Morgan Sonderegger
morgan@cs.uchicago.edu

LSA 2009 Annual Meeting
January 8, 2009

Rhyme graphs, sound change, and perceptual similarity

1 Introduction

- Rhyming data in phonology:

  Synchronic:
  - Half-rhymes tend to be between "similar" segments in lyrical forms:[1]
    - Poetry: Jakobson (1960); Holtman (1996); Hanson (2002, 2003); Steriade (2003), ...
    - Lyrics: Zwicky (1976); Holtman (1996); Kawahara (2007); Katz (2008)
  - Ties into recent interest in similarity in phonology (e.g. Steriade, 2001; Frisch et al., 2004).
  - But: Statistical picture incomplete – only focusing on half-rhymes which *do* occur, problem of relative frequency of full rhymes; also, individual rhymes could have social rather than phonological motivation.[2]

  Diachronic:
  - Rhyming data often crucial to reconstruction: little normative judgment compared to grammarians, less individualized than naive spellings (Wyld, 1923).
  - But: Don't know a priori *which* rhymes are full – potentially v. problematic.
- Issue in both cases ≈ that sequences of (≥ 2) rhymes unreliable in isolation.
- Current solution: Make generalizations based on ensemble rather than individual rhymes (synchronic), compare non-rhyme data (diachronic).
- Potential additional tool: The higher-level structure of rhyming corpus – *rhyme graph*.[3]
- Claim, via two examples: Rhyme graph structure reflects phonological change, similarity.

2 Preliminaries

Rhyme corpus

- 52k rhymes from English poetry (App. A).
- Poets born ≈ 1555-1890. All British, mostly English, chosen by birth date, online availability.
- 6 subcorpora: 1600, 1650, 1700, 1750, 1800, 1900
- 1900 is reference subcorpus where we assume know pronunciation, ≈ CELEX (Baayen et al., 1993).
- Rhymes coded by hand semi-automatically: stricter rhyme schemes ⇒ faster coding.
- Conservative definition of rhyme domain: primary-stressed nucleus and all following segments of *last word* of line.
- Approximation for ease of analysis: misses rhymes such as *see her*:*be her*.

Rhyme graphs

- An (undirected, weighted) *graph* $G = (V, E, W)$ consists of vertices V, edges E between vertices, and positive weights W for edges, written $e = (v_1, v_2)$, $w(v_1, v_2) \geq 0$.[4]

[1] More comprehensive list in Kawahara (2007)

[2] For example, there is a small class of frequently-used half-rhymes in English verse (e.g. between *-ove* words) based on spelling and tradition rather than phonological similarity (Wyld, 1923). A conclusion about the perceptual similarity of FOOT and STRUT class words based only on rhyming pairs would thus overestimate.

[3] To my knowledge, Joyce (1977, 1979) are the first and only studies on rhyme graphs. Joyce uses directed graphs (rather than undirected, used here), and gives several strong components from the graph for *Pearl*, a Middle English poem.

[4] If weights are not restricted to be positive, some basic notions (such as "shortest path") are no longer well-defined.

