Perceptual Salience in English Reduplication

C. Scott Ananian

May 2001

Contents

1 Introduction

Reduplication in English, although not fully productive, is, however, more common than usually supposed. Words such as "backpack," "chitchat," "cookbook," "crackerjack," "criss-cross," "grandstand," "hobnob," "hodgepodge," and "humdrum" (to name just a few) illustrate the language's tendency toward reduplicative compounds. Recent work by Moira Yip (1999) points to constraints mandating such rhyme as the causatory agent of reduplication across languages. Yip also shows, in Yip (1998), how identity avoidance can motivate the variation seen in the reduplicant.

Yip (following McCarthy and Prince 1997) claims that the unfaithful segments in the reduplicant will allow the emergence of the unmarked, and that markedness constraints are the primary factor in the selection of such segments. In contrast, Steriade (2001) claims that an underlying perceptual map (P-map) is responsible for resolving phonotactic violations (such as identity avoidance). In Steriade's view, the unfaithful segments will still attempt to remain perceptually similar to their base segment, via an ordering of correspondence constraints dictated by the P-map.

This report will attempt to discriminate between these two theories of reduplication using computer analysis of English language texts. One set of programs will extract rhyming pairs from a large database of poetry, and use these rhymes to compute (an approximation to) the P-map, according to claims made by Steriade. Another set will identify examples of English word reduplication from entries in a pronouncing dictionary. Comparing the P-map-driven word changes in the rhyming pairs to the internal variation in the reduplicants will, we claim, allow us to determine whether the reduplication process is relying on the P-map, as Steriade would claim, or on underlying markedness and simpler faithfulness constraints, as Yip and McCarthy would claim.

We will note that there is also a null hypothesis being tested: it is possible that the process of reduplication relies on neither markedness nor a perceptual map. The variation may be completely random, it may rely on innate properties of sounds,[1] or some other explanation. Careful examination of the data will show the extent to which the null hypothesis is supported.

The paper is organized as follows: in the next section we will review various analyses of reduplication within Optimality Theory, roughly chronologically. We will start with the original analysis of McCarthy and Prince (1993) in section 2.1, explaining how it obtains the *emergence of the unmarked*. In sections 2.2 and 2.3 we will turn to the reformulation by Yip (1999), in which the influences of rhyme, alliteration, and identity avoidance are shown. Yip's analysis also predicts an emergence of the unmarked, but provides a more reasonable model for describing the non-productive English reduplication data which we will be examining. In section 2.4, we will discuss the perceptual maps of Steriade (2001), and their expected effect on reduplication behaviors.

[1] It is often claimed that [i] has an "inherent" meaning of "small," for example (Jespersen 1933; Taylor and Taylor 1962), in words like *tick-tock*.

Section 3 will shift to our experimental technique, also describing the types of English data we will be interested in. In section 4 we will present and analyse our results, and in section 5 we will draw some conclusions about the behavior of reduplication in English.

2 Optimality Theoretic Analyses of Reduplication

The first treatment of reduplication using optimality theory appeared in McCarthy and Prince (1993). As described in McCarthy and Prince (1997), a unified theory of correspondence and base-reduplicant identity was motivated by several key parallels between reduplication and standard input-output processes:

- **Completeness of mapping.** Total reduplication corresponds to an output identical to the input; partial reduplication corresponds to phonological deletion.
- **Dependence on input (base).** Fixed (non-dependent) segments in the reduplicant correspond to epenthetic (non-dependent) features in the output.
- **Contiguity of mapping.** The reduplicant is usually a contiguous substring of the base; likewise the output is usually a contiguous substring of the input.
- **Linearity of mapping.** The linear order of the base is usually preserved in the reduplicant, and of the input in the output. However, (limited) metathesis of each can occur in certain circumstances.
- **Anchoring of edges.** Alignment constraints are well known in input-output correspondence; reduplicants likewise usually contain elements from one edge of the base.
- **Featural identity.** Input-output and base-reduplicant segments are normally identical to each other, but may differ for phonological reasons.

Motivated by these correspondences, McCarthy and Prince proposed the "Full Model" of base-reduplicant correspondence, with the three bidirectional correspondence relations illustrated by the following diagram:

(1)
$$\begin{tikzcd} \text{Input:} & /\text{Affix}_{\text{RED}} & + & \text{Stem}/ \arrow[dll, shift right, "\textit{I-R Faithfulness}"'] \arrow[d, shift left, "\textit{I-B Faithfulness}"] \\ \text{Output:} & \text{R} \arrow[urr, shift right] \arrow[rr, shift left] & & \text{B} \arrow[ll, shift left, "\textit{B-R Identity}"] \arrow[u, shift left] \end{tikzcd}$$

In actual practice, however, *I-R Faithfulness* plays a subsidiary role due to a "universal meta-condition on ranking" which ensures that faithfulness on the stem domain always dominates faithfulness on the affixal domains. Thus the "basic model" of base-reduplicant correspondence is often used (and will be used in this paper):

(2)
$$\begin{tikzcd} \text{Input:} & /\text{Affix}_{\text{RED}} & + & \text{Stem}/ \arrow[d, shift left, "\textit{I-B Faithfulness}"] \\ \text{Output:} & \text{R} \arrow[rr, shift left] & & \text{B} \arrow[ll, shift left, "\textit{B-R Identity}"] \arrow[u, shift left] \end{tikzcd}$$

The base-reduplicant identity correspondence relations were set up to parallel the input-output relations, resulting in a six-member family of constraints (from McCarthy and Prince (1997)):

- The Max constraint family.

  (3) Max-BR

  Every segment of the base has a correspondent in the reduplicant.
  (Reduplication is total.)

  (4) Max-IO

  Every segment of the input has a correspondent in the output.
  (No phonological deletion.)

- The Dep constraint family.

  (5) Dep-BR

  Every segment of the reduplicant has a correspondent in the base.
  (Prohibits fixed segments in the reduplicant.)

  (6) Dep-IO

  Every segment of the output has a correspondent in the input.
  (Prohibits phonological epenthesis.)

- The Ident constraint family.

  (7) Ident-BR

  Reduplicant correspondents of a base [γF] segment are also [γF].

  (8) Ident-IO

  Output correspondents of an input [γF] segment are also [γF].

This approach to reduplication taken by McCarthy and Prince fits well with a view of optimality theory based on essential conflicts between markedness constraints $\mathbb{M}$ and faithfulness constraints $\mathbb{F}$. In particular, reduplicants are postulated to have *exactly the same markedness constraints* as are operative in the base; the only difference between base and reduplicant is the different set of faithfulness constraints active. This allows for the *emergence of the unmarked.*

2.1 Emergence of the unmarked

"Emergence of the unmarked" refers to the emergence of phonologically unmarked structure in reduplicated forms, even though it is not required in the language as a whole. The following skeletal ranking in the system of McCarthy and Prince is sufficient to produce this effect:

(9) $\{$I-O FAITHFULNESS$\} \gg$ PHONOLOGICAL-CONSTRAINT $\gg$ B-R IDENTITY, $\{$*M$\}^2$

Because every relevant I-O faithfulness constraint dominates the phonological constraint, its effects are not seen in the language generally. The phonological constraint can never compel a violation of input-output faithfulness. However, it *can* compel a violation of B-R Identity, to force the reduplicant to obey this constraint *even though the base does not.*

As an illustrative example, we will consider a case from the Philippine Austronesian language Balango (McCarthy and Prince 1997). The Balangao reduplicant copies the first two syllables of the base, minus the final coda:

(10) /RED-tagtag/ → *tagta-tagtag*

The language as a whole possesses final coda consonants, but the constraint against them emerges in reduplication.

/RED-tagtag/	MAX-IO	NO-CODA	MAX-BR
a. tag.ta.-tag.ta.	*!	**	
b. tag.tag.-tag.tag.		****!	
☞ c. tag.ta.-tag.tag.		***	*

The tableau shows how MAX-IO must dominate NO-CODA to prevent deletion of the final consonant of the base, and how NO-CODA must dominate MAX-BR in order to properly delete the final consonant of the reduplicant. Of course, this is not a complete tableau—no constraint is shown, for example, which will prevent us from losing the medial [g] from the reduplicant. McCarthy and Prince invoke the constraints CONTIG-BR, LINEARITY-BR, and UNIFORMITY-BR to complete the picture, but these details are not relevant for the present investigation.

The crucial point is that ranking I-O correspondence above B-R correspondence allows structural constraints to be active on the reduplicant which are not active on the base.[3] This ought to produce more *unmarked structure* in the reduplicant. In sections 3 and 4 we will attempt to experimentally confirm this prediction.

[2]Following McCarthy and Prince (1997), we use {*M} here to denote "every relevant structural constraint that militates against the desired Phonological Constraint."

[3]The opposite ranking (B-R correspondence over I-O correspondence) produces "over-application," where exact base-reduplicant correspondence is attained even if additional (otherwise unmotivated) structural constraints must be applied to the base in order to do so. For our present investigation, it suffices that *additional structural constraints apply to reduplicated forms* (even if such constraints also spread to the base), such that the net result is (in any case) more highly *unmarked* structure in reduplicants.

