Construction and lexical class effects in the history of *do*-support

Aaron Ecay

University of York

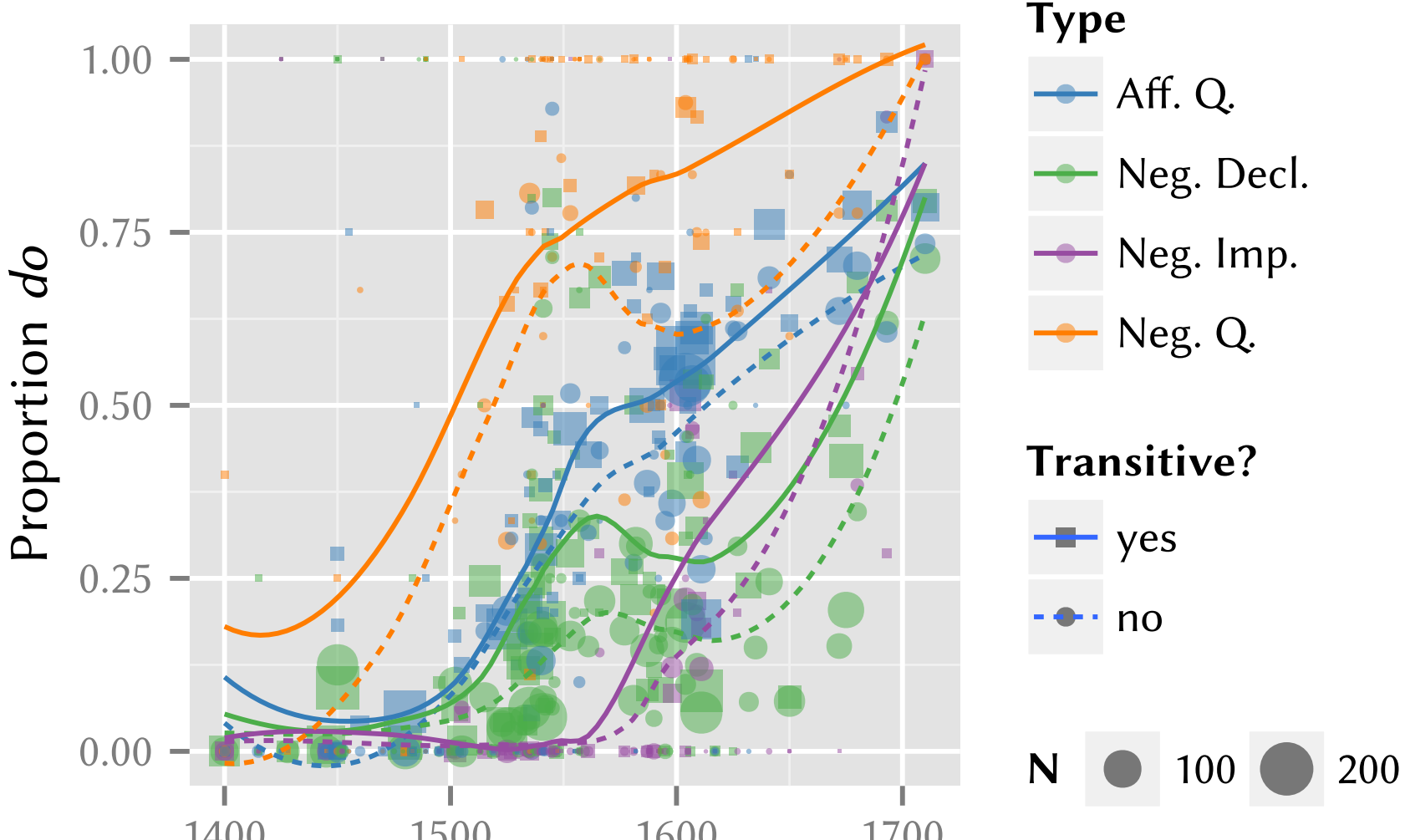
May 30, 2015

Introduction

- Goals of this talk
  - Demonstrate the utility of a large database of *do*-support (and EME text generally)
  - Look for lexical class effects and idiosyncratic constructions in this data
  - Use this data for mathematical modeling of the emergence of *do*-support

Argument structure and lexical class

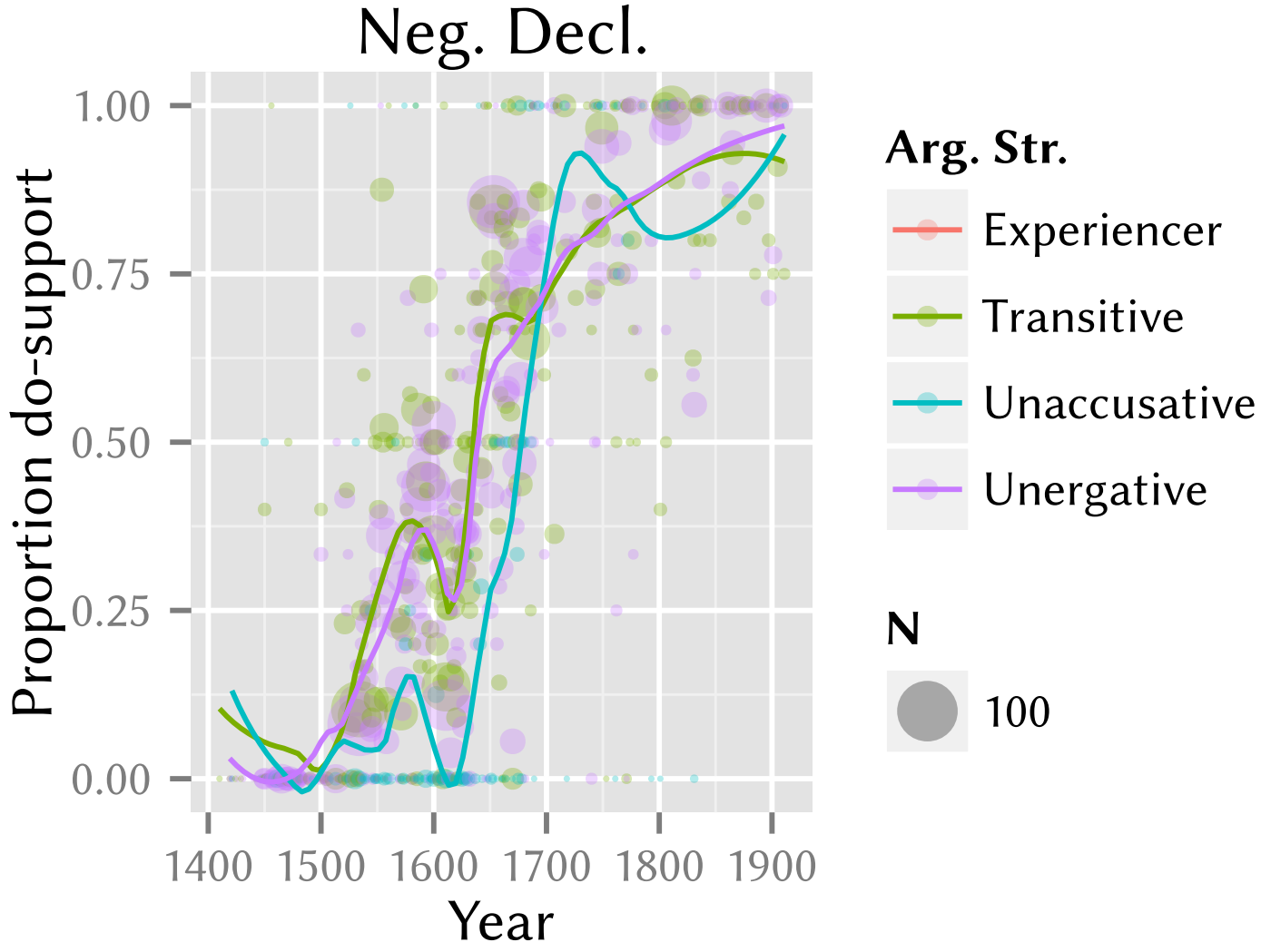
- Argument structure and lexical class effects on *do*-support were noticed by Ellegård 1953



