Available online at www.sciencedirect.com

ScienceDirect

Computer Speech and Language 25 (2011) 655–678

COMPUTER SPEECH AND LANGUAGE

www.elsevier.com/locate/csl

Applications of graph theory to an English rhyming corpus

Morgan Sonderegger*

University of Chicago, Department of Computer Science, 1100 East 58th Street, Chicago, IL 60637, USA

Received 16 October 2009; received in revised form 10 March 2010; accepted 7 May 2010
Available online 21 May 2010

Abstract

How much can we infer about the pronunciation of a language – past or present – by observing which words its speakers rhyme? This paper explores the connection between pronunciation and network structure in sets of rhymes. We consider the *rhyme graphs* corresponding to rhyming corpora, where nodes are words and edges are observed rhymes. We describe the graph **G** corresponding to a corpus of ∼ 12000 rhymes from English poetry written c. 1900, and find a close correspondence between graph structure and pronunciation: most connected components show community structure that reflects the distinction between full and half rhymes. We build classifiers for predicting which components correspond to full rhymes, using a set of spectral and non-spectral features. Feature selection gives a small number (1–5) of spectral features, with accuracy and *F*-measure of ∼90%, reflecting that positive components are essentially those without any good partition. We partition components of **G** via maximum modularity, giving a new graph, **G**′, in which the "quality" of components, by several measures, is much higher than in **G**. We discuss how rhyme graphs could be used for historical pronunciation reconstruction.

Keywords: Rhymes; Graph theory; Complex networks; Poetry; Phonology; English

1. Introduction

How can we reconstruct what English sounded like for Pope, Shakespeare, or Chaucer? Pronunciation reconstruction traditionally involves triangulation from several sources; one crucial type of data is rhyming verse (Wyld, 1923). Because rhymes are usually between words with the same endings (phonetically), we might infer that two words which rhyme in a text had identically pronounced endings for the text's author. Unfortunately, this reasoning breaks down because of the presence of "half" rhymes. Consider the following rhymes, from poetry written by William Shakespeare around 1600.[1]

(a) But kept cold distance, and did thence *remove*,
To spend her living in eternal *love*.

(b) And deny himself for *Jove*,
Turning mortal for thy *love*.

* Tel.: +1 773 702 9110; fax: +1 773 702 8487.
E-mail address: morgan@cs.uchicago.edu

[1] "A Lover's Complaint" (a,e), *Love's Labour Lost* IV.3 (b), "The Rape of Lucrece"(c), "Venus and Adonis" (d,f).

0885-2308/$ – see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.csl.2010.05.005

(c) But happy monarchs still are fear'd for *love*:
With foul offenders thou perforce must bear,
When they in thee the like offences *prove*:
If but for fear of this, thy will *remove*;

(d) And pay them at thy leisure, one by *one*.
What is ten hundred touches unto thee?
Are they not quickly told and quickly *gone*?

(e) Which fortified her visage from the *sun*,
Whereon the thought might think sometime it saw
The carcass of beauty spent and *done*:
Time had not scythed all that youth *begun*,

(f) What bare excuses makest thou to be *gone*!
I'll sigh celestial breath, whose gentle wind
Shall cool the heat of this descending *sun*:

We write $x{:}y$ when a rhyme is observed between x and y, and $x{\sim}y$ if x and y have the same ending (in a sense made more precise below). One intuitively knows, from experience with songs or poetry, that if $x{\sim}y$ then it is possible to rhyme x and y, and that usually, $x{:}y$ implies $x{\sim}y$. If we assume that $x{:}y \Rightarrow x{\sim}y$ always, then from Examples (a)–(c):

love∼Jove∼remove∼prove

and from Examples (d)–(f):

one∼gone∼sun∼done∼begun

However, it turns out that not all words in the first group of rhymes were pronounced the same for Shakespeare, while all words in the second group were.[2] Because of the uncertainty in the implication $x{:}y \Rightarrow x{\sim}y$, in pronunciation reconstruction rhyming data is only used together with other sources, such as grammar manuals and naive spellings (Wyld, 1923). But these sources are expensive and limited, while rhyming data is cheap and plentiful. If we could somehow make the implication stronger, rhyming data could stand on its own, making reconstruction significantly easier.

This paper attempts to strengthen the implication in two ways: first, by building classifiers to separate half (e.g. (a)–(c)) from full (e.g. (d)–(f)) groups of rhymes, based on the groups' *rhyme graphs*; second, by breaking groups of rhymes into smaller and more full groups, based on the structure of their rhyme graphs. Although the long-term goal of this project is to infer historical pronunciation, this paper uses recent poetry, where the pronunciation is known, to develop and evaluate methods. We first (Sections 2 and 3) introduce rhyme graphs, outline the corpus of poetry used here, and describe its rhyme graph, **G**. In Section 4, we build classifiers for components of **G**, using a set of features which reflect components' graph structure. We then (Section 5) partition components into smaller pieces, giving a new graph **G**′, and evaluate the quality of rhymes in **G**′ versus **G**.

2. Data

2.1. Rhyming corpora

Rhyming corpora have traditionally been used in two ways by linguists interested in phonology. In diachronic phonology, collections of rhymes are traditionally a key tool for pronunciation reconstruction (e.g. Kökeritz, 1953; Dobson, 1968; Wyld, 1936 for English); in this case the focus is on full rhymes, which indicate identity between (parts of) words. In synchronic phonology, rhyming corpora have been used for Japanese song lyrics (Kawahara, 2007), Romanian poetry (Steriade, 2003), English song lyrics (Zwicky, 1976; Katz, 2008), and English poetry (Holtman, 1996; Hanson, 2003; Minkova, 2003).[3] In these cases, the focus is on half rhymes (see below, Section 2.2), which reflect speakers' intuitions about phonological similarity.

[2] General sources on pronunciation around 1600 are (in order of accessibility) (Lass, 1992; Kökeritz, 1953; Dobson, 1968); contemporary phonetic transcriptions (e.g. Danielsson, 1955–1963; Danielsson and Gabrielson, 1972; Kauter, 1930) provide direct evidence.

[3] However, none of the English poetry corpora are electronically available.