Language Change or Evolution?

Matilde Marcolli

MAT1509HS: Mathematical and Computational Linguistics

University of Toronto, Winter 2019, T 4-6 and W 4, BA6180

Language Change or Language Evolution

- languages and genes both transmitted to next generations with possibility or changes in replication
- two general mechanisms: drift versus selection
- is language change due to drift, selection, or a combination of the two?
- in biological evolution by natural selection some fitness function
- or a MaxEnt view of evolution
- drift is modelled by stochastic processes

Some recent work on drift vs selection in language change

- Mitchell G.Newberry, Christopher A. Ahern, Robin Clark, Joshua B.Plotkin, *Detecting evolutionary forces in language change*, Nature, Vol.551 (2017) 9 November, 223–226.
- Christopher A. Ahern, Mitchell G.Newberry, Robin Clark, Joshua B.Plotkin, *Evolutionary forces in language change*, arXiv:1608.00938.

Inference based on large corpora of text (English language between 12th and 21st century)
Examples looked at:

1. regularization of past-tense verbs (spilt $\rightarrow$ spilled)
2. the rise of the periphrastic 'do' (ate not $\rightarrow$ did not eat)
3. syntactic variation in verbal negation (Old English "Ic ne secge" $\rightarrow$ Middle English "I ne seye not" $\rightarrow$ Early Modern English "I say not")

- language change as competition between linguistic forms (sounds, morphemes, syntactic structures)
- selection forces: language internal, cognitive, social
- fitness towards learnability or efficient communication (?)
- to detect selection compare to stochasticity (drift) and see if significant deviation from a background stochastic process
- stochastic drift is *null-hypothesis* in population genetics; adopt as null-hypothesis also for language change

How to model stochastic drift?

- Wright–Fisher diffusion
  - James F.Crow, Motoo Kimura, *An introduction to population genetics theory*, Harper & Row, 1970
  - Florencia Reali, Thomas L.Griffiths, *Words as alleles: connecting language evolution with bayesian learners to models of genetic drift*, Proceedings of the Royal Society of London B: Biological Sciences, 277(2010) N.1680, 429–436.

• discrete time model for a population with constant size N and two types

• X_n number of type 1 individuals at time n

• at generation $n+1$ binomial sampling with probability $p = X_n/N$ (current empirical probability of type 1): each individual of the $n+1$ generation pick their parent randomly from individuals of generation n

• Markov model for X_n with state space $\{0, \dots, N\}$ and transition probabilities

$$\mathbb{P}(X_{n+1} = j \,|\, X_n = i) = \binom{N}{j} \left(\frac{i}{N}\right)^j \left(1 - \frac{i}{N}\right)^{N-j}$$

• variant for K different types in the population instead of two types (α_i, β_i how many individuals of each type)

$$\mathbb{P}(X_{n+1} = (\beta_1, \dots, \beta_K) \,|\, X_n = (\alpha_1, \dots, \alpha_K)) = \frac{N!}{\beta_1! \cdots \beta_K!} \left(\frac{\alpha_1}{N}\right)^{\beta_1} \cdots \left(\frac{\alpha_K}{N}\right)^{\beta_K}$$

• fixation phenomenon:
elimination of all but one types in finite time

Fixation in Wright–Fisher models

- measures δ_j on type j are absorbing states (once entered cannot be left)
- two types heterozygosity $H_n = 2\frac{X_n}{N}(1 - \frac{X_n}{N})$

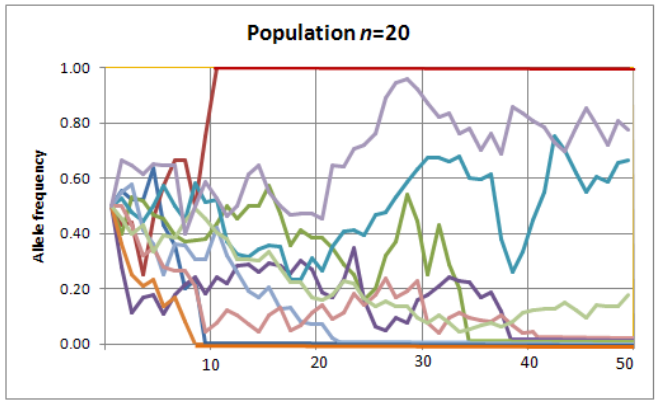
$$\mathbb{E}(H_{n+1}) = \frac{2}{N^2}\mathbb{E}(X_{n+1}(N - X_{n+1})) = \frac{2}{N^2}(N\mathbb{E}(X_{n+1}) - \mathbb{E}(X_{n+1}^2))$$

