UC Berkeley
Proceedings of the Annual Meeting of the Berkeley Linguistics Society

Title

Anti-markedness Patterns in French Epenthesis: An Information-theoretic Approach

Permalink

https://escholarship.org/uc/item/94s0d1vp

Journal

Proceedings of the Annual Meeting of the Berkeley Linguistics Society, 37(37)

ISSN

2377-1666

Authors

Hume, Elizabeth
Hall, Kathleen Currie
Wedel, Andrew
et al.

Publication Date

2013

Peer reviewed

Anti-markedness patterns in French epenthesis: An information-theoretic approach
Author(s): Elizabeth Hume, Kathleen Currie Hall, Andrew Wedel, Adam Ussishkin, Martine Adda-Decker, and Cédric Gendrot
Proceedings of the 37th Annual Meeting of the Berkeley Linguistics Society (2013), pp. 104-123
Editors: Chundra Cathcart, I-Hsuan Chen, Greg Finley, Shinae Kang, Clare S. Sandy, and Elise Stickles

Anti-Markedness Patterns in French Epenthesis: An Information-theoretic Approach[1]

ELIZABETH HUME[a], KATHLEEN CURRIE HALL[b], ANDREW WEDEL[c], ADAM USSISHKIN[c], MARTINE ADDA-DECKER[d,e], and CÉDRIC GENDROT[d]
[a]*The Ohio State University;* [b]*City University of New York-College of Staten Island & The Graduate Center;* [c]*University of Arizona;* [d]*Laboratoire d'Informatique pour la Mécanique et les Sciences—Centre national de la recherche scientifique, Université Paris-Sud 11;* [e]*Laboratoire de Phonétique et Phonologie—Centre national de la recherche scientifique, Université Paris 3*

Introduction

Cross-linguistically, certain vowel types tend to be used to break up otherwise ill-formed consonant clusters in a given language: they are generally non-low, non-round and either front or central. Such epenthetic vowels are commonly referred to the language's *default* vowel. For example, the default vowel in Maltese is [i], in Spanish it is [e], in Korean it is [ɨ], in German, Dutch and Finnish it is [ə], and [ə] or [ɪ] in English. One might assume, then, that these vowels have certain properties that make them particularly good candidates for being the epenthetic vowel. One commonly used means of predicting the quality of the epenthetic vowel has been to draw on markedness. In this approach, default vowels are considered unmarked either universally or on a language-specific basis (e.g. Archangeli 1984; Pulleyblank 1988; Rice 1999, 2000). Indeed, Rice (2000) points to epenthesis as a diagnostic for identifying the unmarked member of an opposition, proposing that the unmarked member is more likely to be inserted (though cf. Rice 2007). While such approaches are successful in predicting the most common patterns involving front or central unrounded vowels, they are less successful when the vowel involved is not obviously unmarked, as in the case of French. The default vowel in French, while commonly referred to as schwa, is a front or centralized *rounded* vowel, realized phonetically as similar or identical to the mid-front rounded vowels [ø] or [œ], depending on speaker and variety. The

104

French pattern is anomalous given the assumption that roundness is typologically marked in non-back vowels (e.g. Chomsky & Halle 1968, de Lacy 2006). In this view, one of the front unrounded vowels such as [i, e, ɛ], also present in the French vowel inventory, would be expected to serve as the default vowel as they are arguably less marked.

The observation that not all vowel epenthesis patterns involve traditional unmarked vowels has led some researchers to exclude vowel epenthesis as a criterion for determining the markedness value of a sound (de Lacy 2006, Rice 2007). De Lacy (2006), for example, proposes that vowel epenthesis belongs to "performance-based markedness," predicted by performance factors, e.g. frequency and phonetics. On the other hand, vowel deletion belongs to "competence-based markedness," and would thus be predicted by the grammar. Interestingly, in many languages including English, Dutch, French, and Brazilian Portuguese, the vowel that epenthesizes has the same quality as the vowel that deletes, suggesting the need for a unified account.

Regardless of whether or not vowel epenthesis fits neatly into a markedness account, the above studies share a common approach: they start from the assumption of a prior distinction between marked and unmarked segments and their associated patterns, and then ask what properties distinguish them. In this paper we turn the causal arrow around. Taking as a starting point the well-established assumption that the properties of segments in a system influence their patterning, we ask what properties are typically associated with unmarked versus marked segments (see also, among others, Lass 1975; Comrie 1983; Menn 1983; Blevins 2004; Hume 2004, 2008; Bybee, to appear). Knowing what properties are associated with unmarked segments can then allow us to re-examine the French data, looking to see which vowels in the French system also share the *properties* of default vowels, regardless of whether or not they are generally assumed to be unmarked.

More specifically, we think that "marked" and "unmarked" are simply labels and do not provide much insight into phonological patterns. Further, in order to fully understand the factors that influence phonological systems, one cannot divorce the phonological component from the larger system in which it occurs, including the system's larger role in communication. Along these lines, identifying the *function* of epenthesis should be the first step in attempting to determine what properties would make a vowel a good candidate for being involved in such a process. Those properties, then, are what we should look for in identifying likely epenthetic vowels in a particular language. These vowels will tend to fit the profile of being "unmarked" vowels, but we claim that this is an artifact of sharing similar properties *vis à vis* the system in question, rather than being an explanatory characteristic itself.

Identifying the desirable characteristics of default epenthetic vowels and looking for those characteristics among French vowels reveals that the front

rounded vowels do in fact emerge as being vowels that would be good candidates for epenthesis. While there is no one property that seems to uniquely determine which vowel would make the "best" candidate to break up an ill-formed consonant sequence, it is clear that the front rounded vowels are not as anomalous as they might otherwise seem from a universalist markedness perspective.

This paper has three objectives. The first is to situate the properties associated with French default vowels with more typologically common epenthesis patterns. The second is to make use of an approach to predicting the quality of the default vowel that uses the function of epenthesis within a system of communication as a starting point. Finally, we address the issue of how to quantify the properties associated with epenthetic vowels, for which we use the Generalized Context Model (Nosofsky 1988) and tools from Information Theory (Shannon 1948), in particular, entropy and information content.