2.2 The RHYME and ALLITERATE constraints

The reduplication analysis of McCarthy and Prince requires an abstract input morpheme, RED, which upon passage through GEN results in a set of output candidates where RED is realized as a full or partial copy of the base.[4] Yip's (1999) analysis replaces the abstract affix with RHYME and ALLITERATE constraints. She suggests that the real aim of reduplication is to produce sequences that rhyme and alliterate, rather than to realize some abstract affix. She defines her constraints as:

(11) ALLITERATE (Yip 2000)

The output must contain at least one pair of adjacent syllables with identical onsets.

(12) RHYME (Yip 2000)

The output must contain at least one pair of adjacent syllables with identical rhymes.

Input segments may have two output correspondents (violating INTEGRITY) under the influence of these constraints; there is no privileged notion of "base" or "affix" in the traditional sense. For an input /pati/ there will be two outputs which fully satisfy both RHYME and ALLITERATE, [*pa*-pati] and [pati-*ti*], analogous to the processes occuring in Tagalog ([pag-*la*-lakad]) and Chamorro ([bunita-*ta*]), respectively.

Markedness constraints such as *COMPLEX-ONSET or NO-CODA interacting with the RHYME and ALLITERATE constraints may produce "undercopying," preventing whole syllables from being copied, as in Tagalog [*ta*-trabaho]. Likewise, generalizing RHYME and ALLITERATE to operate on different prosodic units (feet, stressed syllables) can produce larger reduplicants or target different elements in the input.

The following simple example shows how total reduplication of the input /$b_1u_2i_3$/ is obtained, perfectly satisfying RHYME and ALLITERATE (while violating INTEGRITY):

/$b_1u_2i_3$/	MAX-IO	RHYME	ALLITERATE	IO-INTEGRITY
☞ a. $b_1u_2i_3$ $b_1u_2i_3$				***
b. $b_1u_2i_3$ u_2		*!	*!	*
c. b_1u_2 b_1u_2	*!			**

Yip adheres to the general schema for obtaining the emergence of the unmarked:

(13) FAITH-IO ≫ MARKEDNESS ≫ FAITH-BR

but she replaces FAITH-BR with positional markedness constraints, since there is no longer any notion of base or affix.

As an example of the emergence of the unmarked within her system, she presents the following tableau for the input /bui/:

[4] Diagrams (1) and (2) mentioned this abstract affix, although we didn't comment specifically on it at the time.

/bui/	MAX-IO	RHYME	*LABIAL	ALLITERATE
☞ a. bui lui			*	*
b. bui bui			**!	
c. lui lui	*!			

Introducing the constraint *LABIAL above ALLITERATE forces the replacement of /b/ with the less-marked /l/. The MAX-IO constraint ensures that we have *at least one* correspondent for the input segment /b/, preventing the replacement of both instances of /b/. Positional faithfulness (Beckman 1996) ensures that the word-initial /b/ is preferentially maintained, blocking the additional candidate *[lui bui].

The predictions of this theory with regard to the emergence of the unmarked are essentially the same as for McCarthy and Prince (1997). Markedness constraints which outrank RHYME or ALLITERATE but are outranked by MAX-IO will only affect reduplicated forms. However, we can now draw on positional faithfulness to mediate between "base" and "reduplicant"; according to Yip, the positionally-more-salient features will appear as the "base," i.e. maintain stronger identity. In the section 2.4 we will show how one might increase the work done by "salience" still more.

2.3 Identity avoidance

In an earlier study (Yip 1998), Yip described the common linguistic preference for and avoidance of identical repetition as a conflict between two opposed constraints, REPEAT and *REPEAT. The REPEAT constraint requires identical structure, and for our purposes is subsumed by the finer-grained RHYME and ALLITERATE constraints of Yip (2000). The *REPEAT constraint penalizes identical structures. Yip shows these constraints at work in examples from Mandarin, English, Classical Greek, Hindi, and Javanese. She argues that the English "echo" words,[5]

	table	shmable
	book	shmook
(14)	fantastic	shmantastic
	apple	shmapple
	strike	shmike

are an example of "perfect" reduplication being defeated by the *REPEAT constraint, which forces some slight flaw to defeat total identity.[6]

A more detailed example comes from Turkish, which reduplicates the first CV of an adjective to form an emphatic form. The CV addition is followed by a coda consonant from the set /p,s,m,r/, subject to the condition that this coda consonant cannot be "too similar" (and

[5] Data from Kenstowicz (1994).

[6] Alderete *et al.* (1997) further claims that this is an example of melodic overwriting, which is why the fixed segments [šm] are not unmarked.

certainly not identical!) to any consonant in the base.

(15)

a.	kap-kara	'jet black'		ap-aik	'wide open'
	cep-cevre	'very much around'		sap-sari	'fully yellow'
b.	sim-siki	'extremely tight'		bem-beyaz	'snow white'
	göm-gök	'sky-blue'		bum-burusuk	gloss unknown
c.	kas-kati	'extremely hard'		bes-belli	'unmistakably obvious'
d.	ter-temiz	'spotless'		sir-siklam	'wet through'
	tor-top	'fully round'			

The precise choice of consonant depends on many factors: for example, coda consonants and close consonants exert more influence than more distant ones. There is some degree of freedom involved, but it is clear that avoidance of repetition is the primary constraint.

The constraints in this system are very similar to those at work in some of the English data we will examine: reduplicative compounds of the type *mish-mash*, *criss-cross*, and *tut-tut*. As with the Turkish data, there is some amount of freedom involved: usually one half of the compound is in the lexicon, and the other is modified in some way to satisfy *Repeat. Yip's environment for reduplication seems to fit our case much better than positing an underlying abstract RED morpheme, as per McCarthy and Prince: it appears obvious that her Rhyme, Alliterate, and *Repeat constraints are driving the process of word formation. What is less clear is the system of constraints constraining the mutations forced by *Repeat.

2.4 Perceptual salience

Donca Steriade (2001) proposes a method to address just such uncertainties. She notes that the repair strategies invoked in response to constraint violations (such as the violation of *Repeat, above) display much less variation than the present Optimality Theoretic mechanism permits.[7] She proposes that a distinct grammatical component, the "P-map," projects correspondence constraints and determines their ranking. Her claim is that a more rigid faithfulness constraint ranking made in accordance with the absolute and relative perceptibilities of various phonological contrasts (as encoded in the P-map), would serve to capture all known linguistic behavior, without the wide range of unattested repair strategies allowed under the current theory.

[7] She calls this the "Too-Many-Solutions" problem.

To illustrate, Steriade advances the following hypothetical fragment of a P-map:[8]

(16)

Obstruent Voicing	**V_V**	**C_V**	**V_R**	**V_]**	**V_T**	**C_T**
p/b	p/b	p/b	p/b	p/b	p/b	p/b
t/d	t/d	t/d	t/d	t/d	t/d	t/d
k/g	k/g	k/g	k/g	k/g	k/g	k/g
s/z	s/z	s/z	s/z	s/z	s/z	s/z

In the P-map as diagramed, every row corresponds to some contrast (in this diagram, rows correspond to contrasts in obstruent voicing) and the columns correspond to environments where that contrast may occur. The relative size of the entry in each cell refers to the relative *distinctiveness* of the contrast. This fragment is thus taken to mean that obstruent voicing contrasts are most distinctive intervocalically, less distinctive word-finally, and least distinctive when preceded by a consonant and followed by an obstruent. Each row is identical, meaning that in this case the voicing contrast is equally perceptible for all obstruent pairs. Data justifying this P-map fragment can be found in the works of Steriade.

The precise claim of Steriade (2001) is that "distinctiveness" orderings embodied in the P-map can be *directly* translated to orderings on appropriate correspondence constraints:

(17) **P-map projects correspondence constraints** (Steriade 2001:26)

For any two P-map cells, x-y/$\_K_i$ and w-z/$\_K_j$, associated with different [phonological contrasts],[9] there are distinct sets of correspondence [constraints],[10] CORRESP(X-Y/$\_K_i$) and CORRESP(W-Z/$\_K_j$).

(18) **Ranking correspondence constraints by relative distinctiveness** (Steriade 2001:28)

For any two P-map cells, x-y/$\_K_i$ and w-z/$\_K_j$, if

$$\text{x-y}/\_K_i \rhd \text{w-z}/\_K_j$$

(where $A \rhd B$ indicates A is *more distinctive* than B) then any correspondence constraint referring to x-y/$\_K_i$ outranks any parallel constraint referring to w-z/$\_K_j$.

[8] In the environments in the diagram, **R** refers to a sonorant,] to a word-boundary, and **T** to an obstruent.