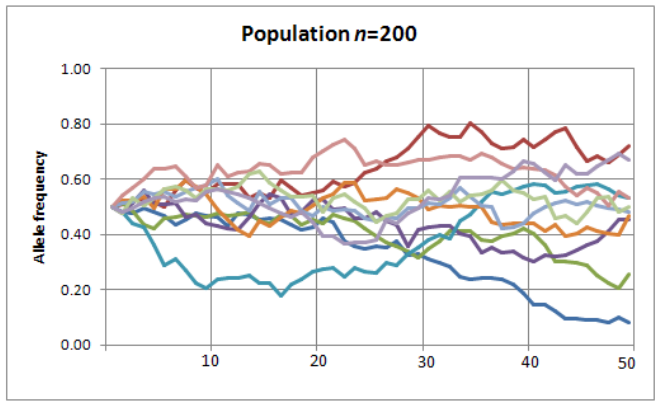
$$= \frac{2}{N^2}(N\mathbb{E}(X_{n+1}) - \mathrm{Var}(X_{n+1}) - \mathbb{E}(X_{n+1})^2) =$$

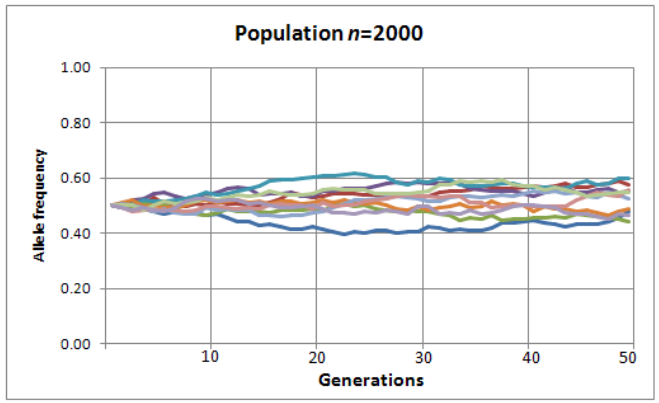
$$\frac{2}{N^2}(NX_n - X_n + \frac{X_n^2}{N} - X_n^2) = H_n(1 - \frac{1}{N})$$

- this gives inductively $\mathbb{E}(H_n) = H_0(1 - 1/N)^n \sim H_0 e^{-n/N}$
- $X_n \to X_\infty$ either 0 or 1 so X_n eventually constant 0 or 1
- can also compute approximate time to fixation $\tau(p) = -2N(p\log(p) + (1-p)\log(1-p))$ for $X_0 = pN$

Fixation and population size (ten simulations, $p = 1/2$, $n \leq 50$)





