Superset grammars and failed changes

Daniel Lassiter and Rob Truswell,
University of Edinburgh

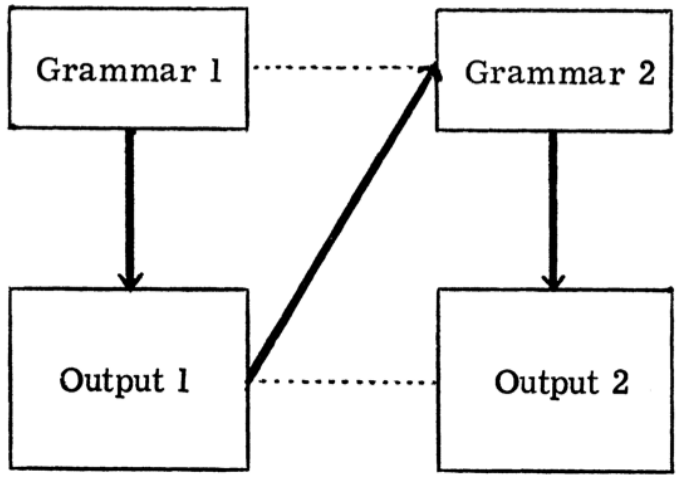
Meertens Instituut, 15/5/2023

Outline

- Many grammatical changes fail.
- Even innovations which first spread to a significant extent can still fail.
- In some cases, the innovation which ultimately fails is more expressive than the grammars it is competing with.
  - We will call this a superset grammar.
  - (Even though the relationships between generative capacity can be slightly more complex than subset–superset.)

Why is this interesting?

- Consider the iterated learning model from Andersen (1973).



- Four components influence the character of Grammar 2:
  1. The character of Grammar 1;
  2. The output produced by Grammar 1 (contingent on communicative intentions, expressivity, style, …);
  3. The learning process based on Output 1;
  4. 'Extrinsic' factors (e.g. social structure).
- Failed superset grammars are interesting because a learner, at some point, induces a grammar (Grammar 2) which does not generate all of Output 1.

Why is this interesting?

- Failed changes impose sufficiency criteria on models of change:
  1. It is possible for changes to fail.
  2. It is possible for superset grammars to fail.
  3. If this cannot be captured in terms of 1–3, we should pursue 'extrinsic' explanations.
- Yang's (2002) model of grammar change, incorporating the Linear Reward–Penalty reinforcement learning model of Bush & Mosteller (1951), cannot capture 1 or 2 intrinsically.
- We investigate an alternative derived by replacing the Linear Reward–Penalty schema with a Bayesian model of grammar learning, within the general framework developed by Yang.
  - This avoids the problematic predictions of Yang's model w.r.t. 1–2.
  - However, it doesn't explain *when* changes fail.
  - And if there is a link between superset grammars and failed changes, it doesn't capture the link.
- This leads us to consider extrinsic motivations for failed changes.

Roadmap

1. Grammar competition and failed changes
2. Yang's model and superset grammars
3. A failed superset grammar in the *Cursor Mundi*
4. Bayesian learning and failed changes
5. Beyond parsing success

Section 1

Grammar competition and failed changes

Themes from Kroch

- Grammars are discrete systems.
- An individual knows, and uses, multiple grammars.
- Part of an individual's knowledge of language is a distribution over grammars.
- Acquisition is not a question of learning a single 'target' grammar, but learning a probability distribution over grammars.

