Stratified sampling biases models towards nonlinearity

The case of Kallel (2007)

Aaron Ecay

The University of Pennsylvania

June 3, 2011

Introduction

- This talk's theme: non-linearities in corpora
- Plan:
  - Background on modeling language change
  - A concrete example of a non-linearity discovered in a corpus
  - Ways of understanding competing models of this non-linearity
  - What does this tell us about diachronic syntax and how to study it?

Table of Contents

History

- The observation that linguistic changes have a characteristically S-shaped pattern is an old one, dating from at least the mid-20th century
- The identification of the S-shaped curve with the logistic function dates from the early 1980s (Bailey 1973; Kroch 1989; Weinreich, Labov, and Herzog 1968)

History

- The logistic belongs to the family of Lotka-Volterra models, named after their discoverers: Lotka (1925) and Volterra (1926)
- The logistic equation in particular describes the dynamic of two species, one of greater fitness than the other, competing for finite ecological resources

Meaning

- The use of a logistic function as the model of the time course of syntactic change answers a fundamental question about what language change is:
  - The **species** are grammars: abstract mental objects that structure linguistic competence
  - The **resource** for which they compete is use in linguistic expression
  - Language change is **competition among grammars for use**

Use

- The logistic regression procedure fits a logistic function of various predictors to a body of binary data
- It does this via a mathematical transformation ("logit") that maps logistic-shaped curves onto straight lines and a linear model (slope(s) and intercept(s))
- It has become a standard tool in the analysis of syntactic change: see Santorini (1993), Frisch (1997), and many others

Misuse

- The use of the logistic regression for the study of language is not statistically "pure"
- The logistic regression model assumes that every data point is absolutely independent of all the others
- Linguists sample from the same speaker more than once
  - Two tokens from the same speaker are not independent of each other!

Other views on language change

- Kallel (2005, 2007) explores the change in Middle English from a grammar with negative concord (NC) to one which lacks NC:

  (1) I would *not* for *no* good (The Lisle Letters,Vol.V:305)

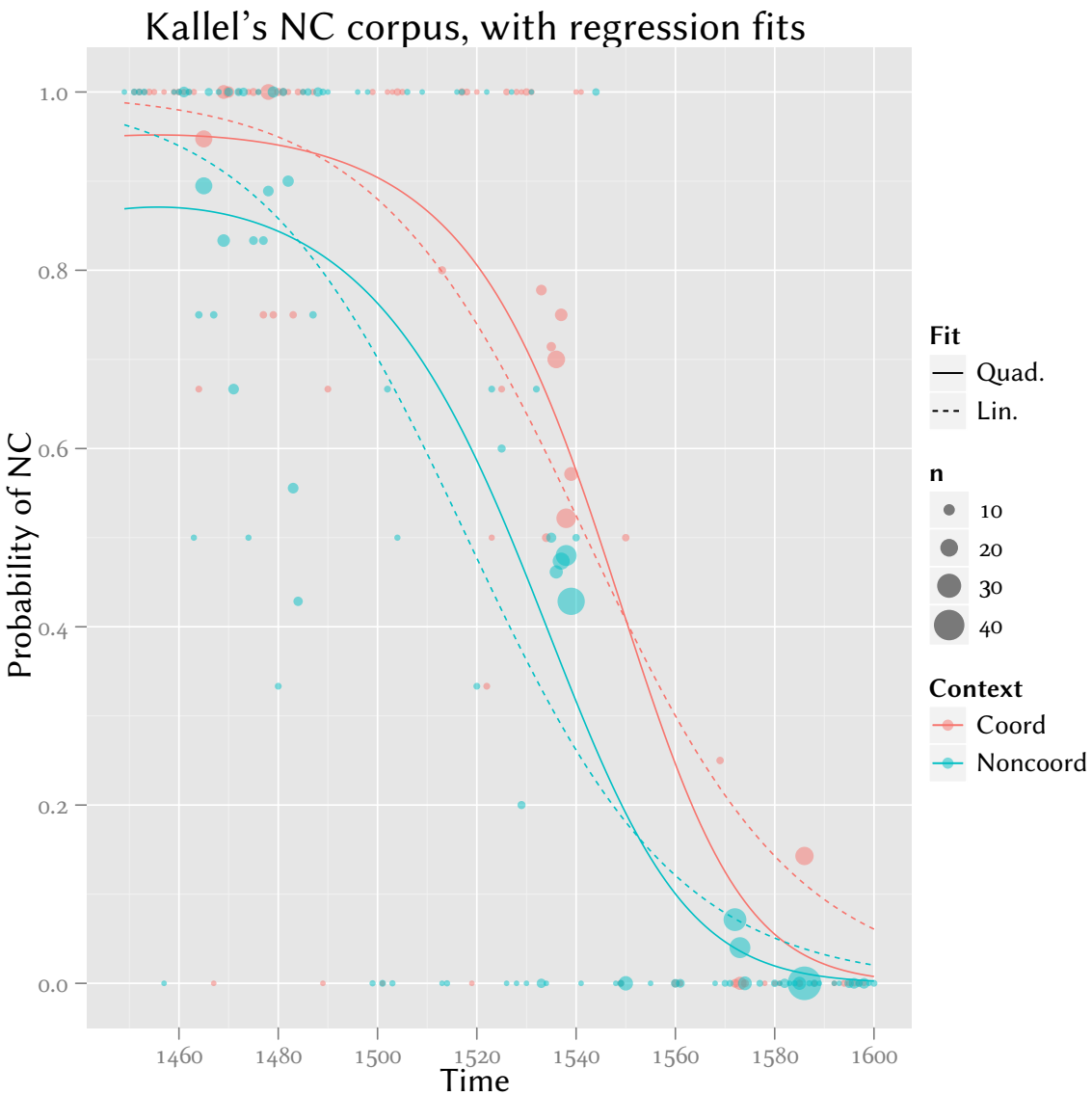
  (2) we cannot certaynelye larne by any meanes (The Letters of Sir Francis Hastings L.3)