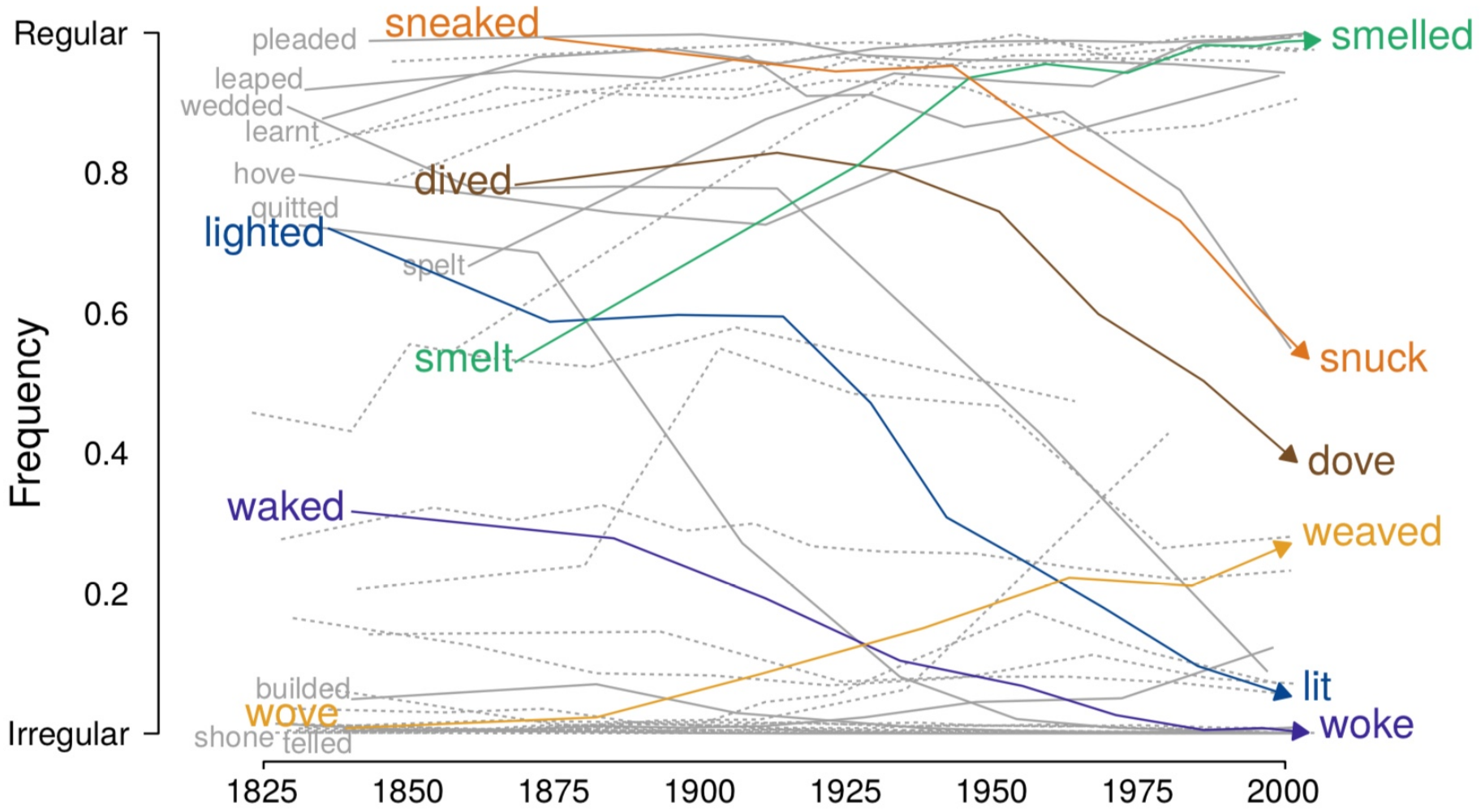


Language Change and Drift model

- How to test the drift null-hypothesis?
  - population size N unknown so to infer selection need observed linguistic changes inconsistent with neutral drift for any arbitrary N
  - statistical test: Frequency Increment Test (compare frequency changes observed between sampled time points to what expected under drift)
  - Alison F.Feder, Sergey Kryazhimskiy, Joshua B.Plotkin, *Identifying signatures of selection in genetic time series*, Genetics, 196(2014) N.2, 509–522.

