

**The Comparative Optimality of Hebrew Roots: An Experimental Approach to
Violable Identity Constraints**

*Daniel L. Everett
Department of Linguistics
University of Pittsburgh*

and

*Iris Berent
Department of Psychology
Florida Atlantic University*

The Comparative Optimality of Hebrew Roots: An Experimental Approach to Violable Identity Constraints¹

1. Introduction

The research results reported here suggest an answer to a specific variant of the following oft-asked question: Why are some structures rare or absent from a grammar, while other forms, apparently only slightly different, are quite common in everyday speech (either as types or tokens)? The specific form of this question we address here is this: Why are trilateral roots in Modern Hebrew (henceforth Hebrew) without repeated initial consonants much more frequent in the lexicon than roots whose first and second consonants are identical? We consider various answers to this question here, each of potential significance to our understanding of the nature of speakers' knowledge of their grammar. We conclude that MH speakers share an active mental constraint against the concatenation of forms that are in some sense identical, i.e. a version of the familiar *Obligatory Contour Constraint* (OCP). Moreover, we also argue that speakers know the constraint violations of any given optimal output and that they are able to use this as a basis for ranking different forms. We refer to this knowledge as the *Comparative Optimality* of distinct forms.

Our discussion of these questions is based on results from two experiments conducted by Berent and Shimron (1997) (described in detail in section 2 below). These were originally designed to test and probe the relevance of the OCP, in particular the version of it developed in McCarthy (1986). McCarthy (1986:208) states the OCP as in (1):

(1) *Obligatory Contour Principle*

At the melodic level, adjacent identical elements are prohibited.

As originally applied to Semitic, the OCP constrains both inputs and outputs of morphophonological processes. Potential output violations of the OCP trigger a set of repair strategies which McCarthy labels "Antigeminatio". McCarthy observes that the scope of the OCP seems wider than the morphophonology, however. It also appears to constrain lexical structures. For example, in Hebrew, there are almost no lexical items which violate McCarthy's statement of the OCP. On the other hand, the evidence for OCP effects in the lexicon, at least in Modern Hebrew, comes largely from silence. The fact that the Hebrew lexicon lacks words which violate the OCP might result from any number of historical factors, including coincidence, and is not necessarily evidence for a synchronic restriction against OCP-violating lexemes.

Our principal goal in this paper is to offer experimental evidence that the dearth of OCP violations in the lexicon is not merely accidental, i.e. that the OCP is indeed a synchronically active constraint on Hebrew root forms. However, the results of the experiments reported on here lead us to refine our understanding of the OCP, requiring a reinterpretation of it as a family of violable constraints prohibiting the occurrence of identical consonants in different positions and configurations within Hebrew roots. By

reinterpreting constraints on Hebrew roots as violable constraints we offer support for and insight into constraint-based approaches to grammar, e.g. Optimality Theory (OT) and we solve certain problems inherent in the derivational approach of McCarthy (1986).

We are not the first to notice that the avoidance of identity in natural language is the result of a family of violable constraints - Yip (1995) argues in detail for a similar proposal. Still, our results are interesting and novel in three ways: they offer (i) independent support for an articulated division of identity avoidance responsibilities among different constraints; (ii) an interesting new source of evidence for the OCP from psychological experimentation in Modern Hebrew; (iii) evidence for speaker knowledge of constraints via Comparative Optimality. To restate these points more specifically, we argue below for the following points related to the OCP (identity avoidance) and speaker knowledge of constraints:

(2) The OCP (identity avoidance):

- a. The OCP *does* play an active role as a mental constraint on Hebrew root forms and is not, for example, an accidental artifact of the diachronic accretion of the Hebrew lexicon.
- b. The OCP is a family of related constraints on identity avoidance.
- c. The OCP constraint family is best understood in a framework of violable, ranked constraints, e.g. *Optimality Theory* (OT).

(3) *Comparative Optimality*:

When native speakers rank words according to their relative well-formedness, they compare the relevant portions of the Optimality-Theoretic tableaux associated with those words.

The paper is organized as follows. First, we discuss two experiments conducted by Berent and Shimron (1996) to test native speaker intuitions on the OCP in Hebrew, considering these experiments' design, implementation, and results. Next we offer a linguistic interpretation of these facts, in terms of ranked, violable constraints, as per OT. In particular, we argue that speakers compare outputs when asked for absolute rankings but that when asked for relative rankings, they consider the comparative optimality of the relevant forms, information derived from the *optimal line* of the Optimality-Theoretic tableaux associated with each form. In conjunction with our proposal, we also consider and reject two alternative analyses, one based exclusively on a particular statistical model (section 4) and the other based on auxiliary hypotheses incorporated into a derivational model (section 3.2.).

2. Experiments

2.1. Background

McCarthy (1986) provides essentially two types of evidence in support of the lexical OCP: (a) distributional asymmetry: Gemination is more frequent at one or another edge of a given type of phonological unit; (b) the 'integrity' of tgemminates: tautomorphic gemminates manifest integrity with respect to various phonological processes.² The attribution of these observations to the OCP relies on two assertions. The positive assertion is that all known cases of gemination are attributable to a double linking of a

single melodic element. The ubiquity of such 'true' geminates is implicitly taken as evidence for their well formedness. Importantly, the inference of 'no lexical gemination' crucially relies also on negative evidence: The absence of fake geminates, and specifically, the absence of fake geminates at the left edge of a morphological or phonological unit. Because left edge geminates cannot result from left to right spreading, their representation must specify the geminates lexically. The rarity of left edge geminates is thus strongly compatible with the idea that gemination is highly marked, if not entirely absent, in the lexicon. Furthermore, the absence of such forms is taken as evidence for their ill-formedness: the attribution of statistical rarity to a linguistic principle assumes that the rare element must be ill formed, hence, undesirable.

