Learning constraints for morphophonological classification

Colin Wilson (colin@cogsci.jhu.edu)
JHU

AMP
NYU

09-16-17

Chomsky & Halle (1968) on lexical classes:

> "In the phonology proper, we also find quite commonly that rules apply in a selection fashion and thus impose an idiosyncratic classification on the lexicon. Often there is a historical explanation for this idiosyncratic behavior, but this is obviously irrelevant as far as the linguistic competence of the native speaker is concerned. **What the speaker knows is, simply, that a given item or set of items is treated differently from others by the phonological component of the grammar."**
>
> (p. 373, emphasis added)

Similar to declension/conjugation classes, lexical strata, …

Diacritic features: "third declension", "[+Slavic]", "[−rule *n*]"

Lexical classes
but not “simply” those

Form-based properties can be predictive of class membership and known by speakers

- Readjustment rules (Chomsky & Halle 1968)
- ‘Patterned exceptionality’ (Zuraw 2000)
- ‘Predicting the unpredictable’ (Ernestus & Baayen 2003)
- Sublexicon phonotactics (Becker & Gouskova 2016)

Lexical classes with form-based predictors

Phonological properties can be predictive of:

- Phonological alternation (ex. Dutch voicing alternation)
- Allomorph selection (ex. Hungarian dative, Russian diminutive)
- Gender, declension, conjugation, screeve, …
- Grammatical category (ex. Noun vs. Verb)
- Semantic properties (ex. Abstract vs. Concrete)

Ex. Sakapultek (Mayan) possessive allomorphy

(data from DuBois 1985, transcriptions based on Inkelas 2014)

ak	‘chicken’	w-aːk	‘my chicken’
c’eˀ	‘dog’	ni-c’iːˀ	‘my dog’
ab’ax	‘rock'	w-ub’aːx	‘my rock’
mulol	‘gourd’	ni-muluːl	‘my gourd’
oč’	‘possum’	w-oč’	‘my possum’
am	‘spider’	w-am	‘my spider’
weʔ	‘head hair’	ni-weʔ	‘my head hair’

[$_{\text{Base}}$C → /ni-/$_{\text{1s}}$ [$_{\text{Base}}$V → /w-/$_{\text{1s}}$

Modeling form-class relations

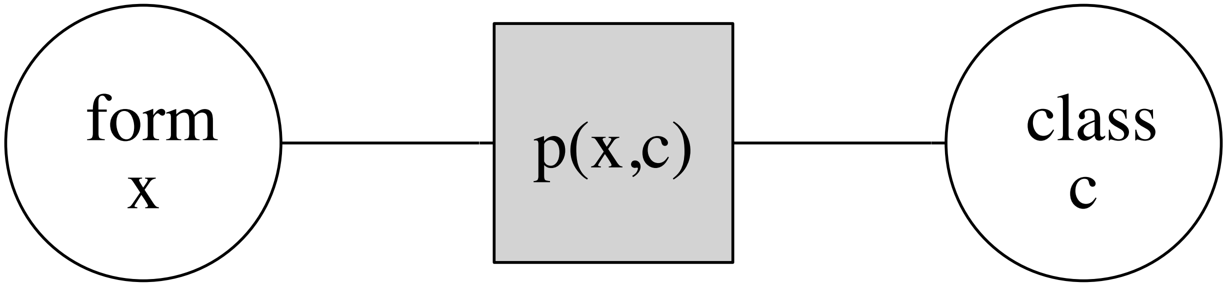
- Learn predictive phonological properties
- Compatible with deterministic and variable, binary and multi-way classification patterns
- Produce explicit, interpretable grammars
    - Compare with hand-written analyses
    - Contribute to empirical typology

Probabilistic model of form-class relations

Form (x) Phonological representation
(abstract or surface, basic or derived, source or product)