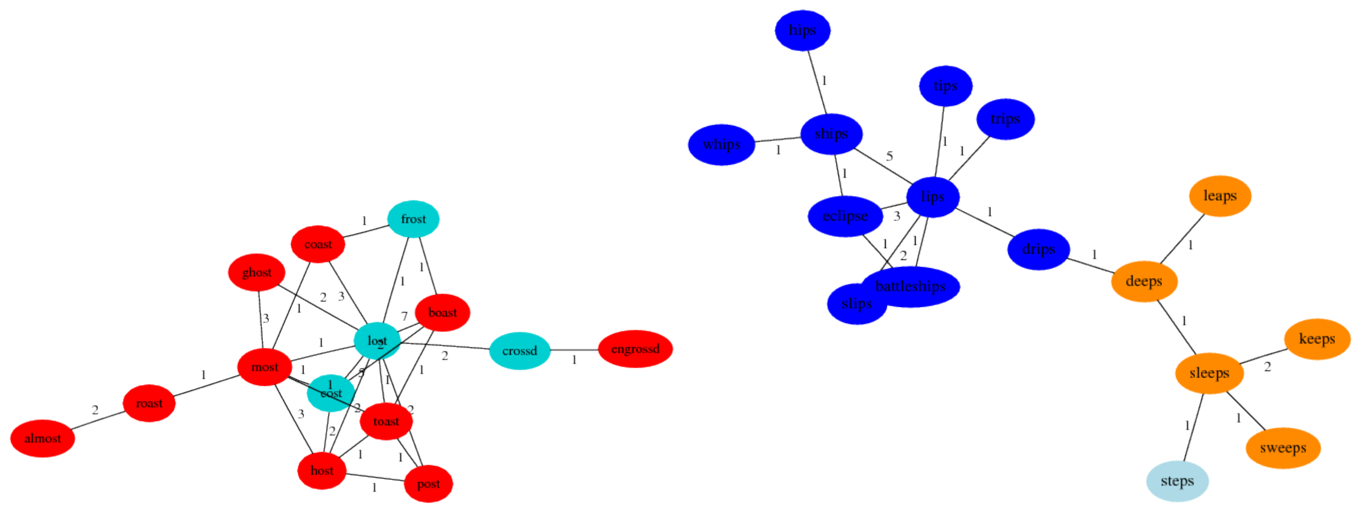

Figure 1: Sample components of rhyme graph from 1700 subcorpus (L), 1900 subcorpus (R). Colors indicate current pronunciations (from CELEX) of final syllable vowels. a necessary but not sufficient condition for full rhyme. Edge labels are number of observed rhymes.

- The *connected component* ("component") of $v \in V$ is the maximal subgraph of G containing v. Intuitively, "piece" of graph.
- The connected components partition the graph.
- For a rhyme corpus, V=words, $(v_1, v_2) \in E \Leftrightarrow$ rhyme between v_1 and v_2 occurs, $w(v_1, v_2) = \#$ of rhymes.
- Fig. 1 shows sample rhyme graph connected components.
- Rhyme data parsed using Perl scripts, $\rightarrow$ diagrams using GraphViz,[5] an open source graph visualization suite.

3 Plotting the *meet*/*meat* merger

Summary

- Present-day English (PDE) vowel /iː/ is (roughly) the result of a merger between Late Middle English (LME) /eː/ and /iː/, the "*meet*/*meat* merger."
- Write LME vowels as $^*e$, $^*i$, class of current /iː/ words usually follows from spelling:
  - $^*e$: *meat*, *complete*, *receive*
  - $^*i$: *meet*, *believe*, *me*
- Question: What was the trajectory of this merger?
- Know $^*e$, $^*i$ largely distinct in 1600 (Lass, 1992), merged today.
- Strategy: Use 1600, 1900 graphs as endpoints, compare relative likelihood of hypotheses that a given poet's rhymes consistent with each graph, using Bayes factors.

Bayesian inference, Bayes factors

- *Bayesian inference*: Statistical approach in which knowledge about $\boldsymbol{\theta}$, the variable of interest, is probabilistic.

[5]`http://www.research.att.com/sw/tools/graphviz/`

- Combine prior knowledge about $\boldsymbol{\theta}$ with observations of data $\boldsymbol{x}$ whose probability distribution is dependent on $\boldsymbol{\theta}$.
- Formally, from the *prior distribution* $P(\boldsymbol{\theta})$, data $\boldsymbol{x}$, and *conditional distribution* $P(\boldsymbol{x} \mid \boldsymbol{\theta})$, the *posterior distribution* is given by Bayes' rule:

$$P(\boldsymbol{\theta} \mid \boldsymbol{x}) = \frac{P(\boldsymbol{x} \mid \boldsymbol{\theta})P(\boldsymbol{\theta})}{P(\boldsymbol{x})} = \frac{P(\boldsymbol{x} \mid \boldsymbol{\theta})P(\boldsymbol{\theta})}{\int P(\boldsymbol{x} \mid \boldsymbol{\theta}')P(\boldsymbol{\theta}')\, d\boldsymbol{\theta}'} \tag{1}$$

