John Benjamins Publishing Company

This is a contribution from *The Mental Lexicon 2:3*

Semantics and phonology constrain compound formation

Mark S. Seidenberg[1], Maryellen C. MacDonald[1], and Todd R. Haskell[2]
[1]University of Wisconsin-Madison-Madison / [2]Western Washington University

Berent and Pinker (2007) presented five experiments concerning the formation of compounds, especially the apparent restriction on the occurrence of "regular" plurals as modifiers (as in *RATS-EATER). Their data were said to support a "words and rules" approach to inflectional morphology, and to contradict the approach developed by Haskell, MacDonald, and Seidenberg (2003) in which multiple probabilistic constraints, mainly involving semantic and phonological properties of words, determine degree of acceptability. We examine Berent and Pinker's studies and show that a) their experiments tested hypotheses that are incorrectly ascribed to our theory, and b) their data are actually compatible with our account. Contrary to the words and rules approach, there are phonological effects on modifier acceptability that cannot be subsumed by a grammatical rule.

Berent and Pinker (2007) have offered a response to our article (Haskell, MacDonald, & Seidenberg, 2003), which provided evidence against the theory that Pinker (1994, 1999) has proposed as an explanation of the apparent bias against regular plurals as modifiers in compounds (e.g., MICE-EATER is acceptable to many English speakers, whereas *RATS-EATER is not).[1] This analysis of compound formation has been repeatedly presented as supporting broader conclusions about the nature of grammatical knowledge, the distinction between words and rules, and the need for innate constraints to explain how such knowledge can be acquired (e.g., Gordon, 1985; Pinker, 1994, 1999). This account is based on level-ordering, a conceptual framework that arose within phonological and morphological theory some years ago (Kiparsky, 1982; Siegel, 1974) and which drew distinctions between regular and irregular inflectional processes.

Since its introduction, the level-ordering account has been subject to extensive criticism (Booij, 1989, 1993; Lardiere, 1995; Nicoladis, 2003, 2004, 2005). Haskell et al. (2003) also summarized various descriptive problems with the

The Mental Lexicon 2:3 (2007), **287–312**.
ISSN 1871–1340 / E-ISSN 1871–1375 © John Benjamins Publishing Company

level-ordering account in the context of offering an alternative approach to the plurals in compounds phenomena. In that work, we presented corpus analyses showing that regular plurals occur more often and irregular plurals less often than the level-ordering theory predicts; presented new behavioral data consistent with the corpus data; proposed a probabilistic constraint-based account of the relative acceptability of the different modifier forms, consistent with the corpus and acceptability results; showed that the theory correctly generalized to two additional cases (pluralia tantum in compounds such as PANTS LEG, voicing change plurals such as WOLVES); proposed an account of how the primary semantic and phonological constraints could be learned from naturalistic input; performed a corpus analysis providing evidence for the phonological constraint; implemented a neural network that discovered phonological properties that differentiate modifiers from nonmodifiers; and showed that a model-derived measure of modifier goodness accounted for significant variance in participants' ratings.

In short, we provided evidence that the well-formedness of compounds varies in degree as a function of multiple interacting probabilistic constraints, as is true of other aspects of morphology (Haskell & MacDonald, 2003; Seidenberg & Gonnerman, 2000) and syntax (Allen & Seidenberg, 1999; Bresnan, in press). We discussed two major constraints (specific semantic and phonological properties of the modifiers) and noted that others are probably involved as well. We presented this research as supporting broader conclusions about the nature of linguistic knowledge, the relevance of semantics and phonology rather than words and rules in explaining these and other facts, and how such linguistic knowledge can be acquired without innate grammatical knowledge such as level-ordered application of rules.

Berent and Pinker (hereafter BP) took issue with our data and conclusions. They presented a series of studies in which participants rated the goodness of noun compounds that varied in phonological or morphological structure. The results were offered as evidence that the singular versus plural morphological status of a prenominal modifier, rather than its phonological properties, strongly affects the goodness of the noun compound, thereby supporting the words and rules approach.

We don't agree. Of the five experiments that BP presented, Experiments 1–3 did not test valid hypotheses derived from our theory and are methodologically flawed to a degree that vitiates the results. Experiments 4–5 yielded results that are similar to ones we have already reported, and they are consistent with our theory. Thus, the validity of the constraint-based theory is unaffected, and the problems with level-ordering based theories remain. We consider their experiments in turn.

BP's Experiment 1

We claim that, among other factors, specific phonological and semantic properties of the modifier affect the acceptability of compounds. BP's account emphasizes that the major phenomena turn on the distinction between words and rules, where rules are ordered via assignment to different levels of lexical strata (see Pinker, 1999). Semantic factors were added to explain the many examples that violate the theory's core principles, such as PARKS DEPARTMENT and NEUROSCIENCES PROGRAM. We discuss the role of semantics below. However, BP's experiments focus on our claims about phonology, attempting to provide evidence that the grammatical distinction between singular and plural modifier is relevant rather than phonology.

In the course of Haskell et al. (2003), we developed the following hypothesis:

> HMS: people prefer modifiers that sound like other modifiers they have heard. They therefore disprefer modifiers that sound like regular plurals, because few modifiers have this phonological property. This property interacts with other constraints (e.g., semantic ones) to determine well formedness.

The most prominent phonological characteristic of plurals is that they end in /s/, /z/, or /iz/ as in BACKS, BINS, and BUSES, respectively (the allomorphs of the plural inflection that is spelled -s or -es; we will refer to these forms collectively as "plural phonology"), and so such forms are dispreferred. We also showed that more subtle phonological properties of modifiers affect judgments of compound acceptability in a graded manner. Thus, people experience many modifier-head constructions, of which compounds are a subset; they learn that few of them have plural phonology; compounds that contain modifiers with regular plural phonology therefore are literally deviant and rated as less acceptable than modifiers lacking regular plural phonology, other factors aside.