Class (c) Member of a set C of classes

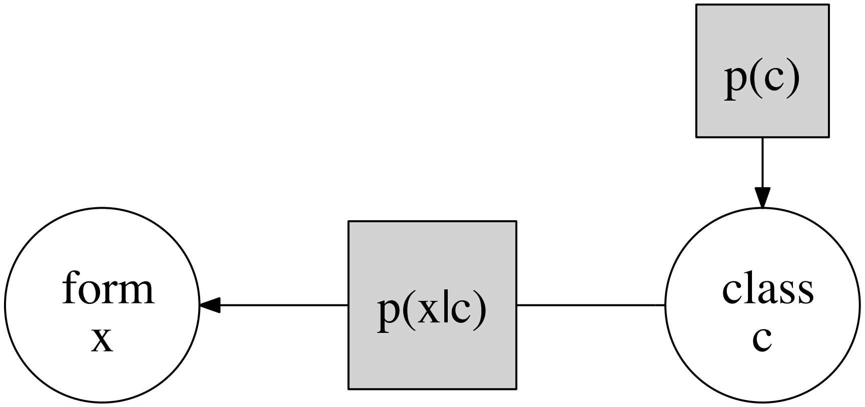
Joint distribution p(form=**x**, class=c)



Probabilistic model of form-class relations

Two ways of rewriting the joint distribution

❶ p(form=**x**, class=c) = p(**x** | c) • p(c)



Calculating p(**x** | c) requires summing over the exponential/infinite set **X** of all possible forms

Probabilistic model of form-class relations

❶ ~ Sublexical morphophonological learner

(Allen & Becker 2015, 2017, see also Gouskova et al. 2015; Becker & Gouskova 2016)

$p(\mathbf{x} \mid c)$ learned class-specific phonotactics

$p(c \mid \mathbf{x}) \propto_{\text{Bayes}} p(\mathbf{x} \mid c) \bullet p(c)$

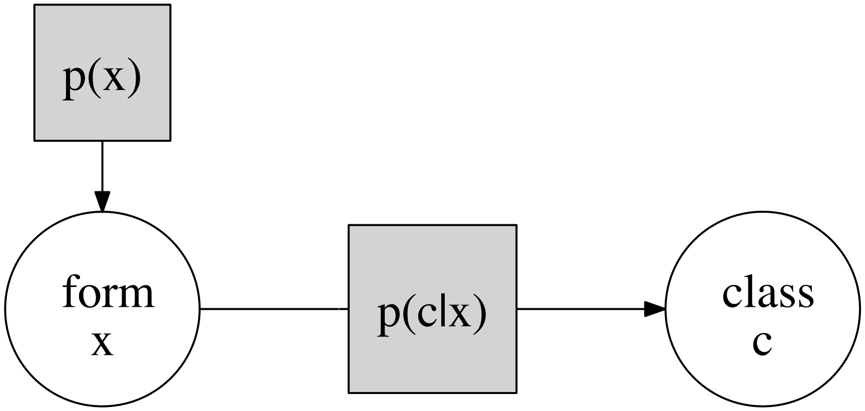
But Bayes relation not consistently used

ex. Gouskova et al. (2015) simplify to $H(\mathbf{x} \mid c) \overset{?}{=} 0$

Probabilistic model of form-class relations

Two ways of rewriting the joint distribution

❷ p(form=**x**, class=c) = p(c | **x**) • p(**x**)



Calculating p(c | **x**) requires summing over only the set C of possible classes (min. {[+F], [−F]})

Probabilistic model of form-class relations

❷ Morphophonological classifier

(Ernestus & Baayen 2003; Hayes, CLS 50; see also Jurafsky & Martin 2009, inter alia)

p(c | **x**) learned form-based predictors

Typical *wug*-test method makes novel base **x** 'observed', so p(**x**) term not needed

How to *parameterize* and *learn* p(c | **x**) ?

Maximum entropy (MaxEnt) morphophonological classifier

$$p(c \mid \mathbf{x}) = \exp(-\textstyle\sum_k w_k\, f_k(\mathbf{x},c)) \,/\, Z(\mathbf{x})$$

$$Z(\mathbf{x}) = \textstyle\sum_{c \in C} \exp(-\textstyle\sum_k w_k\, f_k(\mathbf{x},c))$$

Classifier defined by set of constraints f_k each with learned weight w_k (here assume $w_k \geq 0$). Constraints instantiate a template, here

$$f_k = *\langle \text{tier}_k, \text{pattern}_k, c_k \rangle$$

Ex. Sakapultek (Mayan) possessive allomorphy

Two classes C = {/ni-/$_{1s}$, /w-/$_{1s}$}
(or C = {[+F], [–F]})