1 Epenthesis

1.1 The Function of Vowel Epenthesis

As noted above, vowel epenthesis is typically used to break up sequences of multiple consonants (see Hall 2011 for relevant discussion). This may be a response to difficult or unfamiliar consonant sequences, with one result of epenthesis being that such sequences can be produced and processed faster and more accurately (e.g. Kuipers et al. 1996; Davidson 2006). The epenthetic vowel, then, should be one that is easy to produce and process, so as to make difficult sequences easier, and it should be one that is expected to occur in the system, so as to facilitate production and perception. Furthermore, in languages where there is a single epenthetic vowel in the system, that vowel should be flexible enough to co-occur with a wide range of consonants. Finally, it should furthermore be one that is not likely to create new words in the language when added, as the purpose of epenthesis is to somehow clarify the intended message, and creating a new word could counteract that function.

These desiderata, not surprisingly, lead to characteristics that are commonly associated with unmarked vowels. Note that these are made up of phonological, phonetic, and usage-based factors.

- *High frequency (low information content/surprisal)*: A segment that occurs with high frequency and thus is highly practiced and expected in the language (e.g. Greenberg 1966; Eddington 2001; Hume & Bromberg 2006; Cristófaro-Silva and Almeida 2008; Bybee, to appear).
- *Weak perceptual contrast*: A segment with weak phonetic cues due to inherent nature and/or contextual factors (e.g. Steriade 2009, Riggs, to appear).

- *Weak lexical contrast (low functional load)*: A segment that does not distinguish a large number of lexical items in the language (see e.g. Lass 1975).
- *Wide phonological distribution*: A segment that can co-occur with many different consonants (e.g. Trubetzkoy 1939; Hockett 1955; Greenberg 1966).

We will consider each of these qualities in turn, evaluating the (oral) vowels in the French system along the various dimensions. As will be seen, each quality in isolation would pick out a different set of vowels as being the best choices for epenthesis. The mid front rounded vowels, however, consistently emerge in almost all cases as being among the best candidates, and no other vowels do so. Thus, the quality of the French epenthetic vowel does not appear as anomalous when one considers the function of epenthetic vowels and what qualities would characterize a "good" epenthetic vowel in any language.

1.2   French Epenthesis

The vowel system of Continental French, the variety examined in this paper, is comprised of a series of nasal vowels, [õ œ̃ ɛ̃ ɑ̃] (for most speakers [œ̃] has been merged with [ɛ̃]) and ten oral vowels, [i e ɛ y ø œ a² u ɔ o]. The French "schwa", not shown in (1), can vary between the closed, tense [ø] and more open, lax [œ], based on speaker and dialectal factors. Adda-Decker et al. (1999) found it to be "between the open /œ/ and the closed /ø/...the pronunciation /œ/ appears to be preferred." Fougeron et al. (2007), on the other hand, compared it to non-alternating [ø] and [œ], and found it tended toward [ø]. Despite the phonetic variability associated with the default vowel, we will for simplicity consistently transcribe it as [œ] throughout this paper.

A French default vowel typically occurs, as in (1), to avoid a three-consonant sequence, or word-finally following a consonant. Note there are restrictions on the types of consonants involved that we do not go into here (see, e.g. Grammont 1914, Carton 1999).

(1)   Default vowel occurrence:
  a. [œ̃kõtakt(œ)penibl(œ)] *un contact penible* 'a painful contact' (Noske 1993)
  b. [bjɛ̃ syʀ(œ)] *bien sûr!* 'certainly' (Carton 1999)

Given the French vowel inventory, we might expect one of the front unrounded vowels, [i, e, ɛ], to be used as the default vowel. However, as noted above, the default vowel is a mid front or centralized *rounded* vowel, even though it is often

---

² Some speakers distinguish [a] from a more back vowel [ɑ].

transcribed as an unrounded schwa (Jenkins 1971; Adda-Decker et al. 1999; Côté & Morrison 2004; Fougeron et al. 2007). It should be noted that the vowel is rounded is not uncontroversial. Indeed, Féry (2003) notes that speakers of French can have clear intuitions that the vowel that epenthesizes (referred to as French "schwa") differs in quality from the mid rounded vowels, particularly in not being rounded. She considers this, however, to be mostly "a consequence of the orthography and the distributional facts" (Féry 2003: 253-4): the symbol *e* is used to write French "schwa" (*je* 'I'), while [ø] is written as *eu* (*jeux* 'game'), and [œ] as *oe* (*soeur* 'sister') or *eu* (*abreuvoir* 'trough'). Similarly, Landick (1995: 125) states that "what is called schwa is...realized as /œ/ (or as /ø/, depending on the dialect)....[W]hen the reflex of schwa is /œ/ or /ø/, the schwa is, of course, not distinguishable phonetically from the /œ/ of *seul* or the /ø/ of *deux*...."

We now evaluate the French oral vowels against properties commonly associated with default vowels: high frequency/low information content, weak perceptual contrast, weak lexical contrast, broad distribution.

2 High Frequency, Low Information Content

It has been claimed that the epenthetic, or default, vowel in a language (the "unmarked" vowel) is one that tends to have a high frequency of occurrence (e.g. Greenberg 1966; Eddington 2001; Bybee, to appear), or, in information-theoretic terms, low information content or surprisal (Hume and Bromberg 2006; Cristófaro-Silva and Almeida, 2008; Hume and Mailhot, to appear). This makes sense from the perspective of the function of epenthesis if we consider its communicative purpose. As noted above, vowels that are highly frequent will be more practiced in terms of production and more expected from both production and processing perspectives. Given that the epenthetic vowel should be one that makes processing the intended message easier, without interrupting the lexical content of the message, a highly frequent vowel would be desirable.