Lexical class effects in grammatical analysis

- I have argued that these effects should be understood as the relics of *do*'s past as a causative (in ME)
- In Earliest Modern English (EstME), *do* is a marker of agentivity, bleached from a causative
  - It occurs in affirmative declaratives at a frequency of up to ~10%
  - It almost never occurs with unaccusatives, even in negatives or questions
- This suddenly changes in ~1600 when V-to-T raising is lost; *do* is reanalyzed into its modern syntax
  - *do* with affirmative declaratives ebbs away
  - On the other hand, it begins to appear with unaccusatives at rates which quickly equalize with those in transitives

Lexical class effects, a graphical representation



Establishing lexical class effects

- Original evidence for these claims based on PPCEME + PCEEC (Kroch, Santorini, and Delfs 2005; Taylor et al. 2006)
- Only 6 unaccusative verbs counted
  - Dominated by *come*, and to a lesser extent *go*
- Experiencer-subject verbs such as *know* could not be reliably measured

Establishing lexical class effects

- Original evidence for these claims based on PPCEME + PCEEC (Kroch, Santorini, and Delfs 2005; Taylor et al. 2006)
- Only 6 unaccusative verbs counted
  - Dominated by *come*, and to a lesser extent *go*
- Experiencer-subject verbs such as *know* could not be reliably measured
- Do these effects generalize?

A new corpus

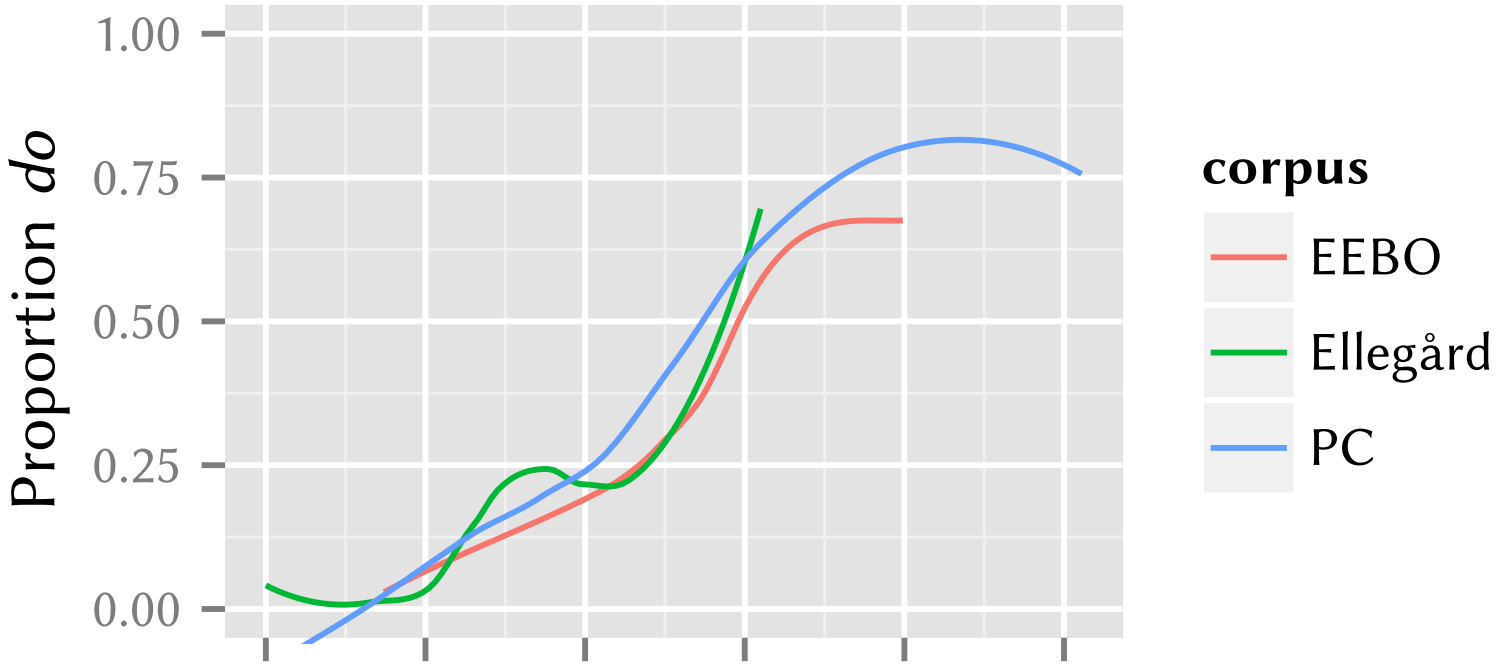
- In order to investigate this question, I created the PYCCLE-TCP corpus
  - Penn-York Computer-annotated Corpus of a Large amount of English
  - Based on the Text Creation Partnership (TCP)
- One billion (10^9) words of English text, printed 1473–1800
  - 900 million before 1700
- Automated POS tagging with PPCEME as training data
- Available today: `http://aaronecay.com/pyccle/`

Gathering data from PYCCLE-TCP

- No syntactic information in the corpus!
- Difficult to find all tokens of (potential) *do*-support
- Solution: focus on declarative clauses with pronoun subjects
  - No long-distance dependencies in the verbal domain, unlike S-Aux inversion sentences
  - Pronouns have morphology which (almost) unambiguously identifies subjects
  - Search for sequences like
    - Subj Do Not Vb
    - Subj Vb Not