(**x**,c)	*⟨C/V, #V, /ni-/$_1$s⟩	*⟨C/V, #C, /w-/$_1$s⟩	
mulol, /ni-/$_1$s	0	0	$p(c \mid \mathbf{x}) \approx 1$
mulol, /w-/$_1$s	0	1	$p(c \mid \mathbf{x}) \approx 0$
am, /ni-/$_1$s	1	0	$p(c \mid \mathbf{x}) \approx 0$
am, /w-/$_1$s	0	0	$p(c \mid \mathbf{x}) \approx 1$

Learning weights

Given data set D = $\{\langle \mathbf{x}_i, c_i \rangle\}$ and constraints $\{f_k\}$, weights $\{w_k\}$ learned by regularized ML

minimize: $-\sum \log p(c_i \mid \mathbf{x}_i) + \lambda_2 \sum w_k^2 + \lambda_1 \sum |w_k|$

- Data can be deterministic or variable (i.e., same form paired with multiple classes)
- Other (non-MaxEnt) ways to set weights

Learning constraints

Greedily induce constraints one at a time

(Della Pietra et al. 1997; Perkins et al. 2003; as in Hayes & Wilson 2008 for phonotactics)

Given current classifier, seek new constraint that can best increase $p(D) = \sum \log p(c_i|\mathbf{x}_i)$

- Many other ways to learn constraints consistent with MaxEnt (random search, simulated annealing, genetic algorithms, …)

Ex. Dutch voicing alternation

(Ernestus & Baayen 2003)

- Not fully predictable

  verwij[t] ~ verwij[d]en ‘widen’ ~ ‘widen-inf’
  verwij[t] ~ verwij[t]en ‘reproach’ ~ ‘reproach-inf’

- Mostly predictable

  - place, manner of final obstruent (e.g., f > p)
  - length of preceding vowel (e.g., long > short)
  - type of preceding segment (e.g., son > obstr)

Ex. Dutch voicing alternation

Stems gathered from CELX and classified as ±alt(ernating) as in Ernestus & Baayen (2003)

Seven learned constraints (on default tier)

*⟨[-cont,-voice]#, +alt⟩$_{1.48}$ ‘[p t k] disprefer to alternate’

*⟨[+long,+stress][]#, –alt⟩$_{1.36}$

*⟨[-phonetic.long][-approx,+cont,+cor]#, +alt⟩$_{1.78}$

*⟨[+long,-diphthong,+stress][-son,-cont,+lab]#, +alt⟩$_{1.91}$

*⟨[+cons,-del.rel,-nasal,-lateral][-approx,+cont,+cor]#, +alt⟩$_{2.14}$

*⟨[+son,+cor][-voice]#, –alt⟩$_{0.98}$ ‘[n ɾ l] induce alternation’

*⟨[+back,+long,-diphthong][+cont,-voice]#, –alt⟩$_{1.61}$

Ex. Dutch voicing alternation

- Learned classifier captures alternation behavior of 79% (1338 / 1694) lexical stems

    Conditional information content of [±alt] < 1 bit

- Classifier matches majority human response for 85% of Ernestus & Baayen (2003) *wug*-stems

    cf. 72% – 91% for hand-written constraints

Alternation and allomorphs

Similar level of performance for other cases of semi-predictable alternation and allomorphs

- Turkish laryngeal alternation (Becker et al. 2011)
- Hungarian vowel harmony in suffixes (Hayes & Londe 2006, Hayes et al. 2009)
- Russian diminutive allomorphy (Gouskova et al. 2015)
- Romanian plural allomorphy (Grosu & Wilson 2016)

Equals or approaches hand-written / UG-biased models

Ex. English Noun-Verb prediction

Phonological correlates of Noun vs. Verb category

(Sereno 1986; Kelly & Bock 1988; Sereno & Jongman 1990; Davis & Kelly 1997; Cassidy & Kelly 1991, 2001; Kelly 1992; Monaghan et al. 2003, 2005, 2007, 2010; Arciuli & Cupples 2004; Albright 2008; Fitneva et al. 2009; Farmer et al. 2011, 2015; Smith 2016 — see also Walker 1984; Becker 2003; Bobaljik 2008; Jaber 2011; and especially Smith 2011, 2016 on cross-linguistic patterns)

❶ Phonotactic approach: p(**x** | gramcat)

Add gramcat as form-level feature to Hayes & Wilson 2008?