Measurements were calculated from a subset of the ESTER (Évaluation des Systèmes de Transcription Enrichie d'Émissions Radiophoniques) corpus which consisted of 24 hours of radio-broadcasted news produced by a total of 574 speakers (Galliano et al., 2005). Articulation remains quite distinct so that speech can be understood by a broad audience. Such speech cannot therefore be described as fully spontaneous, but rather as prepared speech: only a few hesitations, repetitions, and word fragments are observed and syntactic structures often remain close to written language. It also has to be noted that some phoneme frequencies might be dependent on this choice of corpus since differences in lexical items' frequencies can be found (e.g. the use of "tu", used in informal speech, is replaced by "vous" in broadcast speech). At the time of writing, a similar corpus in size with spontaneous speech was not available in the speech community.

The IRISA speech transcription system (Institut de Recherche en Informatique et Systèmes Aléatoires) was used for corpus alignment. Orthographical transcriptions were used by the alignment system to locate phoneme boundaries, to choose among potential pronunciation alternatives, and to discard silences and other noise segments (see Buerki, Gendrot, Gravier, Linares and Fougeron 2008 for further details). The resultant labelling is thus best considered phonemic rather than phonetic. One further note about transcriptions is that French "schwa", i.e. orthographic 'e', is transcribed as [œ] in the corpus.

Using measurements from the above-mentioned corpus, the information content of vowels is measured using Shannon information (or surprisal; Shannon 1948), which more directly reflects the communicative function of being highly frequent than raw frequency counts would. Information content is the negative log probability of frequency, so high frequency corresponds to low information content.

The figure in (2) shows the information content for French vowels with the two phonetic realizations of the default vowel, [œ, ø] indicated by dark bars. Is it noteworthy that [œ], one of the mid front rounded vowels, is indeed of fairly low information content in French, and thus emerges as being a good candidate for epenthesis. Interestingly, [ɛ, e, i], all of which are common "unmarked" epenthetic vowels cross-linguistically, appear alongside [œ] as having relatively low information content. Yet, as these results show, information content alone, when based on unigram token frequency, is not sufficient to predict the default vowel in French given that [ɛ], [e], [i], [a] also have low information content. Indeed, were token frequency the sole factor relevant in predicting the quality of a language's default vowel, we might expect [a] to be the default vowel in French. However, as discussed above, frequency is only one of several factors surmised to be relevant to being a good epenthetic vowel. We now consider contrastiveness.

(2)   Unigram information content for French vowels (negative log probability)
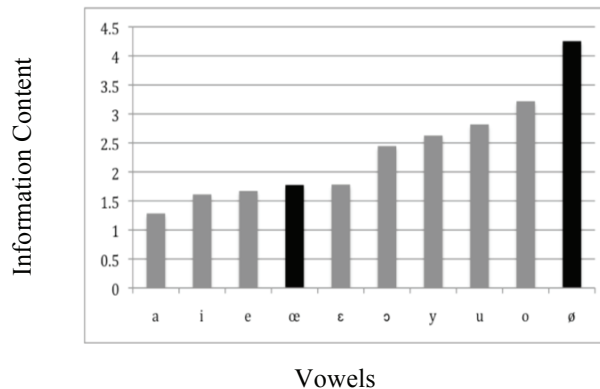
Vowels

3   Weak Contrastiveness

Another property commonly associated with epenthetic vowels is that of weak contrastiveness. With respect to phonological contrastiveness, it has been proposed that default (unmarked) segments lack distinctive feature structure underlyingly (Abaglo and Archangeli 1989; Rice and Avery 1993; Rice & Causley 1998; Causley 1999). In some approaches, phonological contrast is used to determine whether or not feature structure is present (e.g. Rice and Avery 1993, Clements 1988), such that only features that serve to minimally distinguish other sounds in the inventory are specified. An unmarked segment, i.e. one lacking structure, would thus not need to be distinguished from another sound in the inventory by a single feature. However, as later pointed out by Rice (1999), using minimal contrast as a diagnostic for markedness does not work for all languages. French is one such language because [œ, ø] can minimally contrast in roundness, backness, and height with other vowels in the language.

 Nonetheless, since weak contrastiveness seems to characterize epenthetic vowels in some languages, it is worth considering what it is about contrastiveness that could make a vowel be susceptible to epenthesis. As mentioned above, we suggest that there are actually two points that are relevant. First is the concept of phonological contrastiveness, and second, *minimality* of contrast, interpreted here as a measure of perceptual similarity.

3.1   Contrastiveness

First, consider the role that phonological contrast plays in language: it serves as a way to keep words distinct in the lexicon. A vowel that does a lot of work in distinguishing words might, therefore, be a poor candidate for being an epenthetic vowel since it could be detrimental to communicating a message if added material caused a listener to think that a new word had been made. We might then expect that an epenthetic vowel is one that is less contrastive in the system, i.e. it does less work in distinguishing words than other vowels.

 A common way of measuring the work that a particular contrast does in distinguishing words in a language is *functional load* (e.g. Martinet 1955; Hockett 1955, 1966; Surendran and Niyogi 2003; Wedel and Branchaw 2011). This measure, however, indicates how much work a *pair* of sounds does; what is more relevant for predicting the quality of an epenthetic vowel is how much work an *individual* sound does. This is a measure we dub the *relative contrastiveness* of a particular sound, and is essentially equivalent to the average functional load of a segment across all the possible pairs of segments it could occur in.

 To measure relative contrastiveness, we draw on a tool of information theory, *entropy*. Entropy is a measure of the uncertainty associated with selecting among possible outcomes, each occurring with a particular probability. Suppose, for example, that you had to make a guess about which vowel out of all the vowels in

an inventory would occur in a particular word. Entropy provides a measure of how much uncertainty you would have about your guess: the higher the entropy value, the greater the uncertainty about the guess. As shown in (3), entropy is measured as the sum over all possible outcomes of the negative log probability of a given outcome, weighted by the probability of that outcome's occurring. The base of the log is generally taken to be 2 and so entropy is measured in *bits*.

