Instance-based acquisition of vowel harmony

Frédéric Mailhot
Institute of Cognitive Science
Carleton University
Ottawa, ON, Canada
fmailhot@connect.carleton.ca

Abstract

I present a nonparametric regression-based model that induces a generalised and productive pattern of vowel harmony—including opaque and transparent neutrality—on the basis of simplified formant data. The model quickly learns to generate harmonically correct morphologically complex forms to which it has not been exposed.

1 Explaining phonological patterns

How do infants learn the phonetic categories and phonotactic patterns of their native languages? How strong are the biases that learners bring to the task of phonological acquistion? Phonologists from the rationalist tradition that dominated the past half-century of linguistic research typically posit strong biases in acquisition, with language learners using innately-given, domain-specific representations (Chomsky and Halle, 1968), constraints (Prince and Smolensky, 2004) and learning algorithms (Tesar and Smolensky, 2000; Dresher, 1999) to learn abstract rules or constraint rankings from which they can classify or produce novel instances.

In the last decade, however, there has been a shift toward empiricist approaches to phonological acquisition, use and knowledge. In this empiricist literature, *eager* learning algorithms (Aha, 1997), in which training data are used to update intensional representations of functions or categories, have been the norm.[1] However, research in related fields—particularly speech perception—indicates that speakers' knowledge and use of language, both in production and comprehension, is at least partly episodic, or instance-based (Goldinger, 1996; Johnson, 1997). Additionally, motivation for instance-based models of categorisation has a lengthy history in cognitive psychology (Medin and Schaffer, 1978), and these methods are well-known in the statistical and machine learning literature, having been studied for over half a century (Cover and Hart, 1967; Hastie et al., 2009). Consequently, it seems a worthy endeavour applying an instance-based method to a problem that is of interest to traditional phonologists, the acquisition and use of vowel harmony, while simultaneously effecting a *rapprochement* with adjacent disicplines in the cognitive sciences.

[1] Daelemans et al. (1994) is a notable exception.

2 Vowel harmony

Vowel harmony is a phonological phenomenon in which there are co-occurrence constraints on vowels within words.[2] The vowels in a language with vowel harmony can be classified into disjoint sets, such that words contain vowels from only one of the sets. The Finnish system of vowel harmony illustrated in Table 1 provides a standard example from the literature (van der Hulst and van de Weijer, 1995).

	output form	*gloss*
a.	tuhmasta	'naughty' (elative)
b.	tühmästä	'stupid' (elative)

Table 1: Finnish backness harmony

Note, crucially, that elative case is realised with a front or back vowel—as *-stä* or *-sta*—depending on whether the stem has front {*ü*, *ä*} or back {*u*, *a*} vowels.

2.1 Neutral vowels

In most languages with vowel harmony, there are one or more vowels that systematically fail

[2] "Word" is used pre-theoretically here; harmony can occur over both larger and smaller domains.

to harmonise. These are called *neutral vowels*, and are typically further subclassified according to whether or not they initiate a new harmonic domain.

3 Instance-based models

Lazy instance-based learning algorithms, also called memory-based, case-based models, and exemplar-based models, have their modern origins in psychological theories and models of perceptual categorisation and episodic memory (Medin and Schaffer, 1978; Hintzman, 1986; Nosofsky, 1986). The earliest explicit discussion seems to be from Semon (1921); a theory of memory that anticipates many features of contemporary models.

Instance-based models were introduced to linguistics via research in speech perception suggesting that at least some aspects of linguistic performance rely on remembered experiential episodes (Goldinger, 1996). The models implemented to date in phonetics and phonology have largely focused on perception (*e.g.* speaker normalisation in Johnson (1997)), or on diachronic processes (*e.g.* lenition in Pierrehumbert (2001) or chain shifts in Ettlinger (2007)), leaving the types of phenomena that typically interest "traditional" phonologists, *viz.* productive, generalised patterns, comparatively neglected.[3] I present here a simple lazily-evaluated computational model of phonological acquisition and knowledge which learns a pattern of phonological alternations that closely mimics the phenomena that characterise stem-controlled vowel harmony, including learning about opaque and transparent vowels.