- Now suppose there are two hypotheses, H_1 & H_2, which define priors $P(\boldsymbol{\theta} \mid H_2)$, $P(\theta \mid H_2)$. The data's *Bayes factor* is its relative probability under the hypotheses:

$$BF(\boldsymbol{x}; H_1, H_2) = \frac{P(\boldsymbol{x} \mid H_1)}{P(\boldsymbol{x} \mid H_2)} = \frac{\int P(\boldsymbol{x} \mid \boldsymbol{\theta})P(\boldsymbol{\theta} \mid H_1)\, d\boldsymbol{\theta}}{\int P(\boldsymbol{x} \mid \boldsymbol{\theta})P(\boldsymbol{\theta} \mid H_2)\, d\boldsymbol{\theta}} \tag{2}$$

Methodology

- Compute Bayes factor for each of 19 test poets (highest # rhymes), born 1536–c.1890 (see App. A).
- For poet i, consider only the subgraph consisting of /iː/words, and define
  - Data: $\boldsymbol{x}_i = (x_{i,1}, \ldots, x_{i,6})$, where $x_{i,j} =$# observed type j rhymes for poet i.
  - Parameters: $\boldsymbol{\theta}_i = (\theta_{i,1}, \ldots, \theta_{i,6})$, where $\theta_{i,j}$ = underlying fraction type j rhymes for poet i.

  and
  - $j = 1$: Same nucleus ($^*e$), same coda.
  - $j = 2$: Same nucleus ($^*i$), same coda.
  - $j = 3$: $^*e$ and $^*i$ nuclei, same coda.
  - $j = 4$: Same nucleus ($^*e$), different codas.
  - $j = 5$: Same nucleus ($^*i$), different codas.
  - $j = 6$:$^*e$ and $^*i$ nuclei, different codas.
- Given θ_i, $\boldsymbol{x}_i$ is multinomially distributed: $\boldsymbol{x}_i \sim \mathcal{M}(\theta_i, n_i)$, where $n_i = \sum_{i=1}^{6} x_{i,j}$, i.e.

$$P(\boldsymbol{x}_i \mid \boldsymbol{\theta}_i) = \binom{n_i}{x_{i,1}, \ldots, x_{i,6}} \prod_{j=1}^{6} \theta_{i,j}^{x_{i,j}}, \qquad \theta_{i,j} \in [0,1],\ \sum_{j=1}^{6} \theta_{i,j} = 1 \tag{3}$$

Priors

- The corresponding data vectors for the 1600 and 1900 graph are $\boldsymbol{x}_{1600} = (148, 691, 8, 2, 3, 1)$, $\boldsymbol{x}_{1900} = (132, 503, 320, 4, 1, 4)$.
- V. few $^*e$:$^*i$ rhymes in 1600, as expected.
- Must choose how to translate $\boldsymbol{x}_{1600}$, $\boldsymbol{x}_{1900}$ data into priors on $\boldsymbol{\theta}$. Use *Dirichlet priors*, a common choice for multinomial data (Gelman et al., 2004):

$$P(\theta \mid 1600) = \mathcal{D}(\theta; \boldsymbol{x}_{1600} \cdot r), \quad P(\theta \mid 1900) = \mathcal{D}(\theta; \boldsymbol{x}_{1900} \cdot r) \tag{4}$$

  where $r > 0$ is a scaling factor which changes the importance of the prior relative to the data.[6]

[6] Alternatively, r changes the variance of each $\boldsymbol{\theta}_i$: smaller $r \Rightarrow$ larger variance. r is included for *sensitivity analysis*, a check of how its value influences the result. Sensitivity analysis (not shown here) showed little change in the results for all $r > 0.1$.