(3)   Entropy: $H = - \sum p_i \log_2 p_i$

We use the term *entropic contribution* to refer to the contribution that any one outcome makes to the total uncertainty, as shown in (4). It can also be calculated by computing the total entropy of a system with and without a particular outcome present in the system, and subtracting the latter from the former.

(4)   Entropic contribution: $H_c = - p_i \log_2 p_i$

To calculate relative contrastiveness, we begin by calculating the entropy of our French corpus, based on the type frequencies of words, with all vowels included. The amount of uncertainty in the system is x bits, meaning that it takes an average of x bits of information, or binary choices, to guess the identity of a particular vowel occurring in a word. With this as a basis, we merge two vowels in the corpus, e.g. [i] and [e], and recalculate the entropy. Note that the overall frequency counts of the corpus stay the same, but the type frequencies of each word will change since the frequencies of any words contrasting for the two vowels will be summed. The entropy of the new system is then subtracted from the entropy of the original system. To give the proportional change in entropy, we divide by the entropy of the old system. This gives the functional load of that particular pair of sounds. These calculations are done for all possible mergers for a given vowel and averaged to get the relative contrastiveness of that vowel. The equation for relative contrastiveness (RC) is given in (5), where H_1 is the entropy of the system with no merger, H_2 is the entropy of the system with a merger, M is the set of all possible mergers, m, involving a particular vowel, and |M| is the cardinality of M. Similar calculations are done for each vowel in the inventory. Vowels that do little work in distinguishing word meaning will have a low relative contrastiveness.
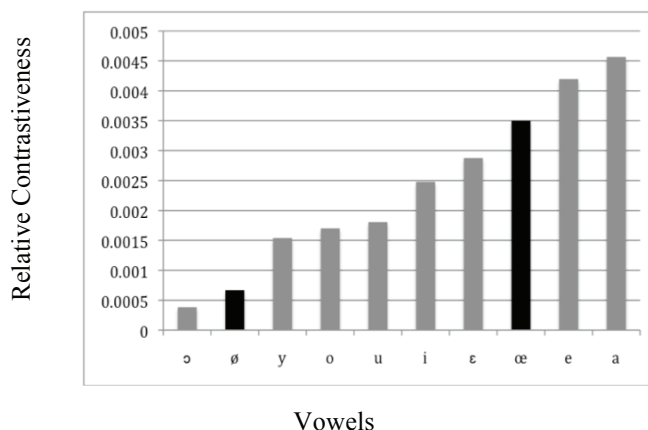
(5)   Relative contrastiveness

$$RC = \frac{\displaystyle\sum_{m \in M} \frac{H_1 - H_2}{H_1}}{|M|}$$

The average relative contrastiveness of a given vowel given its possible vocalic contrasts is shown in (6) below where it can be seen that [ø] has low

relative contrastiveness. Purely from the perspective of relative contrastiveness, then, it is clear that rounded vowels in general and the front rounded vowel [ø] in particular, would be good candidates for being the epenthetic vowel, as they contribute relatively less to making lexical distinctions in French. In particular, both front rounded vowels contribute less than any of the typologically less marked vowels [i, e, ɛ] in CVC sequences, and [ø] contributes less than the unmarked vowels in the word calculations.[3]

(6) Average relative contrastiveness of French vowels

Vowels

3.2   Perceptual similarity

The other aspect of weak contrastiveness that may be relevant is that of perceptual similarity. Pairs of sounds that are "minimally" contrastive are those that differ by exactly one feature and are thus relatively similar. Sounds that are similar to many other sounds in the system would make good candidates for epenthesis because they are less "noticeable" (see, e.g. Battistella 1990, Dupoux et al., 1999, 2011; Rice 2000; Steriade 2009; Riggs, to appear).

We modeled the acoustic distinctiveness of French vowels as a function of miscategorization probability. For this, we make use of the Generalized Context Model (GCM, Nosofsky 1988), a method for assessing categorization patterns, taking into account both acoustic similarity and frequency of occurrence. The assumption is that the more overlap there is in a vowel's acoustic space with those of other vowels in the system, the higher the probability that the vowel will be miscategorized. That is, a high degree of overlap is correlated with poor perceptual distinctiveness. The GCM assumes that categories are mentally represented as

---

[3] It is worth noting that French 'schwa', occurring in common function words such as *je* 'I', *le* 'the, masc.sg.', *que* 'that', is transcribed as [œ] in the corpus and thus the higher relativeness contrastiveness of this vowel is not surprising.

labeled exemplar tokens—in other words, every exemplar consists of a mapping between a category label (e.g. /i/) and a position in a perceptual map (e.g. formant values). In deciding how to map a new percept to a category, the perceptual similarity (π) between that percept and all exemplars in memory is calculated as in (7b), and these similarities are summed for each category. Categories with higher similarity scores are more likely to be identified with the percept. The probability that a percept will be identified with a particular category, $P(C_m|x_i)$, shown in (7c), is then the total similarity score of the percept for that category divided by its total similarity to every category; this general decision algorithm is known as the Luce Choice Rule. The similarity between a percept and an exemplar is calculated from the Euclidean distance (D) in some space, shown in (9a).

(7)　Applying the GCM

a.　Distance (x_i, x_j) = D = $\sqrt{(F1(x_i) - F1(x_j))^2 + (F2(x_i) - F2(x_j))^2 + ...}$

b.　Similarity (x_i, x_j) = π = $e^{(-s \times D)}$

NB: *s* is an empirically determined scaling factor that defines the effective range over which similarity contributes to the outcome; D is the distance measure from (7a).

c.　Probability of Correct Categorization = $P(C_m|x_i)$ = $\dfrac{\sum\limits_{x_j \in C_m} \pi(x_i, x_j)}{\sum\limits_{x_k} \pi(x_i, x_k)}$

NB: C_m is a category labelled *m*; x_i is the percept at hand, x_j are all exemplars stored in category *m*; x_k are all exemplars stored in any category, including category *m*. $\pi(x_i, x_j)$ is the similarity measure from (7b).