Much of the criticism against the OCP has been directed towards the data supporting these inferences. Specifically, Odden (1986) has presented counterevidence to the OCP, suggesting that it fails to apply universally, although the force of this argument is somewhat diminished by the view of phonological principles as violable constraints (e.g., Prince & Smolensky, 1993). The violation of a linguistic principle thus no longer requires its rejection. Yet regardless of whether the evidence for the OCP is based on the absence of counterevidence or merely their rarity, it still heavily relies on negative evidence. It is this rationale of negative evidence which is the center of our investigation. The absence or rarity of counter evidence OCP cannot constitute evidence for the structural ill-formedness of such forms nor can it demonstrate the well-formedness of OCP-compatible structures.

Consider, for instance, the structure of the root morpheme in Semitic. McCarthy showed that root-initial geminates are extremely rare, and that root-final geminates maintain their integrity. The statistical rarity of SSM-type roots is unlikely random, but its cause may not necessarily be synchronically active. Hence, it is at least possible that forms such as SSM may not be undesirable synchronically. Conversely, the frequent, SMM-type roots are not necessarily synchronically acceptable, e.g. for producing new or nonce forms. In short, the inference of the undesirability of gemination from its absence is uncertain. The present investigation was stimulated by a desire to determine whether positive evidence for the acceptability of geminates in surface root forms could be obtained. Our evidence comes from Modern Hebrew (MH). Several observations make MH an interesting case study for the acceptability of lexical geminates. First, much of the support for the OCP comes from the study of Biblical Hebrew, a language that has not been actively used in a language community for hundreds of years. MH differs from its Biblical antecedent in substantial aspects of its phonetics. Thus, even if the evidence from Biblical Hebrew was due to a diachronically active mental principle, such a principle may no longer be present in MH. Second, MH presents phenomena that challenge the OCP. It violates the prohibition against root-initial geminates in two highly productive and frequent surface root forms (e.g., **mmn** and **mmsh**)³.

2.2. Experimental evidence for the acceptability of geminates

In this section we report the outcome of an experimental investigation of the status of lexical geminates (Berent & Shimron, 1997). These experiments examine whether the

statistical distribution of geminates indeed corresponds to their acceptability. Rather than inferring the existence of such a preference from the absence of evidence, these studies directly probe for its presence by means of an experimental manipulation comparing the acceptability of geminates at root-initial position with that of root-final geminates and no geminates controls. These data reflect an active rejection of root-initial gemination. Although such data are compatible with McCarthy's (1986) proposal, several aspects of the findings suggest revisions in its formulation. We next consider an OT account for the preference against gemination.

According to McCarthy (1986), root final gemination in surface forms is due to left to right spreading from an underlying biconsonantal representation. Because root-initial gemination cannot surface from left to right spreading, it must reflect lexical gemination. The rarity of such forms is normally attributed to the unacceptability of lexical geminates. Conversely, the frequency of surface forms with root-final gemination is taken as evidence for their desirability, and, specifically, the acceptability of multiple linking (as opposed to lexical geminates). As noted above, however, the assumption that the distribution of these surface forms reflects their acceptability has not been supported by any positive empirical evidence. The experimental investigation reported next compares the acceptability of lexical vs. nonlexical geminates. Specifically, these experiments address two questions. (1) Is root initial gemination unacceptable? (2) Is root final gemination acceptable.

2.3. Experiment design

The acceptability of gemination was assessed by asking native Hebrew speakers to rate words derived from three types of roots. To ensure that the ratings do not reflect familiarity with specific lexical tokens, Berent and Shimron (1997) used nonroots, combinations of three consonants that do not correspond to any existing Hebrew root. For convenience, we hereafter refer to these consonant combinations as roots, although neither the roots nor the resulting words exist in Modern Hebrew.

The primary question examined in the study is the sensitivity of Hebrew speakers to the location of geminates in the root. The critical items manifested root initial gemination (e.g. SSM). If the rarity of such forms reflects their undesirability, then derivations of SSM roots should be considered unacceptable. Although the rejection of SSM roots is compatible with the OCP, it may also stem from extraneous reasons. The rejection of a word may reflect a general bias, rather than sensitivity to its structure. Even if speakers' judgments are structure sensitive, the rejection of roots like SSM may be due unrelated to the location of geminates. For instance, the rejection of root-initial gemination may be due to the mere presence of gemination, rather than to its location. Alternatively, the rejection of an SSM roots might stem from the unacceptability of the second and third consonants, rather than the initial two, the geminates. To ensure that the acceptability of SSM type roots reflects structure sensitivity to the location of geminates, it is thus necessary to compare them to some other roots which control for these factors. Accordingly, the critical, initial gemination root was compared to two matching control roots. One control manifested root-final gemination (e.g., SMM⁴). This control maintained the same

geminate and nongeminate radicals as in the initial-gemination root SSM, but reversed their order. If the rejection of SSM type roots is due to the location of geminates, rather than merely to their presence, then SSM type roots should be rated lower than SMM type controls. Another potentially extraneous reason for the rejection of SSM type roots is idiosyncratic ill-formedness in the combination of the second and third radicals. This possibility was assessed by means of a no gemination control (e.g., PSM). These controls maintained the second and third radicals of the critical SSM type root, but replaced the initial radical with another (nongeminate) consonant. Together, the set of three matching root types, the critical, initial gemination roots, and their final gemination and no gemination controls, permitted us to assess the acceptability of root initial gemination. A consistent rejection of SSM type roots relative to both SMM and PSM type controls would suggest that root-initial gemination is unacceptable.