4 LIBPHON

LIBPHON is a **L**azy **I**nstance-**b**ased **Phon**ological learner whose purpose (in the context of the simulations described here) is to model an exemplar-based approach to the core aspects of the acquisition of vowel harmony. In the idealised context of the model, there is a single agent which receives its training data, *viz.* linguistic input, from an "oracle" teacher, who has perfect knowledge of its language.

4.1 Decisions & mechanisms

As discussed in Johnson (2007), there are some decisions that need to be made in implementing an instance-based model of phonological knowledge involving the basic units of analysis (*e.g.* their size), the relevant type of these units (*e.g.* discrete or continuous), and the mechanisms for similarity-matching and activation spread in the lexicon.

Units I am persuaded by the arguments given by Johnson (2007) and Välimaa-Blum (2009) for the "word-sized" experience of language (rather than *e.g.* features or segments), and by the arguments in Johnson (1997) from *auditory scene analysis* (Bregman (1990)) for the unity of auditory streams, and hence take words, *qua* meaning-bearing chunks, to be the correct basic unit of analysis in instance-based language models (*a fortiori* in *LIbPhon*).[4]

Feature type Having determined the a size of LIBPHON's basic unit, I move now to its construction. As we are in the phonetic/phonological realm, distinctive features present themselves as candidate dimensions. Since the middle of the 20^{th} century (*ca.* Chomsky and Halle (1968)), phonological theories have nearly all supposed that lexical representations are stored in terms of articulatory features (*cf.* Halle (1997) for explicit discussion of this viewpoint). Coleman (1998), citing evidence from the neuroscientific and psycholinguistic literatures on lexical representation claims that evidence for this position (*e.g.* from speech perception and phoneme monitoring experiments) is weak at best, and that lexical representations are more likely to be acoustic than articulatory. Recognising that the issue is far from resolved, I take the LIBPHON's instance space to be acoustic, and for the purposes of the simulations run here use formant values as the embedding dimension. Vowels are specified by the values at their midpoint of the first four formants, and consonants are specified by so-called "locus" values which can be identified by inspecting the trajectories of consonant-vowel transitions in speech (Sussman et al., 1998). Since the particular phenomenon we address is palatal harmony, and $F2$ magnitude is the primary acoustic correlate of vowel front/backness, I omit $F3$ and $F4$, restricting LIBPHON's representations to $(F1, F2)$

[3] Kirchner and Moore (2009) give a model of a synchronic lenition process.

[4] The assumption that word-level segmentation of the speech signal is available to the language learner prior to acquisition of phonological phenomena is relatively uncontroversial, given evidence from acquisition studies (Jusczyk, 1999), although modelling work shows that the problem is far from solved (Brent, 1999).

trajectories. Hence the basic unit of analysis in LIBPHON is a word-sized sequence of $(F1,F2)$ values.

Similarity I take basic Euclidean distance to be LIBPHON's similarity (or rather, dissimilarity) function.[5]

Fixed-rate representations For the simulations described here, I use fixed-rate trajectories, in which consonants and vowels are represented in a temporally coarse-grained manner with single $(F1,F2)$ tuples. Evidently, consonants and vowels in actual human speech unfold in time, but modelling segments at this level introduces the problem of temporal variability; repeated tokens of a given word—both within and across speakers—vary widely in duration. This variability is one of the main obstacles in the development of instance-based models of speech production, due to the difficulty of aligning variable-length forms. Although algorithms exist for aligning variable-length sequences, these require cognitively implausible dynamic programming algorithms, *e.g.* dynamic time warping (DTW) and hidden Markov models (Rabiner and Juang, 1993). Even as proofs of concept, these may be empirically inadequate; (Kirchner and Moore, 2009) use DTW to good effect in an instance-based production model of spirantisation using real, temporally variable, speech signals. However, their inputs were all the same length in terms of segmental content, and the model was only required to generalise within a word type. It remains to be seen whether DTW could function as a proof of concept in a problem domain like that addressed here, which involves learning about variably-sized "pieces" of morphology across class labels.

4.2 Perception/categorisation

LIBPHON's method of perception/categorisation of inputs is a relatively standard nearest-neighbour-based classification algorithm. See 1 for a description in pseudocode.

If LABEL is not empty, LIBPHON checks its lexicon to see whether it knows the word being presented to it, *i.e.* whether it exists as a class label.