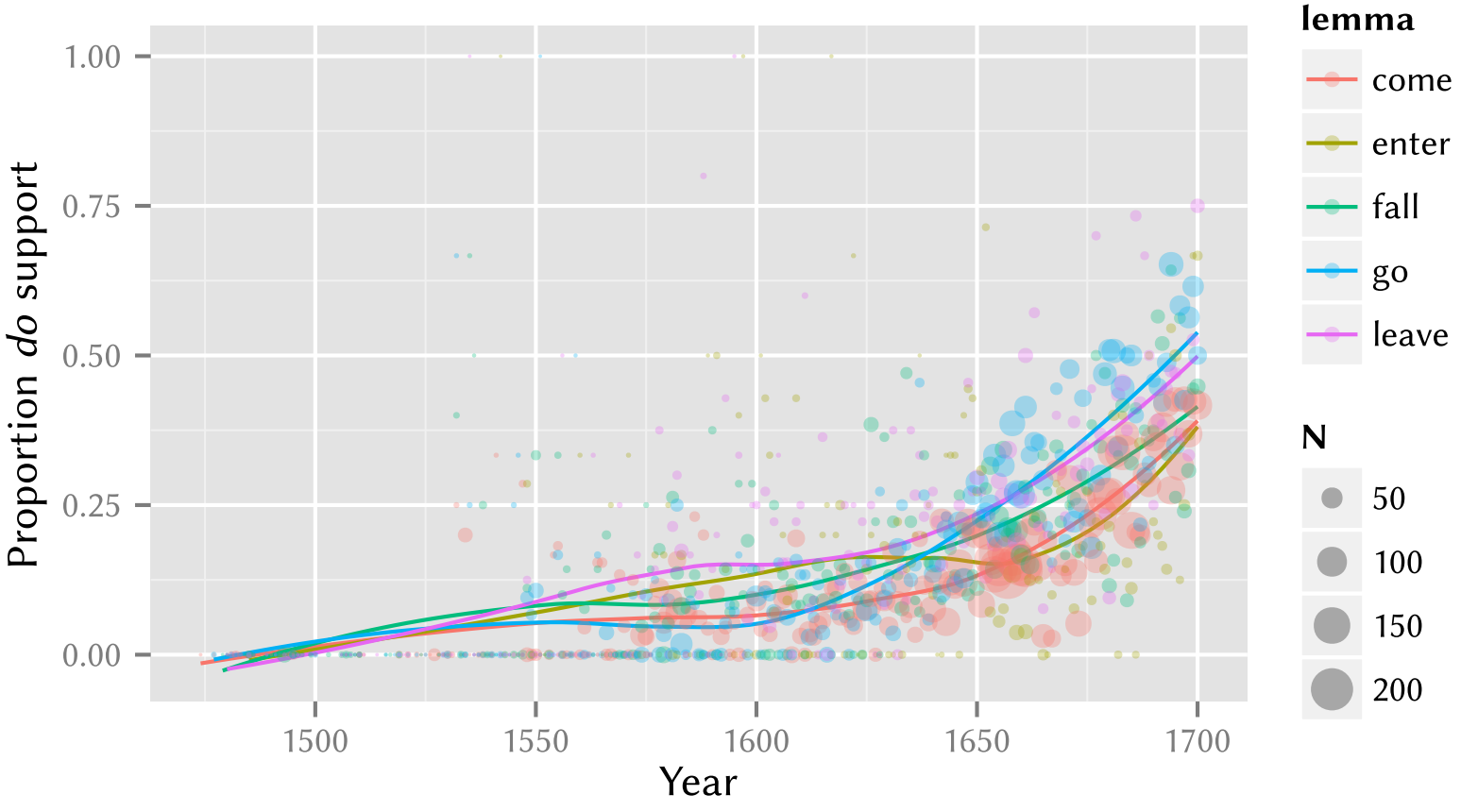
Data from PYCCLE-TCP

- Result: 60ok negative declaratives and 6M affirmative declaratives
  - ~100x larger than dataset derived from PPCHE
  - Alternatively: the most frequent 10 lexical verbs in the EEBO corpus each have more tokens than the entire PPCHE corpus combined
  - 81 lexical verbs have >1000 tokens
- Highly similar results



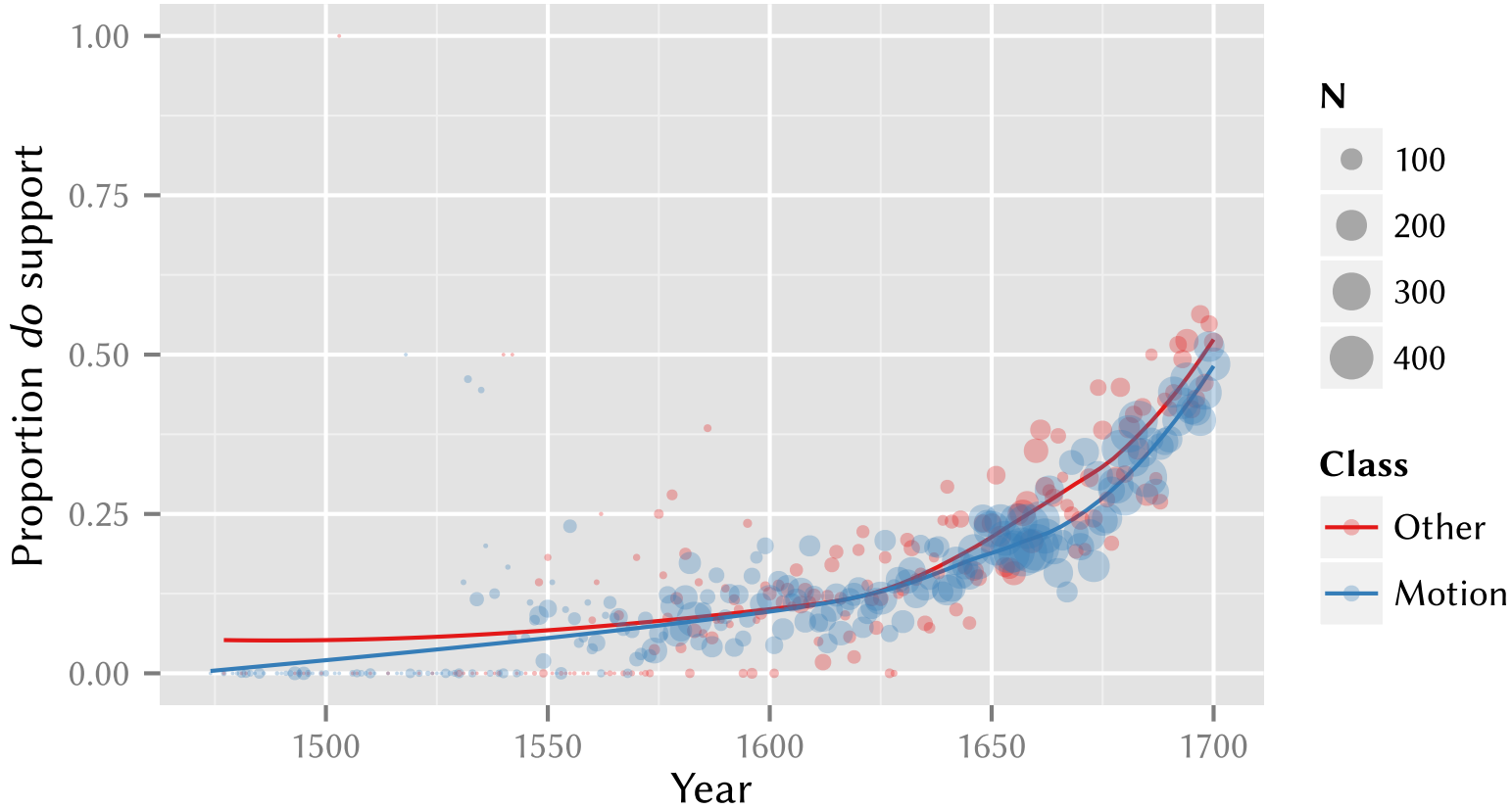
Unaccusative verbs of motion

- Good news: the earlier results driven by *come* and *go* generalize to a larger class
  - In the sense of Levin (1993): class 2, "verbs of inherently directed motion"