Acoustic similarity used as input to the GCM was measured using the first three formant values of the ten oral vowels of 17 native speakers of Continental French from the ESTER corpus. As a starting point (lacking any evidence to do otherwise), we assume an equivalent weighting to information from each formant. Given that each subject has a different vocal tract size and therefore a different set of absolute formant values, probabilities of correct categorizations (PCCs) were calculated separately for each person in the database.

Token frequencies of the ten vowels were also taken from the corpus. To populate the exemplar-based categories that the GCM uses, each person's data was randomly sampled (with replacement) for each vowel a number of times proportional to the frequency of the vowel in the database. The final number of exemplars in each category ranged from ~665 (for [ø]) to ~1290 (for [a]). This represents a near exhaustive sampling from each person's data, as this is several times the number of actual data points for each vowel. The final PCCs are averages over the PCCs from each person's data; the relative PCCs for each person are quite similar to the average.

The figure in (8) shows the results of the GCM (scaling factor = 50). The most confusable oral vowels in the corpus are [ɛ], [e] and [ø]. This indicates that again, [ø] appears among the set of vowels that would be good choices for epenthesis, being a vowel that is relatively similar to other vowels in the system.

We reiterate that it is not the case that this one characteristic by itself uniquely determines the epenthetic vowel in French; rather, we can use it in combination with the other measures to see a trend appearing. In terms of relative contrastiveness, [ø] was a good candidate along with other round vowels; typologically unmarked vowels [e, ɛ, i] were not good candidates. In terms of acoustic similarity, [ø] is again a good candidate, this time along with [e] and [ɛ]. On average, then, [ø] is emerging as a top candidate looking across the different possible desirable characteristics. Further, in terms of frequency, [œ], the other epenthetic vowel, emerged as a good candidate for epenthesis along with [a, i, e, ɛ] (for more on why [ø] and [œ] might be patterning together, see §5.)

(8) Correct categorization probability of French vowels
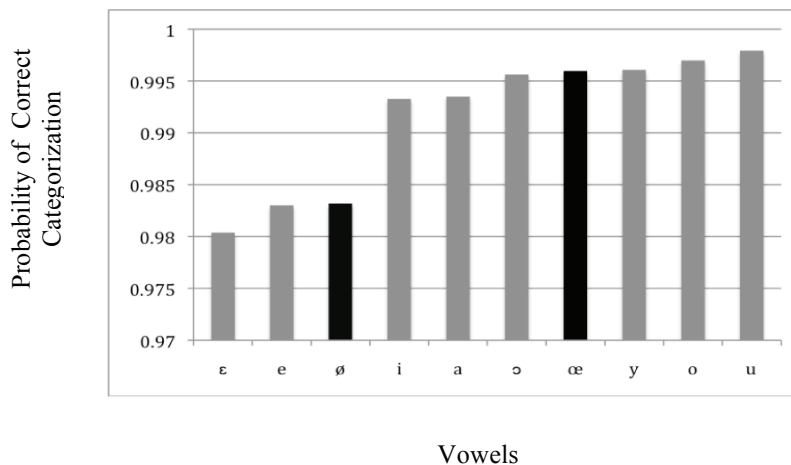
Vowels

4   Distribution

In this section, we consider the distribution of vowels across the lexicon as a means of predicting the quality of the epenthetic vowel. It has been proposed that the unmarked segment (or feature) in a language is more widely distributed than its marked counterpart (see, e.g., Trubetzkoy 1939; Hockett 1955; Greenberg 1966; Battistella 1990; Stemberger 1992; for related discussion, see Rice 1999). Being widely distributed would be a desirable property for an epenthetic vowel when we consider the function of epenthesis; it is important for the selected vowel to be able to break up a broad range of phonotactically illicit sequences. Conversely, a vowel with a narrow distribution in the language would be relatively unexpected in certain contexts and potentially hinder processing or be difficult to

produce in the context (note the similarity between distribution and frequency in this regard).

In traditional phonological accounts, a sound's distribution is assessed by identifying all the different environments in which the sound may occur, with each context having the same weight as all others. Information theory provides a tool that allows for a more nuanced approach to discovering the distribution of sounds, again using the concept of entropy. Recall from the discussion of relative contrastiveness that entropy is a measure of the uncertainty associated with the selection of one out of a set of possible candidates. We can conceptualize the phonological distribution of a particular vowel as the degree of uncertainty about which other segments that vowel can co-occur with. For example, a vowel with a very limited distribution—occurring, say, next to only one or two consonants—will be associated with a very low uncertainty about which consonant it occurs with in any particular word, while a vowel with a wide distribution, occurring next to many consonants, will be associated with a higher uncertainty. Thus, we can measure the entropy of the set of consonants that each vowel occurs adjacent to (which we term that vowel's *distribution entropy*) in order to get a measure of the width of that vowel's distribution. With respect to epenthesis, then, a vowel with a high relative entropy value is one that can occur in a wide distribution and thus would be better as a candidate for epenthesis than a vowel with a low relative entropy value.

The distribution entropy of a vowel can be calculated for both the set of consonants that precedes it (CV) and the set of consonants that follows it (VC) and then averaged to get the overall relative entropy measure for that vowel. The distribution entropy for CV is calculated as in (9), where V is the vowel in question and C is the set of consonants that precedes that vowel; p(c) is the probability of a particular consonant occurring in a CV context, relative to the other possible consonants that could appear in that context. The distribution entropy of VC is calculated similarly.

(9)   Distribution Entropy

$$H_V = -\sum_{c \in C|CV} p(c) \log_2(p(c))$$

Note that this measure is based on the probability of occurrence of each particular consonant. These probabilities can be calculated in several different ways. One way, which is most similar to the standard phonological interpretation of distribution, would be to assign each consonant that can occur in the relevant position in at least one word of the language the same probability as every other consonant that can appear in that position. We term this calculation the *type occurrence* of each consonant, and it is essentially a categorical measure of the distribution of the consonants adjacent to each vowel across the lexicon of the language. (10) shows the average CV and VC distribution entropy values for French vowels based on type occurrence.