[5]Often the measure of similarity in an instance-based model is an exponential function of distance, $d(x_i,x_j)$ of the form $e^{-cd(x_i,x_j)}$, so that increasing distance yields decreasing similarity (Nosofsky, 1986). The Euclidean measure here is sufficient for the purpose at hand, although the shape of the similarity measure is ultimately an empirical question.

Algorithm 1 PERCEIVE(*input*, k)

Require: *input* as (LABEL $\in$ [LEX](PL)[NOM | ACC], `instance` $\in \mathbb{Z}_{2x8,10,12}$), $k \in \mathbb{Z}$

```
if LABEL is not empty then
    if LABEL ∈ lexicon then
        Associate(instance, LABEL)
    else
        Create LABEL in lexicon
        Associate(instance, LABEL)
    end if
else
    neighbours ← k-nearest neighbours of instance
    LABEL ← majority class label of neighbours
    Associate(instance,LABEL)
end if
```

If so, it simply appends the input acoustic form to the set of forms associated with the input meaning/label. If it has no corresponding entry, a new lexical entry is created for the input meaning, and the input trajectory is added as its sole associated acoustic form.

If LABEL is empty, LIBPHON assigns `instance` to the majority class of its k nearest neighbours.

4.3 Production

In production, LIBPHON is provided with a label and has to generate a suitable instance for it. Labels are compositional, and signal both an arbitrary "lexical" meaning, and a case marker from {NOM,ACC}. Thus, there are several different possibilities to consider in generating output for some queried meaning.

In the two simplest cases, either the full queried meaning (*viz.* lexical label with case marker) is already in the lexicon, or else there are no lexical entries with the same lexical label (*i.e.* the agent is essentially being called on to produce a word that it doesn't know). For the former case, a token acoustic output is (uniform) randomly selected from the list of acoustic forms associated with the queried label, the entire set of acoustic forms is used as the cloud, and an output is generated by taking a distance-weighted mean over the 5 nearest neighbours. In the latter case, the same

procedure is carried out, but with a randomly selected lexical entry (since the queried one is unknown), and the queried form and generated output are stored in the lexicon.

In the more interesting cases, LIBPHON has a LABEL in its lexicon with the same lexical meaning, but with differing CASE and/or PL specification. Consider the case in which LIBPHON knows only the (singular) NOM form of the queried label, but has to produce the PL ACC form. A seed instance is (uniform) randomly selected from the set of trajectories associated to the NOM entry in the agent's lexicon, as this is the only entry with the corresponding lexical meaning, and it is a variant of this meaning that LIBPHON must produce. The analogical set, in this case, is composed of the seed's nearest neighbours in the set of all trajectories associated with LABELs of the form [LEX PL ACC]. Once again, the output produced is a distance-weighted mean of the analogical set.

This general procedure (*viz.* seed from a known item with same lexical meaning, analogical set from all items with desired inflection) is carried out in parallel cases with all other possible LABEL mismatches, *e.g.* a singular LABEL queried, but only a plural LABEL in the lexicon, a NOM query with only an ACC form in the lexicon, *etc.* In the cases where the lexicon contains multiple entries with the same lexical meaning, but not the query, the seed is selected from the LABEL with the closest semantic match. 2 gives pseudocode for LIBPHON's production algorithm.

Algorithm 2 PRODUCE(LABEL, k)

Require: LABEL $\in$ [LEX](PL)[NOM | ACC], $k \in \mathbb{Z}$

if LABEL $\in$ lexicon **then**
  seed $\leftarrow$ uniform random selection from `instances` associated to LABEL
  cloud $\leftarrow$ all `instances` associated to LABEL
else if $\exists$ LABEL$'$ $\in$ lexicon s.t. lex(LABEL$'$) = lex(LABEL) **then**
  seed $\leftarrow$ uniform random selection from `instances` associated to LABEL$'$)
  cloud $\leftarrow$ all `instances` associated to plural(LABEL) $\cup$ case(LABEL)
else
  pass
end if

neighbours $\leftarrow$ k-nearest neighbours of seed in cloud
return distance-weighted mean of neighbours

5 The languages

In order to get at the essence of the problem (*viz.* the acquisition of vowel harmony as characterised by morphophonological alternations), and in the interests of computational tractability/efficiency, the artificial languages learned by LIBPHON are highly simplified, displaying only enough structure to capture the phenomena of interest.