[9] Steriade's (2001) original has "confusability indices." We avoid the use of "confusability" to describe the P-map because of the uncertain relationship between it and "similarity," which Steriade also uses to describe the relations in the P-map. As her current draft states, there is *not* an exact relationship between confusability and similarity; in section 2.4.1 we will discuss this problem in more depth. In this paper we will consistently use the (unfortunately rather awkward) term "distinctiveness" to avoid taking sides on the confusability/similarity debate. Our restatement here loses some of the sense of the original, namely that contrasts which are equally "distinct" project but a single correspondence condition, but it is sufficient for the present work.

[10] Reads "conditions" in the original.

From P-map fragment in (16), it then follows that:

(19) IDENT[VOICE]/V_V ≫ IDENT[VOICE]/V_] ≫ IDENT[VOICE]/V_T

Steriade also proposes the following distinctiveness ranking (Steriade 2001:18):

(20)

D vs. T / V_]	Least distinctive	D is a voiced stop
D vs. N / V_]		T is a voiceless stop
D vs. G / V_]		N is a nasal
C vs. ∅ / V_]		G is a glide or lateral
V vs. ∅ / C_]	Most distinctive	

In particular, she cites results from Zwicky (1976) and Walden and Montgomery (1975), among others, to justify her claim that:

(21) D-N/V_] ▷ D-T/V_]

and,

(22) D-G/V_] ▷ D-T/V_]

Likewise, data from Fleishhacker (1999) supports:

(23) V-∅/C_] ▷ C-∅/V_] ▷ D-T/V_]

From these rankings, she concludes (Steriade 2001:28) that:

(24) DEP(V VS. ∅) ≫ MAX(C VS. ∅) ≫ IDENT[SON]/V_] ≫ IDENT[VOICE]/V_]

Steriade notes that, although the conceivable grammatical responses to the markedness constraint *[+VOICE,+OBS]/_] ("voiced obstruents do not occur at the end of the word") include:

(25)

Change in output	**Corresponding constraint ranking**
a. Devoicing: /tob/ → [top]	*[+VOICE,+OBS]/_] ≫ IDENT[VOICE]
b. Nasalization: /tob/ → [tom]	*[+VOICE,+OBS]/_] ≫ IDENT[NASAL]
c. Lenition to glide: /tob/ → [tow]	*[+VOICE,+OBS]/_] ≫ IDENT[CONSONANTAL]
d. C-Deletion: /tob/ → [to]	*[+VOICE,+OBS]/_] ≫ MAX-C
e. V-Insertion: /tob/ → [tobə]	*[+VOICE,+OBS]/_] ≫ DEP-V
f. Segment reversal: /tob/ → [bot]	*[+VOICE,+OBS]/_] ≫ LINEAR(ROOT)
g. Feature reversal: /tob/ → [dop]	*[+VOICE,+OBS]/_] ≫ LINEAR[VOICE]

when *[+VOICE,+OBS]/_] is ranked with respect to (24), only two outcomes are possible: identity (violating the markedness constraint), or final devoicing. Steriade claims that only the devoicing (25a) is actually attested, and so the constrained correspondence ranking obtained by the use of the P-map better matches real typology than the free ranking allowed by conventional Optimality Theory.

In sections 3 and 4 we will put P-map rankings (16) and (20) to the test. As the P-map rankings project the correspondence constraint ranking in (24), if experiment shows that actual P-maps ought to differ from (20), then Steriade's claim that correspondence constraint rankings can be read directly from the P-map is suspect.

2.4.1 Determining perceptibility

Although we had said that P-maps reflect relative phonological "distinctness," we have deferred discussion on what, precisely, constitutes distinctness and how it might be measured. In Steriade (2001), both "similarity" and "confusability" are used as analogs. The intuitive notion is that "distinctness" as encoded in the P-map related to how perceptually "similar" two phonological strings are to each other. An obvious means of assessing similarity might be to determine which features in the strings under comparison are shared, and which differ, and apply some sort of trivial function (for example, take the ratio) of these two values. However, a little thought makes it obvious that such simple-minded feature counting is likely to miss the point. Consider, for example:

		Forms compared	Assessment of difference
(26)	a.	/fist/-[fis]	[t] vs. [∅]/t_
	b.	/fits/-[fis]	[t] vs. [∅]/V_s]

The same features—those of [t]—differentiate these, and yet judgements of similarity are very different in the two cases. Obviously, one wants to assess *phonological* similarity, which requires us to take the perceptual processes of the human auditory system into account.

Steriade considers several variations on feature counting before resolving that *confusability* is likely to be an additional correlate to any account of perceptual "distinctness." However, the experimental evidence for confusability yields conflicting results: voicing distinctions, for example, score as very similar when subjects are asked to rate pairs of stimuli on some numerical similarity scale. However, voicing is also among the *least* confusable contrasts (Steriade 2001:12). Obviously we cannot simply say that confusable contrasts are similar, and vice-versa. Yet it seems obvious that confusabilty *ought* to be a factor in deciding "distinctness."

Steriade has not yet resolved the issue; she resorts to allowing a range of possible data to back claims of perceptual "distinctness" or lack thereof:

> To substantiate a claim of relative [distinctness][11] one can (a) rely on speakers' direct judgements of similarity; (b) use similarity judgements implicit in rhyming

[11] We have substituted "distinctness" for "similarity" here, as throughout this paper, in an attempt to distinguist the experimentally-derived metric of "similarity" (through subject rankings, for example) with the theoretical notion of "distinctness" (which is not necessarily limited to the empirical similarity metric, and may encompass, for example, notions of "confusability").

> practices; (c) rely on confusion studies to show that one of the contrasts is more perceptually robust than the other; or (d) reason from the observation that in the position being considered, one contrast misses an essential acoustic correlate while the other does not. (Steriade 2001:19)

We will not attempt to determine a definitive "precise" definition of phonological distinctness, nor arbitrate between proposals. Instead, we will advance the hypothesis that the production of pairs of rhyming words in poetry corpora and instances of reduplication in the lexicon both spring from constraints based on a common perceptual map. By examining contrast frequencies in these productions we will infer relative "distinctness"; we can then check the derived "distinctness" rankings from reduplication against those from the poetry corpora to verify the hypothesis that both reduplication and poetry emerge from the same underlying constraint system. We can also check our derived "distinctness" measures against the hypothetical P-map fragments advanced by Steriade.[12]

It may be mentioned at this point that Steriade makes no claim of P-map universality, although she suggests that "to first-order" the P-map is identical across languages. The confusability and similarity experiments she cites consistently show metric variation across languages. To mitigate this effect, all the experimental data in this paper is taken from a single language, English.

2.4.2 Consequences of P-map theory

P-map theory makes different predictions from the "standard" faithfulness constraints; it also predicts less uniform behavior than syllable-based positional faithfulness (Beckman 1996) theories. P-map theory allows for constraint distinction based on perceptibility, which means, for example, that onset faithfulness may well depend on the properties of the coda of the preceding syllable.

More important for this investigation, P-map theory also predicts different epenthetic segments than does standard markedness theory. Steriade's formulation allows for the possibility that the epenthetic segment will be the one *which affects nearby segments the least*; i.e., which creates an output most perceptually similar to the output without epenthesis.[13] For example, comparing inputs VV to outputs VCV, we may find the least distinctive outputs to be VʔV, VhV, or VGV (where G is homorganic to V). Standard faithfulness/markedness theory would predict the most unmarked segment in the inventory (say, [t] or [p]) would always be the optimal epenthetic candidate regardless of context.

[12] A quick review of rationale: Steriade's claim (b) in the quote on this page follows from her assertion that poets access the P-map directly in order to select the "best rhyme" from among the entries in the lexicon with the necessary semantics. Yip proposed that reduplication is triggered by the action of a RHYME constraint, which forces self-rhyme in the reduplicant. She further proposes that exact reduplication is prohibited by the *REPEAT markedness constraint. The repair strategy for a *REPEAT violation must act in accordance with the ranking on faithfulness constraints given by (18), so it is expected that the action of the P-map will be visible in these forms, as well.

[13] This is in some ways similar to McCarthy's (1998) Sympathy theory.

This has significance for our analysis of reduplication. If Yip's RHYME and *REPEAT constraints motivate the reduplication, one would expect that the segmental modification to satisfy *REPEAT would always create a segment *less marked* than the input. On the other hand, base-reduplicant correlation should *not* (solely) be correlated with markedness in Steriade's theory: the modifications should instead be closely related to relations embodied in the P-map. Of course, the relations in the P-map might be ordered in such a way that B-R variation *appears* to prefer unmarked structure—in these cases we must be more careful to determine proper causation. In other words, a preference for unmarked structure does not disqualify a P-map-based theory from consideration, but if we do *not* find a preference for the unmarked, then we *must* look beyond the standard theory.

3 Experimental Technique

We desire to collect statistics about phonological processes occurring among rhyme pairs chosen by poets and within reduplicative compounds created by some productive process of the phonology. Our experimental technique breaks down into three parts: first, we will extract rhyme pair data from several large archives of English poetry; second, we will extract examples of reduplication from our lexicon data; and finally, we will reconstruct correspondence relations from our rhyme pair and reduplication data to allow us to compute process statistics.