❷ Classifier approach: p(gramcat | **x**)

not "how likely would a new Noun (vs. Verb) have form **x**"

but "how likely would form **x** be a Noun (vs. Verb)"

Ex. English Noun-Verb prediction

Phonological forms gathered from CMU dictionary (carefully edited by Bruce Hayes)

Merged with SUBTLEX-US (Brysbaert & New 2009) for classification as Noun vs. Verb ‘dominant’

10-fold cross-validation: randomly partition lexicon into ten parts – train on 9 test on 1

Ex. English Noun-Verb prediction

~ 30 constraints (min 29, max 33) learned on segmental, stress, and C/V tiers

Several constraints found consistently

$*\langle$stress, [-stress][+stress], Noun$\rangle$

$*\langle$stress, [+stress][+primary.stress], Noun$\rangle$

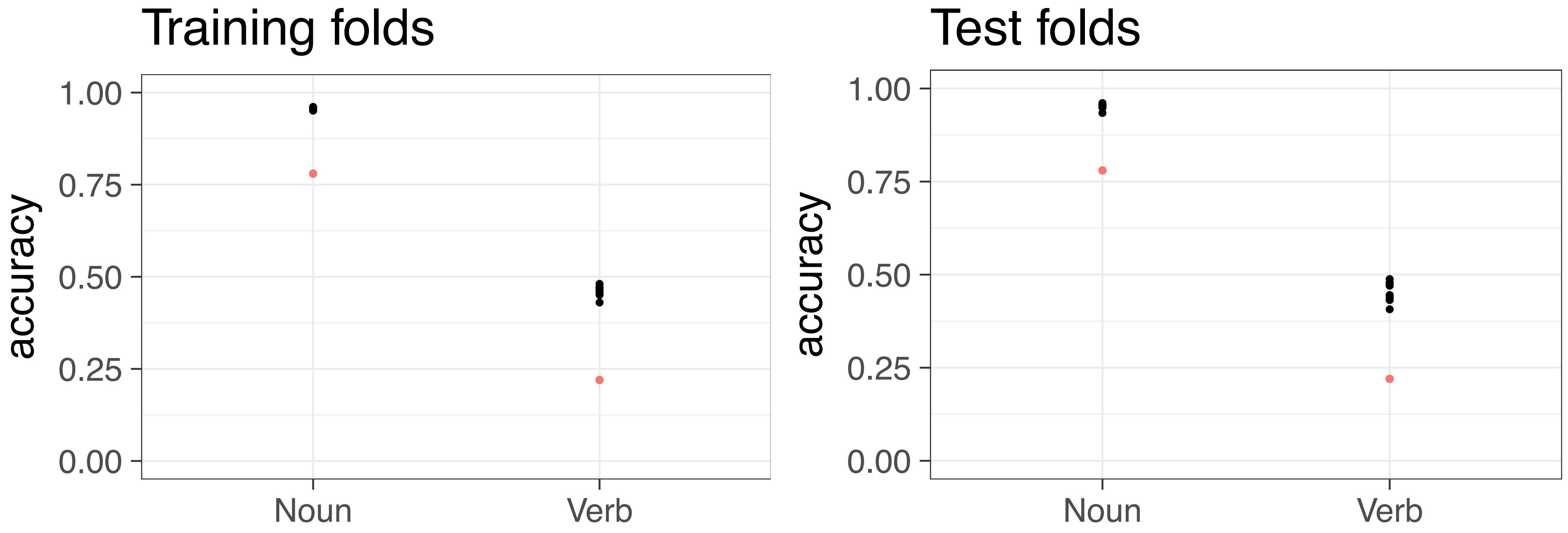
$*\langle$segmental, ə#, Verb$\rangle$

$*\langle$segmental, [+voice,+anterior], Noun$\rangle$

⋮

Ex. English Noun-Verb prediction

Accuracy on lexical items (with baselines)