On the view taken here, phonological knowledge is taken to emerge from generalisation over lexical items, and so the key to acquiring some phonological pattern lies in learning a lexicon (Jusczyk, 2000). Consequently, the languages learned in LIBPHON abstract away from sentence-level phenomena, and the training data simply takes the form of labelled formant trajectories. These trajectories consist of four-syllable "roots" with one or more one-syllable "affixes". All syllables are CV-shaped.

5.1 Phonological inventory

The phonological inventory consists of three consonants and four vowels—two with high $F2$ and two with low $F2$—which we will label /b, d, g/ and /i, e, u, o/, respectively, for convenience.[6] The formant values used were generated from formant synthesis equations in de Boer (2000), and from the locus equations for CV-transitions in (Sussman et al., 1998).

5.2 Lexical items

Tokens of trajectories have associated class labels, which formally are contentless indices. Rather than employing *e.g.* natural numbers as labels, we use character strings which correspond more or less to the English pronounciations of their associated trajectories. We will refer to these metaphorically as MEANINGS. These are compositional, comprising a "lexical meaning" (arbitrary CVCVCVCV string from the phoneme set listed above), one of two obligatorily present "case markers" (NOM|ACC), and an optionally present

[6] LIBPHON's representations lack $F3$, the primary correlate of rounding, so the back vowels might be more standardly represented by /ɤ, ɯ/. Nothing in the results or discussion hinges on this.

"plural marker" (PL). Hence, words in the artificial languages come in four forms, NOM-SG, NOM-PL, ACC-SG, and ACC-PL.

5.3 Harmonizing suffixes

For many generative phonologists vowel harmony is taken to be a *process* characterised by a productive and general synchronic system of alternations, rather than a statistical generalisation over a lexicon. In typical cases of vowel harmony that are seen as diagnostic, we find affix allomorphs which surface with alternating vowels, as in the examples given in section 2.

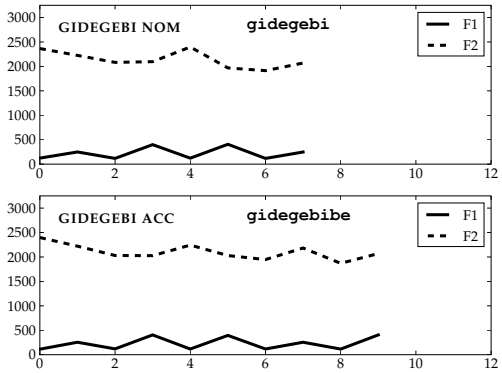

Figure 1: Graphical representation of singular forms of GIDEGEBI, as produced by teacher agent

Figure 1 gives example NOM and ACC-labelled trajectories (for lexical label GIDEGEBI) for the artificial language learned by LIBPHON. Vowels are either all high or all low $F2$, *i.e.* there is pure lexical front/back vowel harmony.[7] The "nominative" label has no morphological realisation, and the LEX-NOM for has a trajectory eight segments long (the "pinches" correspond to the vowels). ACC-labelled forms are realised with a vowel that alternates to match the $F2$ value of the root vowels.

5.4 Neutral vowels

The harmony processes examined thus far are in some sense "local", being describable in terms of adjacency *e.g.* on a hypothesised autosegmental vowel tier (although the presence of intervening consonants still renders the harmony process "nonlocal" in some more concrete articulatory sense, *pace* (Gafos, 1999)). One of the hallmarks of vowel harmony, as discussed in **??**, is the phenomenon of *neutral vowels*. These vowels fail to undergo harmony, and may or may not initiate a new harmonic domain. To introduce a neutral vowel, I added a category label, PL, whose realisation corresponds roughly to [gu], and which is treated as being either opaque or transparent in the simulations described below.

Figure 2 and Figure 3 show graphical representations of the various "inflected" forms of GIDEGEBI. The even-numbered indices on the x-axis correspond to consonants and the odd-numbered indices to vowels. Figure 2 and Figure 3 illustrate the difference between languages with opaque versus transparent PL, as reflected in the realisation of the ACC marker, which agrees in $F2$ with the realised form of the PL root, respectively.