A second question examined in these studies is the effect of word pattern on the acceptability of geminates. The OCP constrains the location of the geminates within the root morpheme. A sensitivity to the OCP thus requires the representation of the root as a mental variable, and its decomposition from the word pattern. These pre-requisites are not universally met by cognitive theories. Eliminative connectionism (e.g., Rumelhart & McClelland, 1986; Seidenberg, 1986; Seidenberg & McClelland, 1989) considers the representation of mental variables, in general, and morphemes, in particular, as obsolete. If the domain of the OCP is eliminated then the OCP is unrepresentable (Berent, Everett & Shimron, 1997). Furthermore, even if the root is represented, it is unclear whether it is routinely decomposed from the word when the morphological structure of the word is opaque. To demonstrate that the domain of the constraint on consonant co-occurrence is indeed the root, it is necessary to dissociate the location of geminates in the root from their location in the word. For this end, the root triplets were conjugated in three classes of root patterns (see Figure 1).

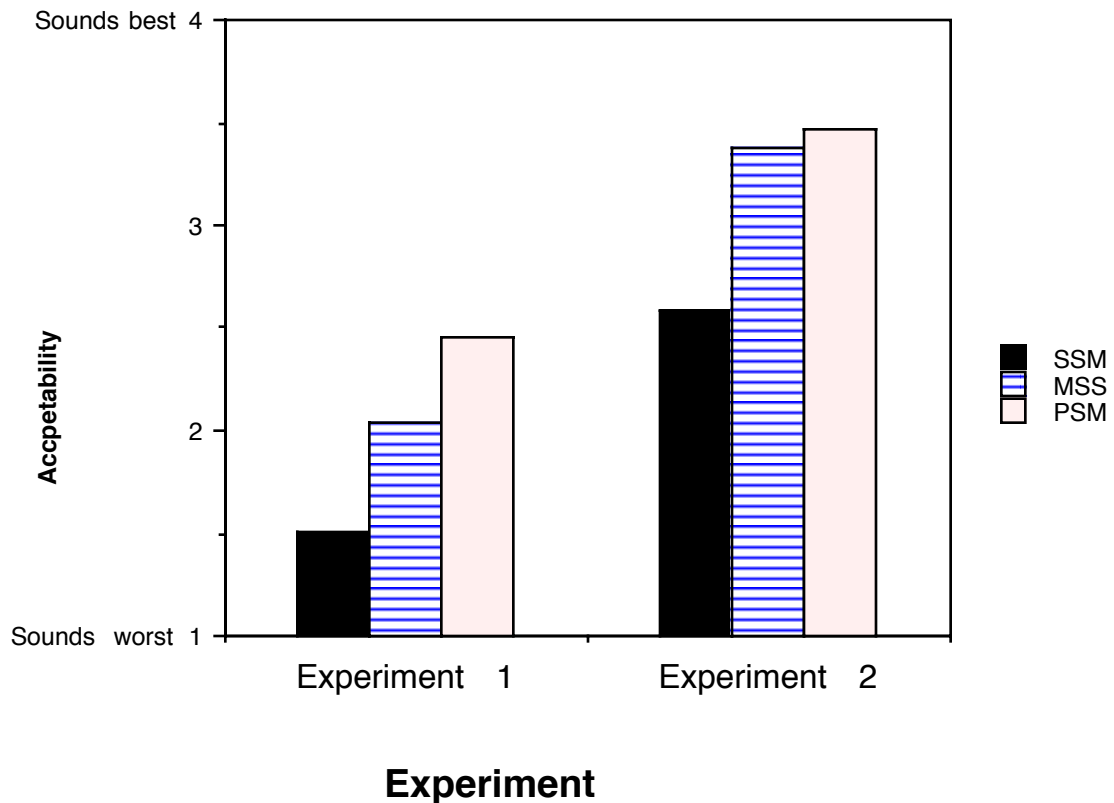
Figure One The structure of the words rated in Experiments 1-2/

	root	class 1	class 2	class 3
Initialgemination	[ssm]	Si-SeM	maS-Si-Mim	hiS-ta-SaM-tem
Finalgemination	[smm]	Si-MeM	maS-Mi-Mim	hit-Sa-MaM-tem
No gemination [psm]	Pi-SeM	maP-Si-Mim	hit-Pa-SaM-tem	

In the first class, the root was unaffixed. The location of the geminates in the root was similar to their location in the word, and the morphological structure of the word was thus highly transparent. These forms included the binyanim *pi'el* and *qal* in the third person singular masculine perfect. The two other word classes were both prefixed and suffixed, hence, the location of geminates in the root differs from their location in the word, and the morphological structure of the word is opaque. However, the second and third word classes differed with respect to the surface adjacency of the geminates. In the second class of items, root-initial geminates were not separated by a full vowel (e.g., **maS-Si-Mim**⁵). These forms included the binyanim *hif'il* and *nif'al* in the past tense and *beynoni*, and the mishkal *tafil*. Conversely, in the third class, including forms in the *hitpa'el*, the geminates were always separated by at least a full vowel (e.g., **hiS-ta-SaM-ti**). The investigation of the effect of word pattern on the acceptability of root structure thus addressed two important questions (a) Is the expected rejection of SSM-type roots due to a sensitivity to root or word structure? (b) Does the acceptability of geminates depend on their surface adjacency in the word?