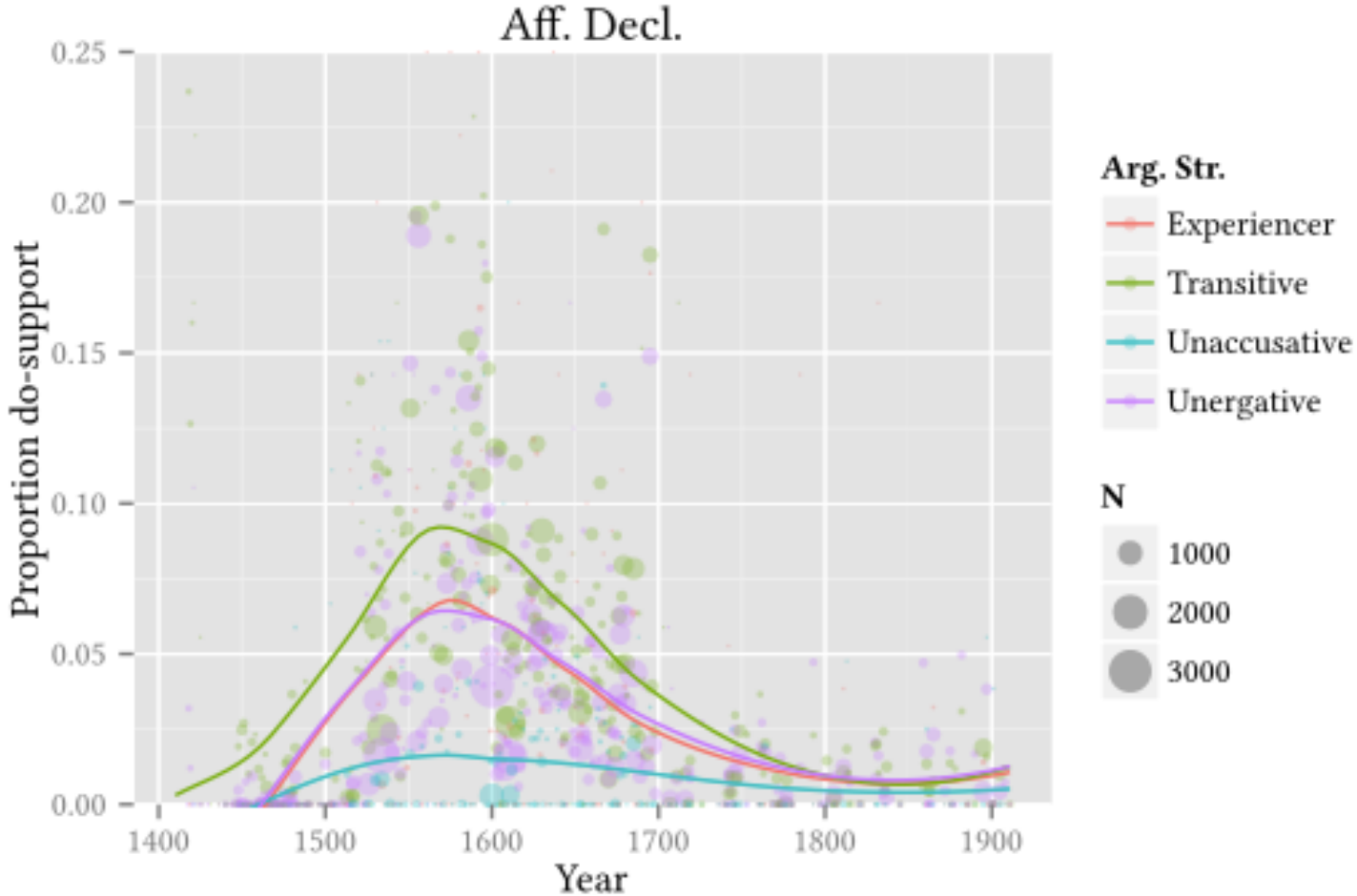
Grammar competition and S-curves

- Kroch linked grammar competition to the S-shaped trajectory of many grammatical changes.
- Blythe & Croft (2012): S-shaped trajectories require replicator selection.

  > The replicator selection mechanisms for propagation proposed by sociolinguists are all social. Other linguists have proposed other mechanisms for propagation, including phonetic biases and morphological analogy favoring some variants over others for sound change, and structural or functional biases favoring some variants over others for grammatical and lexical change (p. 273)

Competing, winning, and losing

- The interest of failed changes stems from their implications for selection.
- Initially, the variant of interest seems to be spreading along an S-shaped trajectory.
- And then, suddenly, it doesn't (figure from Ecay 2015).



Why?

- Adopting Blythe & Croft's terms, one reason why a change can fail is a change in the selectional dynamics of the linguistic community.
- More precisely, this could mean many things:
  - Random change.
  - A sudden change in the composition of the linguistic community (e.g. contact phenomena) engenders changes in the social values associated with competing forms.
  - Concurrent changes affect the status of the competing forms in the grammar.
- All of these are viable explanations for failed changes. We believe that they are all accurate in some cases.
- We want to explore another hypothesis, that some variants fail on their own terms, rather than because of extrinsic changes (in social structure, in other aspects of the grammar, etc.).

Anatomy of a failed change

- Failed changes are typically transitional:
  - The community shifts between two more stable states A and B ...
  - ... via the 'failed' state F.
- In reality A and B may themselves be complex, distributions over multiple grammars. But we'll treat A, B, and F as single grammars.
- The learner is then inducing a distribution over A, B, and F: assign a probability to each grammar in the range $0 \leq p \leq 1$.
  - Initially, $P(A) \approx 1, P(F) \approx P(B) \approx 0$.
  - Transitionally, $P(F)$ increases (while typically staying far below 0.5). $P(B)$ increases (typically $< P(F)$ at first). $P(A)$ decreases.
  - Finally, $P(B) \approx 1, P(F) \approx P(A) \approx 0$.

Questions about failed changes

- In the terms of this general picture, we can ask several questions:
  1. Why does $P(F)$ increase vis-à-vis $P(A)$?
  2. Why does $P(F)$ decrease vis-à-vis $P(B)$?
  3. ...
- In this talk, we're going to focus on Question 2.
- We're also going to concentrate on the case where F is a superset grammar, relative to both A and B.
- We'll also assume, for the sake of the argument, that these *why*-questions have intrinsic answers.

Section 2

Yang's model and superset grammars

Learning and change

- Following Andersen's diagram, Yang (2002) ties learning to change.
- An individual's knowledge of language includes a distribution over grammars.
- Generation n produces utterances, sampling from their distribution over grammars. This is the environment in which Generation $n+1$ learns.
- Grammar learning: updating a distribution over grammars in response to the linguistic environment.
- Grammar change: Generation $n+1$ induces a distribution which differs from Generation n.
  - Interesting grammar change: the differences between Generations n and $n+1$ are nonrandom.