From a distribution perspective, vowels at the left end have a broader distribution and thus a higher degree of uncertainty than those at the right end. Five vowels appear to the right end of the scale: the typologically unmarked front unrounded vowels, [i, e, ɛ], the low vowel [a], and the mid front rounded vowel [œ]. Insofar as distribution is a relevant factor in predicting the quality of the epenthetic vowel, any of these five vowels would make good candidates. Of particular relevance for this study is the observation that the mid front rounded vowel [œ] is included in this group.

(10)   Average CV/VC distribution entropies in French (type occurrences)
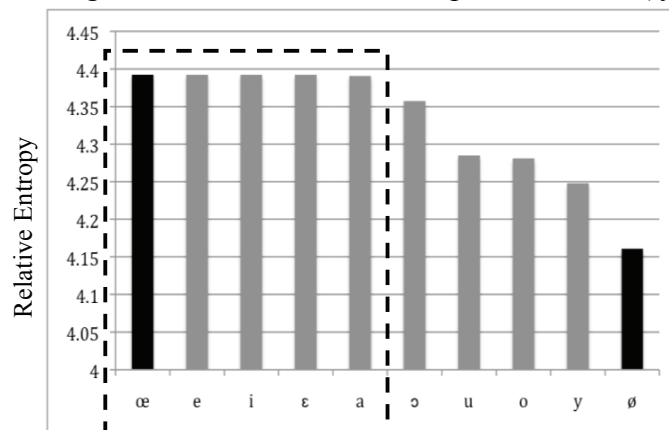
5   Why both [ø] and [œ]?

One point that we have not yet addressed concerns the observation that the quality of the epenthetic vowel is variable between the closed, tense [ø] and more open, lax [œ]. In the criteria used above, [ø] often appears as the best or one of the best candidates for epenthesis, but in the cases of frequency and relative entropy, [œ] was a better candidate than [ø]. Why these specific vowels are variable with each other is an interesting question that most likely relates to their phonetic similarity.

In addition to the inherent nature of the sounds, however, the phonological relationship between sounds can also impact their similarity: two sounds that are allophonic in a language are in many cases perceived as being more similar than two sounds that are contrastive (e.g. Jaeger 1980; Ohala 1982; Dupoux et al. 1997; Harnsberger 2001; Peperkamp et al. 2003; Kazanina et al. 2006; Pruitt et al. 2006; Boomershine et al. 2008). It has been hypothesized that the reason for this difference in perception is linked to the predictability of the distribution of such sounds (Hall 2009, 2011). Specifically, sounds that are more predictably distributed (e.g. in complementary distribution, as is the case for allophony) are perceived as more similar than ones that are less predictably distributed (perhaps because being predictably distributed would mean that the acoustic cues to differentiating the sounds are less important for their identification). Interestingly, this analysis extends to pairs of sounds that are somewhere between perfectly

predictable distribution and perfectly unpredictable distribution. For example, a pair of sounds that is generally contrastive (unpredictably distributed) but neutralized in some context (predictable in that context) tends to be perceived as being more similar than a pair of sounds that is contrastive in all contexts (e.g. Trubetzkoy 1969 [1939]; Hume & Johnson 2003).

The precise predictability of distribution of two sounds can be measured using information-theoretic tools (Hall 2009). This can be conceptualized as a measure of entropy; here, the system whose uncertainty is being measured consists of two sounds, A and B, and so entropy will range from 0 to 1. An entropy of 0 indicates that there is no uncertainty about the choice between the two sounds (they are perfectly predictably distributed), while an entropy of 1 indicates that there is maximal uncertainty about the choice (they are perfectly contrastively distributed). The entropy is calculated as a function of either the type or the token frequency of occurrence of each of the two sounds in question in all of the environments that at least one of the two can occur in, weighted by the frequency of occurrence of those environments.

In the case of [ø] and [œ], this measure shows that there is an entropy of 0.5 (type-based) or 0.59 (token-based) between these sounds in the French system. This means that these two vowels are squarely in the middle of the continuum between predictably and unpredictably distributed. These numbers reflect the observation that they are generally in complementary distribution with the tense [ø] occurring in open syllables, e.g. *peu* [pø] 'few', and lax [œ] occurring in closed syllables, e.g. *peur* [pœʁ] 'fear', though there are some exceptions to this otherwise regular distribution where the vowels contrast, *jeûne/jeune* [jøn]/[jœn] 'fasting/young', *veûle/veulent* [vøl]/[vœl] 'spineless/they want'. Further, word-internally, there is a great deal of variability, much of which may be dictated by vowel-harmonic assimilation, such that non-word-final mid vowels assimilate in tenseness to the stressed (final) vowel of the word (Fagyal et al. (2006), e.g. *abr<u>eu</u>voir* [abʁœvwaʁ] 'trough' vs. *abreuvée* [abʁøve] 'watered'. Given their distribution patterns, we might therefore expect that they would be perceived as being similar to one another, and thus prone to confusability and variability.

6   Conclusion

In this paper we have attempted to situate the properties associated with French epenthetic vowels with more typologically common epenthesis patterns, by considering the role of epenthesis in a system of communication. We have presented a number of arguments for thinking that the front rounded vowels in French make good candidates for being the epenthetic vowel in that language, given (a) the desiderata of any epenthetic vowel in any language, and (b) the ways in which the front rounded vowels line up with those desiderata in French. Specifically, at least one of the mid front rounded vowels in French was shown to be among the most frequent, the least lexically contrastive, the most perceptually

similar, and the most widely distributed vowels in French, all characteristics that make such vowels good choices for being epenthesized. No single characteristic points to the front rounded vowels as being the best, but these vowels do consistently emerge as good candidates, while other possible candidates are good matches for one criterion but poor matches for another. Thus, despite being typologically marked, [ø] and [œ] seem to be exactly the vowels we should *expect* to see as the epenthetic vowels in French, given the ways in which they pattern in the system. Furthermore, the fact that there is variability between the two vowels is to be expected given both their phonetic similarity and their relatively predictable phonological patterning with respect to one another.