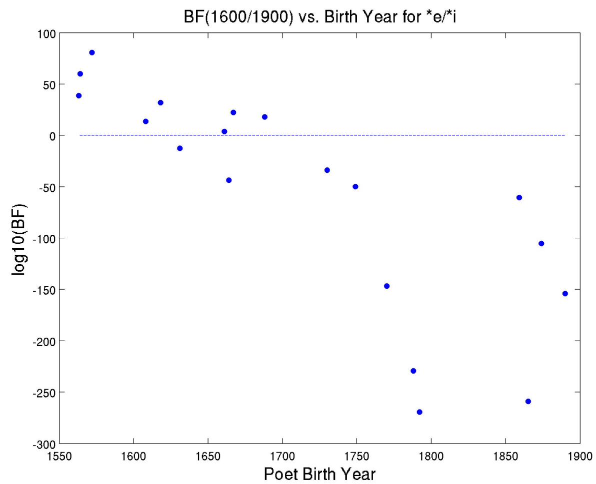

Figure 2: Bayes factors (1600 vs. 1900 reference graphs) for /iː/subgraphs of test poets. See text.

Calculating Bayes factors

- For poet i, the Bayes factor for the hypotheses H_{1600} and H_{1900} is

$$\begin{aligned} BF(\boldsymbol{x}_i, r) &= \frac{p(\boldsymbol{x}_i \mid 1600, r)}{p(\boldsymbol{x}_i \mid 1900, r)} \\ &= \frac{\int p(\boldsymbol{x}_i \mid \boldsymbol{\theta}_i, r) p(\boldsymbol{\theta}_i \mid 1600, r) d\boldsymbol{\theta}_i}{\int p(\boldsymbol{x}_i \mid \boldsymbol{\theta}_i, r) p(\boldsymbol{\theta}_i \mid 1900, r) d\boldsymbol{\theta}_i} \end{aligned} \tag{5}$$

- Eqn. 5 is not analytically solvable (when the expressions from Eqns. 3, 4 are substituted in). The Bayes factor for each poet was thus evaluated via Monte Carlo simulation ($N = 10000$).

Results

- Fig. 2 shows Bayes factors for poets as a function of birth year.
- Note y-axis on log scale.
- $BF < 1$ after $\approx$ birth year 1700: Can interpret as $^*e/^*i$ distinction disappearing in 1725 poetry.
- Agrees with Lass (1992): "The change-over begins in the late 17th century," and by $\approx$ 1750 distinct $^*e/^*i$ rare and seen as rustic.
- Biggest outlier is Matthew Prior (at $\approx (1660, -45)$), vowels merged unusually early. May be because Prior's family was "lower middle class,"[7] often the leading social group in linguistic change (Labov, 2000).

4 Graph structure and perceptual similarity

Component goodness

- Given a rhyme graph G, components $C_1, \ldots, C_n$, need a measure of component "goodness." Want:
  - *Best*: All vertices have same rhyme stem.

[7]This is of course anecdotal and an anachronistic use of class labels, but agrees with Labov's (1994, Ch. 10) speculations on the contemporaneous *meat*/*mate* (near)-merger. Prior came from a family of joiners and tavern owners. He went to the elite Westminster school, but was not a boarder.

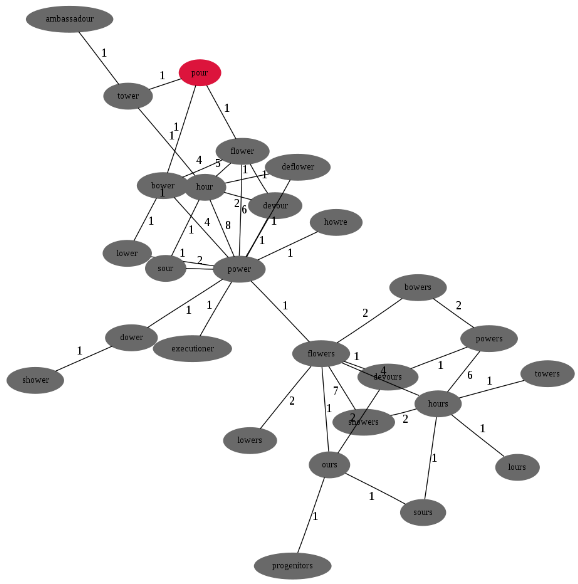

Figure 3: Component from 1600 subcorpus containing a good partition corresponding to a half-rhyme.

