The origin of *do*-support in English: a learning-based account

1. Introduction. There has been much work on the competition between *do*-support and V-to-T movement (Ellegård 1953, Kroch 1989), which took place in the centuries spanning Early Modern English. The present work focuses on the period immediately preceding this competition.

My goal is to provide a learning-based account of a series of changes in *do* in this early period, which set up its subsequent entrance into competition with V-to-T movement. To my knowledge, while some have made similar observations about this early period of *do* (e.g. Ecay 2015), there have been no learning-based, causal explanation of these changes. Moreover, I deviate from existing accounts in that I do not take competition with V-to-T movement as a contributor to the early rise of *do*-support, which aligns with recent work suggesting a later and more gradual loss of V-to-T movement than previously assumed (Haeberli & Ihsane 2016).

2. Data and methods. The data I use are from PPCHE2 (2025 release), which includes PPCME2 (Kroch & Taylor 2000), PPCEME (Kroch et al. 2004), and PPCMBE2 (Kroch et al. 2016). The do-sentence condition of Figs. 2–4 comprises the union of three subconditions, which identify three structures: affirmative declarative (do-V), negative declarative (do-not-V), and subject-verb inversion (do-inversion). Their token frequencies in the whole dataset over time are shown in Fig. 1. Most *do*-sentences in our period of interest (c. 1400–1600) occur in the do-V condition, exemplified by (1). The competition between *do*-support and V-to-T movement, as indicated by do-neg-V and do-inversion, starts after the peak of do-V. The long tail of do-V corresponds to emphatic *do* in recent times.

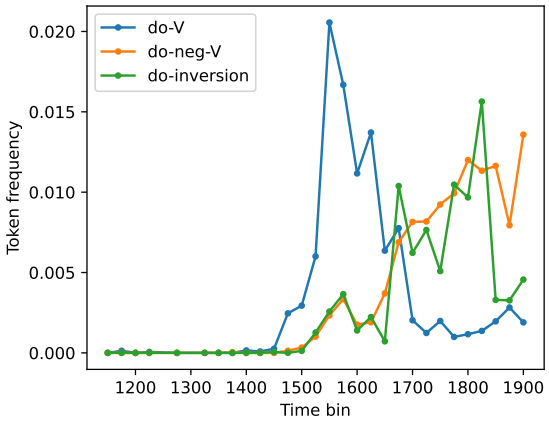

Fig. 1: Token frequency of *do* subtypes over time.

(1) ... no man **doth** blame hym.
(1497, PPCME2; CMINNOCE-M4)

My argumentation is based on learnability: I postulate what kinds of grammars learners at each point in time may plausibly acquire, given their input evidence. As such, Figs. 2–4 count type rather than token frequency; i.e., the number of unique lexical verb lemmas that are attested with the target condition. Type counts are restricted to the most frequent $N = 50$ verbs in the dataset at a given time, in consideration of a typical child's vocabulary size (Yang 2016, Kodner 2020) and data sparsity.

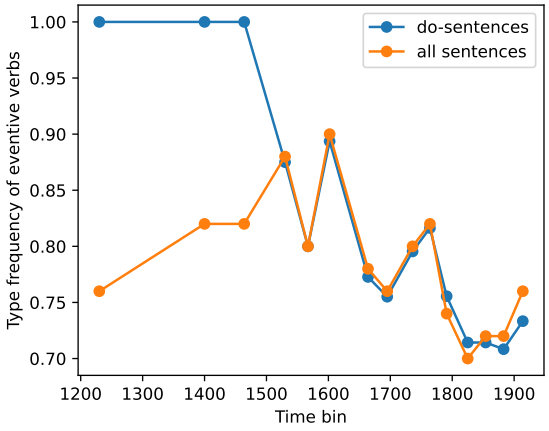

Fig. 2: Type frequency of eventive verbs in *do*-sentences vs. all sentences. Time bins chosen for equal number of sentences per bin.

3. The beginnings: lexically conditioned *do*. Previous work has proposed that early *do* occurred in semantically restricted environments, unlike in PDE. For example, Ecay (2015) argues that *do* (in EME pre-1575) is a light verb that marks agentivity. I will call this early function of *do* "eventivity marking" for convenience.