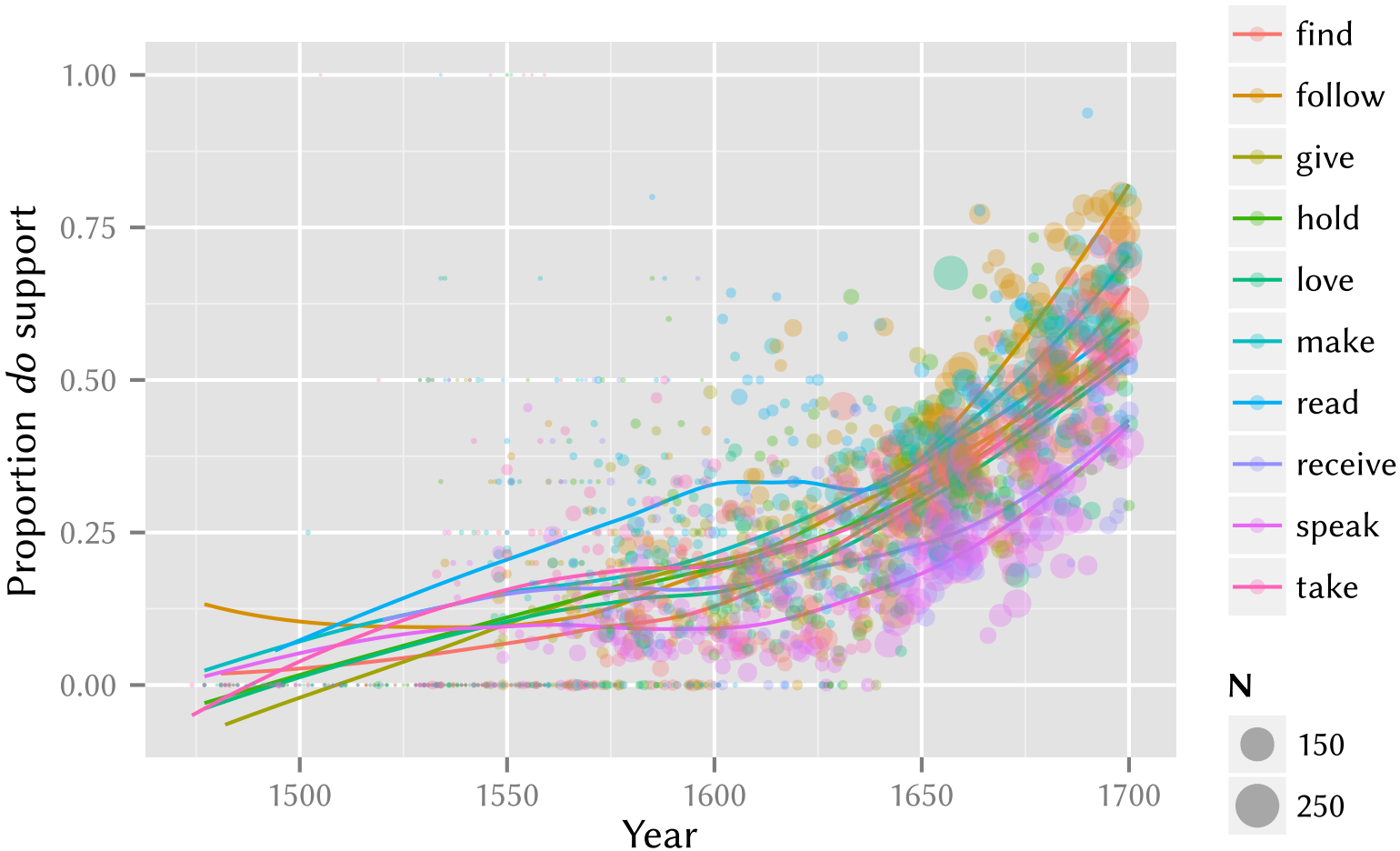
Unaccusatives generally

- ... and to unaccusatives in general
  - "Other" class: *live, die, stay, perish, prevail, depend, remain*