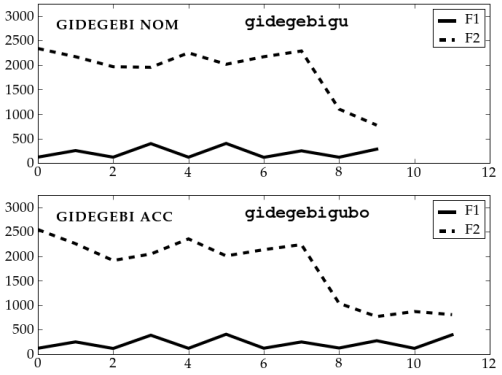

Figure 2: Graphical representation of plural forms of GIDEGEBI, as produced by teacher agent, with opaque PL realisation.

6 The experiments

Assessing successful learning/generalisation in a computational model requires some measurable outcome that can be tracked over time. Because *LIbPhon* is a production-oriented model, its classification of inputs is a poor indicator of the extent to which it has learned a productive "rule" of vowel harmony. In lieu of this measure, we have opted to pursue two difference courses of evaluation.

For the harmony cases, LIbPhon is queried on previously unseen MEANINGs and its output is compared to the mean value of the teacher's output for the same representation. In particular, given

[7] This is unrealistic, as most vowel harmony languages have some exceptionally disharmonic forms in their lexicons.

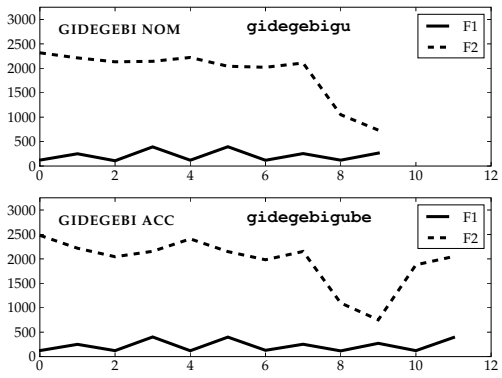

Figure 3: Graphical representation of plural forms of GIDEGEBI, as produced by teacher agent, with transparent PL realisation.

some label which the learner has not seen productions of, we can query the learner at various stages of acquisition (*viz.* with lexicons of increasing size) by having it produce an output for that label, and track its increasing performance over time.

The actual measure of error taken is the root-mean-square differences between the learner's output, y and the teacher's mean output , t, for some label, l, over all of the consonants and vowels within a word, average across the N remaining unseen items of the teacher's lexicon:

$$RMSE = \frac{1}{N} \sum_{l \in lex} \sqrt{\frac{\sum_i (t_i - y_i)^2}{len(t)}}$$

Figure 4 shows clearly that error drops as the lexicon grows, but it is not a terribly informative measure. From a linguistic point of view, we are interested in what LIBPHON's outputs look like, *viz.* has it learned vowel harmony? Figure 5 and Figure 6 show that vowel harmony *is* learned, and moreover quite quickly, after going through a brief initial phase of spurious outputs. In Figure 5, LIBPHON is being asked to produce outputs for all forms of the label GUBOGOBU. For the particular run shown here, at the 10-word stage (*i.e.* when LIBPHON had seen tokens from 10 labels), the only tokens marked PL-ACC were from high $F2$ ("front") trajectories. Hence the nearest neighbour calculation in the production algorithm resulted in a fronted form being output. Although acquisition research in vowel harmony languages is relatively rare, or inaccessible to us due to language barriers, what research there is seems to indicate that harmony is mastered very quickly, with virtually no errors by 2 years of age, hence it is unclear what status to assign to output patterns like the one discussed here. Moreover, given the well-known facts that (i) comprehension precedes production, and (ii) infants avoid saying unfamiliar words, it is unlikely that an infant could be coaxed into producing an output form for such an early-stage class.