Components of Yang's theory

- In these terms, an intrinsic theory of change boils down to:
  1. An initial state
  2. A specification of what a grammar is
  3. A production theory
  4. A learning theory

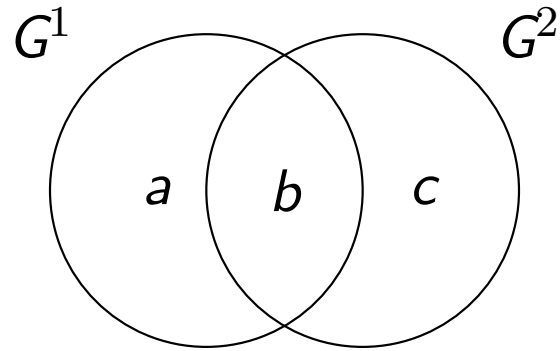
  Anything that can't be explained in these terms must have an extrinsic explanation.
- We have nothing interesting to say about 1. today.
- We will shy away from 2., although we believe that the approach here will ultimately be informative in this respect (work in progress with Richard Blythe and Simon Kirby).
- We'll follow Yang w.r.t. 3.:
  - An agent x samples a grammar G_i with probability $p(G_i)$, determined by the weight that x assigned to G_i during learning.
  - x produces an utterance U_i generated by G_i. Note that U_i could in principle be generated by other grammars as well (grammars overlap extensionally).
- We will cast a critical eye on 4., the learning theory adopted by Yang.

The Linear Reward–Penalty learning theory

- Yang adopts a reinforcement learning theory from Bush & Mosteller (1951). Originally:
  1. Choose an action at time t with probability p_t.
  2. If the action is rewarded: $p_{t+1} = p_t + \alpha(1 - p_t)$, and normalize by reducing probability of other actions.
  3. If the action is punished: $p_{t+1} = p_t - \alpha(1 - p_t)$, and normalize by increasing probability of other actions.
- In our case:
  - the 'action' is choosing a grammar to parse a sentence.
  - the 'reward' is parsing successfully.
  - the 'punishment' is failing to parse.

Parsing success and fitness advantage

- An entailment of this application of LRP: parsing more sentences $\rightarrow$ higher p_t (fitness advantage). To see why, consider the following.

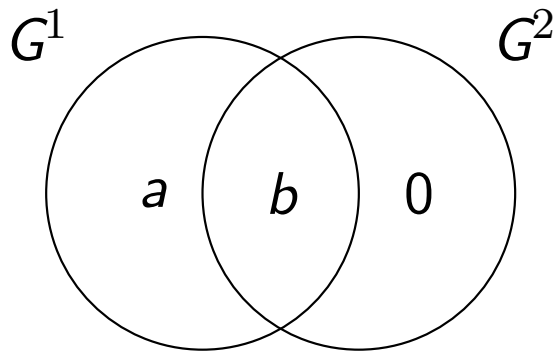


- The linguistic environment consists of three types of utterance:
  - a: parsable only by G^1;
  - b: parsable by G^1 or G^2;
  - c: parsable only by G^2.
- If there are more as than cs, G^1 will be rewarded (and there will be even more as in the next generation). Other way round if there are more cs than as.

$$p^1_\infty = \frac{a}{a+c} \qquad p^2_\infty = \frac{c}{a+c}$$

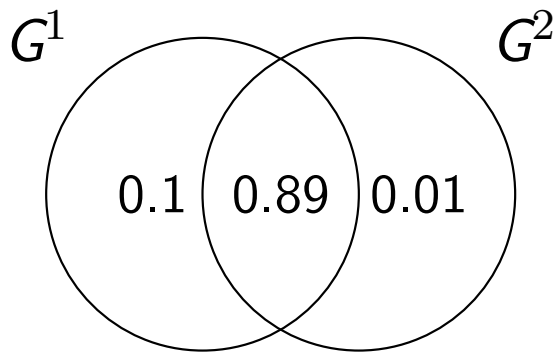
The LRP and superset grammars

- When G^1 is a superset grammar and G^2 is a subset grammar:



$$p^1_\infty = \frac{a}{a+0} = 1 \qquad p^2_\infty = \frac{0}{a+0} = 0$$

- The cases of interest to us aren't 'pure' superset–subset. Rather G^1 is an approximate superset grammar.



In this case, $p^1_\infty = \frac{0.1}{0.1+0.01} = 0.91$. $p^2_\infty = 0.09$.

Yang's 'fundamental theorem'

- The LRP rewards only parsing success.
- **The fundamental theorem of language change**
  G_2 overtakes G_1 if ... the advantage of G_2 is greater than that of G_1.
- **Corollary**
  Once a grammar is on the rise, it is unstoppable (p. 131)
- Any endogenously caused failed change falsifies the corollary (and therefore the theorem).
- Any intrinsically failing superset grammar directly falisifies the theorem.

Detour

- Both of these predictions follow only from the choice of LRP as learning model. This choice is not extensively discussed by Yang (see p. 29).
- LRP was originally developed as a theory of learned behaviour.
- But picking a grammar to parse a sentence is an atypical type of behaviour.
- All the more so when we consider that this behaviour is coupled to a second, more concrete type of behaviour (production), which has complex interdependencies with parsing.
- With echoes of Chomsky (1959): what a learner is actually learning is a mental representation.
- By exclusively rewarding parsing success, the LRP introduces a bias toward superset grammars.
- This is very different from rewarding targetlike representations.
- Hold that thought ...