In their Experiment 1, BP tested the following "phonological familiarity" hypothesis, which they ascribed to us:

> BP: People prefer modifiers that sound like other words in their vocabulary. It follows that they should disprefer modifiers containing sound patterns that do not occur in the language.

To test this hypothesis, they compared "legal" nonwords such as LOONK, LEENK, and LOONKS to "illegal" nonwords like LOOVK, LEEVK, and LEEVKS. This legality factor, on which we have more to say below, was manipulated at the level of bigrams: the legal items contained bigrams that occur in other words in English, the illegal items contained bigrams such as VK, which do not occur in other English words. People should then prefer the legal items in compounds (e.g. LEENK-

EATER) compared to the illegal ones (e.g., LEEVK-EATER), on this version of our theory. BP's main finding was a dispreference for regular plurals in compounds, regardless of legality (BP's Table 1). BP argued that this effect is not phonological, because the legality manipulation had no impact, and it is not semantic because regular plurals were rated worse than the semantically-similar irregular plurals. Hence they inferred that the effect must be due to the grammatical distinction between singular and plural, supporting their theory and contradicting ours.

The problem with BP's conclusion is that our hypothesis is different than the one that they ascribed to us. BP's is about words that are phonologically deviant relative to *all other words in the language*; ours is about words that are phonologically deviant relative to *other modifiers*. Clearly our claim could not be about frequencies of occurrence in the language as a whole because many words end in /s/ or /z/: the CMU pronouncing dictionary shows that /z/ is the second most common word-final phoneme, and /s/ the fifth. Thus, mere "phonological familiarity" would predict that plurals, being highly familiar, should be highly acceptable modifiers. Our theory explicitly makes the opposite prediction; hence BP's characterization of it is incorrect.

The reasoning behind BP's error can be found in their footnote 3, which begins:

> There are multiple ways to define phonological familiarity. Haskell et al. first define their phonological constraint in terms of "whether a potential modifier has the phonological structure typical of a regular plural" (p. 131). As stated, this is difficult to evaluate, since by their own hypothesis the learner has no access to the category "regular plural," and hence cannot use it as a criterion to partition phonological space.

There is a confusion here. The *learner* does not have access to the concept "regular plural" *but we do*. The learner cannot use this term to partition phonological space, but researchers can use such terms to describe phonological patterns. Thus our use of the term "regular plural" is merely a way to describe particular phonological properties of the stimuli, not a claim about how this information is learned or represented. There are different theories about how such words are learned, produced and comprehended, and how they relate to so-called "irregular plurals" (we think that "regular" and "irregular" pick out regions on a continuum, and that they are processed using the same lexical network, encoding phonology, semantics and, in literate individuals, orthography). This confusion was avoidable given our article's extensive discussion of how the child might acquire this knowledge via statistical learning, without knowledge of rules or the regular-irregular distinction.

BP's footnote continues:

> Fortunately, their operational definition hints at a different hypothesis but one more consistent with their overall theory: "The phonological constraint arises from the fact that exposure to nouns and adjectives provides information regarding the phonological structure of modifiers" (p. 139). This in turn can be interpreted in two ways: that speakers learn the phonological properties of nouns and adjectives in the language as a whole, and infer that the same properties apply to nonhead members of compounds [i.e., modifiers], or that speakers learn the phonological patterns that specifically discriminate acceptable versus unacceptable nonheads. Our materials test both of these interpretations: Our phonologically unfamiliar nonheads are phonologically infrequent in the language as a whole as well as in compounds, specifically.

This passage requires clarification for several reasons. First, our account is not about rarity in the language as a whole. It states that children learn about the phonological properties of words that serve as modifiers. This knowledge includes the fact that whereas phonological plurals are common in the language, they rarely occur as prenominal modifiers. Children are then displeased (at a young age; Gordon, 1985) by compounds containing plural modifiers (although see Nicoladis, 2003, 2004), and adults continue to disprefer them. Our computational model learned about phonological properties that distinguish modifiers from non-modifiers, and we used this modifier-specific knowledge to predict the goodness of various compounds. Second, BP's text suggests that their stimuli provide a way of testing both our claim and the hypothesis about rarity in the language as a whole. However, in order to differentiate between these alternatives, the stimuli must vary with respect to these dimensions. BP's stimuli do not; they are phonologically atypical *both* as modifiers and in the language as a whole; the two properties are confounded. A sound pattern like LEEVK does not occur very often in the language and therefore necessarily also doesn't occur very often as a modifier. Hence such stimuli cannot distinguish between the two hypotheses, and thus they do not bear on our theory. In fact, our theory explicitly predicts that LEEVKS and LOONKS should both be poor modifiers because they have the plural phonology that is atypical of modifiers. In short, BP acknowledged the important distinction between typicality as a modifier vs. in the language as a whole, but their stimuli do not instantiate it and therefore do not allow a test of the alternative hypotheses.

Even if the stimulus manipulation had been appropriate, BP's experiment could not have yielded a test of our account, because the experimental design was structured so that BP's morphology hypothesis predicted a null effect of phonological legality. This null result, like most, is not informative; the lack of a difference between legal and illegal nonwords could occur for reasons unrelated to the validity of the theory under test, and would remain so even if the legality manipulation had been a relevant one.