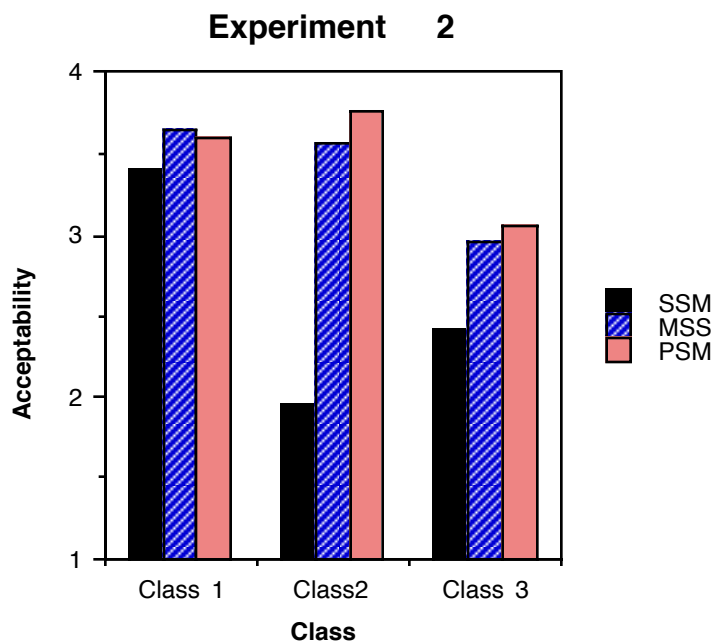
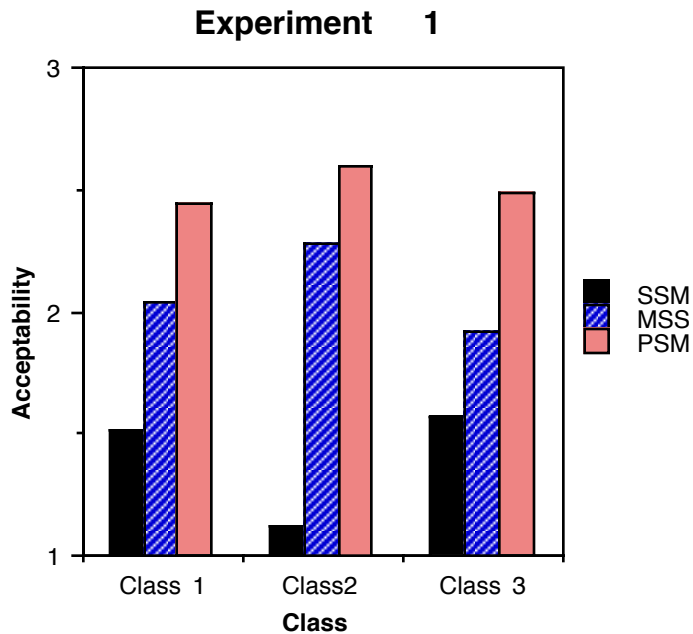
These questions were investigated in two studies. In these studies, native Hebrew speakers were asked to rate the acceptability of 216 words formed by conjugating 24 root triplets (e.g., SSM, SMM, PSM) in each of the three word classes. The two studies differed in the method of rating used. Experiment 1 obtained a rating of the members of the triplets *relative* to each other: subjects were presented with a printed list of 72 word triplets, and asked to determine the acceptability of each member of the triplet relative to the other members using a 1-3 scale. Experiment 2 obtained *absolute* ratings: Subjects were presented with a randomized printed list of the entire set of 216 words and asked to determine the acceptability of each word separately on a scale of 1 to 5. The words in each of these experiments were presented with all their vowels specified using diacritic marks. There were 18 participants in Experiment 1 and 15 participants in Experiment 2.

Figure Two Acceptability ratings in Experiments 1-2 as a function of root type



The principal findings of these studies are the following (a complete description of the method and results may be found in Berent & Shimron, 1997). First, root-initial gemination was considered ill formed⁶. Words with root initial gemination, (e.g., **SaSaM**), were rated significantly lower compared to final gemination controls, (e.g., **MaSaS**). Because these roots differed only in the location of the geminates, these findings must indicate the unacceptability of root-initial gemination, rather than merely the rejection of gemination. Similarly, the rejection of SSM-type roots cannot be attributed to some ill-formedness in the second and third radicals: Root initial gemination was rated significantly lower than no gemination controls, PSM, which were equated with regards to their second and third radicals. Hence, root initial gemination is unacceptable.

Figure Three Acceptability rating as a function of root type in Experiment 1-2



Root-initial gemination was unacceptable even when word structure was opaque, at the second and third word classes. SSM roots were rated significantly lower relative to either root-final gemination or no-gemination controls in each of these two classes. These findings suggest that subjects are sensitive to the location of geminates in the root, rather than merely in the word. However, the acceptability of root-initial gemination was modulated by word class. The disadvantage of SSM roots was strongest in the second

class, in which the geminates were not separated by a full vowel (e.g., **maS-Si-Mim**). Thus, the unacceptability of root initial gemination depends also on surface adjacency: The surface adjacency of root initial geminates increases their ill-formedness.

Interestingly, however, Experiment 1 revealed a general bias against gemination. Although subjects were clearly sensitive to the location of gemination, preferring root-final over root initial gemination, root-final gemination was nevertheless rated significantly lower than no gemination controls. The rejection of root-final gemination was replicated in a second study using the same rating procedure (see Berent & Shimron, 1997; Berent, Everett & Shimron, in preparation). In each of these studies, root-final gemination was rated significantly lower than no gemination controls in each of the three word patterns. Conversely, the open ended rating procedure used in Experiment 2 did not reveal such a bias (a finding replicated also by Berent, Everett & Shimron, in preparation).

In summary, the findings of the two experiments suggest that (i) root initial gemination is unacceptable; (ii) the rejection of SSM type roots increases when they are truly adjacent. (iii) root-final gemination is undesirable as well.