  - *Worst*: All vertices have different rhyme stems.
  - *Nonlinear measure*: Two distinct rhyme stems much worse than 1, slightly better than 3...
- Simple measure: C has k distinct rhyme stems, define *component goodness* as $\gamma(C) = 1/k$.
- Note "similarity" measure between rhyme stems binary: no segmental information.

Graph structure

- How does a component's structure reflect its goodness? Some qualitative hypotheses from observed components.
  - *Connectedness*: Components which are "more connected" are less likely to contain half-rhymes. ("If a rhyme can occur, eventually it will.")
  - *Existence of good partition*: A component consisting of two well-connected groups of vertices, with few connections between the groups, likely indicates a half-rhyme (Fig. 3).
  - *Size*: Bigger components are more likely to contain half-rhymes.
- Want quantitative measures corresponding to these quantities for a given component (with N = # obs. rhymes): *graph structure variables* (GSVs)
  - *Connectedness GSVs*:
    1. NORMALIZED MAXIMUM CLIQUE SIZE: Same, but divided by $|V|$ (the maximum possible clique size).
    2. EDGE RATIO: Ratio of observed to possible edges.
    3. DIAMETER: Maximum shortest-path distance between words, where the length of a path is the number of edges traversed. (No information used about the *number* of rhymes associated with each edge.)
  - *Partition GSV*:
    1. EIGENVALUE GAP: A measure of how easily V can be cut in two. Between 0 and 1, higher = more cuttable.[8]

[8]Formally, $1 - \lambda_2$, where λ_2 is the second largest eigenvalue of the Laplacian of G.

– *Size GSVs*:
  1. LOG(SIZE): $\log(|V|)$
  2. LOG(# RHYMES): $\log(N)$
– *Other GSVs*:
  1. RHYMES PER VERTEX: $N/|V|$.
  2. MAXIMALLY-CONNECTED VERTEX: Maximum of degree$(v)/|E|$.

Results

- Calculate correlation coefficient r, significance $(p(r))$ of component goodness with each variable.
- Used 1900 subgraph, rhyme stem pronunciations from CELEX, only components with ≥ 4 words included (135 components).[9]
- Results: [10]

Variable	r	$p(r)$
`max_clq_nzd`	0.67	0
`edge_rat`	0.67	0
`diameter`	-0.56	0
`log(size)`	-0.69	0
`log(rhymes)`	-0.65	0
`ev_gap`	-0.48	$< 10^{-3}$
`maxdegs_nzd`	0.55	0
`rhy_per_vert`	-0.41	$< 10^{-2}$

- GSV/goodness example plots in Fig. 4.
- All correlations highly significant, covariance high for GSVs with same motivation (not shown).
- Connectedness, size best predictors of component goodness
- Eigenvalue gap r good considering "existence of good partition" $\approx$ binary.
- Component graph structure reflects similarity of rhyme stems.

5 Discussion

Summary

- Rhyme graphs give linguistic evidence at different level from individual rhymes: ensemble. Made up of individual rhymes, but has distinct higher-level structure.
- RG also useful for visualization of corpus.
- Have shown (for this corpus) rhyme graph structure reflects:
  – *Change*: Similarity to pre- vs. post-merger rhymes $\rightarrow$ merger trajectory.
  – *Similarity*: Component structure $\leftrightarrow$ (binary, rhyme-stem) similarity of included words.

Potential applications, future work

- Reconstruction for languages with less data, non-alphabetic/syllabic: much more important than for English to know which rhymes reliable.
- Graph spectrum $\rightarrow$ can say *what* good partition is, reconstruct pre-merger classes.
- (Synchronic) phonological significance of components, vs. individual rhymes?

[9]The lowest cutoff such that all graph variables used make sense. A component of two words has no higher-level structure, and partitioning into two sub-graphs only makes sense for a component of ≥ 4 words.
[10]$p(r) = 0$ means $< 10^{-4}$.