Section 3

A failed superset grammar in Northern Middle English

Introduction

- We document a failed change involving a superset grammar (the CM grammar) in Northern Early Middle English.
- The Edinburgh manuscript of the *Cursor Mundi* (early 14th century) shows highest frequencies (still low!) of sentence types associated with this grammar (Truswell 2021).
- It can parse almost everything parsable by competing grammars of the time, and more.
- The grammar didn't spread: it is primarily detectable in northern texts, in a very short time window (admittedly confounded by scarcity of early northern texts).
- This is the kind of pattern which cannot be explained by Yang's model in intrinsic terms.

Background

- The CM grammar must be seen in the context of at least three other English grammars:
  1. Northern EME is a classic V2 grammar (V-to-C movement). Although the clearest evidence for this is a later 14th century text (prose rule of St. Benet), Kroch & Taylor (1997) argue that this grammar was stably present since OE.
  2. Southern EME is a mainly-V2 grammar that has obligatory XSV orders in certain cases including subject pronouns. Analysed following Haeberli (2000) in terms of verb movement to an intermediate head (F) whose specifier is filled by only the relevant subject classes.
  3. Late ME is an SVO language with no V2, but (transitionally?) with V-to-T movement.
- The most important contact languages at the time are Old Norse (this being inside the Danelaw) and Old French (source for much of the *Cursor Mundi* text).

The CM grammar

- The CM defines a left periphery where:
  - The verb moves to F like southern EME;
  - Spec,FP and Spec,CP are both A′-positions; fillable by the same phrase or different phrases.
  - A complementizer occupies C in embedded finite clauses; Spec,FP remains an A′ position.
- Predictions:
  - A full range of V2 and V3 orders with and without inversion in matrix clauses;
  - Embedded V2 following complementizers.

Examples: matrix V3

(1) [$_{PP}$ Of þis t^{o}wþe] [$_{AP}$ hard] es t^{o}wþe to find
of this truth hard is truth to find

'Of this truth, truth is hard to find.' (edincmat.180)

(2) [$_{AP}$ Sa brad] [$_{PP}$ of hir blis] es þe wai
so broad of her bliss is the way

'The way of her bliss is so broad.' (edincmat.1090)

(3) [$_{NP}$ A clud] [$_{PP}$ again hī] sau þai liht
a cloud against him saw they alight

'They saw a cloud descend towards him.' (edincmat.152)

Examples: embedded V2

(4) Men wat [$_{CP}$ þat [$_{AdvP}$ fur ner] es somꝰ comād]
men know that full near is summer coming

'Men know that summer is drawing near.' (edincmbt.258)

(5) He wenes [$_{CP}$ [$_{Obj}$ his mak] mai naman find]
he believes his match may no.man find

'He believes that no one may find his equal.'
(edincmbt.553)

(6) ... [$_{CP}$ Þat [$_{PP}$ al to pecis] sal tai brist]
that all to pieces shall they burst

'... that they shall burst all to pieces.' (edincmat.131)

These orders reflect a single underlying grammar

- Strong correlation between matrix XYVS and embedded XVS in ME (Spearman's $\rho = 0.54$, $p < 0.001$).

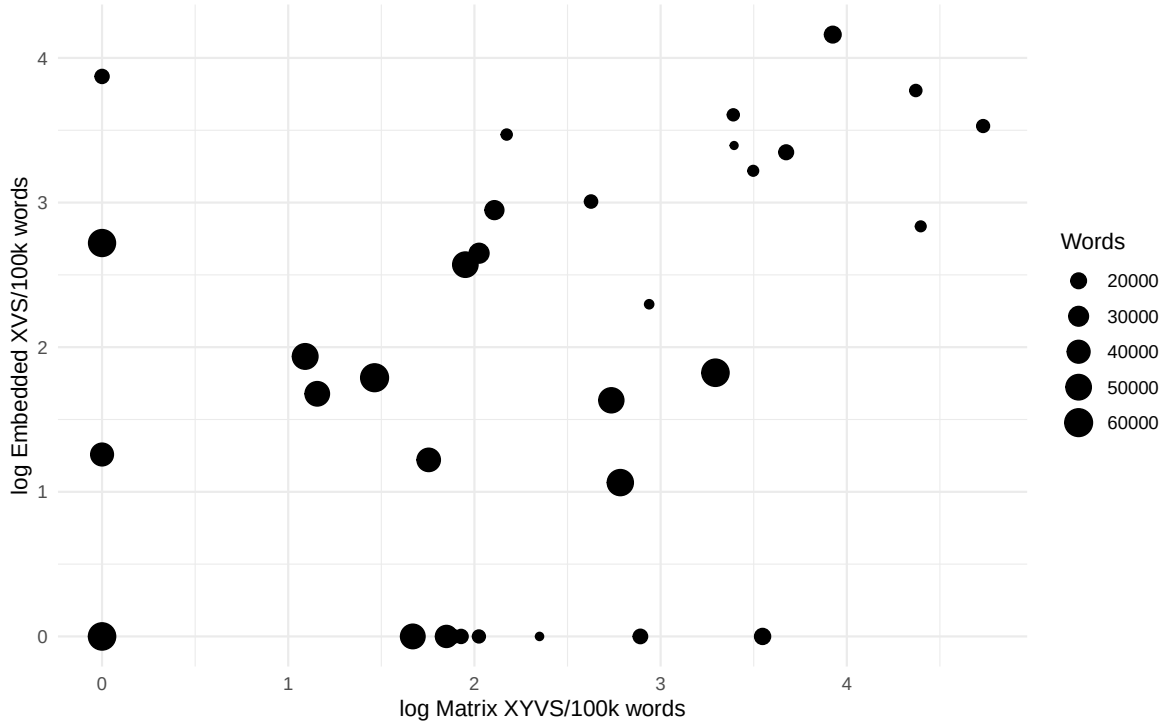


Figure 1: Frequency of matrix XYVS and embedded XVS orders in Middle English texts