Other views on language change

- In the cited publications, Kallel explores the hypothesis that the logit-transformed shape of NC loss is not a linear function of time, but rather a quadratic
- Kallel accepts the hypothesis

Source	Num DF	Den DF	F Value	Pr > F	Chi-Square	Pr > Chi-Square
Stage	1	19	3.86	0.0644	3.86	0.0496
Type	1	19	14.43	0.0012	14.43	0.0001
Stage * Stage	**1**	**19**	**22.93**	**0.0001**	**22.93**	**<.0001**
Function	1	19	0.57	0.4592	0.57	0.4500

The data


Kallel's NC corpus, with regression fits
Probability of NC
1.0
0.8
0.6
0.4
0.2
0.0
1460
1480
1500
1520
1540
1560
1580
1600
Time
Fit
Quad.
Lin.
n
10
20
30
40
Context
Coord
Noncoord

Other views on language change

- If Kallel's conclusion is to be believed, it poses a serious question to historical syntax
- "Constancy across time [= linearity of regression fit – AE], this study shows, is not a requirement; it may or it may not be obtained" (Kallel 2007, p. 47)
  - Linear logistic regression came to be used because language change was thought to be like ecological competition (in the relevant sense). If there is no linear logistic in general, then what is language change?
  - Put another way, years are a plausible regression predictor of language usage because usage patterns change over time. If we accept years-squared as a predictor also, we owe ourselves an explanation of what makes them meaningful in the context of language

Back from the brink

- I will argue that Kallel's conclusion is not warranted
- Further statistical analysis of the models used by Kallel fails to show a maeningful difference between the empirical power of the linear and quadratic models
  - By Ockham's Razor, the traditional linear model is preferred
- The statistical patterns in the data noticed by Kallel may instead attributable to familiar sources of non-homogeneity in language change: inter-speaker and inter-dialect variation
  - The effects of this variation on the model could come about in several ways

Why not higher-order models?

- Accepting *arguendo* Kallel's hypothesis that the quadratic logistic is a better fit for NC-loss, it is possible to ask whether a higher-order function might yield a still-better fit
- An ANOVA test with significance levels derived from the χ^2 test (analogous to Kallel's procedure) reveals the following results:

Model	Resid. Df	Resid. Dev	Df	Deviance	P(>|Chi|)
Linear	928	724.56	—	—	—
Quadratic	927	701.85	1	22.70	0.0000
Cubic	926	697.45	1	4.40	0.0359
Quartic	925	694.64	1	2.81	0.0938
Quintic	924	693.07	1	1.57	0.2098

Why not higher-order models?