In fact, although our primary concern is that the legality manipulation did not test a valid prediction of our theory, we note as an aside that the manipulation itself was extremely weak. In particular, both the "legal" and "illegal" nonwords contained illegal letter and phoneme clusters. Whereas BP manipulated bigrams (e.g., legal NK vs. illegal VK), many of the "legal" items contain larger units that are unattested in English (e.g., OONK, EENK). The stimuli in the legal and illegal conditions are also similar with respect other aspects of lexical structure such number of neighbours (because they have so few) and rime frequency (because they are extremely low; see BP's Table 2). The one factor on which the two groups differ, bigram frequency, is known to have little effect on measures of lexical processing: In a large-scale study of factors that influence word recognition, Balota, Cortese, Sergent-Marshall, and Yap (2004) explicitly excluded bigram frequency from their reported regression analyses, noting that it had no predictive value in preliminary analyses and that "there have been repeated failures to demonstrate an influence of this variable (see, e.g., Andrews, 1992; Treiman, Mullennix, Bijeljac-Babic, & Richmond-Welty, 1995)." (p. 285).[3]

In summary, BP were correct in stating that we hold that phonological properties affect the acceptability of modifiers in compounds. However, a valid test of this theory would have to assess the impact of *relevant* phonological properties. BP examined a structural difference between the stimuli that does not differentiate good and poor modifiers, and so has no effect. To compound the error (so to speak) they manipulated bigrams, a property that has little impact on lexical processing. The outcome was a meaningless null result.

Computational evidence

Further evidence that BP's data are compatible with our theory is provided by an assessment of their stimuli using the computational model presented in Haskell et al. (2003). We used a neural network as a tool to discover phonological properties that distinguish adjectives from nonadjectives and to derive a graded measure of relative modifier goodness. Adjectives modify nouns and so by definition instantiate phonological patterns characteristic of modifiers. The model was trained to classify words as adjectives or nonadjectives based on phonological information alone. Because there is overlap between the phonological properties of adjectives and other types of words, the model cannot learn to perform this task perfectly. However, in the course of training, it discovers phonological characteristics that probabilistically differentiate the categories (see Kelly, 1992, 2004, for discussion of phonological correlates of grammatical categories, and Cassidy, Kelly, & Sharoni, 1999, for a similar use of a connectionist model). Words varied with respect to how close they came to being classified correctly by the model, which we used as a measure of prenominal modifier goodness. The model was then tested on the modifiers that had been used in our studies of compound acceptability. Although the model had not been trained on these items, the derived measures of relative goodness were significantly correlated with subjects' ratings of compound acceptability (Haskell et al.'s Figure 7). These findings provided additional evidence that phonological properties of modifiers affect the well-formedness of compounds, including "regular" ones, contrary to the words and rules account. Phonology is not expected to account for all of the variance because phonological cues are probabilistic rather than absolute, and because other factors (e.g., semantics) also affect modifer well-formedness.

We used this same model to test the nonwords from BP's Experiment 1. The stimuli were run on our adjective classifier model, which of course had not been exposed to their nonwords in training. Results for the experiment and model are presented in Table 1. The modeling results indicate no effect of the legality factor; the model treats legal and illegal nonwords alike, as did BP's experiment participants. Also like the human raters, the model strongly dispreferred the nonwords with plural phonology. Thus, varying the legality of an irrelevant part of a stimulus has little impact. In contrast, the presence of the phonological plural had a major effect, for the reason we stated: although this sound pattern is common in the language as a whole, it is uncharacteristic of modifiers. In summary, BP's experiment was said to provide data incompatible with our account and therefore to require a

Table 1. Model classification values for berent & pinker's experiment 1 nonword stimuli

Nonword Phonology	Singular	Irregular Plural	Regular Plural
Haskell et al. (2003) Adjective Classifier Model			
Legal	0.9993	0.9974	0.0295
Illegal	0.9991	0.9976	0.0306
Noun Classifier Model (Phonology only)			
Legal	0.9358	0.9347	0.1925
Illegal	0.9162	0.9151	0.1927
Noun Classifier Model with High Noise on Semantic Units			
Legal	0.9371	0.8724	0.1035
Illegal	0.9202	0.8450	0.1031
Noun Classifier Model with Moderate Noise on Semantic Units			
Legal	0.9786	0.4982	0.0357
Illegal	0.9766	0.4652	0.0349

Values in the Adjective Classifier Model range from 0–1, with 1 being the most adjective-like. Thus high values predict items that should be well-suited to the prenominal modifying position in compounds. In the Noun Classifier model, values shown are output of the singular output node, with high values indicating items to be judged more singular. The model with high noise on the semantic units weighs phonology more heavily than does the model with moderate noise on the semantic units.

morphological rule, but our model, working only with phonological information, captures both the absence of an effect of legality and the strong dispreference for plural phonology.

More About Modeling. Having used the model from Haskell et al. (2003) to address BP's data, we should discuss the fact that BP did not find the original use of the model convincing and so discounted Haskell et al.'s results, suggesting they were artifactual. Specifically, they questioned why our model was trained on adjectives, given that the relevant behavioral data concerned noun modifiers. There were three reasons for this choice. First, adjectives are one of the main sources of information about the phonological properties of words that modify nouns. Second, adjectives are prominent examples of modifying expressions in the utterances that children hear in learning a language. Although children hear examples of noun compounds (e.g., CAR SEAT), they hear many more tokens of adjectival modification. Third, we hypothesized that adjectives and nouns were phonologically similar with respect to properties that affect prenominal modifier acceptability in compounds. Adjectives and nouns do not have identical phonological distributions; for example the syllables -FUL and -EST are more common in adjectives (COLORFUL, BIGGEST) than in nouns. However, the question is whether they tend to share properties that affect modifier acceptability. This empirical question was clearly answered by our results: a model trained on adjectives and then tested on nouns used as modifiers in our experiments accounted for significant variance in the ratings of noun compounds.[4]

In short, we take it as a significant *finding* that phonological properties of adjectives are systematically related to the acceptability of noun modifiers. But, to address BP's skepticism, we replicated the model using only nouns as the training set, with the task changed to classifying the input as singular or plural (see Appendix for details). Following training we tested the model on BP's nonword stimuli from Experiment 1. As shown in Table 1, this model also captures the critical findings from BP's experiment: there is again no effect of the legality manipulation, but there is a large difference between the conditions with plural phonology and the others. Again, we did not attempt to fit BP's data exactly, because the classification task does not incorporate all aspects of BP's rating task.