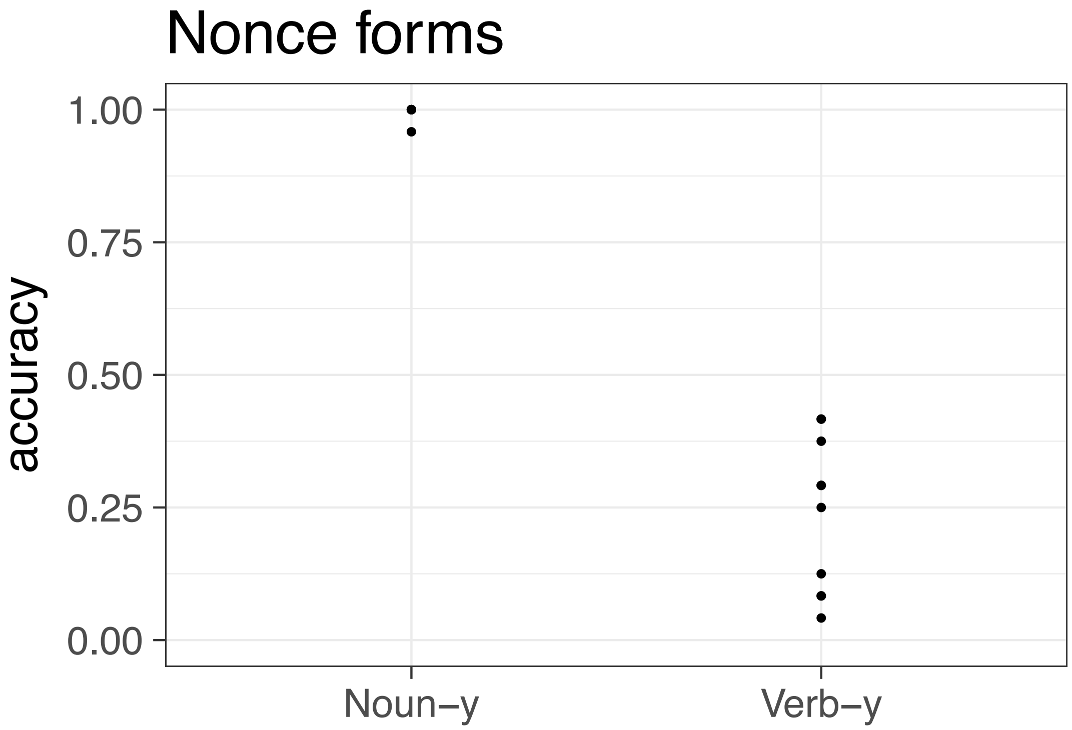

Training folds

Test folds

Ex. English Noun-Verb prediction

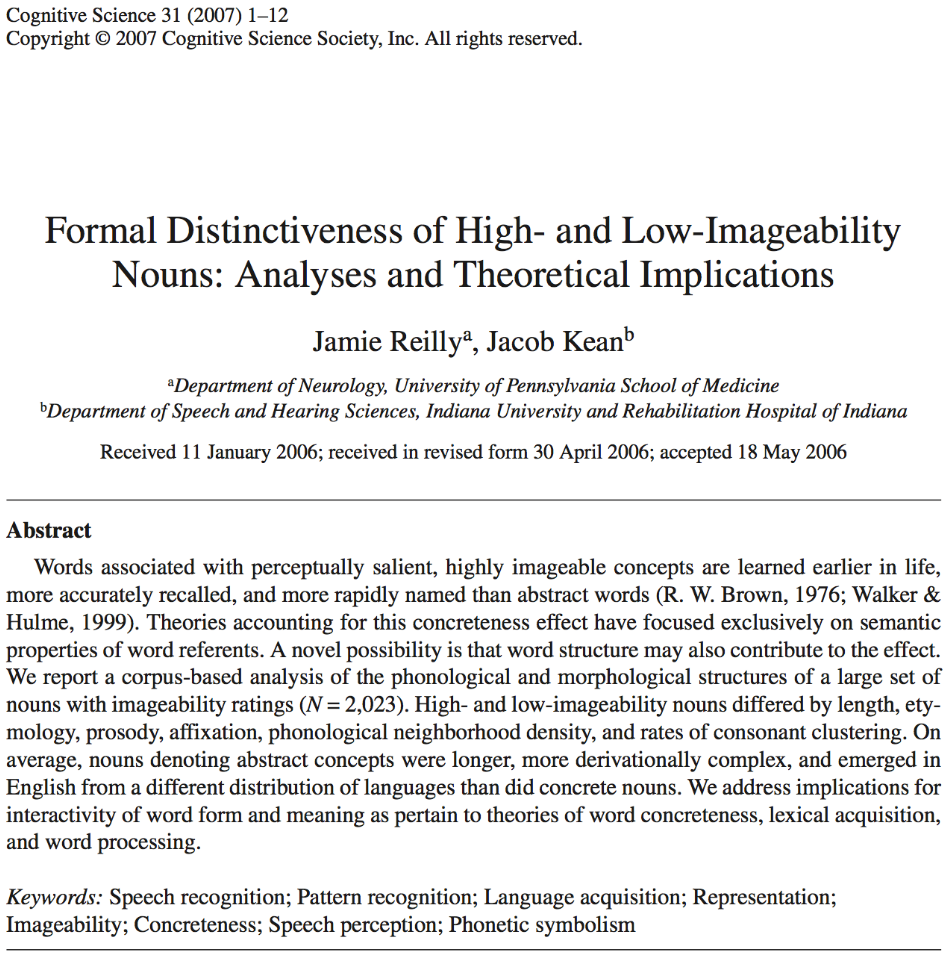
Distinguishing between segmentally more Noun-y vs. more Verb-y nonce words from Smith 2016