- A different model-comparison techique, the AIC, yields similar results:

Linear	Quadratic	Cubic	Quartic	Quintic
730.56	709.85	707.45	706.64	707.07

- The AIC is a measure of the *likelihood* of the model, penalized by the *number of parameters*. Of a pair of models, the one with a lower AIC is taken to be a better fit
- This is an even more worrying result than Kallel's original: we have further increased the explanatory tension between out statistical procedure and our scientific hypothesis about language change

Structured datasets

- In the area of sociolinguistics, where rich data about speaker background is available, attempts have been made to determine the effect of corpus structure on regression results (Gorman 2009; Johnson 2009)
  - Structure could mean sampling over-extensively from one social class, gender, ethnicity, etc.
  - But it also means sampling more than once from the same speaker's idiolect, *simpliciter*
- These authors conclude that structured datasets can wreck statistical inference, leading to spurious results
- In the present context, we can understand the variation in the data as composed of two parts:
  - that which is generated by the model
  - that which is generated by unmodeled structure

Simulating structure

- The "deviance" of a model measures how much sample variance a model fails to capture
- We can quantify the structure in Kallel's data set with a simulation experiment:
  1. Start by generating a dataset from a model (either linear or quadratic)
  2. Add some noise to this data
  3. Fit the model to the noisy data, measure the deviance
  4. The amount of noise that most closely approximates the deviance of the model fit to real data will be a quantitive measure of the amount of structure in said real data

Simulating structure

- Here are the results of the simulation experiment, with 1000 simulated datasets per σ:
- The deviances of the data-fitted models are 724 for the linear model, and 698 for the quadratic.
- These are most closely approximated by 0.55 and 0.1 respectively.

σ	Linear	Quadratic
0.00	699.71	698.88
0.05	696.01	695.29
0.10	698.50	697.58
0.15	701.59	700.58
0.20	703.40	702.20
0.25	704.37	703.73
0.30	707.05	706.38
0.35	706.73	705.30
0.40	712.39	711.92
0.45	718.00	716.62
0.50	719.71	718.52
0.55	723.15	722.02
0.60	727.21	725.95
0.65	734.34	733.25
0.70	740.18	739.58

Simulating structure

- What these results suggest it that the quadratic model has almost completely soaked up the structural variance in the model
- We know from the sociolinguistic studies that structured variation exists in similar corpora
- So, Kallel's quadratic term is modeling real variation, but in the wrong way
  - It is not that years-squared are important in themselves
  - Rather, it is necessary to look for the true causes of structured variation in the data

Crossvalidation

- Another paradigm for comparing the linear and quadratic models is to ask the question "How well do the models predict usage?"
- One way to answer this question: collect new data, compare to model predictions
  - ...maybe there isn't any more data
  - ...maybe data collection methodologies differ
- Easier answer: split the data set in two
  - **Training**: fit the model to one subset of the data
  - **Test**: evaluate the model's predictions on the other set

Crossvalidation

- Over 100 iterations, with a test set of size 10% of the data, the linear model makes an average of 7.1 errors, whereas the quadratic model makes on average 6.5.
- By comparison, a straight line (not S-curve) fit to the data makes 11 errors; guessing a random probability in $(0, 1)$ results in a mean of 23 errors. The total number of tokens in the training corpus was 93
- Given the other problems with the quadratic model, this is not impressive enough a result to justify the model's adoption.

Nonlinearities in corpora

- The regression procedure used by Kallel (2007) was successful in finding a non-linearity in the corpus
- The analysis of this fact must take into account the understanding of competition in a population that underlies the statistical technique.
- The non-linearity might be due to:
  - Dialect variation (Ingham 2006)
  - Properties of the corpus used? ...