In addition to examining the predictors of epenthesis, we have explored ways of measuring these diagnostics. We find tools from Information Theory, especially information content and entropy, to be particularly promising for quantifying properties that are well established in the phonological literature.

References

Abaglo, Poovi, and Diana Archangeli. 1989. Language-Particular Underspecification: Gengbe /e/ and Yoruba /i/. *Linguistic Inquiry* 20 (3):457-480.

Adda-Decker, Martine, Philippe Boula de Mareuil, and Lori Lamel. 1999. Pronunciation Variants in French: Schwa and Liaison. *Proceedings of ICPhS*. 2239-2242.

Archangeli, Diana. 1984. Underspecification in Yawelmani phonology and morphology. Ph.D. dissertation, MIT.

Battistella, Edwin L. 1990. *The Evaluative Superstructure of Language*. Albany: State University of New York Press.

Battistella, Edwin L. 1996. *The Logic of Markedness*. Oxford: Oxford University Press.

Blevins, Juliette. 2004. *Evolutionary Phonology: The Emergence of Sound Patterns*. Cambridge: Cambridge University Press.

Boomershine, Amanda, Kathleen Currie Hall, Elizabeth Hume, and Keith Johnson. 2008. The Influence of Allophony vs. Contrast on Perception: The Case of Spanish and English. In P. Avery, B. E. Dresher and K. Rice, eds., *Contrast in Phonology: Perception and Acquisition*, 145-171. Berlin: Mouton.

Buerki, A., Gendrot, C., Gravier, G., Linares, G., & Fougeron, C. (2008). Alignement automatique et analyse phonétique: Comparaison de différents systèmes pour l'analyse du schwa. *Traitement Automatique des Langues* 49:165–197.

Bybee, Joan. To appear. Markedness: Iconicity, Economy, and Frequency. In J. J. Song, ed., *The Oxford Handbook of Language Typology*. Oxford: Oxford University Press.

Causley, Trisha. 1999. *Complexity and Markedness in Optimality Theory*. PhD thesis, University of Toronto.

Chomsky, Noam, and Morris Halle. 1968. *The Sound Pattern of English*. New York: Harper & Row.

Clements, G. Nick. 1988. Towards a substantive theory of feature specification. In J. Blevins and J. Carter, eds., *Proceedings of NELS 18*, 79-93. Amherst, MA: GSLA.

Côté, Marie-Hélène, and Geoffrey Stewart Morrison. 2007. The Nature of the Schwa/Zero Alternation in French Clitics: Experimental and Non-Experimental Evidence. *French Language Studies* 17:159-186.

de Lacy, Paul. 2006. *Markedness: Reduction and Preservation in Phonology*. Cambridge: Cambridge University Press.

Dupoux, Emmanuel, Christophe Pallier, N. Sebastian, and J. Mehler. 1997. A Destressing 'Deafness' in French? *Journal of Memory and Language* 36 (3): 406-421.

Dupoux, Emmanuel, Erika Parlato, Sonia Frota, Yuki Hirose, and Sharon Peperkamp. 2011. Where Do Illusory Vowels Come From? *Journal of Memory and Language* 64:199-210.

Eddington, David. 2001. Spanish epenthesis: Formal and performance perspectives. *Studies in the Linguistic Sciences* 31: 33-53.

Fagyal, Zsuzsanna, Douglas Kibbee, and Fred Jenkins. 2006. *French: A Linguistic Introduction*. Cambridge: Cambridge University Press.

Féry, Caroline. 2003. Markedness, Faithfulness, Vowel Quality, and Syllable Structure in French. *Journal of French Language Studies* 13 (2):247-280.

Fougeron, Cecile, Cedric Gendrot, and A. Bürki. 2007. On the Phonetic Identity of French Schwa Compared to /ø/ and /oe/. Paper read at 5emes Journees d'Etudes Linguistiques (JEL), at Nantes, France.

Galliano, S., Geoffrois, E., Mostefa, D., Choukri, K., Bonastre, J. -F., & Gravier, G.(2005). ESTER Phase II evaluation campaign for the rich transcription of French Broadcast News. In *Proceedings of Interspeech*. Lisbon, Portugal: 1149–1152.

Grammont, Maurice. 1913. *Le Vers Français, Ses Moyens D'expression, Son Harmonie*. Paris: H. Champion.

Hall, Kathleen Currie. 2009. *A Probabilistic Model of Phonological Relationships from Contrast to Allophony*. PhD thesis, The Ohio State University, Columbus, OH.

Hall, Kathleen Currie. 2011. Cognitive Reflexes of Linguistic Sound Patterns. University of Canterbury, Christchurch, NZ.

Hall, Nancy. 2011. Vowel Epenthesis. In Marc van Oostendorp, Colin Ewen, Elizabeth Hume and Keren Rice, *Companion to Phonology*. Oxford: Wiley-Blackwell 1576-1596.

Harnsberger, James. 2001. The Perception of Malayalam Nasal Consonants by Marathi, Punjabi, Tamil, Oriya, Bengali, and American English Listeners: A Multidimensional Scaling Analysis. *Journal of Phonetics* 29 (3):303-327.

Hockett, Charles F. 1955. A Manual of Phonology. *International Journal of American Linguistics* 21 (4).

Hockett, Charles F. 1966. The Quantification of Functional Load: A Linguistic Problem. *U.S. Air Force Memorandum RM-5168-PR*.

Hume, Elizabeth. 2004. Deconstructing Markedness: A Predictability-Based Approach. *Proceedings of the Annual Meeting of the Berkeley Linguistics Society* 13:182-198.

Hume, Elizabeth. 2008. Markedness and the Language User. *Phonological Studies* 11:295-310.