We do not expect BP to accept these modeling results, given the comments of BP and their colleagues on connectionist models to date. Their objections include arguments that every model is inadequate, failing to capture some aspect of the phenomena; that plugging the leaks in one model invariably creates new leaks in the successor model; that no model since Rumelhart and McClelland's (1986) ambitious but doomed effort is worth serious consideration; that the models use artificial tasks that don't correspond to reality; that the models merely implement the higher level symbolic theory; and so on.

Two brief comments on the modeling issue, which has been discussed elsewhere (e.g., McClelland & Patterson, 2002; Seidenberg & Plaut, 2006). First, every computational model *is* false in the sense of failing to capture some aspect of the data; that is inherent in the methodology. The models are tools, not complete instantiations of the human mind. Their value is in providing powerful ways to explore theoretical concepts and methodological assumptions. It is therefore necessary to thoughtfully analyze whether the limitations of a given model have any bearing on the theoretical claims it is meant to clarify. In the past, insignificant limits of specific implemented models have been erroneously taken as imposing limits on the entire class of such models (e.g., Pinker, 1999; Pinker & Prince, 1988; see MacWhinney & Leinbach, 1991, for an earlier discussion of this point). The second comment is that these tools are also available to BP. Instead of merely asserting without evidence that our model only accounted for acceptability rating data because it was trained to classify adjectives, they could have determined for themselves whether similar results would obtain using a model trained on other types of words, e.g., nouns, or to perform other kinds of tasks. Any good faith effort to understand the phenomena would find that the results presented in Haskell et al. (2003) reflect general properties of the language, which can be revealed using modeling and other quantitative tools; they are not an artifact of the particular models we implemented.

The role of semantics. There is one interesting difference between the behavioral results in BP and the results from both adjective and noun models shown in Table 1. In the BP study, irregular plurals were rated as intermediate in acceptability, better than regular plurals but worse than singulars. In contrast, the models grouped together the singulars and irregular plurals, both of which were more acceptable than the regular plurals. The fact that irregular plurals did not pattern with singulars in the human ratings requires some post hoc rationalization from within the words and rules perspective; since both forms are stored in the lexicon rather than generated by rule, they should be equally available for compounding and thus equally acceptable. BP's explanation appeals to task demands of the experiment: Participants were always confronted with a choice between singular and plural forms, as established both by the story contexts that were provided and by the rating task used in all five BP experiments, in which participants rated the goodness of the compound twice, once with the singular and once with the plural modifier. This comparison process was said to draw participants' ratings away from values predicted by the theory. For example, whether LOOVK was the singular form of an irregular plural (LEEVK) or a regular plural (LOOVKS) was established by context. LOOVK was rated as more acceptable in a compound when compared to LOOVKS than when it was compared to LEEVK. Thus, ratings in this task do not reflect only properties of an individual nonword; rather, they also reflect comparisons between the two alternatives provided on each trial. Haskell et al. (2003) also discussed the comparison phenomenon in such rating tasks.

Our classification models utilized only phonological information and had no access to the discourse contexts that established the nonwords' semantics. We hypothesized that adding a simple semantic component, indicating whether the meaning of the word was singular or plural, would allow a model to integrate effects of phonology and semantics. Specifically, we predicted that with a semantic component, the LEEVK-type items, which lack regular plural phonology but were given plural semantics in the contexts provided to the raters in BP's study, should begin to decrease in acceptability compared to singulars. Because adjectives do not have number, we used the noun model described above as a test of this hypothesis. We added a semantic component to this model (see Appendix for details), and the results are also summarized in Table 1. The model was run twice, varying the strength of the semantic information compared to phonological. Adding semantics has the expected effect; it moved the ratings of "irregular plural" items to intermediate values. The size of the effect depended on how semantics and phonology were weighed. When semantics is weighed more heavily than phonology, the irregular plurals pattern more closely with regular plurals. If phonology is weighed more heavily, the irregular plurals pattern more closely with singulars. Variation in how the different types of information are weighed will be affected by experiment specific factors such as the proportions of items of different types and instructions to the participants.

To summarize, the computational modeling provides additional evidence that the legal-illegal manipulation in the BP experiment was ineffective. Our original model encoded only phonological information, and so singulars and irregular plurals patterned together, and both were better than regular plurals. When simple semantic information was added, the acceptability of the irregular plurals dropped, yielding the graded effect predicted by our theory. The discourse contexts on each trial in BP's study established whether a stimulus was intended to be semantically singular or plural. This combination of singular phonology and plural semantics yields the intermediate ratings that they observed for LEEVK-type items. This result requires no appeal to morphology and is entirely consistent with our account.