- This is expected if the two orders are products of a single underlying grammar.

The CM grammar has a fitness advantage over all competing grammars, everywhere

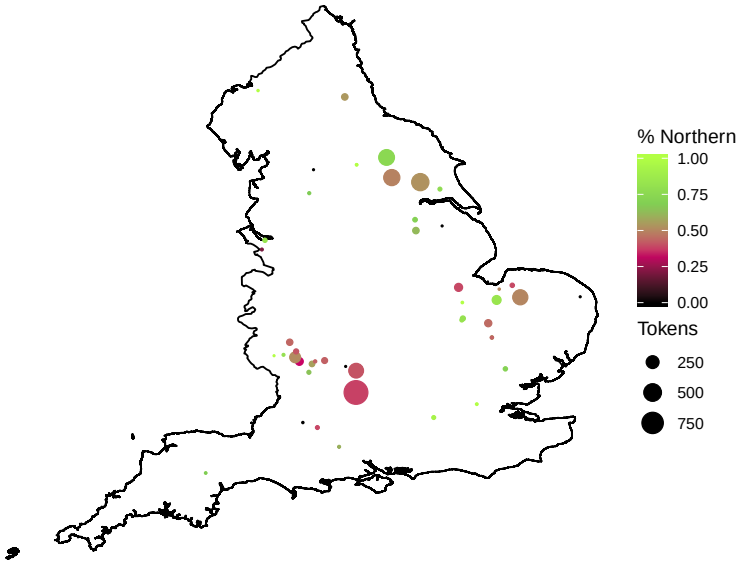
Region	# clauses	Northern V2	Southern V2	SV	CM
SW	2,074	863 (42%)	1,384 (67%)	1,365 (66%)	1,490 (72%)
North	1,412	828 (59%)	840 (59%)	797 (56%)	1,045 (74%)
E. Anglia	696	383 (55%)	462 (66%)	446 (64%)	514 (74%)
Total	4,251	2,126 (50%)	2,739 (64%)	2,657 (63%)	3,107 (73%)

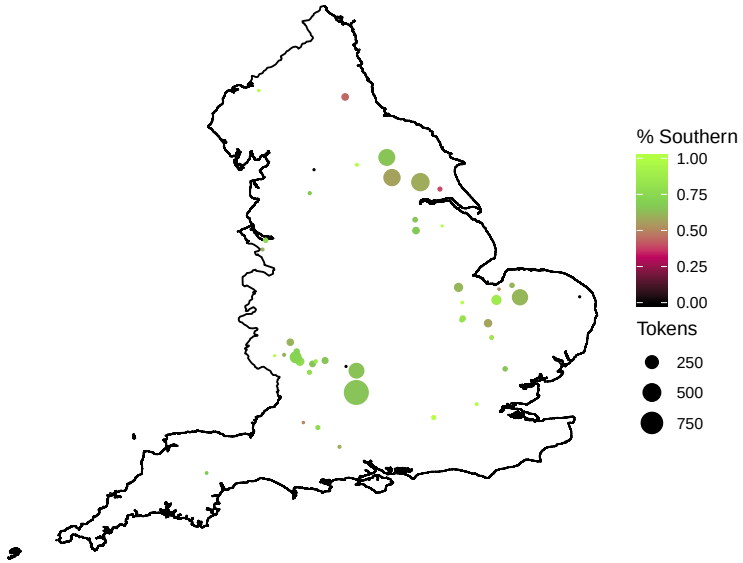
Table 1: Parsing success of four competing grammars in three different regions

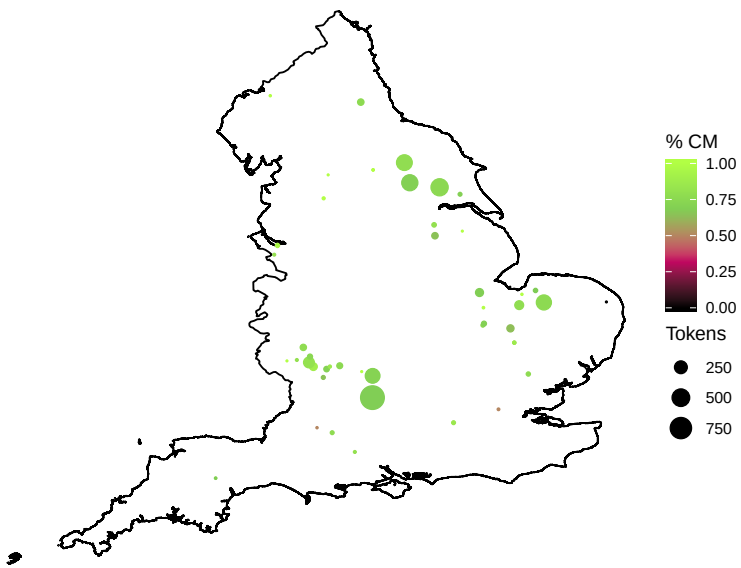
Clause type	# clauses	Northern V2	Southern V2	SV	CM
Matrix	3,132	1,307 (42%)	1,915 (61%)	1,838 (59%)	2,510 (80%)
Embedded	1,119	819 (73%)	819 (73%)	819 (73%)	597 (53%)

