

The acquisition and loss of subject inversion in the history of English

Ans van Kemenade and Charles Yang

Radboud University Nijmegen, University of Pennsylvania

This paper models the development and loss of V2 in English from the perspective Yang's Tolerance principle (Yang 2005, 2010), a formal model for how the learner detects productive processes (or lack thereof) under inconsistent input.

It is well-known that V2 was lost in English in the 15th-17th centuries. Van Kemenade and Westergaard (2012) and van Kemenade (2012) discuss the nature of the variation in Middle English, and the loss of V2 in various contexts across the Middle English and early Modern periods. They analyse this variation in terms of an interplay between syntactic and information structural factors. Of the two types of V2 environment in Old and Middle English, type 1 involves V to C movement (following an initial XP that is *wh-*, *ne*, or an adverb of what we call the THEN-group (comprising then, thus, now, cf. Warner 2007). Type 2 involves movement to a functional head below C, which we call then non-THEN (~THEN) group. Van Kemenade and Westergaard show that that V2 variation in Middle English is keyed to three factors: 1) the type of first constituent, 2) the type of finite verb (auxiliary, unaccusative, transitive/unergative intransitive), and the type of subject (pronominal vs nominal, or more precisely, discourse-given vs. discourse-new, generic or focused).

We investigate how language acquisition of the variable data in the early period of ME may account for the observed changes. We focus on a specific subset of the data, involving V2 variation in contexts with adverbs, in which we distinguish between adverbs of the THEN-group (the first type of V2 involving V to C), and the ~THEN group (the second type of V2). We distinguish, within these contexts, two types of finite verb (unaccusative henceforth *unacc* and transitive/unergative intransitive; henceforth *unerg*), and analyse the development of inversion from the perspective of Yang's Tolerance Principle (Yang 2005, 2010).

The Tolerance Principle guides the learner to learn productive processes which may have exceptions, and places an upper bound on the amount of exceptions that a productive process can tolerate. Suppose a generalization in principle applies to N lexical items, out of which some exceptional members do not follow the generalization. The Tolerance Principle holds that the generalization is productive if the number of exceptions does not exceed $N/\ln N$. The model has been widely applied to the study of phonological, morphological and syntactic acquisition.

We examine the Middle English and early Modern English data in the Penn parsed corpora for Middle English (Kroch and Taylor 2000). For each of the seven periods, we extracted patterns of inversion and non-inversion following specific types of adverb, categorized for three finite verb types and according to lexical item (verb lemma). Our goal is to show that the trajectories of change are broadly predictable on the basis of the data from the first stage of Middle English (1150-1250), which may be viewed as an approximation of the data for the learners at the time.

To apply the Tolerance Principle, we examined the properties of verbs—whether it participates in subject inversion or noninversion—the nature of the subject (pronominal or nominal NP), initial topic (THEN or ~THEN) as well as verb class: unaccusative (*unacc*) and transitive/unergative intransitive (*unerg*). We counted the verb type frequencies for all such cases, excluding items that show variation, i.e. verbs used in both inversion and non-inversion were excluded, as they provide conflicting information. The excluded items constitute less than 10% of the verb types. We illustrate with a concrete example. When the initial topic is THEN,

and the subject is a lexical NP, the data contain 16 unacc verbs (14 inversion, 2 non-inversion) and 37 unerg verbs (31 inversion and 6 non-inversion). The tolerated thresholds for the verb classes are $16/\ln 16=6$ and $37/\ln 37=10$. We use the notation such as (16, (4, 2), 10) to combine these quantities. For both verb classes, then, the model predicts inversion to be the productive process. (A productive process exists if either of the numbers in the inner brackets is smaller or equal to the last number, the threshold value.) The results of our quantitative analysis are given below, where the productive generalizations are given in boldface:

- (1) (THEN, pro): (a) **unacc (19, (17, 2), 6)**, (b) unerg (64, (46, 18), 15)
- (2) (THEN, NP): (a) **unacc (16, (14, 2), 6)**, (b) **unerg (37, (31, 6), 10)**
- (3) (~THEN, pro): (a) **unacc (28, (1, 27), 8)**, (b) **unerg (167, (16, 151), 33)**
- (4) (~THEN, NP): (a) unacc (17, (8, 9), 6), (b) unerg (31, (15, 16), 9)

Thus, the model predicts in type 1 V2 (following THEN), NP subjects are productively inverting (2) whereas the noninversion pattern for pronominal subjects in type 2 V2 following (~THEN)—the modern English form—was already firmly established. The latter point is notable, because at this early stage of ME, the statistical frequencies of inverted subjects in all contexts are very high, yet the productivity model predicts its ultimate demise. The result also suggests that word order variation is sensitive to the type of adverb as a function of information structure (Kemenade & Westergaard 2012).

A final piece of the puzzle is: Why did the non-inversion eventually eliminate the inversion option, especially for (1) and (2), where the inversion pattern was productive in early ME? Here we provide a tentative solution again based on the productivity model. If we pool lexical and pronominal subjects together (1 and 2), the inversion pattern ceases to be productive for both verb classes: unacc (27, (17, 10), 8); unerg (58, (37, 21), 14). By contrast, if we pool lexical and pronominal subjects together for the ~THEN case, the noninversion patterns remains productive: **unacc(41, (6, 35), 11)**, **unerg(186, (24, 162), 36)**. Indeed, if we collapse the CP and non-CP level adverbs together, we obtain: unacc (49, (15, 34), 13), unerg (208, (46, 162), 39). While non-inversion process does not reach the theoretically predicted threshold for productivity, the number of exceptions is only slightly higher (15 vs. 13 and 46 vs. 39). With the effect of sampling in linguistic input, this result amounts to a tentative argument that non-inversion was on the brink of productivity—a prediction we can make on purely quantitative basis from the earliest stage of ME.

In sum, we have provided preliminary evidence that a formal model of productivity learning, with the use of parsed syntactic corpora, may offer insight on syntactic change from the perspective of language acquisition. Future studies will expand to the full range of first position constituents for a complete understanding of the inversion phenomenon in the history of English.

References

- Kemenade, Ans van. 2012.** Rethinking the loss of V2. Interfaces section of E. Closs Traugott and T. Nevalainen (eds) *Rethinking the History of English*. OUP. 822-834.
- Kemenade, Ans van; Westergaard, Marit . 2012.** Syntax and Information Structure: V2 variation in Middle English. In: López-Couso, María José; Los, Bettelou; Meurmann-Solin, Anneli (eds.). *Information Structure and Syntactic Change*. OUP. 87-118.
- Kroch, Anthony & Ann Taylor. 2000.** *The Penn-Helsinki Parsed Corpus of Middle English 2. English*. University of Pennsylvania, Dept. of Linguistics.
- Warner, Anthony. 2007.** Parameters of variation between verb-subject and subject-verb order in late Middle English. *English Language and Linguistics* 11.1, 81–112.
- Yang, C. 2005.** On productivity. *Language Variation Yearbook*. 5, 333-370.
- Yang, C. 2010.** Three factors in language variation. *Lingua*. 120, 1160–1177.