Note that the purpose of the additional modeling here is to illustrate the way in which semantic and phonological constraints systematically combine to affect well-formedness, not to model a realistic lexical system and how it is acquired. The idea that well-formedness is determined by evaluating multiple simultaneous constraints has not been fully acknowledged in the debate about the proper treatment of inflectional morphology, and so such modeling demonstrations are instructive. As Pinker and Prince noted in 1988, a model that only represents phonological information (e.g., Rumelhart & McClelland, 1986) cannot differentiate the alternative forms of homophones such as RING-RANG/RING-RINGED. Kim, Pinker, Prince, and Prasada (1991) asserted that merely adding semantic features would not solve this problem, citing the examples SLAP, STRIKE, and HIT, which are semantically similar but form their past tenses differently. However, the *conjunction* of semantic and phonological information produces the correct results, as in the present context.[5]

We must also stress that we have not attempted to fit the data in BP's Table 3 exactly, for principled reasons. First, people's knowledge of modifiers is derived from more than just their experiences with adjectives. Hence we would not expect this model to have captured all the relevant phonological properties. Second, the models *shouldn't* capture the detailed pattern of data because they do not incorporate all aspects of the experimental procedures. The behavioral data reflect experiment-specific factors such as the nature of the instructions, participants' strategies based on what they thought the experiment was about, the examples used to anchor the rating scale, the number of times the patterns were repeated (which increases the familiarity of all stimuli), how raters performed the comparison process discussed above, how they weighed the degree of contrast between the strangeness of the nonwords and the presence/absence of the plural ending, and other demand characteristics. Using the implemented model to fit the observed data more closely could be done but it would be theoretically inappropriate, given that the data, but not the model, reflect these other factors (see Seidenberg & Plaut's, 2006, discussion of overfitting in computational modeling).

BP's Experiments 2–3

These experiments compared words like HOSE, which have the phonological form of a plural (e.g., /hoz/ is the pronunciation of both the word HOSE and the word that is the plural of HOE), to words like PIPE, which do not have any phonological resemblance to a plural. We have provided evidence that people disprefer modifiers that have the phonological form of the regular plural, as exemplified by actual regular plurals. BP's reading of our account is that the HOSE type words should be less acceptable as modifiers because they too have the phonological form of plurals. It turns out they are not, and BP conclude that our account is wrong.

The problem with BP's reasoning is captured in the introduction to their article, where they present a variety of examples that illustrate their claim:

> While there is no strong evidence that the constraint on regular plurals in compounds is due to their rare phonology [*sic*; see above], there is substantial evidence against it. In particular, there are numerous compounds with singular nonheads

> that sound just like regular plurals but which are perfectly acceptable: rose garden, praiseworthy, prizefight, breezeway…Mars probe, box-cutter [many others follow]. Not only is there nothing unnatural-sounding about these compounds (as we will confirm in Experiments 2 and 3), but unlike compounds referring to multiple entities (i.e., the referents of regular plurals), they show no tendency to lose their final -s or -z: compare Beatle records with *Ray Charle records, birdwatcher with *fokhole. (p. 10)

Here is the problem. We have discussed a phonological constraint that combines with others to determine the well-formedness. BP are treating this constraint as a deterministic prohibition against the occurrence of modifiers with plural phonology; that is contrary to our theory, in which the constraint is probabilistic. Constraint satisfaction in connectionist networks is the process of evaluating multiple simultaneous probabilistic constraints (Seidenberg & MacDonald, 1999); the effect of one such constraint is not absolute but depends on what other constraints are also in force. So the question is: are there other constraints that override the bias against regular plural phonology in the cited cases?

There are several obvious reasons why forms such as HOSE INSTALLER or ROSE GARDEN are acceptable. One is that alternative ways to convey the same information are comparatively *worse*. BP must agree with this relative goodness idea because they invoke it in explaining why a form such as LOONK is rated differently depending on the words to which it is compared (see their footnote 6 and discussion above). A similar effect of relative goodness applies in the HOSE INSTALLER cases in Experiments 2–3 and in BP's other examples. It happens that the language does not provide another way of expressing the intended concept without violating other constraints that create worse expressions. So, there is a place called the "Rose Garden," which is its name. Proper names are the standard, felicitous way of identifying a designated entity; that is their essential function. ROSE happens to be homophonous with a plural word (/roz/ is the plural of ROW); what other referential expression would be felicitous but avoid plural phonology? Speakers could utter a circumlocution such as "the garden at the White House that has a lot of roses in it, the one that's not the East Garden or the Children's Garden," but both observation and a good deal of evidence from word choice in language production suggest that such expressions are much less felicitous than the simple prenominal modification (e.g., Bock & Levelt, 1994). Following BP, people could create a nonce form such as RO GARDEN but this would violate the integrity of the word ROSE, which is reinforced by its use in other, non-modifier contexts, which are far more frequent. The same mechanism will also maintain its integrity in ROSE GARDEN when used in the generic sense (referring to any rose garden, not the one at the White House), and similarly for the FOX in FOXHOLE, blocking the neologism FOKHOLE.

The force of BP's examples is that the probabilistic bias against modifiers with regular phonology is so strong it should cause the speaker of the language to temporarily suspend the convention of calling things by their names or maintaining words in the lexicon. However, they have the probabilities backwards: the pressure to maintain entrenched naming conventions and lexical items is far more powerful than a phonological constraint that applies to the restricted set of modifier-noun constructions. Abandoning convention in favor of utterances such as *RO GARDEN illustrates another problem: such utterances create ambiguities (a garden of rows, not of roses) or nonwords (e.g., FOKHOLE). Research in language production suggests that several mechanisms work to avoid such anomalous forms (Ferreira, Slevc, & Rogers, 2005; Postma, 2000). Thus, BP's examples all involve violations of stronger constraints — violating conventions concerning proper names, creating neologisms, creating ambiguities — than the narrower one against using modifiers with plural phonology.