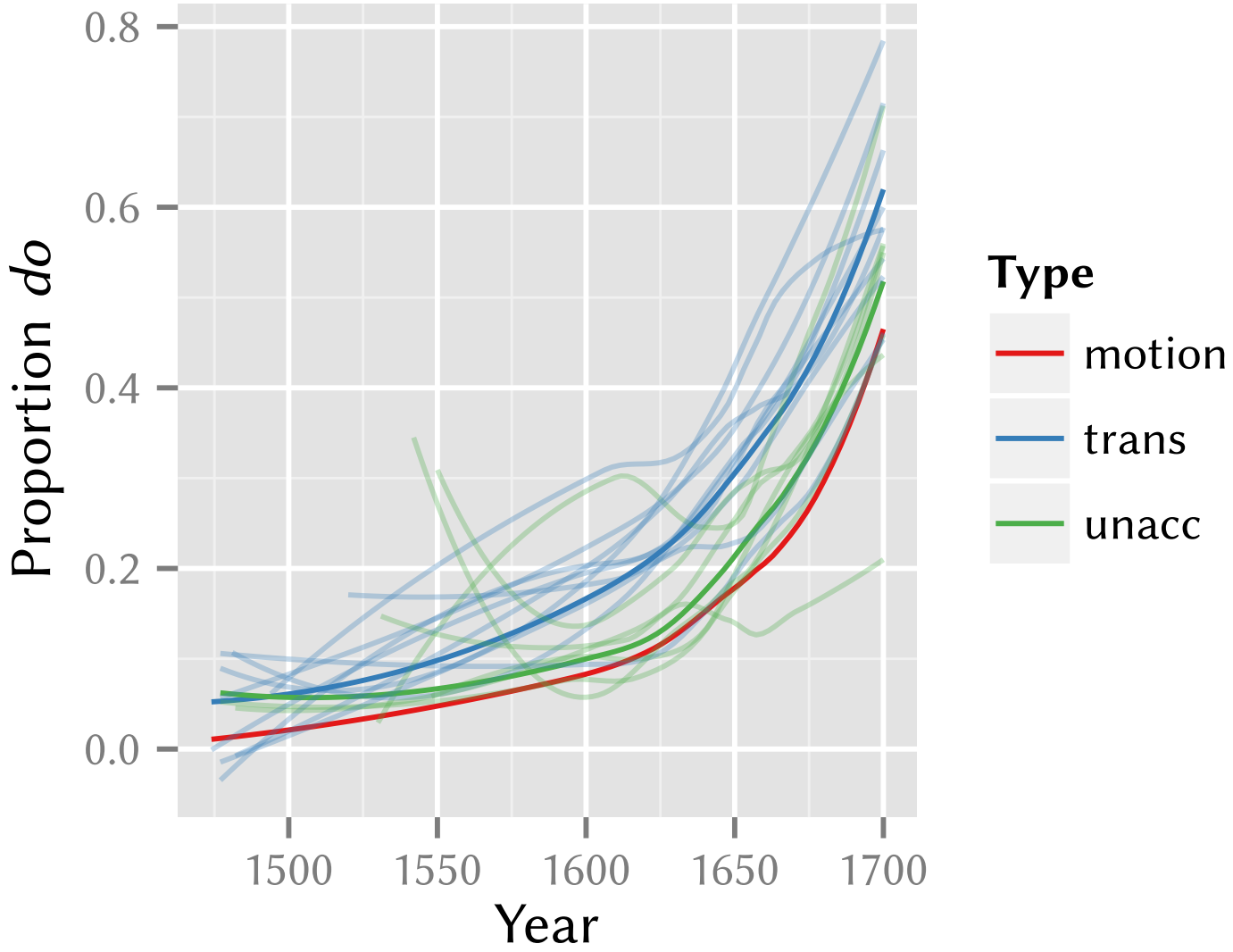


Transitives

- Transitives are different

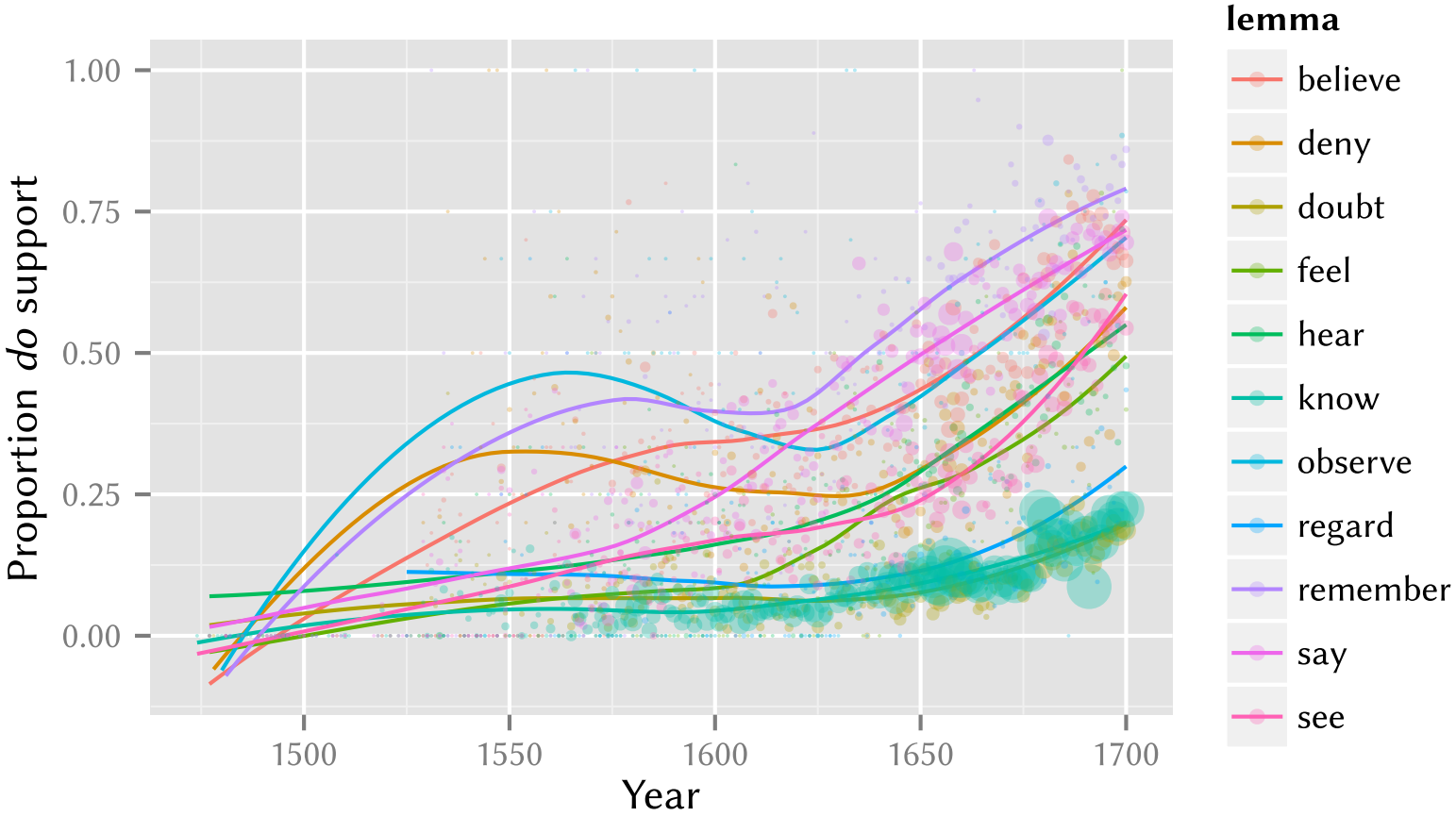


Summary

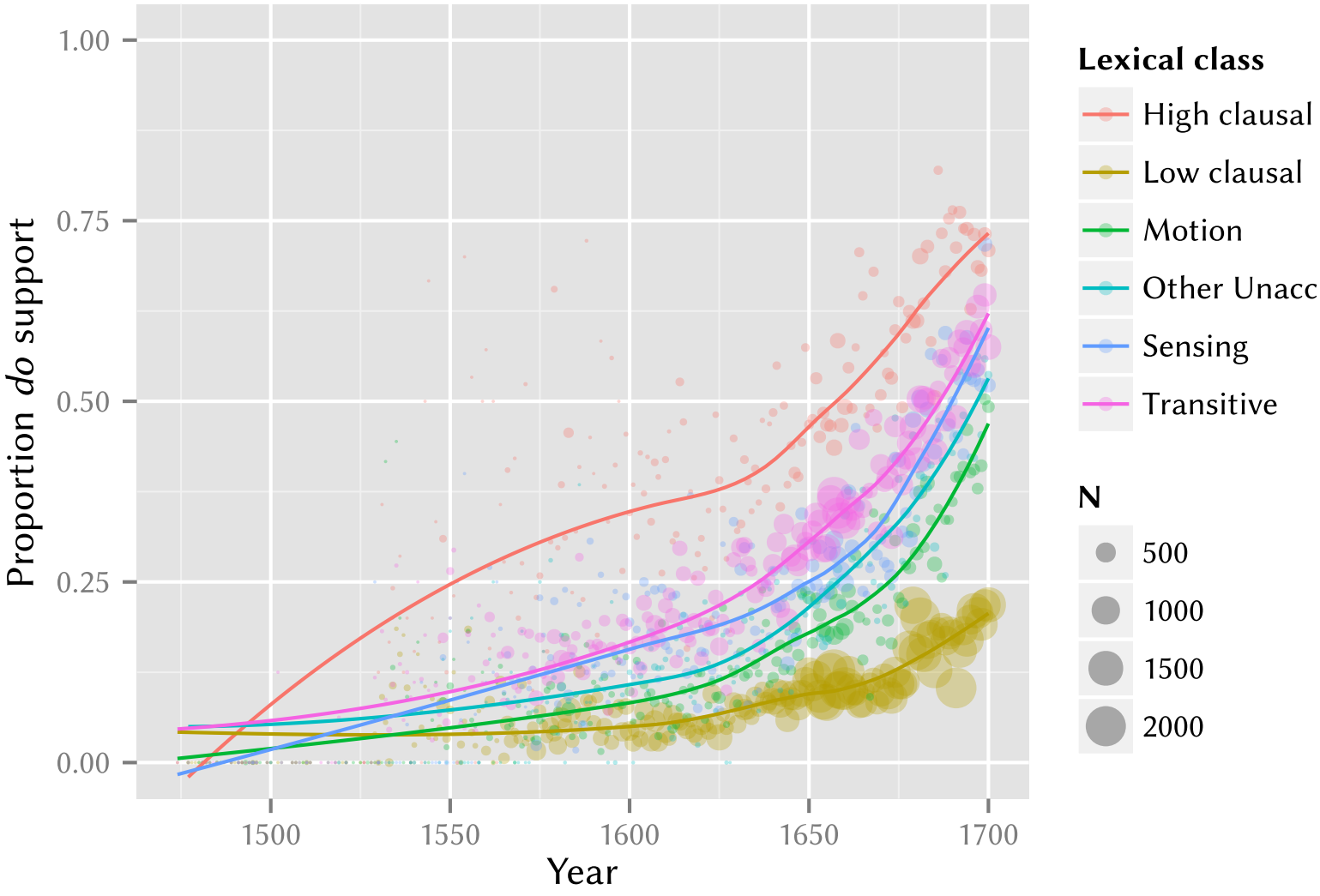

0.8
0.6
0.4
0.2
0.0
Proportion *do*
1500
1550
1600
1650
1700
Year
Type
motion
trans
unacc

Clausal-object verbs

- Beyond the unaccusative-transitive dimension, there are other classes of verb which might be interesting to investigate
- One example of such a class is experiencer-subject verbs
- As a guide, I have looked at verbs that can take a clausal object



Summary


1.00
0.75
0.50
0.25
0.00
Proportion *do* support
1500
1550
1600
1650
1700
Year
Lexical class
High clausal
Low clausal
Motion
Other Unacc
Sensing
Transitive
N
500
1000
1500
2000

Conclusions

- The agentivity effects reported in previous work are quite general, and not driven by lexically-specific effects (e.g. of *come* and *go*)
- However, these effects are dwarfed by other lexically-specific effects, which are not explained on systematic grounds
- Might “constructions” have a role?

Constructions and *do*

- It is often claimed that constructions play a role in the emergence of *do*-support
  - What are constructions?
  - Theory-neutral position: N-grams with special behavior
  - Put another way: violations of contextual independence and the CRH
- This may be the case for some authors early in the (pre)history of the change (Culicover 2008, p. 36)

  > "*[D]o*-support began as a lexically restricted construction. Its development as a rule of English is a consequence of its spread through the population, and through the lexicon, until it generalized free of particular items and contexts."