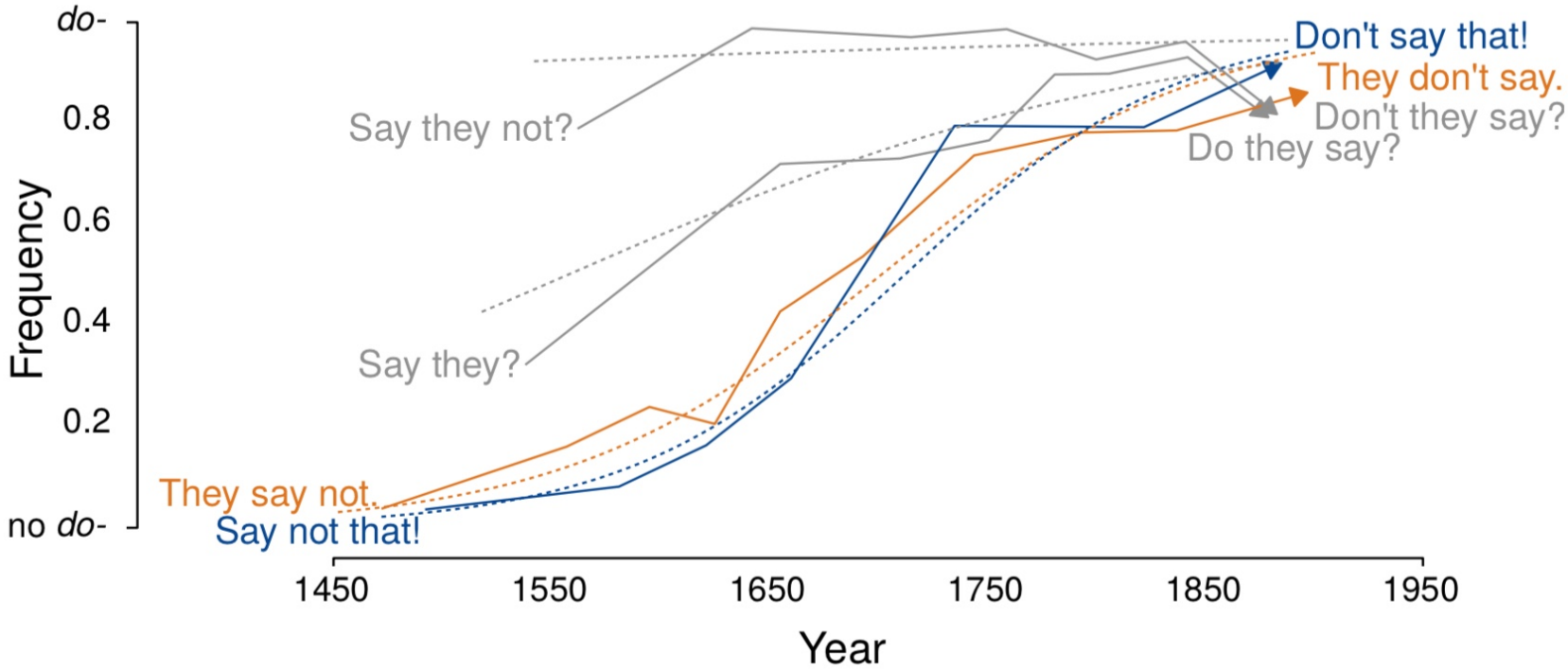
Case N.1: past-tense verb conjugation

- naive expectation: irregular past-tense forms regularize over time
- this truly happens for rare words over long timescales
- but commonly used verbs remain (or become!) irregular
- Data: all past tense verb tokens from Corpus of Historical American English (period 1810–2009)
- retain those with two variants each occurring at least 50 times in the corpus (total 704,081 tokens of regular/irregular past-tense variants of 36 verbs)
- Result: based on Frequency Increment Test some verbs experienced selection (significant deviation from drift) either towards regularization or towards irregularization, others (dotted lines) fit well the drift hypothesis


Regular
Irregular
Frequency
0.8
0.6
0.4
0.2
pleaded
sneaked
smelled
leaped
wedded
learnt
hove
dived
quitted
lighted
spelt
smelt
snuck
dove
waked
weaved
builded
wove
lit
shone
felled
woke
1825
1850
1875
1900
1925
1950
1975
2000

Case N.2: periphrastic 'do' in English

- Data fromYork-Helsinki Parsed Corpus of Early English Correspondence (1400-1700), Penn-Helsinki Parsed Corpus of Early Modern English (1500-1700), Penn Parsed Corpus of Modern British English (1710-1910)
- total 16,072 tokens in affirmative questions, negative questions, negative declaratives, and negative imperatives
- Examples: you asked not $\rightarrow$ you did not ask; asked you a question? $\rightarrow$ did you ask a question?;
- Results: rise of periphrastic 'do' more rapid in negative declarative and imperative statements where significant deviation from drift hypothesis; in interrogative statements explainable by drift
- Proposed explanation: periphrastic 'do' rose first by random drift in interrogative statements then driven by selection in negative declarative and imperative statements (selection for grammatical consistency with interrogative form)


do-
0.8
0.6
0.4
0.2
no do-
Frequency
Don't say that!
They don't say.
Don't they say?
Do they say?
Say they not?
Say they?
They say not.
Say not that!
1450
1550
1650
1750
1850
1950
Year

Case N.3: sentence negation from 12th to 16th century

- Data: Penn Parsed Corpus of Middle English, 5,475 negative declaratives
- pre-verbal negation ("Ic ne secge") $\rightarrow$ bipartite negation ("I ne seye not") $\rightarrow$ post-verbal negation ("I say not")
- Result: both transitions not explainable by drift (selection involved); linguistic hypothesis, favoring of more emphatic forms of negation