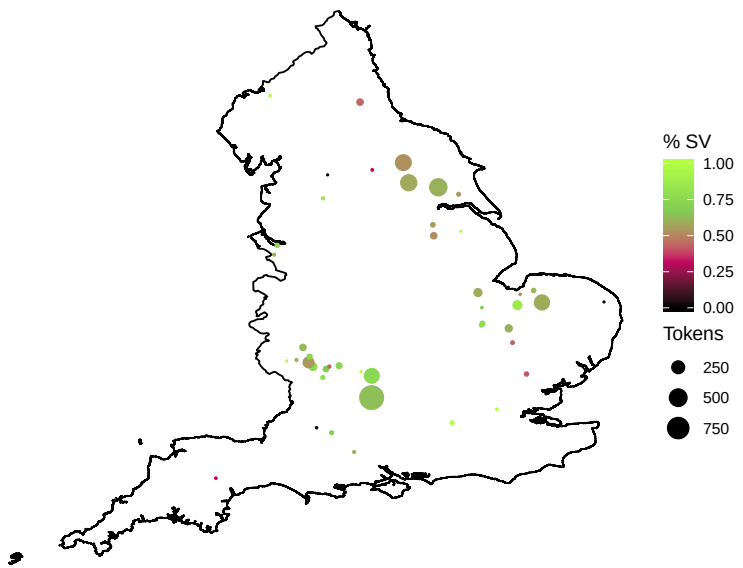
Table 2: Parsing success of four competing grammars in matrix and embedded clauses

Fitness advantage: maps









The CM grammar didn't spread

- We don't know exactly what happened to the CM grammar.
- It is clearly associated with northern texts (*Cursor mundi*, *Rule of St. Benet*), and to a lesser extent East Anglian texts (*Genesis and Exodus*, ?*Ormulum*).
- We don't have many long northern ME texts.
- But we do know that:
  - it was never very visible elsewhere;
  - by the time we do have more plentiful northern material, it's disappeared.
- With low certainty, we can interpolate a failed change, peaking in the 14th century and failing by the late 15th century.

Section 4

Bayesian learning and failed changes

Introduction

- From a LRP perspective, the CM grammar shouldn't have failed:
- The CM grammar is very expressive, and the LRP model rewards expressiveness.
- So the LRP model predicts that the CM grammar should be 'unstoppable'.
- Question for this section: can other intrinsic models make better predictions?
- If no, we should endorse the implication that failed superset grammars have extrinsic explanations
  - (even if we develop improved models of learning and change as we go).

A Bayesian alternative

- Bayesian learning models are attractive from this perspective, because they penalize overexpressivity.
- In the general case:

$$P(M \mid D) = \frac{P(D \mid M) \times P(M)}{P(D)}$$

The posterior probability of a model M given data D is a function of:

  - the prior probability of M, and
  - the likelihood of D given M.

- A learner is choosing between competing models of the target language. For any two models M_1, M_2:

$$\frac{P(M_1 \mid D)}{P(M_2 \mid D)} = \frac{P(D \mid M_1) \times P(M_1)}{P(D \mid M_2) \times P(M_2)}$$

- An overexpressive model assigns lower likelihood to the data, and is therefore penalized, all else being equal.

What is a targetlike representation?

Not just a grammar

- A model of language can't just be a grammar as typically understood.
- A grammar G_i generates a set of strings S_i.
- G_i assigns probability 0 to any $s \notin S_i$ (G_i can't generate s).
- In grammar competition, speakers command multiple discrete grammars G_i, G_j generating distinct sets of strings S_i, S_j.
- If a speaker considers only the likelihood of some $s_j \notin S_i$, *given* G_i, $P(s_j \mid G_i) = 0$.
- So the posterior probability of G_i is zero, if the learner encounters a single sentence G_i can't parse.
- Usual Bayesian workaround is to set $P(s_j \mid G_i)$ in such cases to an error term ϵ close to zero.
- We think this does violence to the useful conception of grammars as discrete systems, so we prefer to avoid this workaround.

What is a targetlike representation?