  - Machyn: 216/370 tokens of *do* with *preach*
  - *Polychronicon* (anon. 1450 transl.): 243/816 *do* with *slay*; no finite past tense *slay*
- However, such cases seem to be the exception, not the rule

Searching for constructions with *do*

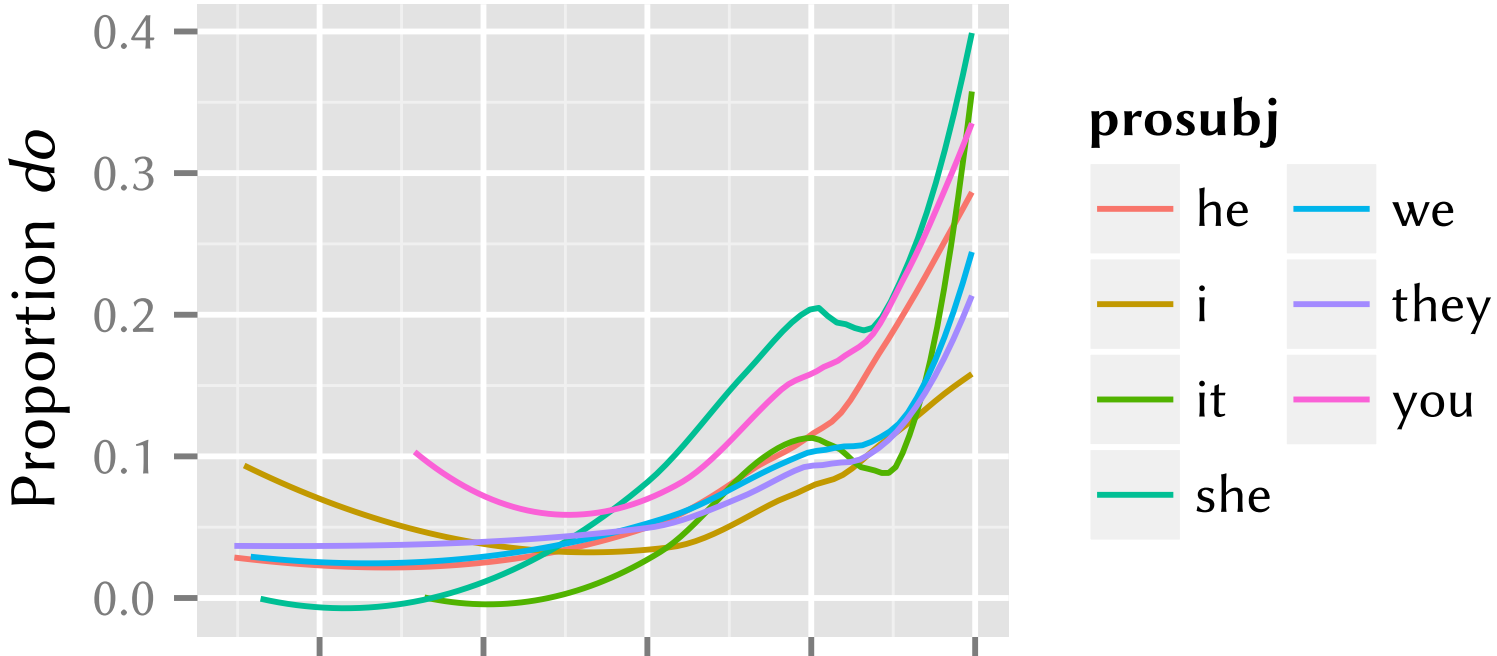
- 89 texts with at least one token of *do*-support before 1535
- 187 tokens of *do*-support
- 120 are hapaxes
- Most common verb (*say*) has 21 tokens of *do*-support

Searching for constructions with *do*

- 89 texts with at least one token of *do*-support before 1535
- 187 tokens of *do*-support
- 120 are hapaxes
- Most common verb (*say*) has 21 tokens of *do*-support
- (Is it possible that "constructions" are a property of the earliest attestations, but not the spread?)

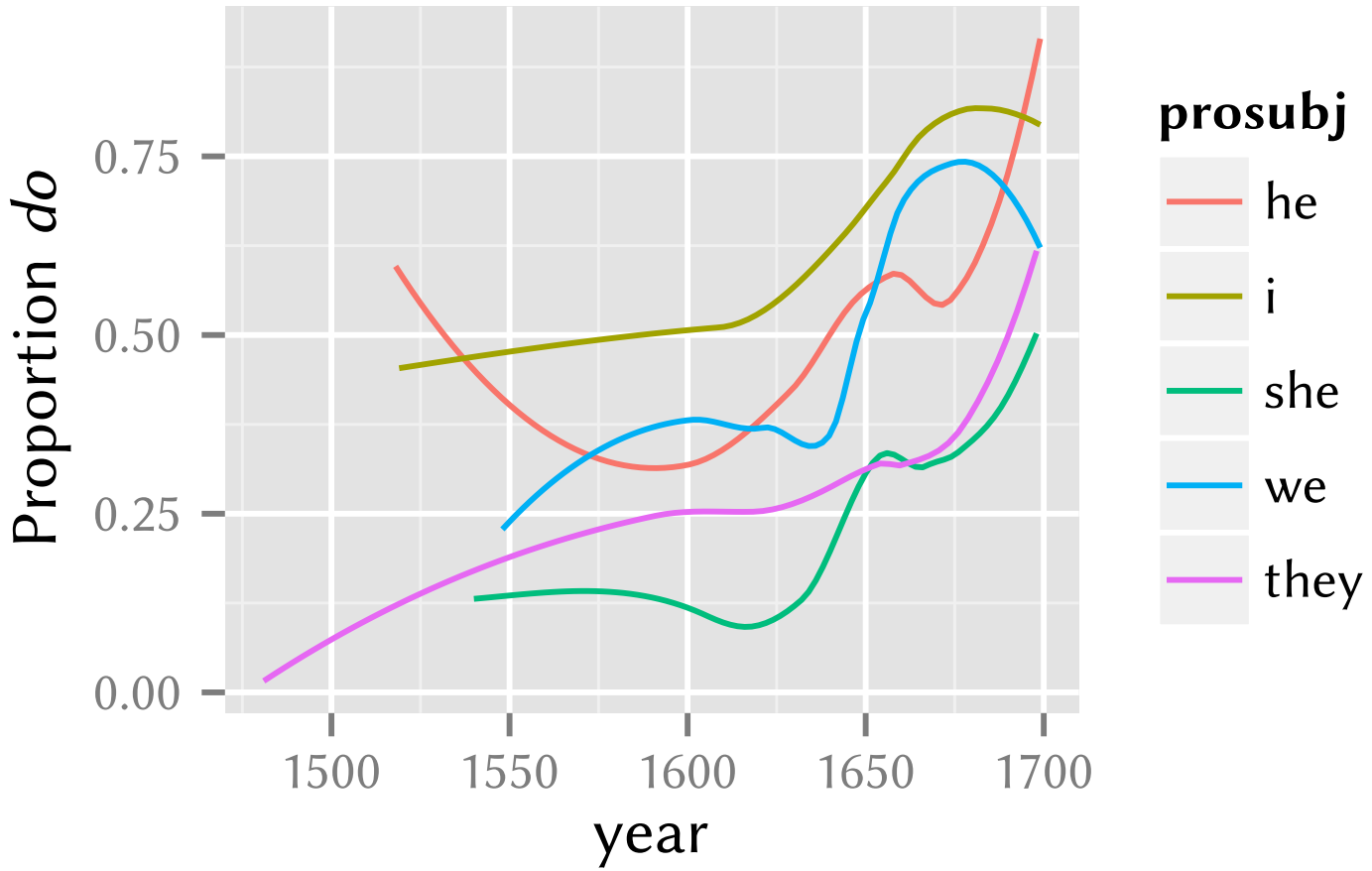
Constructions and aberrant lexical items

- It is sometimes suggested that constructions play a role in regulating the spread of verbs which differ greatly from the overall rate of *do*-support
- Insofar as this is a useful explanation, we should look at pronoun subjects
  - "I know not"
- No difference between "I" and other subjects
  - Perhaps some differences, in the other direction?