The principal finding of Berent & Shimron (1997), namely, the undesirability of SSM type roots clearly supports the lexical OCP. However, the two other findings suggest that preferences regarding gemination must stem from additional sources as well. One source is the surface adjacency of the geminates. The interpretation of this finding is not entirely certain, since phonetically, geminates and nongeminates do not contrast in Modern Hebrew. It is thus possible that the rejection of forms like **maS-Si-Mim** also reflects their ill formedness at a phonetic level. However, if the rejection of second class forms with root initial gemination stems from their phonological representation, then this finding may be explained by the derivational version of the OCP (McCarthy, 1986). More problematic to the OCP is the third finding, showing a rejection of SMM-type roots. Specifically, there are two aspects of this finding that are at odds with the OCP. One is the relative unacceptability of root final gemination. This finding is surprising, since root-final gemination is extremely frequent in modern Hebrew. Evidently, the statistical frequency of a linguistic structure is not necessarily evidence for its psychological desirability: Despite its ubiquity, root final gemination is relatively undesirable. To our knowledge, the unacceptability of root final geminates has not been previously documented. This is an important point, to which we return in section 3. A second interesting aspect of this finding is the sensitivity of this constraint to the nature of the rating procedures: root final gemination was deemed unacceptable in the relative rating technique, but not using absolute rating. Note that this divergence cannot be attributed to random error: The bias against root final gemination is replicable and highly reliable. Yet, the absolute rating procedure showed no hint of the unacceptability of SMM-type forms. The problem raised by this finding, then, is the following: The rejection of root final gemination suggests SMM-type roots violate a mental constraint. Yet, in Experiment 2, the same roots are rated just as acceptable as PSM roots, which do not violate any constraint. If root final gemination is ill formed, then why are SMM-type roots acceptable and frequent in Semitic? It is this question we now address.

3. A linguistic analysis

3.1. Introduction

In this section we argue that the experimental results reported in the previous section are best interpreted in terms of ranked, violable constraints, with no derivations.

In order to better appreciate the specific relevance of the present analysis, we begin with a brief review of the mechanics of McCarthy's (1986) analysis. It turns out that under careful inspection, McCarthy's account shares some significant aspects with an OT-like approach to the problem. This observation will lead us to consider a more explicitly OT approach to the data.

3.2. McCarthy 1986

McCarthy (1986,209) claims that '... the CiVCi sequence is prohibited in stem-initial but not stem-final position.' He analyzes this in terms of the constraints in (4):

- (4) a. Arabic roots are subject to the OCP.
 b. All autosegmental spreading in Arabic is rightward.

These constraints operate together to predict the following grammaticality contrasts:

- | | | |
|---------------|-----------|----------|
| (5) a. *sasam | b. *sasam | c. samam |
| a | a | a |
| /\ | /\ | /\ |
| [CVCVC] | [CVCVC] | [CVCVC] |
| | \ / | \ / |
| s s m | s m | s m |

(5a) violates the OCP. (5b) violates rightward spreading. (5c) violates nothing, according to McCarthy's analysis.

Recall that in both of Berent & Shimron's (1997) experiments subjects were asked to rank individual roots, in order to test the OCP. If we examine Figure 4 above, we see that subjects in the second experiment accepted roots whose final two consonants were identical and roots with no identical consonants equally well, while uniformly rejecting roots whose initial two consonants are identical. These results are exactly what McCarthy's account predicts.

Were these the only results, we would have succeeded in offering support for the OCP as an active preference among native speakers of Modern Hebrew. And this would be a significant result, since there is no other study offering such support, to our knowledge. But the results from Experiment 1, the comparative ranking task, reveal a considerably more complex and, to our minds at least, more interesting picture.

Experiment 1 obtained an ordered ranking: root-initial identical consonants were worst, roots with no identical consonant were best, and SMM forms were ranked in the middle. According to McCarthy's account, this is strange, because they violate no rule nor constraint. So why should speakers find them any worse than roots with no identical consonants? Why was there no evidence of this in Experiment 1? What knowledge do speakers draw on to make this distinction?

Let's first attempt to provide an account for these gradient judgments which is consistent with McCarthy's analysis. For example, we might account for these facts by hypothesizing that they result from an accumulation of constraint violations throughout the derivation. So, for example, assume that the OCP applies whenever its structural description is met, prohibiting geminates at any grammatical level. This is, to be sure, a noninnocuous reinterpretation of the OCP, but let's assume it nevertheless for the sake of discussion. Assume further that the representation is inspected for OCP violations at three levels: a deep structure, an intermediate level and a surface representation. In the first, deep structure, the root corresponds to its lexical representation and it is segregated from vowels and affixes. At the second, intermediate level, the root is still segregated from vowels and affixes. However, in preparation for plane conflation, the doubly linked element reduplicates and attaches to two skeletal slots. Finally, following plane conflation, the surface level represents consonants and vowels on a single plane.

Under this view, we are able to propose a possible reason for why speakers rank SSM forms lower than forms without identical consonants. If SSM outputs correspond to SSM inputs, but SMM-outputs correspond to SM inputs (as they would do, based on markedness and learnability) then the derivations of the relevant forms entail different numbers of violations of the OCP. SSM-type forms violate the OCP at both the deep and the intermediate level, leading to an accumulation of at least two violations of the OCP. Forms such as maSSiMim, where the geminates are also adjacent at the surface level, manifest a third violation of the OCP. Conversely, MSS-type forms do not violate the OCP at the deep level, because their input is MS. However, in preparation for plane conflation, their rightmost radical reduplicates, thereby violating the OCP at the intermediate level. Thus, MSS-type forms accumulate only one violation of the OCP. Finally, according to our analysis to this point PSM forms will not accumulate any violations of the OCP.