Compare these cases to the much-discussed RAT-EATER and *RATS-EATER. With RAT and RATS, the language affords two expressions that do not create neologisms or ambiguity, do not differ greatly in complexity or clarity, do not create off-target verbosity, and do not violate conversational conventions. Each word is supported by its occurrence in other, non-modifier contexts. When the language offers two such alternatives for the modifier, the observed preference for one over the other must be due to other constraints (e.g., against plural phonology). BP's examples are not like this: the alternatives are either a word and a nonword (FOX vs. FOK) or a word conveying the intended meaning vs. one that favors an incorrect interpretation (HOSE vs. HOE).

Thus, it is a mistake to pit a probabilistic phonological constraint involving two legal words against alternatives that would be highly infelicitous for other reasons. BP's prediction only holds if they deny that any other such factors could affect peoples' lexical choices, radically narrowing their focus to the presence/absence of plural phonology. Other factors clearly do affect lexical choices, as abundant research in production attests.

An experiment, then, that compares HOSE FIXER to PIPE FIXER is no more informative than an experiment comparing RAT EATER to GERBIL EATER. In both cases, our theory predicts no difference, and that is what BP observed in their experiments–another null effect. The comparison of interest is between the singular and plural, which are highly similar with respect to discourse/felicity factors but differ prominently with respect to phonology and semantics.[6]

Although the RO GARDEN type examples are trivial, it is interesting to consider why reductions do sometimes occur when pluralia tantum (words such as PANTS or GLASSES) occur in compounds (i.e., PANT LEG, GLASS CASE, with the meaning "case for eyeglasses" not "display case"). Discussions of such reductions in the literature suggest that a variety of factors are involved (e.g., Pinker, 1999; Zimmer, 2007). In particular, there are semantic factors here as well, given the bifurcate character of many pluralia tantum (e.g., pants have two legs, scissors two blades). In modeling these phenomena, the challenge would be to determine whether a system of interacting semantic, phonological and discourse constraints would prevent neologisms like RO GARDEN yet allow an occasional PANT LEG. This project is worth pursuing using realistic natural language input and representations. It seems clear, however, that any account that does not allow multiple interacting soft constraints is going to have a hard time merely achieving descriptive adequacy.

In summary, BP's Experiments 2–3 also do not test a valid hypothesis derived from our theory, because we do not predict that the HOSE INSTALLER type items should be less acceptable than the PIPE INSTALLER type. The observed lack of a difference between the two does not distinguish between the competing accounts. Our appeal to other factors to explain the integrity of proper names and other nouns serves as a reminder that compounds are part of a much larger linguistic system that exerts multiple influences on word formation, including but not limited to ones involving compounds.

BP's Experiments 4–5

BP's last two experiments attempted to provide a strong test of the role of grammatical rather than phonological knowledge in determining modifier acceptability, by keeping the phonological form of the modifier constant but manipulating whether it was interpreted as a regular or irregular plural. The stimuli's critical property was that the regular and irregular plurals were homophones, for example /gliks/. In Experiment 4, the stimuli were presented visually, as in BP's previous experiments; for example, the plural /gliks/ was spelled GLEEKS in the regular condition and GLEEX in the irregular condition. In Experiment 5, stimuli were presented auditorily. As in their other experiments, on each trial a story context provided the setup for producing acceptability ratings for two noun compounds, one with a singular modifier and one with a plural modifier, as in this partial example from the regular plural condition:

> …Fearing an attack on their lives, this report greatly concerned the
> **gleek**-hunters __________
> **gleeks**-hunters __________

Other participants would rate the irregular singular and plural, GLOOX-HUNTERS and GLEEX-HUNTERS, respectively.

For BP, the critical finding in both experiments was that regular plurals were rated as less acceptable than irregular plurals even though they were phonologically identical. Hence a grammatical distinction (whether the plural was rule-governed, as in /glik//gliks/, or an exception, as in /gluks/-/gliks/), determined the results, not phonology. This was taken as consistent with the words and rules theory, which rests heavily on this distinction, and inconsistent with our account, because the phonological and semantic properties of the plurals were identical.

BP's logic here does not hold. In Experiments 4–5, the critical issue is that the difference between the regular and irregular *plurals* must be taken in the context of the results from the other, *singular* conditions, which determined whether a plural such as /gliks/ was interpreted as regular or irregular. In their data, the differences between the homophonous plurals, which are critical for BP, are offset by differences in the opposite direction for the singulars. The tradeoffs between the singulars and plurals cast the results in a completely different light than the one provided by BP.

To clarify, we have replotted the results of Experiments 4 and 5 in Figure 1; the Experiment 1 data, averaged over the (ineffective) legality manipulation, are also included. There are two graphs for Experiment 5 (panels C and D) because BP presented both original rating data (as in other experiments) and data adjusted using a procedure discussed below. The data are plotted so that the two points on any given line reflect the two items that participants rated on a given trial. The curved arrow in each panel identifies the differences in the ratings for two phonologically identical items in different conditions. For example, in Panel C, the arrow identifies the difference in ratings for /gliks/-hunters when it and /glik/-hunters were rated within the same trial vs. when it and /gluks/-hunters were rated. As seen in the figure, the data consist of a series of crossover interactions, and in each case a difference on the left (singular) side of the graph is accompanied by the opposite effect on the right (plural) side. In Experiment 1, the singular items were identical across conditions but yielded varying ratings as a function of the other alternative being rated on the same trial. BP pointed to this tradeoff in accounting for why irregular plurals were not rated as highly as singulars (contrary to the prediction that they should behave alike because both are stored in the lexicon). In their footnote 6, they argue: "Because these regular and irregular singulars were identical, and they were presented in identical contexts, this difference must be due to the acceptability of their plural counterparts–the fact that regular plurals were strongly disliked compared to irregular plurals. Thus, when compared to a highly unacceptable plural (i.e., regular), the same singular is rated more favorably than when compared to a more acceptable one (i.e., irregular)." This comparison process is quite strong, as similar rating tradeoffs were observed in Haskell et al.'s (2003) second experiment, even though participants rated only one item per trial.