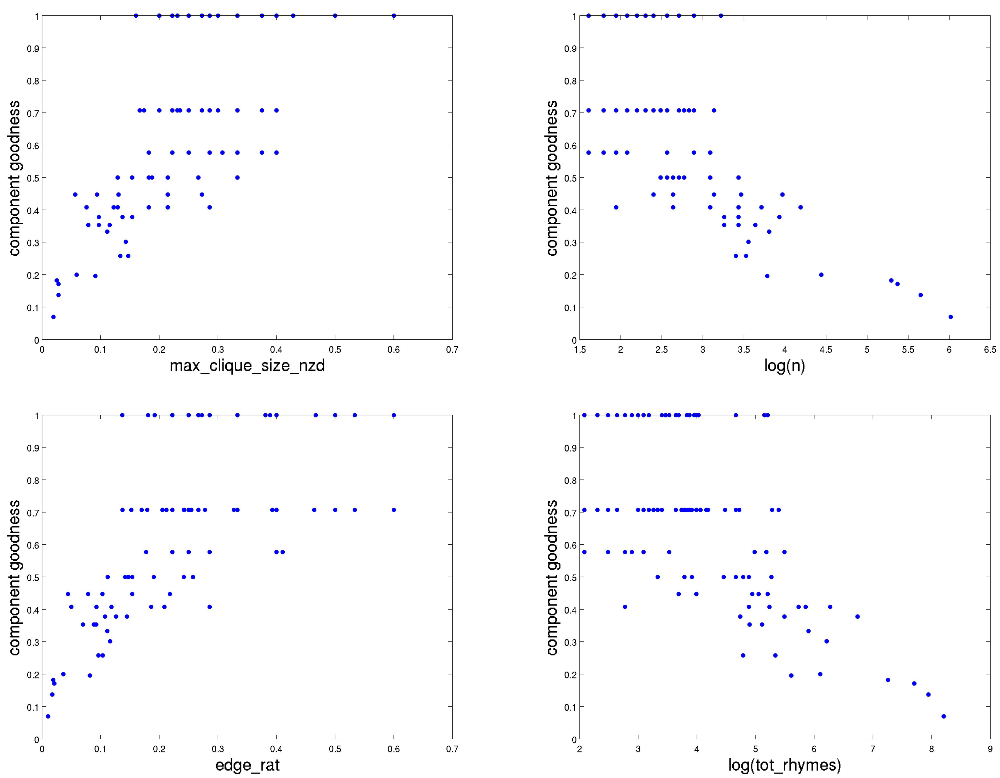

Figure 4: Sample GSV/goodness plots (see text).

- Future work:
  - Binary similarity → phonetic distance, based on natural classes (Frisch et al., 2004) or segments (Levenshtein).
  - Better statistical models: individual components instead of whole graph, reconstruct mergers in different phonetic environments.
  - Harder mergers: *meat/meet* well-separated in 1600, well-merged now. Not as clean for *toe/tow* merger.
  - Splits, e.g. *foot/strut*.
  - Assess relative influence of spelling.

References

Baayen, R., Piepenbrock, R., and van Rijn, H. (1993). *The CELEX lexical database (CD-ROM)*. Linguistic Data Consortium, Philadelphia.

Frisch, S., Pierrehumbert, J., and Broe, M. (2004). Similarity Avoidance and the OCP. *Natural Language & Linguistic Theory*, 22(1):179–228.

Gelman, A., Carlin, J., Stern, H., and Rubin, D. (2004). *Bayesian Data Analysis*. Chapman & Hall/CRC, Boca Raton, 2nd edition.

Hanson, K. (2002). Vowel variation in English rhyme. In Minkova, D. and Stockwell, R., editors, *Studies in the History of the English Language: A Millennial Perspective*, pages 207–229. de Gruyter, New York.

Hanson, K. (2003). Formal variation in the rhymes of Robert Pinsky's the *Inferno* of Dante. *Language and Literature*, 12(4):309–337.

Holtman, A. (1996). *A generative theory of rhyme: An optimality approach*. PhD thesis, Universiteit Utrecht.

Jakobson, R. (1960). Linguistics and poetics. In Sebeok, T. A., editor, *Style in Language*, pages 350–377. MIT Press, Cambridge, MA.