Table One
Violations of the OCP

		Deep Structure	Intermediate Levels	Surface Structure
INPUT	OUTPUT			
SSM	SiSeM, etc.	*	*	
SSM	maSSiMim, etc.	*	*	*
SM	SiMeM, etc.		*	
PSM	PiSeM, etc.			

This derivational account of the forms does, at first blush at least, seem able to distinguish the forms. If speakers have access to the total number of violations of a construction across derivational levels, then we account for the relative judgments observed in Berent & Shimron's experiments. Specifically, this analysis then accounts for the relative 'badness' of SSM and SMM forms, and also for the fact that surface adjacency produces an even worse sense of 'badness' for native speakers.

There are three reasons to question this derivational analysis, however. First it works by applying a rule at multiple levels of analysis. While this is not immediately fatal, it can indicate that a generalization is being missed, as Halle (1962) pointed out long ago in his criticism of the structuralist levels of phonemics and morphophonemics. *Ceteris paribus*, if a constraint or rule can achieve the same results without the need to apply at multiple levels, this would be a more desirable approach. Second, this reinterpretation entails, contra McCarthy (1986), that the final consonants in SMM-type forms are viewed as distinct by the OCP. This would be curious since the OCP needs to distinguish roots with initial geminates from roots with final geminates based on the fact that final geminates do not violate it. Yet here they appear to need to violate it. This is somewhat paradoxical. One might account for the problem by arguing that the consonants are somehow separated prior to Conflation and that at that point they violate the OCP. But this is clearly a complication of the analysis. Finally, and perhaps more importantly, this solution requires the recognition of violable constraints - the OCP is violated with SMM forms, but such forms surface nevertheless. But this means that we not only must allow the OCP to block other rules, as in McCarthy's analysis (e.g. by ranking it above those other rules in some way), but that we must also guarantee that it will not prevent a form which violates it from surfacing, as with the SMM forms in Table One. This is a most curious state of affairs and will only follow if the theory has a well-developed subtheory of the nature of constraints and their interactions. It is exactly this type of situation that has led many linguists to try to develop theories of constraints. Since there is a well-developed theory of constraints 'on the shelf', as it were, namely, Optimality Theory, it is worth pursuing an analysis in that framework, since we have been independently led to consider ranked constraints in our analysis. We therefore turn now to consider an OT analysis of the facts.

3.3. The nature of the constraints

If we are correct that an account of these facts must assume ranked constraints, then what might the relevant constraints be? Consider, as an initial proposal, the constraint ranking in (6):

(6) **OCP** >> ***IDENTITY** >> ***INITIAL IDENTITY**⁷

These constraints are to be interpreted as follows:

- (7) a. OCP: Adjacent identical items are prohibited (McCarthy (1986)).
 b. *IDENTITY: Identical consonants are prohibited within the root.

- c. *INITIAL IDENTITY: The first two consonants of the root are nonidentical.

These are all supported both by relative numbers of root forms in the Hebrew lexicon, and evidence from Berent & Shimron's (1997) experiments: *IDENTITY is justified by the speaker preference for no identical consonants in the root; *INITIAL IDENTITY is justified by the fact that native speakers reject nonce words which begin with initial identical consonants. Since this effect is distinct from and stronger than *IDENTITY, it will rank SSM forms below SMM forms.

We will assume that SMM forms are derived by reduplication from SM inputs, following Gafos (to appear). This is so since, as Gafos points out, an analysis based on Spreading would entail multiple levels of derivation, a consequence unacceptable in an OT framework. Reduplication not only avoids this problem it is, as Gafos argues, more parsimonious and empirically superior to the Spreading account.

We now turn to consider an interesting implication of these findings both for OT and for linguistic theory more generally. We will label this implication *Comparative Optimality*.

3.4. Comparative Optimality

Again, what the experiments show is that even optimal forms may violate constraints which, given the right context, can lead speakers to reject forms which they would otherwise accept. Why might this be? To answer this, let's begin by considering the data in the following tableau. These data represent the optimal lines (i.e. where the 'pointy finger appears in the original tableaux) for the tableaux of the nonce forms. Remember, rows three and four in the tableau below show common surface forms. That is, at the level of individual words at least, they are both 'optimal'.

Tableau

	OCP	*INIID	*FINID
OUTPUT			
SiSeM, etc.		*!	
maSSiMim, etc.	*!	*	
SiMeM, etc.			*
PiSeM, etc.			

Comparing these words, **maSSiMim** violates the greatest number of constraints. Hence, it should be the least acceptable form. Conversely, **PiSeM** violates the smallest number of constraints. It is predicted, therefore, that speakers will prefer words of this shape (i.e. CiCjCk) to both the **SiSeM** and the **SiMeM** forms. However, we need to say a bit more about this prediction since it depends on a particular view of what speakers know about constraints.

Let us assume that given an input **SM**, a speaker of Hebrew can only produce **SiMeM**, based on the effect of higher ranked constraints in the tableau. Does the speaker know that this output, while optimal given the input, still violates high-ranking constraints? Put another way, how does the speaker compare *optimal* outputs? How can s/he tell that one output, although optimal given its input, is 'less optimal' compared to a lexically independent output? The data here indicate that we must allow that speakers have a relative or *omparative* notion of optimality that extends beyond individual tableaux, allowing them to compare *different* words, roots, etc., by comparing their associated tableaux. Notice that they need not compare entire tableaux, since only the optimal lines are relevant.