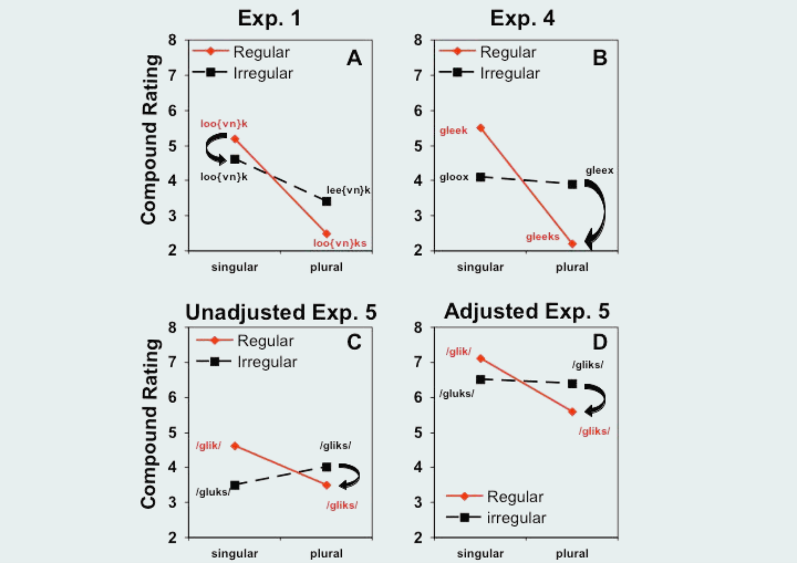

Figure 1. Data from BP's Experiment 1 (Panel A), Experiment 4 (Panel B), and 5 (Panels C-D). The two data points on a line reflect the two types of items rated within a trial. The Experiment 1 data are averaged over the legality manipulation. For example, the data point labeled LOO{VN}K is the average of rating data with compounds containing illegal items LOOVK and legal ones such as LOONK. The curved arrows in each panel indicate the data points for phonologically identical forms that vary as a function of the other item being rated in the trial.

We suspect that at least some of the time Haskell et al.'s raters considered an alternative form before entering a rating, e.g. in rating TOES EXAMINATION they thought about the alternative TOE EXAMINATION.

Having established that all agree that rating tradeoffs exist such that a given phonological form can receive different ratings as a function of the other item rated on a trial, consider the results of Experiments 4–5. The results exhibit the same overall pattern as in Experiment 1. In the earlier experiment, the ratings of phonologically identical singular items varied across conditions. In the later experiments, the ratings of phonologically identical plural items varied across conditions. Thus, people liked /gliks/ less when compared to GLEEK than when compared to GLOOX. This is not surprising given that GLOOX is an odder potential word (e.g., because there are fewer /uk/ and /uks/ words in the language), causing /gliks/ to seem better than when it was compared to GLEEK. The net results of these phonologically-based comparisons are that (a) for the singulars, GLOOX is rated lower than GLEEK, but (b) for the plurals, GLEEX [in the context of GLOOX] is rated higher than GLEEKS [in the context of GLEEK] (with the same pattern observed using auditory presentation). Thus, the differences in plural ratings that

BP observed (marked by curved arrows in Panels B-D in Figure 1) can be wholly attributed to the tradeoff phenomenon that BP themselves invoked in interpreting the results of Experiment 1. Because of this comparison process, the results can be explained by phonological properties of the stimuli *even though the two types of plural were phonologically identical.* The results therefore do not implicate a grammatical rule, or word-rule distinction.

Although Experiments 4–5 do not provide positive evidence for the authors' preferred theory, several additional aspects of the data should be noted. One is that the tradeoff effect is substantially larger in Experiment 4 (Figure 1 Panel B) than in Experiment 5 (Panels C-D). This pattern stems from the fact that the phonologically identical forms were spelled differently in Experiment 4 (GLEEX vs. GLEEKS), whereas in all other studies under discussion here the spelling was identical (Panel A) or the materials were presented auditorily (Panels C-D). Orthography provides another potential constraint on modifier acceptability when the stimuli are presented visually. Many studies of skilled readers indicate that orthographic and phonological information are closely linked (Rayner, Foorman, Perfetti, Pesetsky, & Seidenberg, 2001). The salience of phonemes as units of representation greatly depends on exposure to an alphabetic writing system. Behavioral studies (e.g., Seidenberg & Tanenhaus, 1979) show that it is harder to identify rhymes that are spelled differently (e.g., STONE-FLOWN) than rhymes that are spelled similarly (e.g., STONE-CLONE), even when all stimuli are presented auditorily, as would be true if phonological representations were shaped by orthography. This effect also occurs in computational models of reading (Harm & Seidenberg, 1999), creating a mixed underlying representation that is neither strictly orthographic nor phonological. Neuroimaging evidence suggests that processing in the occipito-temporal area, including the "visual word form area," is penetrated by phonological knowledge (Sandak, Mencl, Frost, Rueckl, Katz, et al., 2004). What this means is that the different spellings of GLEEKS and GLEEX come to push the representations of these "phonologically identical" items apart.