Hume, Elizabeth, and Ilana Bromberg. 2005. Predicting Epenthesis: An Information-Theoretic Account. Paper read at 7th Annual Meeting of the French Network of Phonology, at Aix-en-Provence.

Hume, Elizabeth, and Keith Johnson. 2003. The Impact of Partial Phonological Contrast on Speech Perception. *Proceedings of the Fifteenth International Congress of Phonetic Sciences*.

Hume, Elizabeth, and Frédéric Mailhot. To appear. The role of entropy and surprisal in phonologization and language change. In A. Yu (ed.), *Origins of Sound Patterns: Approaches to Phonologization*. Oxford University Press.

Jaeger, Jeri J. 1980. Testing the Psychological Reality of Phonemes. *Language and Speech* 23:233-253.

Jenkins, Fred. 1971. The Phonetic Value of Mute-e. *The French Review* 45 (1):82-87.

Kazanina, Nina, Colin Phillips, and William J. Idsardi. 2006. The Influence of Meaning on the Perception of Speech Sounds. *Proceedings of the National Academy of Sciences of the United States of America* 103 (30):11381-11386.

Kuipers, Cecile, Wilma van Donselaar, and Anne Cutler. 1996. Phonological Variation: Epenthesis and Deletion of Schwa in Dutch. *Proceedings of the Fourth International Conference on Spoken Language Processing* 1:94-97.

Landick, Marie. 1995. The Mid-Vowels in Figures: Hard Facts. *French Review* 68 (1):88-102.

Lass, Roger. 1975. How Intrinsic Is Content? Markedness, Sound Change, and 'Family Universals'. In D. Goyvaerts and G. Pullum, eds., *Essays on the Sound Pattern of English*, 465-504. Ghent: E. Story-Scientia.

Martinet, André. 1955. *Économie Des Changements Phonétiques* Bern: Francke.

Noske, Roland. 1993. *A Theory of Syllabification and Segmental Alternation. With Studies on the Phonology of French, German, Tonkawa and Yawelmani*. Tübingen: Max Niemeyer.

Nosofsky, Robert M. 1988. Similarity, Frequency, and Category Representations. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 14 (1):54-65.

Ohala, John J. 1982. The Phonological End Justifies Any Means. *Proceedings of the 13th International Congress of Linguistics*:232-243.

Peperkamp, Sharon, Michèle Pettinato, and Emmanuel Dupoux. 2003. Allophonic Variation and the Acquisition of Phoneme Categories. In M. Kehoe et al. eds., *Proceedings of the 27th Annual Boston University Conference on Language Development*, 650-661. Somerville, MA: Cascadilla Press.

Pruitt, John S., James J. Jenkins, and Winifred Strange. 2006. Training the Perception of Hindi Dental and Retroflex Stops by Native Speakers of American English and Japanese. *Journal of the Acoustical Society of America* 119 (3):1684-1696.

Pulleyblank, Douglas. 1988. Underspecification, the Feature Hierarchy, and Tiv Vowels. *Phonology* 5:299-326.

Rice, Keren. 1999/2000. Featural Markedness in Phonology: Variation. *GLOT International* 4.7, 4.8:3-6, 3-7.

Rice, Keren. 2007. Markedness in Phonology. In P. de Lacy, ed., *The Cambridge Handbook of Phonology*, 79-97. Cambridge: Cambridge University Press.

Rice, Keren, and Peter Avery. 1993. Segmental Complexity and the Structure of Inventories. *Toronto Working Papers in Linguistics* 12:131-153.

Rice, Keren, and Trisha Causley. 1998. Asymmetries in Featural Markedness: Place of Articulation. In *21st Generative Linguistics in the Old World Conference*. Tilburg University.

Riggs, Daylen. To appear. Minimal Salience and the Quality of Epenthetic Vowels in Loanwords. *Proceedings of the 41st Meeting of the North East Linguistic Society*.

Stemberger, Joseph Paul. 1992. A Connectionist View of Child Phonology: Phonological Processing without Phonological Processes. In C. A. Ferguson, L. Menn and C. Stoel-Gammon, eds., *Phonological Development: Models, Research, Implications*, 165-189. Parkton, MD: York Press.

Steriade, Donca. 2009. The Phonology of Perceptibility Effects: The P-map and Its Consequences for Constraint Organization. In K. Hanson and S. Inkelas, eds., *The Nature of the Word: Studies in Honor of Paul Kiparsky*, 151-180. Cambridge, MA: MIT Press.

Surendran, Dinoj, and Partha Niyogi. 2003. Measuring the Functional Load of Phonological Contrasts. Available: http://arxiv.org/pdf/cs.CL/0311036.

Trubetzkoy, Nikolai Sergeevich. 1969. *Principles of Phonology*. Translated by C. A. M. Baltaxe. Berkeley: University of California Press. Original edition, 1939.

Wedel, Andrew, and Sherrylyn Branchaw. 2011. Detection of Statistical Relationships between Measures of Functional Load and Probability of Phoneme Merger. Paper read at Linguistic Society of America Annual Meeting, at Pittsburgh, PA.

Elizabeth Hume
The Ohio State University
Department of Linguistics
1712 Neil Ave.
Columbus, OH 43202

evhume@gmail.com

Andrew Wedel
University of Arizona
Department of Linguistics
P.O. Box 210028
Tucson, AZ 85721

wedel@u.Arizona.edu

Martine Adda-Decker
LIMSI - CNRS
B.P. 133
91403 ORSAY CEDEX
FRANCE

madda@limsi.fr

Kathleen Currie Hall
College of Staten Island
Department of English, 2S-218
2800 Victory Blvd.
Staten Island, NY 10314

kathleen.hall@csi.cuny.edu

Adam Ussishkin
University of Arizona
Department of Linguistics
P.O. Box 210028
Tucson, AZ 85721

ussishki@u.Arizona.edu

Cédric Gendrot
LIMSI - CNRS
B.P. 133
91403 ORSAY CEDEX
FRANCE

cgendrot@univ-paris3.fr