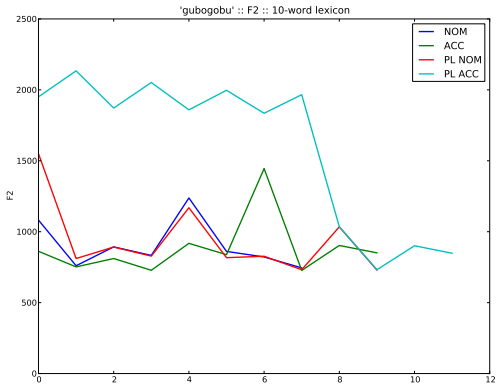

Figure 5: Evolution of gubogobu: 10 word lexicon

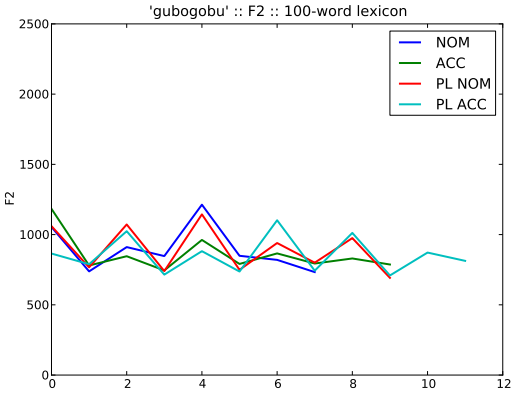

Figure 6: Evolution of gubobogu: 100 word lexicon

7 Discussion

The preliminary experiments discussed here show that on the basis of limited input data, *LIbPhon* learns to produce harmonically correct novel outputs. In particular, it is able to generalise and pro-

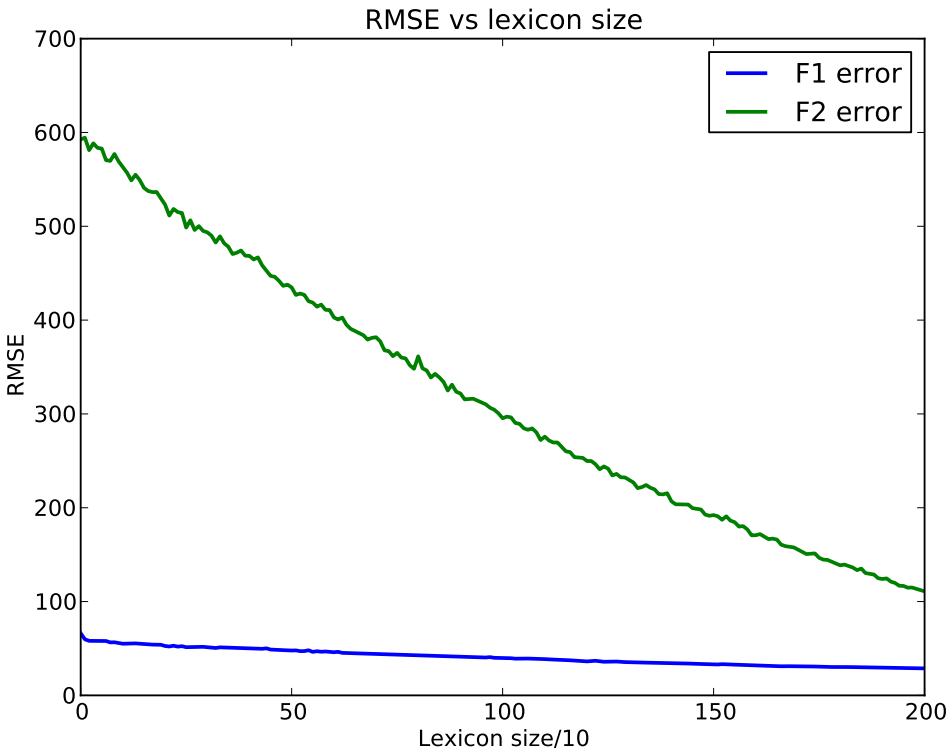

Figure 4: RMSE 2000 word lexicon

duce correct morphologically complex forms to which it has not been exposed in its training data, *i.e.* a previously unseen case-marked form will be output with harmonically correct $F2$, including neutrality (opaque or transparent). In ongoing research we are evaluating LIBPHON's performance with respect to more traditional measures, in particular tracking F-score on held-out data as the lexicon grows.

Acknowledgments

This work carried out with the support of NSERC Discovery Grant 371969 to Dr. Arshia Asudeh. Many thanks to Ash Asudeh, Lev Blumenfeld, Jeff Mielke, Andrea Gormley, Andy Wedel, and Alan Hogue for helpful feedback.

References

[Aha1997] David Aha. 1997. Lazy learning. In *Lazy Learning*, pages 7–10. Kluwer Academic Publishers.

[Bregman1990] Al Bregman. 1990. *Auditory Scene Analysis*. MIT Press.

[Brent1999] Michael R. Brent. 1999. Speech segmentation and word discovery: a computational perspective. *Trends in Cognitive Sciences*, 3(8):294–301.