In accordance with these proposals, Fig. 2 shows that early *do* occurs at a disproportionately high frequency with eventive verbs, compared to the baseline proportion of such verbs in the dataset. In the earliest time bins, all instances of *do* occur with eventive verbs. Thus, learners are able to maintain a restriction on *do*: given N instances of *do*-sentences, all N of them occur with an eventive verb, trivially surpassing any threshold for successful acquisition. This learned restriction can be formulated either distributionally, as *do* being conditioned by a restricted lexical class; or semantically, as *do* contributing a coherent function only compatible with eventive verbs.

4. *Do* becomes unconditionally productive. In the mid-1500s, the distributional evidence for learners changes. In the view of learning underlying the Tolerance Principle (Yang 2016), given sufficient evidence, learners learn the maximally general rule. In the earlier period of *do*, learners only had sufficient evidence to learn *do* as conditioned by a semantic class of verbs, but not compatible with all verbs. However, learners exposed to data generated by this conditioned use of *do* may nevertheless learn *do* as unconditionally productive, if the (semantically restricted) verbs they hear with *do* happen to constitute a sufficiently large proportion of all verbs.

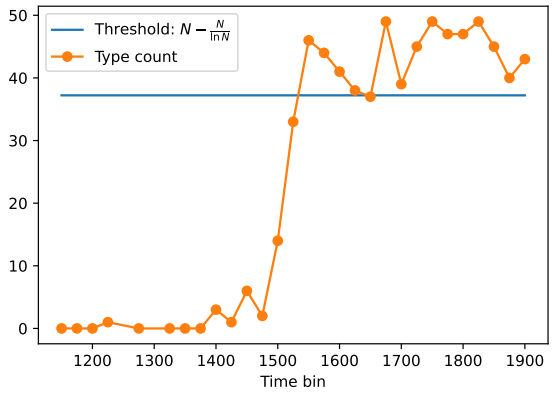

Fig. 3: Type count of verbs attested in *do*-sentences vs. generalization threshold over time. $N = 50$.

As Fig. 3 shows, from the mid-1500s onwards, *do* occurs with enough verb lemmas such that it surpasses the threshold for productivity among all verbs. In the time bin 1550–1575, the number of verbs that occur with *do* ($k = 46$) first exceeds the threshold for successful generalization ($50 - \frac{50}{\ln 50} = 37$ for $N = 50$ total verb lemmas).

Once *do* becomes unconditionally productive, it begins to occur with verbs incompatible with its original function of eventivity marking. For example, *do* appears with the stative verb *know* in 1529, as in (2). Fig. 4 shows the growth of *do*-sentences with non-eventive verbs. Once learners are exposed to sufficiently many non-eventive verbs with *do*, they can no longer learn *do* as an eventivity marker.

(2) ... your power, wych I **do** know is great ...
(c. 1529, PPCEME; WOLSEY-1529-E1-P1)

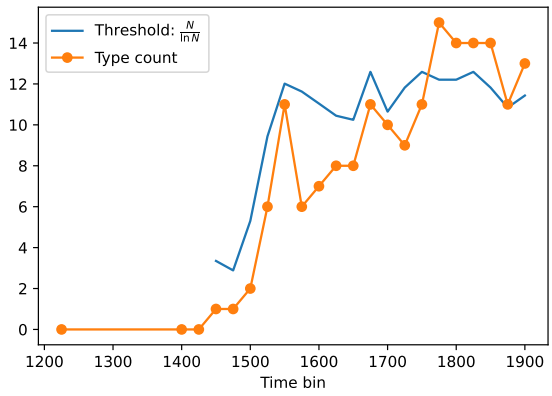

Fig. 4: Type count of non-eventive verbs in *do*-sentences vs. tolerance threshold over time. N_t is the total number of *do*-sentences.

6. The aftermath. Learners must then hypothesize a new rule to explain the observed uses of *do*. Unlike distributionally similar auxiliaries (e.g. modals), *do* does not have a clear semantic contribution. Thus, it is plausible that learners reanalyze *do* as a purely functional element bearing tense for a lexical verb, consistent with its distribution and lack of meaning otherwise. Once this reanalysis happens, *do* becomes an alternative to V-to-T movement in the relevant environments, and thus starts to compete directly with it.

References. Ecay (2015). Ellegård (1953). Haeberli & Ihsane (2016). Kodner (2020). Kroch (1989). Kroch & Taylor (2000). Kroch, Santorini, & Delfs (2004). Kroch, Santorini, & Diertani (2016). Yang (2016).