ex. [toʊb] vs. [teɪb]



Semantic classes

- Concrete vs. Abstract
  ex. *abscess*, *absence*
  10-fold cross-validation
  Concrete .76 (.71 – .81)
  Abstract .64 (.60 – .69)
- High vs. Low imageability
- Semantic richness
- Others with (psycho-) linguistic support ?



Cognitive Science 31 (2007) 1–12

Formal Distinctiveness of High- and Low-Imageability Nouns: Analyses and Theoretical Implications

Jamie Reilly[a], Jacob Kean[b]

[a]*Department of Neurology, University of Pennsylvania School of Medicine*
[b]*Department of Speech and Hearing Sciences, Indiana University and Rehabilitation Hospital of Indiana*

Received 11 January 2006; received in revised form 30 April 2006; accepted 18 May 2006

Abstract

Words associated with perceptually salient, highly imageable concepts are learned earlier in life, more accurately recalled, and more rapidly named than abstract words (R. W. Brown, 1976; Walker & Hulme, 1999). Theories accounting for this concreteness effect have focused exclusively on semantic properties of word referents. A novel possibility is that word structure may also contribute to the effect. We report a corpus-based analysis of the phonological and morphological structures of a large set of nouns with imageability ratings ($N = 2{,}023$). High- and low-imageability nouns differed by length, etymology, prosody, affixation, phonological neighborhood density, and rates of consonant clustering. On average, nouns denoting abstract concepts were longer, more derivationally complex, and emerged in English from a different distribution of languages than did concrete nouns. We address implications for interactivity of word form and meaning as pertain to theories of word concreteness, lexical acquisition, and word processing.

Keywords: Speech recognition; Pattern recognition; Language acquisition; Representation; Imageability; Concreteness; Speech perception; Phonetic symbolism

Theory integration

Morphophonolexicological classifier must be combined with other components that:

- Segment words, identify alternations and allomorphs (and genders, gramcats, etc.)
- Combine morphemes, apply processes (e.g., Albright & Hayes 2003; Cotterell et al. 2015, 2017; Rastogi et al. 2016)
- Determine degree of morphophonological 'polarization' of individual lexical items (Zuraw 2016)