[Chomsky and Halle1968] Noam Chomsky and Morris Halle. 1968. *The Sound Pattern of English*. Harper and Row.

[Coleman1998] John Coleman. 1998. Cognitive reality and the phonological lexicon: A review. *Journal of Neurolinguistics*, 11(3):295—-320.

[Cover and Hart1967] Thomas Cover and Peter Hart. 1967. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13:21–27.

[Daelemans et al.1994] Walter Daelemans, Steven Gillis, and Gert Durieux. 1994. The acquisition of stress: A data-oriented approach. *Computational Linguistics*, 20(3).

[de Boer2000] Bart de Boer. 2000. Self-organization in vowel systems. *Journal of Phonetics*, 28:441–465.

[Dresher1999] B. Elan Dresher. 1999. Charing the learning path: Cues to parameter setting. *Linguistic Inquiry*, 30(1):27–67.

[Ettlinger2007] Marc Ettlinger. 2007. An exemplar-based model of chain shifts. In *Proceedings of the 16th International Congress of the Phonetic Science*, pages 685–688.

[Gafos1999] Adiamantos Gafos. 1999. *The articulatory basis of locality in phonology*. Ph.D. thesis, New York University.

[Goldinger1996] Stephen Goldinger. 1996. Words and voices: Episodic traces in spoken word identification and recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 22:1166–1183.

[Halle1997] Morris Halle. 1997. Some consequences of the representation of words in memory. *Lingua*, 100:91–100.

[Hastie et al.2009] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. 2009. *Elements of Statistical Learning*. Springer Series in Statistics. Springer-Verlag, 2 edition.

[Hintzman1986] Douglas Hintzman. 1986. "schema abstraction" in a multiple-trace memory model. *Psychological Review*, 93(4):411–428.

[Johnson1997] Keith Johnson. 1997. Speech perception without speaker normalization: an exemplar model. In *Talker Variability in Speech Processing*, chapter 8, pages 145–166. Academic Press.

[Johnson2007] Keith Johnson, 2007. *Decision and Mechanisms in Exemplar-based Phonology*, chapter 3, pages 25–40. Oxford University Press.

[Jusczyk1999] Peter Jusczyk. 1999. How infants begin to extract words from speech. *Trends in Cognitive Sciences*, 3(9):323–328.

[Jusczyk2000] Peter Jusczyk. 2000. *The Discovery of Spoken Language*. MIT Press.

[Kirchner and Moore2009] Robert Kirchner and Roger Moore. 2009. Computing phonological generalization over real speech exemplars. ms.

[Medin and Schaffer1978] Douglas Medin and Marguerite Schaffer. 1978. Context theory of classification learning. *Psychological Review*, 85(3):207–238.

[Nosofsky1986] Robert Nosofsky. 1986. Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General*, 115(1):39–57.

[Pierrehumbert2001] Janet Pierrehumbert. 2001. Exemplar dynamics: Word frequency, lenition, and contrast. In *Frequency effects and the emergence of linguistic structure*, pages 137–157. John Benjamins.

[Prince and Smolensky2004] Alan Prince and Paul Smolensky. 2004. *Optimality Theory: Constraint interaction in generative grammar*. Blackwell.

[Rabiner and Juang1993] Lawrence Rabiner and Biing-Hwang Juang. 1993. *Fundamentals of Speech Recognition*. Prentice Hall.

[Semon1921] Richard Semon. 1921. *The Mneme*. George Allen and Unwin.

[Sussman et al.1998] Harvey Sussman, David Fruchter, Jon Hilbert, and Joseph Sirosh. 1998. Linear correlates in the speech signal: The orderly output constraint. *Behavioral and Brain Sciences*, 21:241–299.

[Tesar and Smolensky2000] Bruce Tesar and Paul Smolensky. 2000. *Learnability in Optimality Theory*. MIT Press.

[Välimaa-Blum2009] Riitta Välimaa-Blum. 2009. The phoneme in cognitive phonology: episodic memories of both meaningful and meaningless units? *Cognitextes*, 2.

[van der Hulst and van de Weijer1995] Harry van der Hulst and Jeroen van de Weijer. 1995. Vowel harmony. In John Goldsmith, editor, *Handbook of Phonological Theory*. Blackwell.