The quadratic in other data sets

- Using Ellegård's *do*-support corpus and the parsed corproa of Early Modern English (PPCEME, PCEEC), it is possible to calculate a quadratic regression. Those results are shown here:

Ellegård (sz. corr):	
Intercept	-2.114
	$p : 0.000$
Sent. type: Aff. Q/Neg. Decl	0.683
	$p : 0.000$
Sent. type: Neg. Q/Neg. Decl	1.900
	$p : 0.000$
Time	56.404
	$p : 0.000$
Time^2	-8.520
	$p : 0.029$

The quadratic in other data sets

- Using Ellegård's *do*-support corpus and the parsed corproa of Early Modern English (PPCEME, PCEEC), it is possible to calculate a quadratic regression. Those results are shown here:

Parsed corpora:	
Intercept	-2.013
	$p : 0.000$
Sent. type: Aff. Q/Neg. Decl	0.667
	$p : 0.000$
Sent. type: Neg. Q/Neg. Decl	2.104
	$p : 0.000$
Time	60.437
	$p : 0.000$
Time^2	-7.842
	$p : 0.117$

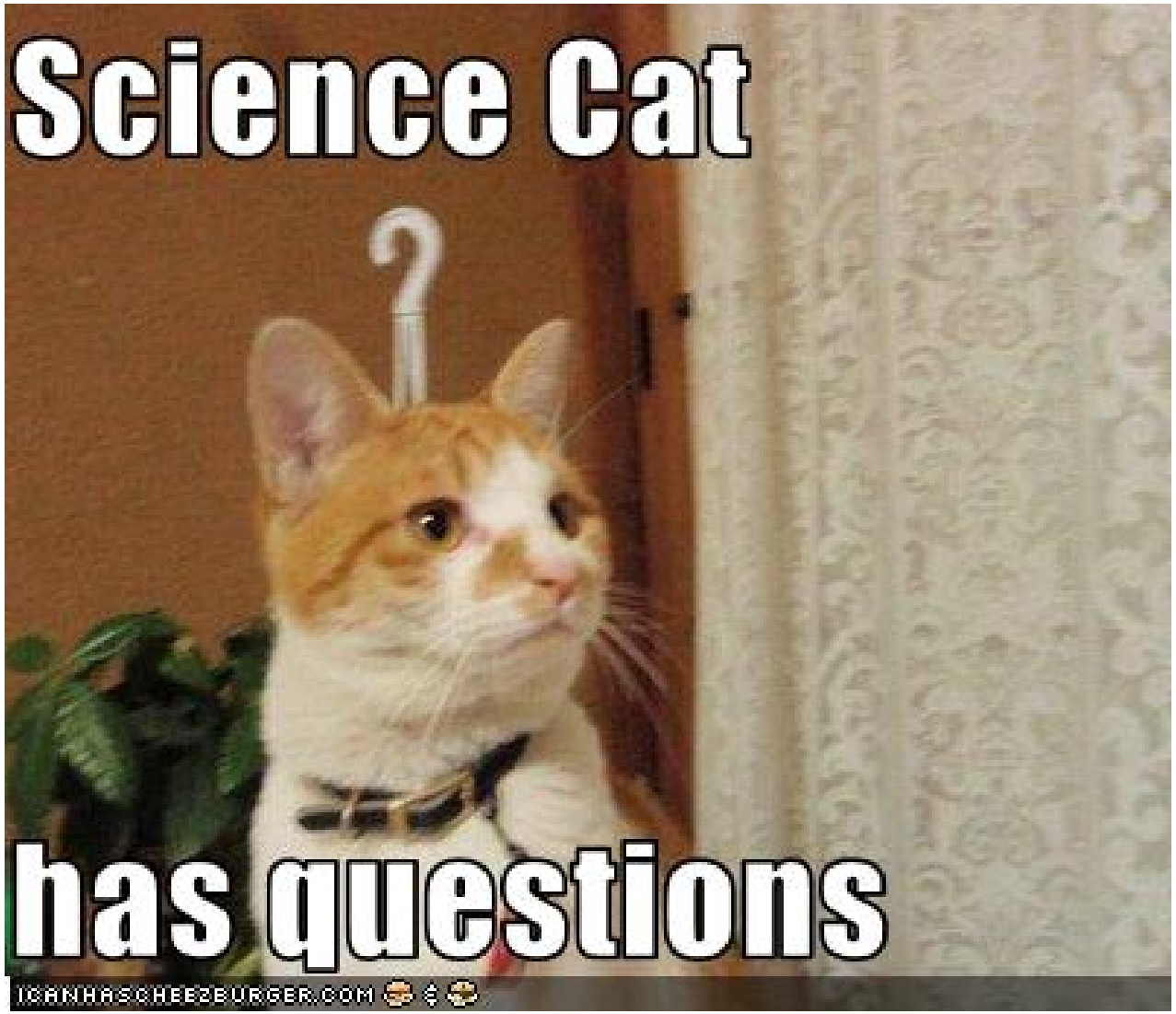
Conclusion

- There are several messages to be distilled from this exploration:
  - An understanding of the logistic function as a model of the behavior of a population is essential to historical syntax
  - Using general-purpose corpora might help minimize the impact of non-independence on statistical procedures
  - It is possible to compare models based on their predictive power, not (only) their abstract statistical properties
- The more data we get, the more likely we are to find non-linearities
  - These are best understood not as a failure of the competition model of change, but rather an invitation to study further the ways in which linguistic and biological change differ

Thanks

- Thanks are due to the following individuals and organizations:
  - Tony Kroch
  - The authors of the parsed corpora used in this study
  - Kyle Gorman, for digitizing Kallel's corpus
  - Anthony Warner, for digitizing Ellegård's corpus
  - Mark Liberman, for helpful discussion
- Remaining lacunae are, naturally, attributable to the author alone

Questions?



Bibliography I

- Bailey, Charles James (1973). *Variation and Linguistic Theory*. Washington: Center for Applied Linguistics.