Joyce, J. (1977). Networks of sound: Graph theory applied to studying rhymes. In *Computing in the Humanities: proceedings of the Third International Conference...*, pages 307–316. University of Waterloo Press, Waterloo.

Joyce, J. (1979). Re-weaving the word-web: Graph theory and rhymes. In *Proceedings of the 5th Annual Meeting of the Berkeley Linguistics Society*, pages 129–141. BLS, Berkeley.

Katz, J. (2008). Phonetic similarity in an English Hip-Hop Corpus. Handout, UMass/Amherst/MIT Meeting in Phonology.

Kawahara, S. (2007). Half rhymes in Japanese rap lyrics and knowledge of similarity. *Journal of East Asian Linguistics*, 16(2):113–144.

Labov, W. (1994). *Principles of linguistic change. Vol. 1: Internal factors*. Blackwell, Oxford.

Labov, W. (2000). *Principles of linguistic change. Vol. 2: Social factors*. Blackwell, Oxford.

Lass, R. (1992). Phonology and morphology. In Hogg, R., editor, *The Cambridge History of the English Language*, volume 3, pages 23–156. Cambridge University Press, Cambridge.

Steriade, D. (2001). The phonology of perceptibility effects: The P-map and its consequences for constraint organization. Ms., UCLA.

Steriade, D. (2003). Knowledge of similarity and narrow lexical override. In Nowak, P. M., Yoquelet, C., and Mortensen, D., editors, *Proceedings of the Berkeley Linguistics Society 29*, pages 583–598. BLS, Berkeley.

Wyld, H. (1923). *Studies in English rhymes from Surrey to Pope*. J. Murray, London.

Zwicky, A. (1976). Well, this rock and roll has got to stop. Junior's head is hard as a rock. In *Chicago Linguistic Society 12*, pages 676–697.

A Corpus information

The table lists poets used in the corpus. Notes:

- Under "Notes": 1xx0=part of 1xx0 subcorpus. T=used as test poet in diachronic analysis.
- "Georgian Poets" are contributors to the *Georgian Poetry* anthologies (1912–22).
- Poetry was obtained from various online sources, mostly Project Gutenberg (`gutenberg.org`), Internet Archive (`archive.org/details/texts`), and Luminarium (`luminarium.org`).

Significant chunks of public-domain rhyming poetry are harder to obtain after c. 1920 for two reasons. Artistically, rhyming verse became less popular with the rise of Modernism. Legally, work by twentieth-century poets is mostly still under copyright.

Poet	Alive	Rhymes (K)	Notes
Thomas Lodge	1558?–1625	0.49	1600
John Fletcher	1579–1625	0.56	1600
Henry Constable	1562–1613	0.94	1600
Samuel Daniel	1562–1619	0.56	1600
Michael Drayton	1563–1631	1.81	1600, T
William Shakespeare	1564–1616	3.41	1600, T
Ben Jonson	1572–1637	3.52	1600, T
Bartholomew Griffin	?–1602	0.44	1600
William Smith	fl. 1600	0.37	1600
John Milton	1608–1674	0.67	1650, T
Richard Lovelace	1618–1657	2.40	1650, T
John Dryden	1631–1700	2.13	1700, T
Anne Finch	1661–1720	1.90	1700, T
Matthew Prior	1664-1721	1.27	1700, T
Jonathan Swift	1667–1745	2.70	1700, T
Alexander Pope	1688-1744	2.29	1700, T
Oliver Goldsmith	1730?–1774	1.43	1750, T
Charlotte T. Smith	1749–1806	1.25	1750, T
William Wordsworth	1770–1850	2.23	1800, T
Lord Byron	1788–1824	3.99	1800, T
Percy Shelley	1792–1822	5.85	1800, T
A.E. Housman	1859–1936	1.52	1900, T
Rudyard Kipling	1865–1936	2.60	1900, T
Walter de la Mare	1873–1956	1.74	1900
G.K. Chesterton	1874–1936	1.29	1900, T
Edward Thomas	1878–1917	0.52	1900
Rupert Brooke	1887–1915	1.05	1900
Georgian Poets	c. 1890	3.07	1900, T