A distribution over grammars

- Instead, a learner learns a distribution over multiple grammars.
- In the simplest case, with two grammars G_1, G_2, the distribution is just a probability $p \in [0, 1]$ assigned to G_1. The learner's model is a distribution over possible values of p.
- Assuming a flat prior over values of p:

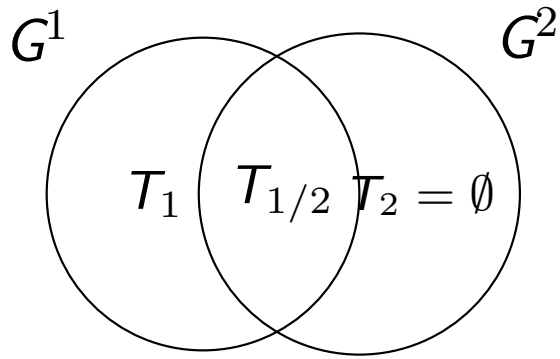
$$P(p \mid D) \propto P(D \mid p)$$

- Assume a simple production algorithm:
  1. Pick G_1 with probability p; otherwise pick G_2.
  2. Sample from the strings generated by the grammar you pick.
- Strings can then be partitioned into three classes:
  - T_1, generated only by G_1;
  - T_2, generated only by G_2;
  - $T_{1/2}$, generated by both grammars.
- The posterior distribution is then determined by how much data falls into each of these classes.

$$P(D \mid p) = \prod_{T_i} \prod_{d \in D : d \in T_i} P(d \in T_i \mid p),$$

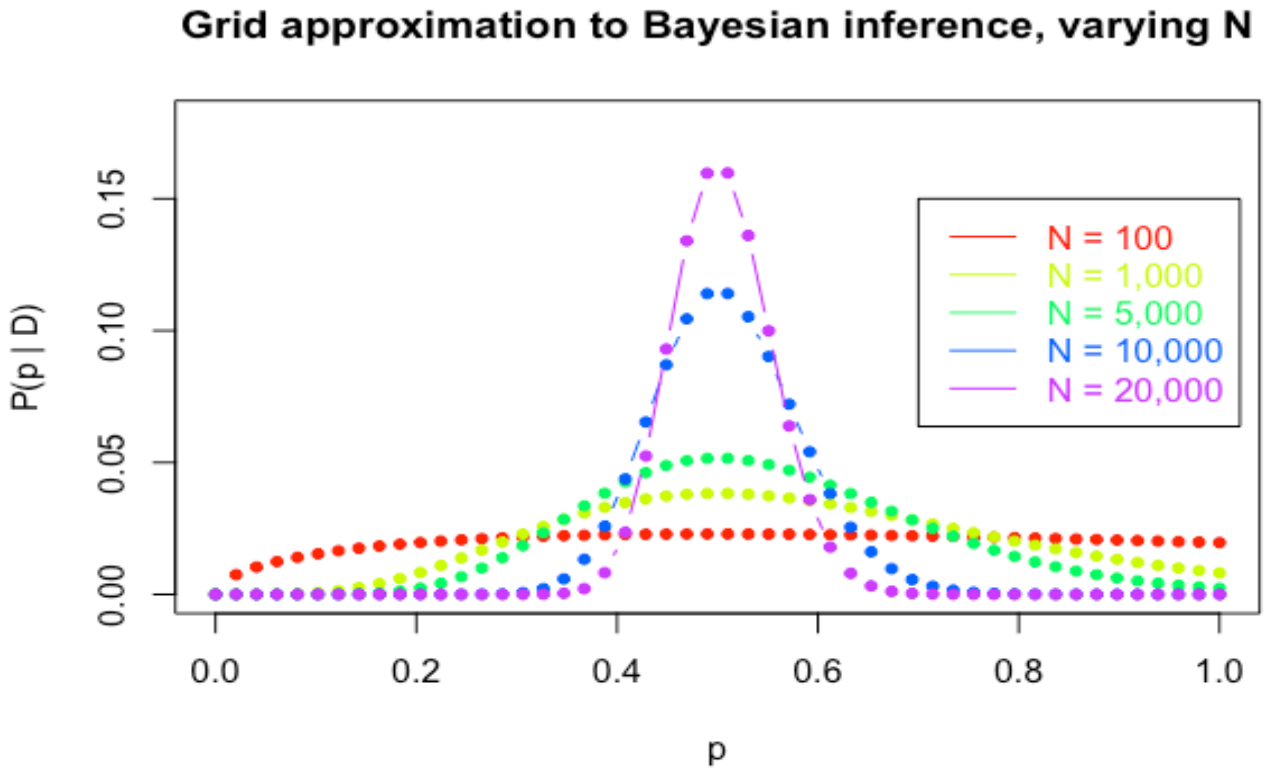
The superset case

- We can model the superset case by stipulating that there are no strings of type T_2.



- The learned distribution of values for p is conditioned by:
  - The actual value of p;
  - The relative size of T_1 and $T_{1/2}$ (generative capacity of G^1 and G^2)
  - The number of tokens N.
- We're still exploring here, but headline result for today: superset grammars aren't unstoppable.

The superset case



- With input $p = 0.5$, 0.5% of input in T_1, and reasonably large N, the learned distribution over values of p is centred on 0.5.
- This demonstrates that there are conditions under which superset grammars do not outcompete subset grammars.

Discussion

- We have developed a simple Bayesian model of grammar competition.
- It models learning as inducing a distribution over grammars.
  - In the 2-grammar case, fixing the value of a single continuous parameter.
- The behaviour of this model is a clear improvement over Yang's LRP-based learning model, in that it doesn't automatically reward more expressive grammars.
- Rather, it assigns a nonzero probability to the subset grammar, reflecting:
  - the amount of data;
  - the amount of evidence in the data which uniquely supports the superset grammar.
- Because of the modular nature of Yang's model, it can be improved by switching out the learning model and keeping everything else constant.
- However, the model doesn't explain how changes can fail ...