- Frisch, S. (1997). The change in negation in Middle English: a NEGP licensing account. *Lingua* **101**.(1-2), 21–64. DOI: 10.1016/S0024-3841(96)00018-6.
- Gorman, Kyle (2009). The Consequences of Multicollinearity among Socioeconomic Predictors of Negative Concord in Philadelphia. *Penn Working Papers in Linguistics* **16**, 66–75.
- Ingham, Richard (2006). On two negative concord dialects in early English. *Language Variation and Change* **18**.(3), 241–266.
- Johnson, D.E. (2009). Getting off the GoldVarb Standard: Introducing Rbrul for Mixed-Effects Variable Rule Analysis. *Language and linguistics compass* **3**.(1), 359–383. ISSN: 1749-818X.

Bibliography II

- Kallel, Amel (2005). "The Lexical Reanalysis of N-words and the Loss of Negative Concord in Standard English". PhD thesis. Reading University.
- Kallel, Amel (2007). The loss of negative concord in Standard English: Internal factors. *Language Variation and Change* **19**.(1), 27–49. ISSN: 0954-3945.
- Kroch, Anthony (1989). Reflexes of grammar in patterns of language change. *Language variation and change* **1**.(3), 199–244. ISSN: 0954-3945.
- Kroch, Anthony, Beatrice Santorini, and Lauren Delfs (2005). *Penn-Helsinki parsed corpus of Early Modern English*. University of Pennsylvania. http://www.ling.upenn.edu/hist-corpora/PPCEME-RELEASE-1/.

Bibliography III

- Lotka, A.J. (1925). *Elements of physical biology*. Williams & Wilkins company.
- Santorini, Beatrice (1993). The rate of phrase structure change in the history of Yiddish. *Language Variation and Change* **5**.(03), 257–283.
- Taylor, Ann et al. (2006). *Parsed Corpus of Early English Correspondence, parsed version.* Compiled by the CEEC Project Team. York: University of York and Helsinki: University of Helsinki. Distributed through the Oxford Text Archive. http://www-users.york.ac.uk/~lang22/PCEEC-manual/index.htm.
- Volterra, V. (1926). *Variazioni e fluttuazioni del numero d'individui in specie animali conviventi*. Accademia Nazionale dei Lincei.

Bibliography IV

- Weinreich, Uriel, William Labov, and Marvin Herzog (1968). *Empirical foundations for a theory of language change*. University of Texas Press.