For linguistic theory more generally, this finding means that native speakers have access to information about the comparative well-formedness of words, a type of knowledge which goes beyond the normally recognized semantic, syntactic, and phonological information present in words. It would appear that speakers also know and have access to the constraints violated at the optimal line of each tableau. There are various ways one might represent this knowledge, but it goes beyond the scope of the present paper to engage this issue.

3.2.6. Representing word-initial geminates

Before we can close our account, however, we need to consider the important, howbeit small, set of words which violate *INITIAL IDENTITY, seen in (8)-(10), these are the only violations we are aware of with word-initial geminates:

- (8) **mimen** 'financed'
- (9) **mimesh** 'realized'
- (10) **nanas** 'dwarf'
- (11) **didah** 'limped'

Recall that *INITIAL IDENTITY violations were uniformly rejected in both ratings experiments as not possible Hebrew roots. How are these forms possible then? Clearly they must be marked in the Hebrew lexicon. We propose that this marking is achieved by entering these words with three consonants in the lexicon, i.e. with inputs consonantally identical to their outputs, as in (11)-(13):

- (12) **mmn** 'finance'
- (123) **mmsh** 'realize'
- (14) **nns** 'dwarf'
- (15) **ddh** 'limp'

This has a couple of advantages. First, it accounts for the fact that the initial consonantal identity is not the result of a morphological process (e.g. Reduplication) and not associated with any binyan or mishkal. Second, it succinctly indicates their markedness by their output form, with no need to resort to abstract diacritics.

How these forms came into the MH lexicon is a secondary issue. Once there, they will be more marked than other roots because their output violates *INITIAL IDENTITY, a higher ranked constraint relative to the OCP family.

4. Alternative analysis - Similarity and statistics

In recent work, Pierrehumbert (1993) and Frisch, Broe, and Pierrehumbert (1997) propose that morpheme structure conditions in Arabic, similar to those discussed above for MH, can be explained in terms of a singled gradient constraint on perceived similarity, rather than identity. In their model, the perceived similarity between root radicals is determined by the relative number of shared natural class features and the distance between the segments. In support of their account, Frisch et al (1997) present data suggesting that perceived similarity can account for the distribution of trilateral root forms in Arabic, and that the fit provided by the model is superior to that of the OCP place constraint proposed by McCarthy (1994).

The model presented by Frisch et al carry two implications for the present proposal. First, it suggests that the constraint on identity may be simply a particular case of a more general constraint on perceived similarity. Put differently, the rejection of *smm* and *ssm* may not be due specifically to their identity, but instead may stem from the full perceived similarity between adjacent identical segments. Second, the validation of their model by the statistical distribution of roots tacitly assumes that token distribution at least reflects well formedness. Conversely, we argue that the equation of well formedness and ubiquity is generally uncertain. In particular, we argued that this account cannot account for the constraint on segment identity, as root final gemination is weakly unacceptable, despite its ubiquity. We believe, however, that the evidence presented by Frisch et al is insufficient to support either of their two claims.

Consider first the attribution of the constraint on identity to perceived similarity.

The principle support provided by Frisch et al to the adequacy of the constraint on perceived similarity is a statistical analysis of the distribution of trilateral Arabic roots (see also Frisch and Zawaydeh 1997 for some pilot experimental data on relative acceptability of root types in Arabic). Crucially, however, they do not include *SMM*-type forms in their count. This is a potentially serious omission since their claim is that the more frequent a form is, the more acceptable it is. The justification for ignoring these facts is not given in their paper, but it is no doubt due to the fact that they adopt McCarthy's analysis of these forms as underlyingly *SM*. However, since we (and they) are concerned about output constraints, we cannot exclude a set of outputs based on a theory of their input forms. This would be circular in the following fairly pernicious way: these outputs violate the predictions but they can be 'excused' since they conform to the theory being tested. A second limitation of their analysis for comparing the constraint on identity with perceived similarity is the failure to specify the location of the segment in the root by collapsing the counts for the initial two radicals with the latter two. Given the strong asymmetry in the distribution and acceptability of identical segments in the root, it is unlikely that a model that disregards their location can adequately account for their distribution. Thus, while we believe that it will be necessary to further investigate the

relationship between OCP and perceived similarity, in general, and OCP-place, in particular, it is still unclear whether the constraint on perceived similarity can adequately account for identity constraints in Semitic.

However, even if perceived similarity could adequately handle the distribution data of root forms in Semitic, it is unclear whether such a model would necessarily account for the well formedness of roots with geminates. We contend that statistical analysis of the lexicon does not necessarily correspond to well formedness. This claim stands at odds with Frisch et al.'s attempt to validate their model by distributional analysis. Underlying this analysis is the assumption that the distribution of tokens at least correlates with well formedness. A stronger version of this view, held by eliminative connectionist theory, assumes that acceptability is merely an artifact of token co-occurrence. On this view, the acceptability of linguistic forms stems from the acquisition of information regarding the co-occurrence of token specific features, rather than the constituent structure of formal types. The rejection of SMM type roots challenges both positions. SMM type forms are highly frequent. In subsequent studies, Berent Everett and Shimron (in preparation) observed a rejection of SMM type roots using a set of materials in which the summed type frequency of adjacent bigrams (C1C2 and C2C3) was significantly higher than that observed for PSM type controls⁸. The dissociation between the statistical frequency of the materials and their acceptability has several important implications to our discussion. First, it suggests that the rejection of SMM type forms cannot be simply attributed to the rarity of the tokens used in the study. Second, this finding stands at odds with eliminative connectionist models that view formal structure as an artifact of token properties. Finally, these findings emphasize the methodological repercussions of our proposal. The dissociation between frequency and acceptability indicates that well formedness cannot be accounted for solely by statistical data. The experimental approach illustrated in this study may provide important converging evidence for linguistic theory.