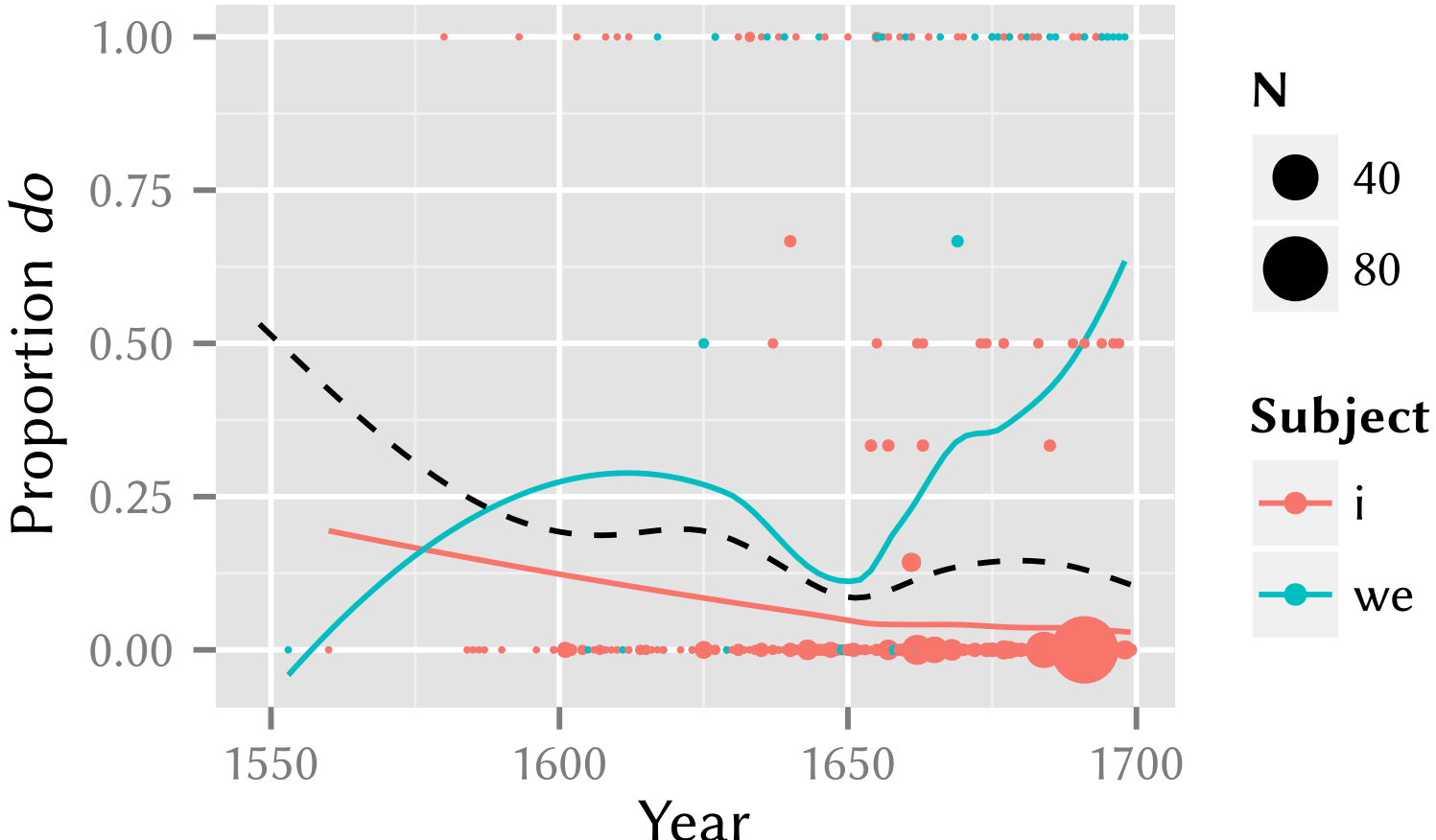
Constructions and aberrant lexical items

- Constructions may play a role for "remember," one of the most consistently advanced verbs



One clear constructional effect

- For the verb "mistake," there is a genuine constructional effect
- "We" does what it should
- "I" gets stuck (and dominates we quantitatively: 1085 tokens vs. 117)
  - Most tokens are "if I mistake not"



Summary

- Constructions do exist (at a descriptive level)
- However, they play a marginal role in the regulation of the spread of *do*-support in EME
- This data probably cannot answer theoretical questions about constructions
  - Stay tuned...

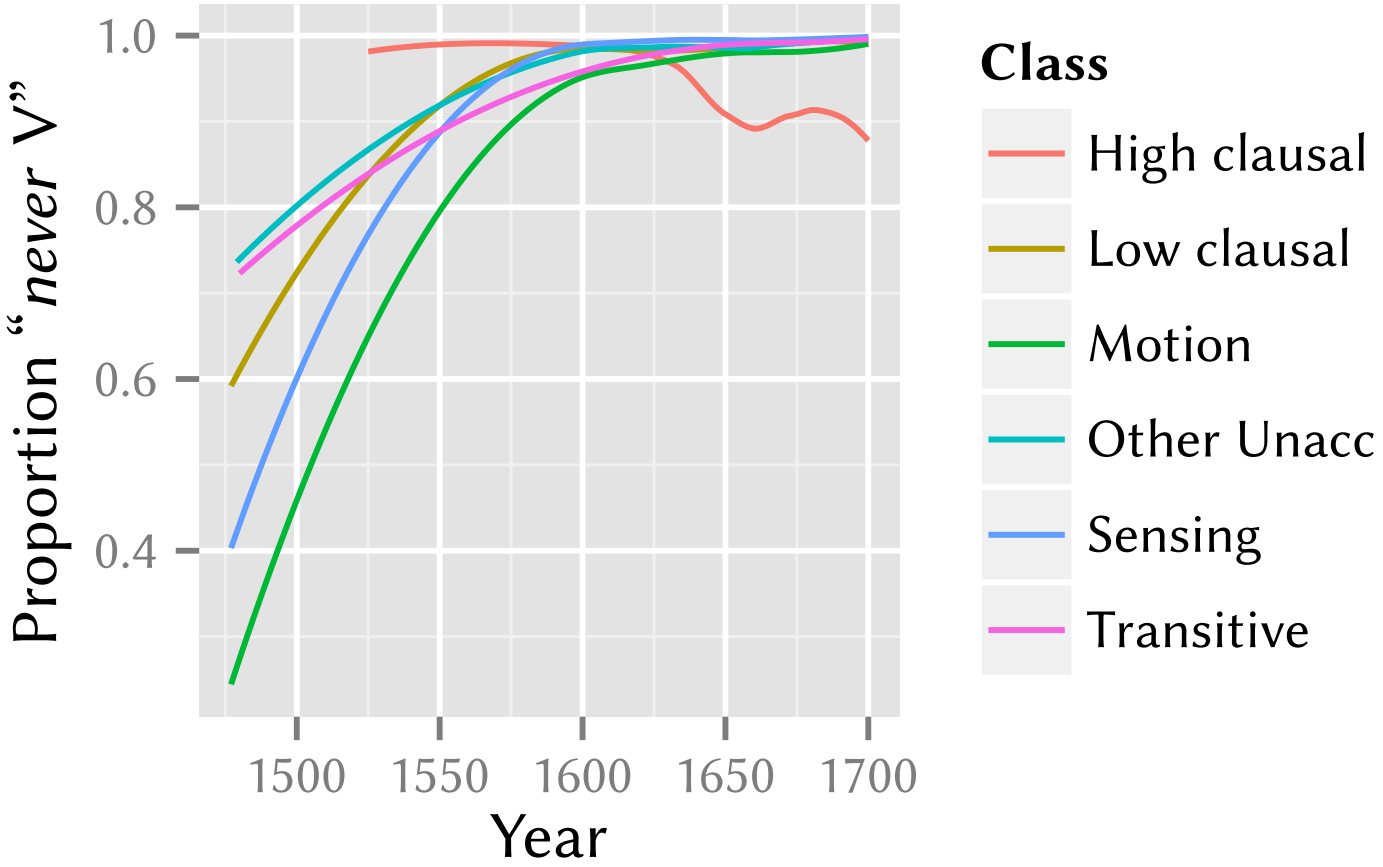
Do-support and the CRH

- Kroch (1989) used *do*-support as one of the case studies to support the Constant Rate Hypothesis
- His finding: *do*-support enters the language at the same rate as verb raising over *never* leaves
  - Conclusion: these are reflexes of the same underlying grammatical change: the loss of V-to-T raising