3.1 Extracting rhyme-pair data

Obtaining rhyme-pair data began with the following collections of English language texts:

- **Project Gutenberg** (`http://promo.net/pg`) Project Gutenberg is a digital library project curated by Michael Hart which has been scanning and publishing public domain texts since 1971. Our snapshot of Project Gutenberg's archives contains 2.73 GB of texts, from a variety of languages and genres, omitting the very large archive of Human Genome Project sequencing data which Project Gutenberg also includes.
- **Internet Wiretap** (`http://wiretap.area.com/`) The Wiretap Online Library was established in 1990 as an etext repository. Its focus is decidedly less formal than that of Project Gutenberg, but it does maintain an archive of "classic" texts. These texts are, helpfully, categorized such that the portion of the archive containing poetry is easy to extract. We use the portion of the library under the path `Classic/Poetry`.[14]
- **Hello Toes! Hello Feet!** (Paul 1998) This children's poem is dense with slant rhymes, and is thus an excellent calibration target for the rhyme-pair extraction engine. (Other

[14] "Classic" is a relative term: this section of the archive includes works by Carl Sandburg, for instance, but not Shakespeare—he gets his own section under `Classic/Shakespeare`. The Bard's English is not much use to us, anyway, as we do not have a reliable pronunciation dictionary for his orthography.

calibration targets included the poetry of Emily Dickenson,[15] and a sonnet beginning "Ignorance shed! Revealed—a world of woe!" whose origin has been lost.) This text was formatted for proper form by the author from text provided by Cheryl Zoll.

After some source-specific preprocessing steps to normalize the inputs and remove standard headers,[16] we form "stanzas" from sections of the text separated by blank lines. Multiple blank lines, or a single-line "stanza" (which is taken to be a title) delimit stanza candidates into "poems". This a heuristic on the nature of poetry formatting in plain-text that has been shown to be very effective on our corpora: the few poem inputs whose stanzas are misparsed by this heuristic are discarded when they are rejected by our classifier as a result.

The poem candidate is then passed to our "poetry classifier" routine which attempts to determine whether it really is a poem or not. It does so by advancing a number of rhyme scheme hypotheses (one per stanza) and determining which hypothesis best fits the poem as a whole.[17] If the best fitting rhyme scheme matches the poem closely enough, than the determination is made that this candidate is, indeed, a poem, and the best rhyme scheme is used to print out line endings which are supposed to rhyme. Otherwise, the candidate is discarded.

We prefer to discard candidates if we are uncertain about proper classification, to avoid polluting the data set with data derived from non-poems. In particular, it will be noted that this classifier does not encompass poems which shift between different rhyme schemes; these will be classified as non-poems.

The heart of this classifier is (a) extraction of rhyme scheme hypotheses, and (b) scoring hypotheses against the poem. When we say two lines rhyme, what is typically meant is, "all segments after and including the lines' final stressed nucleus[18] are similar." This domain is called the *rhyme skeleton*. We will also refer to the *rhyme onset* as the onset of this final stressed syllable. For an "exact" rhyme, the rhyme skeletons of the two lines are identical, although we also permit "inexact" rhymes, where they are merely similar.

Extracting a rhyme scheme hypothesis from a stanza simply entails searching for matching rhyme skeletons. If lines 1, 2, and 5 have identical rhyme skeletons, as do lines 3 and 4 (as might be the case in a limerick), then we say the rhyme scheme is AABBA.

The missing element in doing this programmatically is a mapping from English orthography to pronunciation. We use the Carnegie Mellon University Pronouncing Dictionary (version 0.6d)

---

[15] Files `1mlyd10.txt` and `2mlyd10.txt` in the Gutenberg archive

[16] For Project Gutenberg we also specially filter out texts in German, using our program `identify-german.pl` which searches for "overly-frequent" use of the common German words *ist*, *mit*, *der*, and *ein*. Our poetry classification step rejects poem candidates containing "rhyming words" which it cannot find in its pronunciation dictionary, which suffices to reject most non-English texts in Gutenberg; however, German words seem to have English cognates in the dictionary often enough to confuse the classifier unless this special filtering is done.

[17] We allow stanzas to vary in length, as long as the rhyme scheme provides correct predictions for the additional lines when they are present.

[18] The final stressed nucleus of a line need not belong to the last word of a line: prosodically unstressed words may be used to create feminine rhymes. However, our analysis technique does not attempt to analyze phrasal stress.

Phoneme	Example	Translation	Phoneme	Example	Translation
AA	odd	AA D	L	lee	L IY
AE	at	AE T	M	me	M IY
AH	hut	HH AH T	N	knee	N IY
AO	ought	AO T	NG	ping	P IH NG
AW	cow	K AW	OW	oat	OW T
AY	hide	HH AY D	OY	toy	T OY
B	be	B IY	P	pee	P IY
CH	cheese	CH IY Z	R	read	R IY D
D	dee	D IY	S	sea	S IY
DH	thee	DH IY	SH	she	SH IY
EH	Ed	EH D	T	tea	T IY
ER	hurt	HH ER T	TH	theta	TH EY T AH
EY	ate	EY T	UH	hood	HH UH D
F	fee	F IY	UW	two	T UW
G	green	G R IY N	V	vee	V IY
HH	he	HH IY	W	we	W IY
IH	it	IH T	Y	yield	Y IY L D
IY	eat	IY T	Z	zee	Z IY
JH	gee	JH IY	ZH	seizure	S IY ZH ER
K	key	K IY			

Figure 1: List of phonemes used in the `cmudict` pronunciation dictionary.

from `http://www.speech.cs.cmu.edu/cgi-bin/cmudict` to perform this mapping. The CMU dictionary consists of 129,481 entries which give pronunciations in terms of the phonemes shown in Figure 1. We augment this with 102 "extra" dictionary entries which were added to correct errors or resolve high-frequency omissions, and an algorithmic adjunct (in `prodict.pl`) which provides pronunciations for variants on the dictionary entries; in particular, it handles British spellings, hyphen compounds, the prefix "pre-" and the suffixes "-able," "-ly," "-er," "-ing," "-ship," "-less," "-(i)ness," and "-s." These additions were found necessary during testing with our calibration data, and by reviewing the "no pronunciation found" outputs of our whole-corpus runs.

The data in the CMU dictionary includes primary and secondary stress markings via a numeric suffix on vowel phonemes. From this it is easy to extract the rhyme skeleton from the phonemic transcription of the last words[19] in the stanza lines.

An important issue arises here: if we count as rhymes only lines in which the rhyme scheme is *identical*, we will omit exactly that data which is most relevant to the present investigation: namely, *inexact* rhymes which (according to Steriade's hypothesis) are said to "rhyme" because they are judged sufficiently perceptually similar by means of the P-map. We therefore *do not use*

[19]In actual practice, we use only a single final word. In some (rare) cases this means that our rhyme skeleton is too short, but this has not been seen to have an adverse effect.

consonantal features at all when formulating rhyme scheme hypotheses and scoring scheme fits. We use *only the vocalic features of the skeleton* and for these we *require an exact match.* There are in fact interesting issues regarding vocalic perceptual similarity which lie dormant in the data—especially frequent are American authors using British vowel pronunciation in order to co-erce a rhyme, and vice-versa—but in this investigation we look only at consonant perceptual similarity. We will defer to future work the possibility of flip-flopping to use the consonantal segments to derive rhyme scheme, allowing the investigation of the perceptual similarities of vocalic segments.

This point is very important, so we will repeat it: *no bias for or against correlated consonantal segments is introduced by the rhyme scheme analysis or poetry classifier.*

A sample analysis by the poetry classifier for a short Emily Dickenson poem follows:

```
AVG LINES: 4
SCHEME LINES: 4
BEST SCHEME SCORES: 1 ABCB
VERDICT: A POEM
Our share of night to bear (bear) [EH 0]
Our share of mourning (mourning) [AO 1]
Our blank in bliss to fill (fill) [IH 0]
Our blank in scorning (scorning) [AO 1]
---new stanza---
Here a star and there a star (star) [AA 0]
Some lose their way (way) [EY 0]
Here a mist and there a mist (mist) [IH 0]
Afterwards day (day) [EY 0]
---new stanza---
MOURNING SCORNING DICKENSON/1mlyd10.txt
WAY DAY DICKENSON/1mlyd10.txt
```

Note that the final word and the rhyme skeleton (encoded as the nucleus of the stressed syllable and the number of following non-stressed syllables) are repeated in parenthesis and square brackets, respectively, after each line; and that the (correct) rhyme scheme, ABCB, has achieved a perfect score. After the poem are printed the two rhyming pairs extracted, along with filenames indicating the source of the rhyming pairs in case verification is later needed.