References

- Berent, Iris, & Shimron, Joseph. (1997). 'The representation of Hebrew words: Evidence from the Obligatory Contour Principle', *Cognition*, 64, 39-72.
- Berent, Iris, Everett, Daniel L. & Shimron, Joseph. (1997). 'Do phonological representations specify formal variables? Evidence from the Obligatory Contour Principle', ms submitted for publication.
- Berent, I. Everett D. & Shimron, J. (in preparation). The structure of Hebrew roots: evidence for violable mental constraints.
- Chomsky, Noam and Morris Halle. 1968. *The sound pattern of English*, Harper & Row, New York.
- Frisch, Stefan. 1996. *Similarity and frequency in phonology*, unpublished Ph.D. dissertation, Northwestern University.
- Frisch, Stefan, Michael Broe, and Janet Pierrehumbert. 1995. 'The role of similarity in phonotactic constraints', ms. Northwestern University, Evanston, IL.
- Frisch, Stefan and Bushra Zawaydeh. 1997. 'Experimental evidence for abstract phonotactic constraints', *Research on spoken language processing*, Progress Report No. 21, Indiana University, Bloomington, IN.
- Frisch, Stefan, Michael Broe, and Janet Pierrehumbert. 1997. 'Similarity and phonotactics in Arabic', ms. submitted for publication, Northwestern University, Evanston, IL.
- Gafos, Diamandis. 1997. 'Eliminating long-distance consonantal spreading' to appear in *NLLT* 16.
- Goldsmith, John. 1979. *Autosegmental phonology*, Garland, New York.
- Greenberg, Joseph. 1950. 'The patterning of root morphemes in Semitic', *Word* 5: 162-181.
- Halle, Morris. 1962. 'Phonology in Generative Grammar', *Word* 18: 54-72.
- MacEachern, Margaret R. 1997. *Laryngeal Cooccurrence Restrictions*, unpublished Ph.D. dissertation, UCLA.
- McCarthy, John. 1986. 'OCP effects: gemination and antigemination', *Linguistic Inquiry* 17:2, 207-263.
- McCarthy, John. 1988. 'Feature geometry and dependency: a review', *Phonetica* 43: 84-108.
- McCarthy, John. 1994. 'The phonetics and phonology of Semitic pharyngeals', in: Patricia Keating (ed.), *Papers in laboratory phonology III*, Cambridge University Press, Cambridge, pp. 191-283.
- Odden, David. 1986. 'On the Obligatory Contour Principle', *Language* 62: 353-383.
- Pierrehumbert, Janet. 1993. 'Dissimilarity in the Arabic verbal roots', in *Proceedings of the North East Linguistics Society* 23, pp367-381.
- Rumelhart, David E. & McClelland, Jay L. (1986). 'On learning the past tense of English verbs: Implicit rules or parallel distributed processing?' in: Jay L. McClelland, David E. Rumelhart & The PDP Research Group (eds.). *Parallel distributed processing: Explorations in the microstructure of cognition*, Vol. 2: Psychological and biological models (pp. 216-271). Cambridge, MA: Bradford Books/ MIT Press.

- Seidenberg, M. (1987). 'Sublexical structures in visual word recognition: Access units of orthographic redundancy?' In M. Coltheart, (Ed.), **Attention and performance XII: Reading** (pp. 245-263). Hillsdale, NJ: Erlbaum.
- Seidenberg, M., & McClelland, J. (1989). 'A distributed developmental model of word recognition and naming', **Psychological Review**, 96, 523-568.
- Yip, Moira. 1995. 'Identity avoidance in phonology and morphology', Rutgers Optimality Archive (<http://ruccs.rutgers.edu/roa.html>)

Notes

¹ We would like to thank Stefan Frisch, Sally Thomason, Terry Kaufman, Peggy MacEachern, Diamandis Gafos, Colleen Fitzgerald, and the participants in the Fourth Manchester Conference on Phonology for many helpful comments on this research.

² Our use of the terms 'gemination' and 'geminate' is intended to refer to identical consonants which result from surface or underlying multiple linking. This use is strictly for expositional reasons and has no theoretical status.

³ **sh** is used to represent the voiceless alveopalatal grooved fricative.

⁴ The root MSS is an existing root in Hebrew. We use MSS to illustrate the structure of the appropriate control for SSM. However, neither SSM nor MSS was used in the study. None of the roots used correspond to an existing root.

⁵ To clarify the morphological structure of the root, we indicate root consonants in upper case. No such orthographic distinctions were present in the experiments.

⁶ A detailed description of the statistical analysis may be found in Berent & Shimron (in press). In all findings referred to as significant, $p < .05$ by subjects and items.

⁷ One might reasonably ask if these constraints have any intuitive basis or whether we simply pulled them out of the air. While this is largely speculative, there does seem to be an intuitive basis for morpheme structure constraints generally, of which those in () are but a subset. So, for example, a functional motivation for such constraints might be that MH speakers prefer to avoid identical consonants in the root, especially at the beginning of the root, because the existence of such a constraint makes it easier to identify roots, an important parsing aid. MacEachern (1997), for example, argues that the left edge of words and roots can be an important parsing cue, when aided by different kinds of phonological processes which she investigates. Of course, even if the constraints were functionally unmotivated, they would still be empirically well-supported and thus a crucial component of a theory of native speaker grammatical knowledge.

⁸ The summed type frequency of SMM type root was still higher than PSM, albeit not significantly so when nonadjacent bigrams were included (i.e., summing C_1C_2 , C_2C_3 , C_1C_3)