(1) he speakes neuer a true word

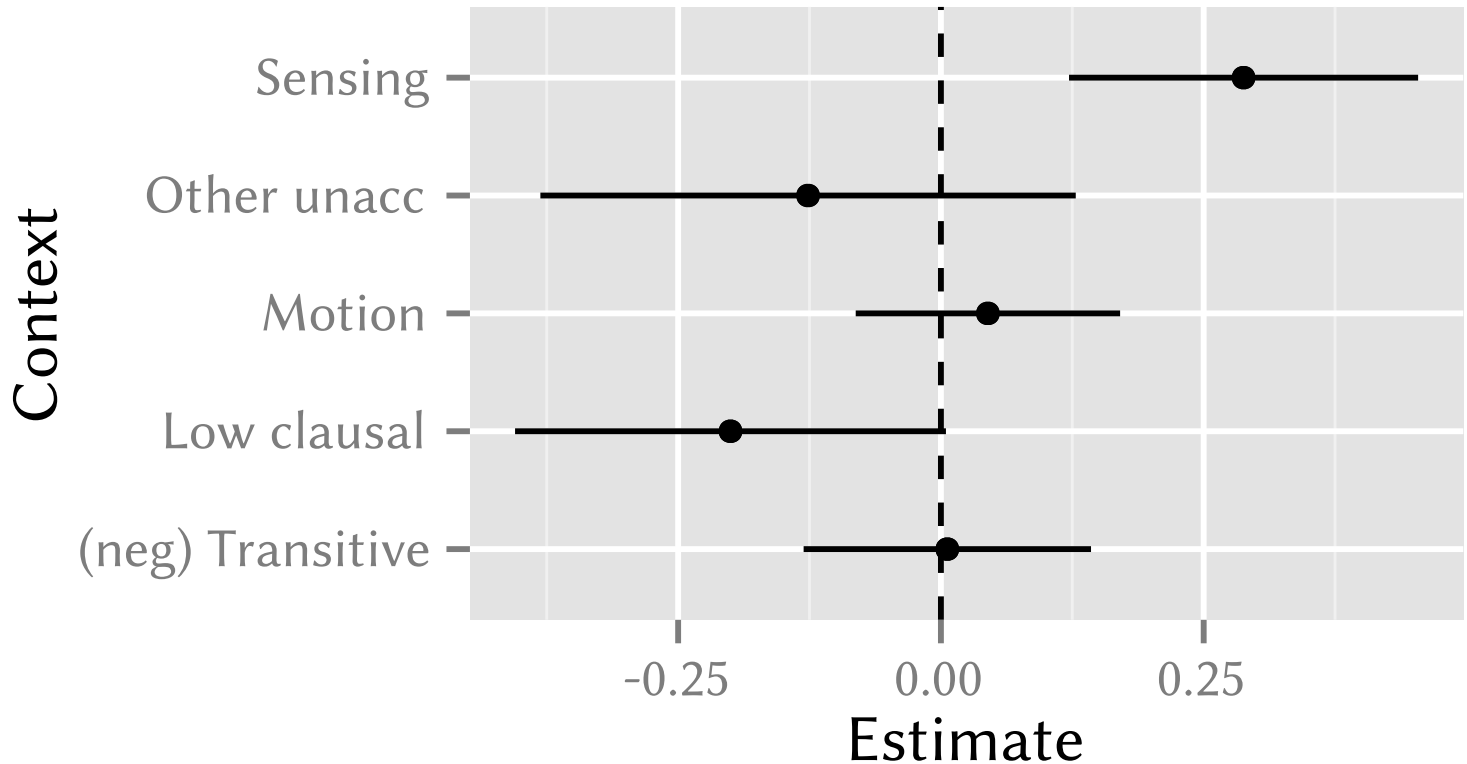
Constraints on *do*-support and raising past *never*

- An argument structural effect is visible in the rate at which verbs raise over *never*
  - Similar (but not identical to) the effect on *do*-support
  - But not predicted!



A CRH effect for *never*

- Fitting a CRH model to this data gives an uncertain result
- The contexts along the unaccusative/transitive dimension have the same slope
- The clausal object verbs differ in their slope
  - High clausal verbs excluded



CRH effects across *do* and *never*

- Fitting CRH models between *do* and *never* across different argument structure contexts gives mixed results
  - (Probably) have a CRH: Transitive, non-motion unacc, high clausal
    - AIC vs. LRT/t-test
  - No CRH: motion verbs, low clausal, sensing
- Collapsing all verbs together: no CRH (clearly)

Questions raised

- Under the account I have given in previous work, the loss of verb raising over *never* triggers a reanalysis to *do*-support, rather than occurring simultaneously to it
  - Ergo, we do not expect a CRH effect
- Yet, in some contexts we seem to find one
  - Accident or pattern?
- The argument structure conditioning of *do*-support possibly is reflected in verb raising over *never*
  - Coincidence?
  - Which is the dog and which is the tail?

Whither the CRH?

- The treatment of the CRH in large datasets is unclear
  - With enough data, every model is false
- Independence assumptions are important to the CRH's success
  - If conditioning factors are independent, violations of the CRH in one dimension will "average out" in others
  - We don't understand (non)independence conditions of syntactic processes well enough
    - Style, priming, other extragrammatical effects
    - Interacting rules (pied-piping/stranding ~ *who(m)*)
    - Intrinsic nonindependence? (Goes to the heart of debates about construction grammar)
- We can separate the insight of the CRH from its mechanics
  - There are structures which underlie changes and span various lexicalizations of these structures
  - Effects reported in the CRH literature still serve to demonstrate this point, even if they prove not to hold up in large datasets
  - We can find evidence of violations of the spirit of the CRH (construction-specific effects), but these are (in this case) quite marginal

Conclusions

- It is now possible to study historical changes in minute lexical detail
  - Charles said yesterday "we can study acquisition to an infinitely greater level of detail than change" (paraphrasing slightly)
  - This is true – but I'm doing my best to keep up!
- Lexical effects on the spread of *do*-support are robustly attested
  - The agentivity dimension is important, and easy to understand
  - Other factors are also at play in clausal object verbs, which cannot yet be explained
- This lexical "big data" continues to support the core intuition of the CRH, even as it fails to fit neatly into the established methodological paradigm
  - New ways to operationalize this insight are needed

Thanks

- This work would not be possible without the help of many people:
  - The creators of the PPCEME, PCEEC, TCP corpora (EEBO/ECCO), and digitized version of Ellegård's corpus
  - Audiences at DiGS, PLC, and CUNY
  - My colleagues at Penn and York
  - Special thanks to Tony Kroch
- (Any imperfections are all mine)

Time for questions

Bibliography I

- Culicover, Peter (2008). The rise and fall of constructions and the history of English *do*-support. *Journal of Germanic Linguistics* **20**:1, 1–52.
- Ellegård, Alvar (1953). *The auxiliary do: The establishment and regulation of its use in English*. Engelska språket.
- 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.
- Levin, Beth (1993). *English verb classes and alternations: A preliminary investigation*. University of Chicago Press.

Bibliography II

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.