A longer Dickenson poem, for which the analysis extracts the correct rhyme scheme but still rejects as a poem:

```
AVG LINES: 6
SCHEME LINES: 6
BEST SCHEME SCORES: 0.666666666666667 AABCCB
VERDICT: NOT POETIC
```

```
T is so much joy 'T is so much joy (joy) [OY 0]
If I should fail what poverty (poverty) [AA 2]
And yet as poor as I (I) [AY 0]
Have ventured all upon a throw (throw) [OW 0]
Have gained Yes Hesitated so (so) [OW 0]
This side the victory (victory) [IH 2]
---new stanza---
Life is but life and death but death (death) [EH 0]
Bliss is but bliss and breath but breath (breath) [EH 0]
And if indeed I fail (fail) [EY 0]
At least to know the worst is sweet (sweet) [IY 0]
Defeat means nothing but defeat (defeat) [IY 0]
No drearier can prevail (prevail) [EY 0]
---new stanza---
And if I gain oh gun at sea (sea) [IY 0]
Oh bells that in the steeples be (be) [IY 0]
At first repeat it slow (slow) [OW 0]
For heaven is a different thing (thing) [IH 0]
Conjectured and waked sudden in (in) [IH 0]
And might o'erwhelm me so (so) [OW 0]
---new stanza---
```

Here either two extremely questionable rhymes—*joy/poverty* and *I/victory*—or a shift in rhyme scheme in the first stanza occurs, which is enough to make the poetry classifier doubt its analysis and discard the poem.

The next Dickenson example illustrates how phrasal stress can alter the rhyme skeleton of a line. As our analysis is limited to word-level stress, we (conservatively) discard this poem as well:

```
AVG LINES: 3
SCHEME LINES: 3
BEST SCHEME SCORES: 0.666666666666667 AAB
VERDICT: NOT POETIC
Some things that fly there be (be) [IY 0]
Birds hours the bumble-bee (bee) [IY 0]
Of these no elegy (elegy) [EH 2]
---new stanza---
Some things that stay there be (be) [IY 0]
Grief hills eternity (eternity) [ER 2]
Nor this behooveth me (me) [IY 0]
---new stanza---
There are that resting rise (rise) [AY 0]
Can I expound the skies (skies) [AY 0]
```

```
How still the riddle lies (lies) [AY 0]
---new stanza---
```

Note that the prosodic footing process shifts the stress on "elegy" from the first to the last syllable, and puts a secondary stress on the final syllable of "eternity." These shifts are invisible to our program.

A final example illustrates this analysis' power to extract even very inexact rhymes by inference from a rhyme scheme.

```
AVG LINES: 4
SCHEME LINES: 4
BEST SCHEME SCORES: 0.75 ABCB
VERDICT: A POEM
I taste a liquor never brewed (brewed) [UW 0]
From tankards scooped in pearl (pearl) [ER 0]
Not all the vats upon the Rhine (Rhine) [AY 0]
Yield such an alcohol (alcohol) [AE 2]
---new stanza---
Inebriate of air am I (I) [AY 0]
And debauchee of dew (dew) [UW 0]
Reeling through endless summer days (days) [EY 0]
From inns of molten blue (blue) [UW 0]
---new stanza---
When landlords turn the drunken bee (bee) [IY 0]
Out of the foxglove's door (door) [AO 0]
When butterflies renounce their drams (drams) [AE 0]
I shall but drink the more (more) [AO 0]
---new stanza---
Till seraphs swing their snowy hats (hats) [AE 0]
And saints to windows run (run) [AH 0]
To see the little tippler (tippler) [IH 1]
Leaning against the sun (sun) [AH 0]
---new stanza---
PEARL ALCOHOL DICKENSON/1mlyd10.txt
DEW BLUE DICKENSON/1mlyd10.txt
DOOR MORE DICKENSON/1mlyd10.txt
RUN SUN DICKENSON/1mlyd10.txt
```

For some (to this author unclear) reason, Dickenson has chosen to place "pearl" and "alcohol" in the locations which the scheme indicates should rhyme. In this case, the notion of perceptual similarity she was using seems to derive solely from the identity of the final liquid.

A final component of this analysis filters the list of rhyming pairs generated to include only inexact rhymes. In our results section, we will present statistics separately for the entire data set and the set of inexact rhymes. From our input corpora, we obtain 62,583 rhyming pairs (21,182 unique), of which 3,904 (3,107 unique) are inexact.

3.2 Compiling lists of reduplicative compounds

The most common reduplicative forms in English are:

- the *tut-tut* type [CVC-CVC] (currently the most productive),
- the *hodge-podge* type [C_1VC-C_2VC], and
- the *tick-tock* type [CV_1C-CV_2C] (least productive). (Minkova 2000) (Schiffman 2001)

According to Minkova (2000), there are "over 500 such compounds in use in present-day English, over 70 of which have been around at least for two centuries, and some considerably longer." As we do not have Minkova's list of compounds, we wrote our own program to search for reduplicative forms among the 129,481 entries in our pronunciation dictionary. Because we were worried about the relative paucity of compound words in `cmudict`, we wrote a program to extract words and pronunciations from Project Gutenberg's copy of the 1913 Webster Unabridged Dictionary. However, Gutenberg's scan of this text frequently omitted or garbled the pronunciation information from the source, so we were only able to extract 5,868 entries from the data.

For each entry in our lexicon, we first obtain the phonetic transcription from `cmudict`+Webster, as we did for words in the poetry corpora. We then syllabify this transcription by greedily building legal onsets for each nucleus. We use the legal onset tables found in Kenstowicz (1994:256-258), with the addition of the onsets /hy-/ and /vy-/, needed for the words *hue* and *view*, respectively. We mark the left edge of each maximal onset as a syllable boundary, and also mark a boundary between two adjacent vowels. Adding a mark word-finally completes the syllabification.[20]

We now have a syllabified phonetic transcription of our candidate word. We first see if it is an instance of the *hurly-burly* type of reduplication. For each primary- or secondary-stressed syllable we construct a template containing the syllable nucleus and all material between it and the end of the word. We then check if this template occurs more than once in the transcription. If it does, then we have found a base-reduplicant pair.

Note that this technique will find reduplicative forms with "extra" syllabic material between the first and second template match, like *abracadabra*. The template contains appropriate syllable boundaries, so word-final consonants in the template will only match other coda consonants; they will not match onsets.

[20]One small additional detail: as Figure 1 on page 15 indicates, `cmudict` often marks *r-colored vowels* with an ER phoneme, in place of supplying an independent R. Unfortunately, when the R occurs intervocalicly, coloring the left-hand vowel, this transcription appears to deprive the right-hand vowel of its onset. A special rule creates an R in the onset position for this case.

The first ten reduplicative forms starting with "B" found by this routine in the `cmudict` lexicon are:

```
BACCARAT : B AA2 : K ER0 R AA1
BACKPACK : B AE1 K : P AE2 K
BACKTRACK : B AE1 K : T R AE2 K
BAJA : B AA1 : HH AA2
BANDSTAND : B AE1 N D : S T AE2 N D
BARGAR : B AA0 R : G AA1 R
BBC : B IY2 B IY0 : S IY1
BEAUDREAU : B OW2 : D R OW1
BEDSPREAD : B EH1 D : S P R EH2 D
BEEBE : B IY1 : B IY2
```

Note that no attempt to parse the candidate into morphemes is done, so the algorithm may classify monomorphemic words like *baja* and *baccarat* as reduplicative "compounds." We do not believe this affects the validity of our analysis.

We also find reduplication of the *tick-tock* and *flip-flop* type. In this case, from every stressed syllable we construct a template by replacing the vowel with a generic vowel pattern. Again, we look for words where a template so constructed matches more than once. But in fact, this criterion is too loose, so we also restrict the number of syllables intervening between the template matches and insist that the template must contain at least one consonant. The reduplicants found by this routine in the `cmudict` lexicon include:

```
BUGABOO : B AH1 : G AH0 B UW2
CHITCHAT : CH IH1 T : CH AE2 T
DADO : D EY1 : D OW2
FLIMFLAM : F L IH1 M : F L AE2 M
KITCAT : K IH1 T : K AE2 T
KNICK-KNACK : N IH1 K : N AE1 K
MEMO : M EH1 : M OW2
MISHMASH : M IH1 SH : M AE2 SH
MOOMAW : M UW1 : M AO2
POPEYE : P AA1 : P AY2
SEESAW : S IY1 : S AO2
SHIPSHAPE : SH IH1 P : SH EY2 P
SO-SO : S OW1 : S OW1
TELLTALE : T EH1 L : T EY2 L
TIT-FOR-TAT : T IH1 T : F AO2 R T AE1 T
ZIG-ZAG : Z IH1 G : Z AE2 G
```

Again, note that "accidental" reduplication such as in *memo* is also found with this technique. It may well be, however, that perceptual "goodness" as an instance of reduplication (including satisfaction of the RHYME constraint) affected the derivation and evolution of words like *memo* and *popeye*.

Note that reduplication of the *tut-tut* type may be found with either of these routines.

