000). *Penn-Helsinki Parsed Corpus of Middle English, second edition.* University of Pennsylvania. URL.

Labov, William (1989). The child as linguistic historian. *Language Variation and Change* **1**, 85–97.

Laka, Itziar (1990). "Negation in syntax: On the nature of functional categories and projections". PhD thesis. MIT.

Levin, Beth (1993). *English verb classes and alternations: A preliminary investigation.* University of Chicago Press.

Matisoff, James (1973). *A grammar of Lahu.* University of California Press.

Ozanne-Rivierre, Françiose (2004). "The evolution of the verb 'take' in New Caledonian languages". In *Complex predicates in Oceanic languages: Studies in the dynamics of binding and boundedness.* Ed. by Isabelle Bril and Françiose Ozanne-Rivierre. Mouton de Gruyter, pp.331–346.

Platzack, Christer (2008). "Cross Linguistic Variation in the Realm of Support Verbs". In *Proceedings of Comparative Germanic Syntax Workshop 23.* URL.

Pollock, Jean-Yves (1989). Verb movement, Universal Grammar, and the structure of IP. *Linguistic inquiry* **20**:3, 365–424. URL.

*POS Tagging (State of the art)* (2014). Association for Computational Linguistics wiki. Accessed August 8, 2014.

Potsdam, Eric (1995). "Phrase Structure of the English Imperative". In *Proceedings of the Sixth Annual Meeting of the Formal Linguisitcs Society of Mid-America.* Ed. by Leslie Gabriele, Debra Hardison, and Robert Westmoreland. Indiana University Linguistics Club, pp.143–154. URL.

Pullum, Geoffrey and Deirdre Wilson (1977). Autonomous syntax and the analysis of auxiliaries. *Language* **53**:4, 741–788. DOI: 10.2307/412911.

Rizzi, Luigi (1997). "The fine structure of the left periphery". In *Elements of Grammar.* Ed. by Liliane Haegeman. Dordrecht: Kluwer, pp.281–337.

Roberts, Ian (1985). Agreement parameters and the development of English modal auxiliaries. *Natural Language and Linguistic Theory* **3**:1, 21–58. DOI: 10.1007/BF00205413.

Roberts, Ian (1993). *Verbs and Diachronic Syntax.* Vol. 28. Studies in Natural Language and Linguistic Theory. Dordrecht: Kluwer Academic Publishers.

Ruppert, David and Matthew Wand (1994). Multivariate Locally Weighted Least Squares Regression. *The Annals of Statistics* **22**:3, 1346–1370. URL.

Sankoff, David and Suzanne Laberge (1978). "Statistical dependence among successive occurrences of a variable in discourse". In *Linguistic variation: models and methods.* Ed. by David Sankoff. Academic Press. Chap. 8, pp.119–126.

Schoenfeld, David and Michael Borenstein (2005). Calculating the power or sample size for the logistic and proportional hazards models. *Journal of Statistical Computation and Simulation* **75**:10 (Oct. 2005), 771–785.

Schütze, Carson (2004). Synchronic and diachronic microvariation in English do. *Lingua* **114**:4 (Apr. 2004), 495–516. DOI: 10.1016/s0024-3841(03)00070-6.

Schütze, Carson (2013). "Superfluous 'do' and comparison of Spell-outs". In *Dummy auxiliaries in first and second laanguage acquisition*. Ed. by Elma Blom, Josje Verhagen, and Ineke van de Kraats. Mouton de Gruyter.

Schwarz, Gideon (1978). Estimating the Dimension of a Model. *Annals of Statistics* **6**:2 (Mar. 1978), 461–464. DOI: 10.1214/aos/1176344136.

Stone, M. (1976). An Asymptotic Equivalence of Choice of Model by Cross-validation and Akaike's Criterion. *Journal of the Royal Statistical Society, Series B (Methodological)* **39**:1, 44–47.

Tagliamonte, Sali (2013). "Comparative Sociolinguistics". In. Wiley, July 2013, pp.128–156. DOI: 10.1002/9781118335598.ch6.

Tamminga, Meredith and Aaron Ecay (2014). *Priming effects in language change as diagnostics of grammatical structure*. Presentation at Penn Linguistics Conference 38. Mar. 2014.

Taylor, Ann, A. Nurmi, Anthony Warner, Susan Pintzuk, and T. Nevalainen (2006). *Parsed Corpus of Early English Correspondence*. Compiled by the CEEC Project Team. Distributed through the Oxford Text Archive. URL.

Væth, Michael and Eva Skovlund (2004). A simple approach to power and sample size calculations in logistic regression and Cox regression models. *Statistics in Medicine* **23**, 1781–1792. DOI: 10.1002/sim.1753.

Van Craenenbroeck, Jeroen (to appear). "VP-ellipsis". In *Blackwell Companion to Syntax*. second. URL.

Viðarsson, Heimir Freyr (2009). *"Sól gerði eigi skína:" stoðsagnir með nafnhætti í fornnorrænu*. MA thesis, University of Iceland. URL.

Visser, Fredericus Theodorus (1963). *An historical syntax of the English language*. E. J. Brill.

Wallenberg, Joel (2013). *A unified theory of stable variation, syntactic optionality, and syntactic change*. Presentation at Diachronic Generative Syntax 15. Aug. 2013.

Warner, Anthony (1993). *English Auxiliaries: Structure and History*. Cambridge University Press.

Warner, Anthony (2005). Why *do* dove: Evidence for register variation in Early Modern English negatives. *Language Variation and Change* **17**:3, 257–280.

Weiner, E. Judith and William Labov (1983). Constraints on the agentless passive. *Journal of Linguistics* **19**:1, 29–58.

Wickham, Hadley (2009). *ggplot2: elegant graphics for data analysis*. Springer New York. ISBN: 978-0-387-98140-6. URL.

Wohlgemuth, Jan (2009). *A typology of verbal borrowings*. Mouton de Gruyter.

Yang, Charles (2000). Internal and external forces in language change. *Language Variation and Change* **12**, 231–250.

Yi, Eun-Young (1994). NegP in Korean. *Cornell Working Papers in Linguistics* **12**.