Section 5

Beyond parsing success

Grammar competition in context

- Both the LRP learning model and our Bayesian alternative assign weights to competing grammars based exclusively on the parsing success (weak generative capacity) of those grammars.
- These models are useful in that they act as baseline models for richer theories of learning and change.
- A full theory of change would have to consider at least:
  - Strong generative capacity
  - Interfaces between grammatical components
  - Use (including pragmatics, processing, communicative success, etc.)
  - Social structure
- If a realistic theory of grammatical learning and change can't capture attested patterns of grammar change, that should motivate the investigation of extrinsic causal factors.
- We don't have anything like a full theory, but in this last section we'll speculate about extrinsic forces involved in the CM grammar failed change.

Emergence of the CM grammar: Contact?

- The distinctive CM word orders already existed, marginally, in Old English.
- But they were apparently more common in Old French.

(7) Car [AdvP ja] [Obj bonne ouevre] ne fera Qui la fin ne
because never good work NEG do.FUT who the end NEG
resgardera
look.FUT
'Because someone who doesn't keep the goal in sight will never do good work.'
(anonyme_alexandrie,.164)

(See Salvesen & Walkden 2013 on embedded V2 in OF).

- It's possible that the emergence of the CM grammar has an extrinsic cause, especially as much of the source material for *Cursor Mundi* is French.
- If so, we wouldn't expect the intrinsic dynamics of grammar competition to capture it.

Decline of the CM grammar

1: Overexpressivity?

- Warning: speculation!
- Part of knowing a grammar is knowing how to use the different possibilities it affords.
- The CM grammar is more expressive than competing grammars.
- It's not clear what (if anything) the communicative function of that surplus expressivity is.

Decline of the CM grammar

2: Social structure?

- Long variationist tradition of documenting the effects of social structure on the propagation of grammatical changes.
- Blythe & Croft (2012) provide a bridge between this research and competition-based modelling.
- The CM grammar is concentrated in the north of England, not long before a London-based standard emerged.
- In principle, the social dynamics in England at the time could militate against the spread of the CM grammar, disregarding any linguistic facts about the grammar itself.
- Some work would be needed to develop this hypothesis, though, because other distinctly northern forms (e.g. 3pl *they* etc.) did spread throughout English.

Decline of the CM grammar

3: Grammatical interactions?

- The loss of the CM grammar was part of a dizzying 200-year period in which:
  - V2 was lost;
  - Verb-raising was lost;
  - Causative *do* initially became an external argument marker (Ecay 2015) and then became a support morpheme.
- From a competition perspective, all of these changes concern patterns of evidence concerning the position of the verb: in C, in F, in T, in *v*/V.
- Evidence for a low position of the verb is evidence against any V2 grammar, including the CM grammar.
- If reanalysis of *do* (grammaticalization, lexical syn/sem) were to increase evidence for a low position of the verb, that may be enough to tip the scales away from V2 grammars with the verb in C/F.

Conclusion

- Grammar change is complex, in many, many respects:
  - Description of changes;
  - Modelling of causal factors;
  - Modelling of interactions.
- The field has done very well by focusing on unusually 'pure' cases:
  - Competition between a couple of variants;
  - Change along S-shaped trajectory.
- There's a gulf between elegant models and messy reality.
- This talk reports on one step in attempting to reduce that gulf:
  - Improve models of grammar learning;
  - Explore the limits of those models;
  - Induce desiderata for 'extrinsic' types of explanation.

References

Andersen, Henning. 1973. Abductive and deductive change. *Language*. 765–793.

Blythe, Richard & William Croft. 2012. S-curves and the mechanisms of propagation in language change. *Language* 88. 269–304.

Bush, Robert R & Frederick Mosteller. 1951. A mathematical model for simple learning. *Psychological review* 58(5). 313.

Chomsky, Noam. 1959. Review of *Verbal Behavior* by B.F. Skinner. *Language* 35. 26–58.

Ecay, Aaron. 2015. *A multi-step analysis of the evolution of English Do-support*. University of Pennsylvania dissertation.

Haeberli, Eric. 2000. Adjuncts and the syntax of subjects in Old and Middle English. In Susan Pintzuk, George Tsoulas & Anthony Warner (eds.), *Diachronic syntax: models and mechanisms*, 109–131. Oxford: Oxford University Press.

Kroch, Anthony & Ann Taylor. 1997. Verb movement in Old and Middle English: dialect variation and language contact. In Ans van Kemenade & Nigel Vincent (eds.), *Parameters of morphosyntactic change*, 297–325. Cambridge: Cambridge University Press.

Salvesen, Christine Meklenborg & George Walkden. 2013. Diagnosing embedded V2 in Old English and Old French. In Éric Mathieu & Robert Truswell (eds.), *Micro-change and macro-change in diachronic syntax*, 168–181. Oxford: Oxford University Press.

Truswell, Robert. 2021. Grammar competition and word order in a northern Early Middle English text. *Languages* 6. Article 2.

Yang, Charles. 2002. *Knowledge and learning in natural language*. Oxford: Oxford University Press.