On the `cmudict` data set, we find 27 examples of the *tut-tut* type, 45 examples of the *flip-flop* type, and 233 examples of the *hurly-burly* type, for a total of 305 reduplicants. An additional four reduplicants are found in the Webster dictionary lexicon, and one more in `extradict` which contains pronunciations for words used in the Dickenson poems which were noted to be missing during development of the poetry classifier.

Note that the method used to identify reduplicants constrains all syllable codas in the rhyme skeleton to be identical. Valid perceptual similarity data can only be extracted from the rhyme onsets.

3.3 Collecting statistics

Once we've obtained our rhyming pairs and reduplicants, we are interested in collecting statistics on the frequency of phonological processes in certain contexts; for example, we would like to determine how often a word-final consonant is deleted in slant rhyme pairs. Our `categorize.pl` program collects these statistics.

To do this, it must first compute a correspondence function between our two inputs.[21] The rhyme skeletons are right-edge aligned to obtain a syllable correspondence. We then compute correspondences within each syllable.[22] Nuclei always correspond. Working out from the nucleus towards the onset and then coda we assume a one-to-one mapping ($C_1C_2C_3V \rightarrow C_1C_2C_3V$) unless (for each segment) skipping a single segment would yield a better match. That is, for "input" $C_{I_3}C_{I_2}C_{I_1}V$ and "output" $C_{O_3}C_{O_2}C_{O_1}V$, we assume that C_{I_1} corresponds with C_{O_1} unless C_{I_2} is a better match for C_{O_1} or C_{O_2} is a better match for C_{I_1}, in which case C_{I_1} or C_{O_1} (respectively) are assumed to be deleted/epenthesized (have no correspondent). For "is a better match" we mean that the two segments are identical and C_{I_1} and C_{O_1} are not, or that the two candidate segments are of the same sonority (vowel, glide, liquid, nasal, obstruent) and C_{I_1} and C_{O_1} are not.

Note that this technique does not assign correspondences to feature geometry below the root node, and that multiple correspondences are never generated. For the domain of this investigation, the algorithm is quite adequate.

Once the `categorize.pl` program has computed correspondences, it examines each corresponding segment and its immediate environment and collects statistics based on the nature of the process (deletion/epenthesis or feature change) and the properties of the context. These

[21]This is a base-reduplicant correspondence in the case of reduplicative inputs, or an output-output correspondence in the case of rhyming pairs.

[22]This means that our technique will never identify a coda consonant as having moved to an onset, for example. As this investigation is not concerned with processes that might cause this sort of syllable reparsing, this simplification is justified.

statistics use the place (labial, coronal, dorsal, guttural), stricture (stop, fricative, approximant), sibilance, voicing, and sonorance of the corresponding segments and context.

4 Results

The tools described in the preceding section were used to test predictions of Steriade and Yip/McCarthy about the structure of correspondence constraints in optimality theory. They were also used to compare rhyme-pair data with reduplication data to determine if similar processes were at work, as Yip claims.

4.1 Word-final contrasts

In section 2.4 we reviewed claims in Steriade (2001:18) about the relative perceptibility of certain word-final contrasts. As an initial check on the validity of her theory, we will use our rhyme-pair data to verify the ranking made in (20) on page 10 of this paper. Specifically, Steriade claims that the following "distinctiveness" rankings hold:

(27) D-N/V_] ▷ D-T/V_]

(28) D-G/V_] ▷ D-T/V_]

(29) C-∅/V_] ▷ D-T/V_]

(30) V-∅/C_] ▷ C-∅/V_]

The "rhyme fitness" measures of poets are seen as a good window into the rankings in the P-map (see quote on page 12). Therefore, it is reasonable to assume that the frequency of slant rhymes differing by one of the contrasts ranked above would be inversely proportional to its "distinctiveness" ranking; that is, that overly "distinct" contrasts would tend *not* to be acceptable to poets in rhyming pairs. Word final contrasts are ideal for this test, as this context is guaranteed to be inside the rhyme domain.

Statistics on contrast frequency in our rhyme-pair data are shown in Figure 2 on the following page; more distinctive contrasts are towards the right. The data show a surprisingly good fit to theory: the "least distinctive" contrast D-T/V_] is indeed the most common in rhyme pairs. Contrasts D-N/V_] and C-∅/V_] are also observed to occur, at lower rates, satisfying predicted orderings (27) and (29). It appears that orderings (28) and (30) are also supported by the data, but we are cautious about making claims based on an absence of data points. In particular, the V-∅/C_] environment is likely excluded from our data set by our heavily syllable-based experimental technique, rather than any inherent perceptibility properties. In addition, there are only twelve words in our 135,000 entry lexicon which end in a glide, which very likely accounts for the lack of D-G/V_] contrast examples.

The one ordering that appears *n*ot to be supported by the data is C-∅/V_] ▷ D-N/V_], but closer examination of (27)–(30) shows that Steriade does not actually make this ordering claim.

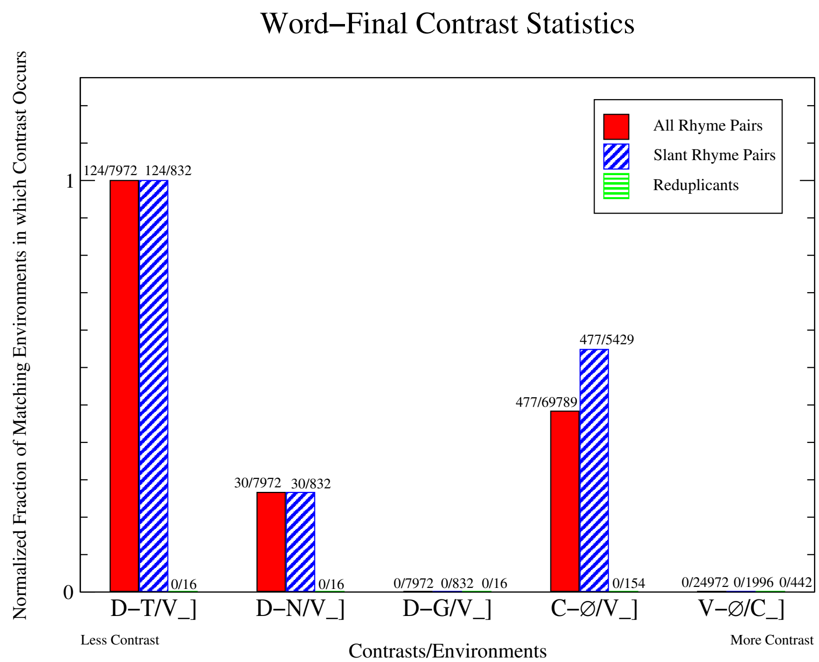

Figure 2: Statistics on word-final contrasts. Each bar has the fraction of contrast examples found over instances of the appropriate environment. Bars are normalized to the value of the D-T/V_] ratio.

Indeed, given the evidence presented here for the opposite ordering, it might be interesting to examine obstruent nasalization processes and determine if the preference for deletion rather than nasalization (which would be predicted by P-map theory based on this perceptability ordering) is borne out by linguistic evidence.

Note that the slant rhyme data in Figure 2 is identical to the rhyme pair data, as the presence of word-final contrast was one of the factors used to extract slant rhymes from among the rhyme pairs. The only difference is a differing C-∅/V_] contrast ratio, due the greater number of VC] environments with the potential to display this contrast among the complete rhyme pair set.

Also note that no word-final contrasts were found among the reduplicated word data. The reason for this lack should be obvious from the description in section 3.2 of the technique used to identify reduplication: inexact base-reduplicant rhymes in were not allowed by the classifier.

4.2 Context-sensitive voicing contrasts

We now turn to testing our hypothesis that reduplication can be explained using Steriade's P-map-based correspondence relations and Yip's RHYME constraint. We focus on voicing contrast,

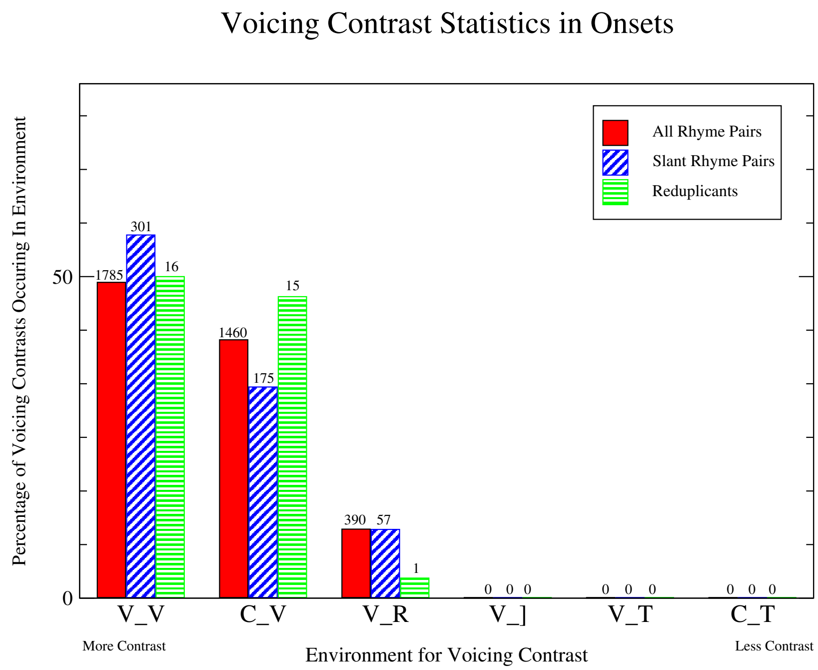

Figure 3: Voicing contrasts in the onsets of rhyming pairs. The actual number of examples of each contrast found is shown above the bar.