Further applications

- Artificial-grammar allomorphy

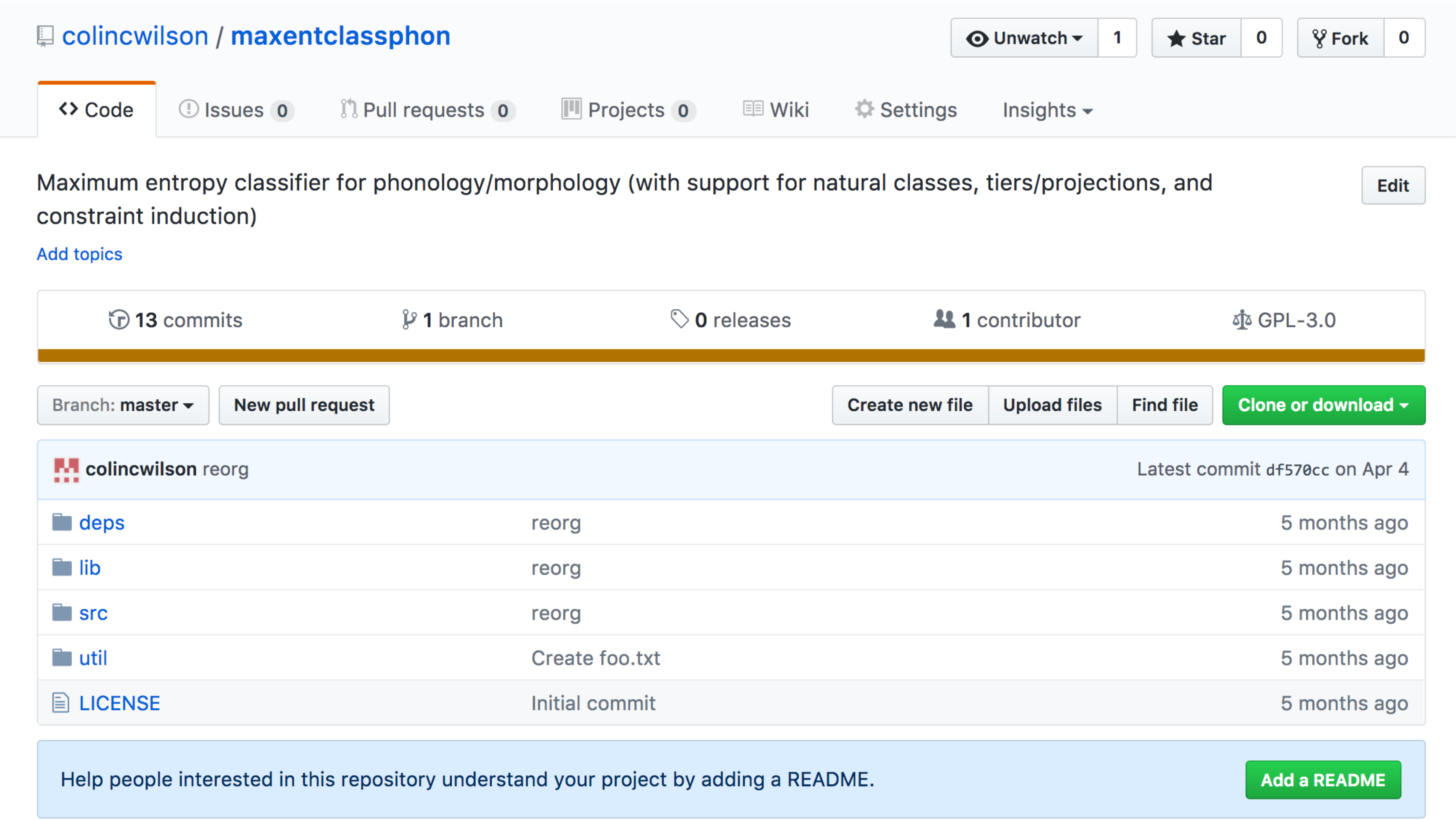
(e.g., Pater & Tessier 2005; Finley & Badecker 2009, Finley 2012, 2015; Baer-Henney 2009)

- Orthographic predictors

(e.g., Arciuli & Cupples 2003, 2006, 2007; Arciuli & Monaghan 2009)

- Any data or experimental materials with 2+ classes of phonological / orthological form

Further applications



colincwilson / maxentclassphon

Maximum entropy classifier for phonology/morphology (with support for natural classes, tiers/projections, and constraint induction)

colincwilson reorg — Latest commit df570cc on Apr 4

deps	reorg	5 months ago
lib	reorg	5 months ago
src	reorg	5 months ago
util	Create foo.txt	5 months ago
LICENSE	Initial commit	5 months ago

Help people interested in this repository understand your project by adding a README.

compatible with Java 8+

Thank you!

Thanks to Michael Becker, Maria Gouskova, Ioana Grosu, Bruce Hayes and Jennifer Smith for discussion and responses to queries

Thanks to all others who made lexical resources and experimental data available