using the P-map fragment proposed by Steriade and duplicated in (16) on page 9. This fragment proposes the "distinctiveness" relation:

(31) D-T/V_V $\triangleright$ D-T/C_V $\triangleright$ D-T/V_R $\triangleright$ D-T/V_] $\triangleright$ D-T/V_T $\triangleright$ D-T/C_T

Again, we assume that "distinctiveness" of contrast can be inferred from frequency of occurrence. We collected statistics on the frequency of voicing contrasts in these environments in both the rhyme-pair and reduplication data. We expected correlation in the rhyme-pair data, especially considering the positive confirmation of Steriade's technique given in the previous section; the real question was: would reduplication correlate? If so, then this would be an argument that a constraint similar to Yip's RHYME was at work, creating the synergy between non-productive poetic processes and productive phonology.

Our first attempts at analyzing the collected data ended in confusion. There didn't seem to be any significant correlation with the theorized P-map rankings at all—not even in the rhyme-pair data. The reason did not become obvious until we separated the processes occuring in the rhyme skeleton and rhyme onset.

Figure 3 shows the contrast frequencies in the rhyme-pair and reduplication *onsets*; more distinctive environments are on the left. Immediately apparent is a nice clean correlation of contrast to frequency, sloping down from the left-hand side. The only catch is that voicing contrast in *more* distinctive environments is *more* common!

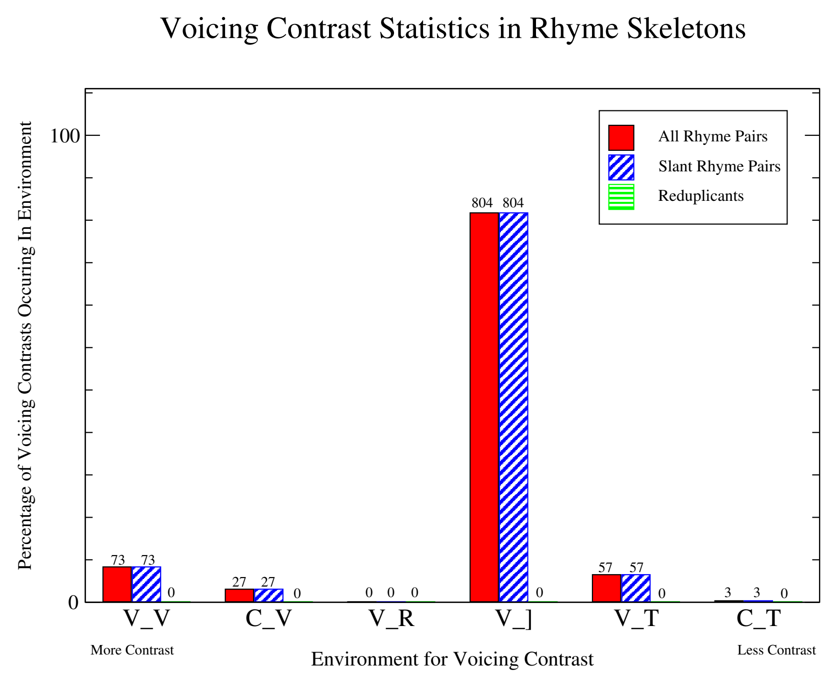

Figure 4: Voicing contrasts in the rhyming skeleton. The actual number of examples of each contrast found is shown above the bar.

This is exactly the opposite of what we expected to see, but it makes sense upon reflection. A phonological dictum against monotony—the same mechanism underlying the OCP and even our *Repeat constraint—attempts to make the *non-rhyming* portion of the word as distinct as possible. A poet doesn't rhyme *dead* with *dead*; and this data suggests even *Ted* is dispreferred to more contrastive rhymes. Gratifyingly, the same process seems to hold for both rhyme pairs and reduplication; the frequency ratios of contrasts in the different environments are comparable. Behavior for slant rhymes was also roughly the same as for rhymes in general, indicating that there is likely no compensation occuring for inexactness of the rhyme skeleton match.

There are no onset voicing contrasts found in the V_], V_T, or C_T environments. Onsets are inconsistent with a V_] environment, explaining the first lapse; and tri-consonant clusters are rare in English word-medially, explaining the last. Confirmation of the ordering V_R ⊳V_T may adequately explain the remaining lack of data, but we hesitate to draw conclusions from absence.

Figure 4 shows the statistics collected on voicing contrast in the rhyming skeleton, which does not include the rhyme onsets just analyzed. Here we see the expected high frequency for a low distinctiveness contrast: D-T/V_]. We also have instances of the lower distinctiveness contrasts D-T/V_T and D-T/C_T. Their relative infrequency may be explained by the paucity of lexicon entries with these environments.

We also have significant instances of high-distinctiveness D-T/V_V and D-T/C_V contrasts.

High frequencies here might be related to the relative prevalence of these environments in the lexicon. More research is required to investigate and understand the processes that lead to these high-contrast distinctions within the rhyme skeleton.

Note that the slant rhyme data is again identical to the full rhyme-pair data; the presence of contrast in the rhyme skeleton is exactly the criterion which is used to pick out slant rhymes from the rhyme-pair list. Likewise, there are no skeleton contrasts among the reduplication data, again due to experimental technique.

4.3 Voicing and markedness

Let us consider the typology arising from the constraints *REPEAT, IDENT$_{BR}$[VOICE], and *[+VOICE] in the reduplication framework of Yip and McCarthy.[23] We will assume that RHYME is ranked high enough to force reduplication. There are three possible behaviors, depending on constraint ranking:

(32)

	Ranking	Outcome	
M_1 :	IDENT$_{BR}$[VOICE] $\gg$ *[+VOICE], *REPEAT	/b/→/b/	/p/→/p/
M_2 :	*[+VOICE] $\gg$ *REPEAT, IDENT$_{BR}$[VOICE]	/b/→/p/	/p/→/p/
M_3 :	*REPEAT $\gg$ IDENT$_{BR}$[VOICE], *[+VOICE]	/b/→/p/	/p/→/b/

Note that rankings M_1 and M_2 are only distinguishable if IDENT$_{IO}$[VOICE] $\gg$ *[+VOICE], which is true for English. These three rankings lead to three different patterns of voicing contrasts in reduplicants, and so we can test the McCarthy/Yip model by seeing if the statistics from our reduplication data match one of these patterns. In order to keep from contaminating our analysis with processes governed by constraints not in our simple typology, we will compare only rhyme onsets.[24] This avoids positional faithfulness effects,[25] since all the compared consonants are in the same position.[26] Further, we will compare only cases in which obstruents in the base correlate with obstruents in the reduplicant, to avoid the effects of constraints like "sonorants-are-voiced" and IDENT[+SON]. Finally, we will consider only cases where *at least one* feature change occurs; that is, the base and the reduplicant must not be identical. This avoids polluting our data with reduplicated words of the *tut-tut* form, which arguably are subject to a different constraint ranking than words of the *hurly-burly* form.[27]

Let us first compute the voicing contrast rates we would expect. Suppose the fraction of voiceless obstruents in the rhyme onset of our bases is x, and the fraction of voiced obstruents in

[23]Yip would substitute a positional faithfulness constraint on [VOICE] for IDENT$_{BR}$[VOICE]; the difference is not significant here.

[24]Defined on page 14.

[25]Or, "ensures that the same positional faithfulness constraint is active" (the one corresponding to the McCarthy IDENT$_{BR}$[VOICE]) for Yip's formulation.

[26]It doesn't avoid the effects of *context*-sensitive faithfulness constraints, as proposed by Steriade, but we are checking McCarthy/Yip's predictions here, not Steriade's.

[27]*REPEAT ought be higher ranked in *hurly-burly* type words than it is in *tut-tut* type words.

the rhyme onset is y, where $x+y=1$. Then, the frequency of combinations of base/reduplicant features for each ranking should be as follows:

(33)

	Base→Reduplicant	M_1	M_2	M_3	R
a.	[-VOICE]→[-VOICE]	x	x	0	$x/2$
b.	[-VOICE]→[+VOICE]	0	0	x	$x/2$
c.	[+VOICE]→[-VOICE]	0	y	y	$y/2$
d.	[+VOICE]→[+VOICE]	y	0	0	$y/2$

The column labelled R in this table represents the null hypothesis: these are the frequencies we would expect if voicing contrast did not follow our typology at all, but instead was completely random.

A complication arises because it is impossible to programmatically distinguish base from reduplicant—a human may be able to claim that *roly* is the base of *roly-poly* by analogy to *roll* and reference to etymology (a *roly-poly* was originally a filled baked good) but a computer (at this time) certainly can not. It is impossible to distinguish between a [-VOICE] base becoming *more marked* by altering to [+VOICE] in the reduplicant, and a [+VOICE] base becoming *less marked* by altering to [-VOICE]. In concrete terms, this means that we cannot tell the difference between rows (33b) and (33c) of our frequency table!

Our frequencies are still distinctive even if we cannot tell these cases apart, however, as the following table shows:

(34)

Voicing features	M_1	M_2	M_3	R
[-VOICE] unchanged	x	x	0	$x/2$
voicing changed	0	y	1	1/2
[+VOICE] unchanged	y	0	0	$x/2$

Note that we used the relation $x+y=1$ to simplify expressions in the table. Each column now has at least one constant, so in the ideal case we could "find the zeros" to distinguish between rankings. However, it is trivial to discover the actual values of x (fraction of voiceless onset obstruents) and y (fraction of voiced onset obstruents). A quick pass over our lexicon (`cmudict` and Webster) yields 42% unvoiced and 58% voiced onset *consonants* (of 343,622 total onset consonants examined), but 67% unvoiced and 33% voiced onset *obstruents* (of 217,723 onset obstruents examined).[28] With our obstruent fractions in hand, we can now evaluate the expressions in table (34) to obtain:

(35)

Voicing features	M_1	M_2	M_3	R
[-VOICE] unchanged	67%	67%	0%	33%
voicing changed	0%	33%	100%	50%
[+VOICE] unchanged	33%	0%	0%	17%

$x=.67, y=.33$

[28]Obviously the inclusion of sonorants was responsible for biasing the consonant count towards [+VOICE].

The statistics we actually obtain from our list of reduplicative compounds are:

(36)

Voicing features	Number of examples	Percentage
[-VOICE] unchanged	41	38%
voicing changed	55	51%
[+VOICE] unchanged	12	11%

for $C_1 \to C_2$ processes where C_1 and C_2 are different obstruents

Unfortunately, this data does not seem to support any of the McCarthy/Yip rankings. In fact, it seems to closely follow the distribution we'd expect if the voicing process was completely random! So our simple markedness/faithfulness model does not appear to be adequate.

In section 2.4.2, we outlined some of the predictions P-map theory makes in regard to reduplication. In particular, P-map theory predicts that epenthetic segments (forced violations of DEP) need *not* necessarily be the most unmarked segments possible. Instead, evaluation of epenthetic candidates is done on the basis of perceptual similarity to the pre-epenthesis form, and is thus highly context-sensitive.

P-map theory will make similar claims about the voicing process we are examining here: assuming that *REPEAT forces a faithfulness violation, the chosen repair strategy will be context-sensitive—the feature identity violated will be related to the perceptibility of that feature *in the context in which it appears.*[29] Since we did not control the *context* of the onset consonants in the statistics collected above, we are seeing the averaged results of constraint evaluation across many different contexts; we might expect that, in the absence of some strong correlation on voicing distinctiveness in these different contexts,[30] the result would be indistinguishable from a random process.

5 Conclusions

The results presented in section 4 support a model of English reduplication based on the RHYME and *REPEAT constraints of Yip, but making reference to the P-map of Steriade. The word-final perceptibility rankings which Steriade required for her explanation of final-consonant voicing repair strategies appear to be the same rankings used by English-language poets to choose good rhyme pairs.

Furthermore, applying Steriade's P-map to our English reduplication data yielded a surprising insight: violations of *REPEAT are repaired in such a way as to *maximize* contrast in the non-rhyming portions of the word, both in English poetry and in reduplication. The parallel processes support the hypothesis that the underlying mechanism is the same. This is also an intriguing support of Yip's Identity Avoidance constraints. Applying the P-map orderings to the

[29] Perhaps counter-intuitively, the results presented in the previous section suggest that the feature modified is chosen to *maximize* perceptibility of the changed feature, rather than the opposite.

[30] An example correlation would be, for example, a general principle that voiced segments are more/less "distinctive" than unvoiced segments across all contexts.

rhyming skeletons of rhyme-pair data showed that in that case “distinctiveness” is minimized, as expected, but more research is necessary to adequately understand outliers in the data.

Finally, an attempt to explain voicing behavior in reduplication using a simpler McCarthy-esque context-free faithfulness model failed to explain the distributions of voicing contrasts seen in our reduplicative forms. This adds more support to the necessity of a revised faithfulness theory; Steriade’s context-sensitive faithfulness constraints were not inconsistent with the distribution of contrasts seen.

These results support a theory of reduplication based on rhyme and perceptual salience, and seem to mark as inadequate approaches to reduplication similar to that originally proposed by McCarthy.

References

ALDERETE, JOHN, JILL BECKMAN, LAURA BENUA, AMALIA GNANADESIKAN, JOHN MCCARTHY, and SUZANNE URBANCZYK. 1997. Reduplication with fixed segmentism. Ms., University of Massachusetts, Amherst. Also available as ROA #226.

BECKMAN, JILL. 1996. Positional faithfulness and the distribution of phonological features. Ms., UMass Amherst.

BROGAN, T. V. F. 1999. *English Versification, 1570–1980: A Reference Guide*. Japan: University of Shizuoka, Hamamatsu Campus, hypertext edition. `http://ham.t.u-shizuoka-ken.ac.jp/versif/Versification.html`.

FLEISHHACKER, HEIDI. 1999. The location of epenthetic vowels with respect to consonant clusters: An auditory similarity account. Master's thesis, UCLA.

JESPERSEN, OTTO. 1933. Symbolic value of the vowel *I*. *Linguistica: Selected Papers in English, French, and German* 283–303. Reference C337 from Brogan (1999).

KENSTOWICZ, MICHAEL. 1994. *Phonology in Generative Grammar*. Malden, Massachusetts: Blackwell Publishers Inc.

MCCARTHY, JOHN. 1998. Sympathy and phonological opacity. Ms., University of Massachusetts, Amherst. Available as ROA #252.

——, and ALAN PRINCE. 1993. Prosodic Morphology I: constraint interaction and satisfaction. Ms., University of Massachusetts, Amherst, and Rutgers University, New Brunswick, NJ. RuCCS TR-3.

——, and ——. 1997. Faithfulness and identity in Prosodic Morphology. In *The Prosody-Morphology Interface*, ed. by Harry van der Hulst, René Kager, and Wim Zonneveld. Cambridge: Cambridge University Press. Also available as ROA #216.

MINKOVA, DONKA. 2000. Reduplication and syllable well-formedness: the criss-crossing of phonology and semantics. Abstract of a talk given at ICEHL2000. Available on-line as `http://www.uvigo.es/webs/h04/jperez/icehl/phono.htm#minkova`.

PAUL, ANN WHITFORD. 1998. *Hello Toes! Hello Feet!*. DK Publishing. Illustrated by Nadine Bernard Westcott.

SCHIFFMAN, HAROLD F. 2001. Echo-word reduplication lexicon page. On the web at `http://ccat.sas.upenn.edu/plc/echo/`.

STERIADE, DONCA. 2001. The phonology of perceptibility effects: the P-map and its consequences for constraint organization. Unpublished Ms.; draft of 20 May 2001. Available from `http://www.linguistics.ucla.edu/people/steriade/steriade.htm`.

TAYLOR, INSUP K., and MAURICE M. TAYLOR. 1962. Phonetic symbolism in four unrelated languages. *Canadian Journal of Psychology* 16.344–356. Reference C353 from Brogan (1999).

WALDEN, BRIAN, and ALLEN MONTGOMERY. 1975. Dimensions of consonant perception in normal and hearing-impaired listeners. *Journal of Speech and Hearing Research* 18.445–455.

YIP, MOIRA. 1998. Identity avoidance in phonology and morphology. In *Proceedings of the Conference on Morphology and its Relation to Syntax and Phonology*, ed. by Steven G. Lapointe, Diane K. Brentari, and Patrick M. Farrell, 216–263, Stanford, CA. Also available as ROA #82.

——. 1999. Reduplication as alliteration and rhyme. *GLOT International* . Also available as ROA #377.

——. 2000. Segmental unmarkedness versus input preservation in reduplication. In *Segmental Phonology in Optimality Theory*, ed. by Linda Lombardi. Cambridge University Press. Also available as ROA #378.

ZWICKY, ARNOLD. 1976. Well, this rock and roll has got to stop. Junior's head is hard as a rock. In *Papers from the 12th Regional Meeting of the Chicago Linguistic Society*, ed. by S. Mufwene, C. Walker, and S. Steever, University of Chicago. Chicago Linguistic Society.