Second, Panel D shows the results of Experiment 5 after a data adjustment procedure that was intended to factor out differences in the "inherent acceptability" of stimulus items. In a separate task, Experiment 5 participants rated the nonword stimuli for their "inherent acceptability" in non-compound constructions (e.g., participants rated the goodness of GLEEKS in a sentence such as "The *New York Times* reported an attack by 20 gleeks."). The adjusted data reflect the original compound ratings minus this second rating, plus a scaling factor of 6. BP intended the adjustment procedure to remove effects of differences between the stimuli with respect to factors such as phonological acceptability, but the procedure makes no sense within the context of their own theory. The essential claim is that a compound such as RATS EATER is poor even though RATS is a perfectly fine lexical item. Thus, the goodness of RATS does not predict the goodness of RATS EATER. Indeed, the main theoretical focus of BP's paper is that the grammatical status of a noun as a regular plural, not its phonology, dictates its goodness as a modifier. Using a rating of GLEEKS to adjust the rating of GLEEKS HUNTER contradicts this view.

In summary, the major concern about Experiments 4–5 is that the data are explained by ratings tradeoffs caused by the phonological factors we have discussed at length. BP's grammatical hypothesis adds nothing to this explanation. The size of the tradeoff varies according to whether the stimuli differ in spelling or not, a result that is entirely consistent with our multiple constraints view and with previous research on the integration of orthography and phonology in literate language users. The adjustment procedure is inconsistent with the authors' theoretical claims; more importantly, it "corrects" the wrong problem. The subtraction process does not eliminate the effects of the comparison process, the fact that in their rating task, participants always rate two alternative forms for the same concept. This task demand encourages participants to give relative goodness ratings for the presented alternatives. Subtracting a separate measure of "inherent acceptability," even were it theoretically sensible, does not remove the effects of these comparisons.

Discussion

Because BP tested hypotheses that are not valid deductions from our theory, overlooked obvious ways in which our theory might account for their results, and failed to consider how demand characteristics of their experiments (e.g., the comparison process) affected all of the results, their experiments do not challenge the account that we proposed in Haskell et al. (2003). Experiment 1 tested a different hypothesis than the one we formulated, and not surprisingly, our own models provide a good account of BP's data, despite the absence of the singular/plural distinction or other grammatical information. Far from requiring a grammatical rule, as BP suggest, the results of Experiment 1 are wholly consistent with our own account. Similarly, Experiments 2–3 are irrelevant because they examined compounds such as HOSE INSTALLER for which there are other powerful pressures to maintain rather than deform or avoid their pseudoplural forms. Finally, Experiments 4–5 are fully explained by the rating tradeoffs that BP invoke to explain their Experiment 1 data. The net result is five experiments that are claimed to demand a morphological rule but instead offer additional support for an approach in which phonological, semantic and other (e.g., orthographic) constraints modulate the goodness of alternative forms.

Beyond these five experiments, BP's article contains a good bit of tangential discussion of linguistic theory[7] and arguments based on amusing but irrelevant examples such as FOKHOLE and RAY CHARLE RECORD. This material has little bearing on any of the empirical or theoretical issues addressed by Haskell et al.[8]

Although most of our comments about BP's article have been negative, we close with some thoughts about the benefits of this exchange and observations concerning areas meriting further investigation. In many ways, BP's studies are an important reminder of the complexity of the phenomena being studied. First, the tradeoff phenomenon is a powerful influence on participants' ratings. These results suggest that the task demands of explicit goodness judgments make these data non-optimal evidence regarding the underlying knowledge and representations that are the subject of debate. It is possible that more implicit measures, such as comprehension latencies or compound production tasks, might prove to be more useful. Second, BP's Experiments 2–3 point to the important role of competing constraints in production choices. The field of language production has intensively investigated the factors underlying speakers' lexical choices. There is essentially universal agreement within word production research that alternative words to convey a concept (e.g. PLAYGROUND vs. PARK) compete for activation during utterance planning (for review see Bock & Levelt, 1994), and that multiple constraints contribute to the relative activation of the alternatives. A natural extension of this view is that these competitive effects also underlie speakers' choices in producing compounds such as PARK COMMISSIONER vs. PARKS COMMISSIONER. This approach has already been extended to other aspects of morphological variation, including gender marking (Mirković, MacDonald, & Seidenberg, 2005), and to singular vs. plural verb morphology (Haskell & MacDonald, 2003, 2005; Thornton & MacDonald, 2003). Whether noun-verb number agreement is best handled by competitive constraint satisfaction processes that we advocate or a symbolic rule application system is a matter of ongoing debate, but this literature suggests alternative methodologies and an opportunity to integrate studies of noun compounds into other research on language use.

Acknowledgement

This research was supported by NIMH grant P50 MH644445, NICHD grant R01 HD047425, and the Wisconsin Alumni Research Fund. We thank Harald Baayen for very helpful comments on an earlier version of this article.

Notes

1. Throughout the article we use the terms "regular" and "rule-governed" interchangeably, and the same for "exception" and "irregular." These terms are used descriptively, not as an endorsement of theories in which they are processed by separate mechanisms.

2. See, for example, Prasada and Pinker (1993, p. 45): "This difference [in the acceptability of regular and irregular forms] is explainable if irregularly inflected items behave just like any other word stored in memory, and hence can feed the compounding process, but regularly inflected items are formed downstream in the information flow from lexicon to syntax, too late to enter the lexical compounding process."

3. Baayen, Feldman, and Schreuder (2006) found that bigram frequency is correlated with many other lexical measures including word frequency, word length, number of orthographic neighbors, and morphological family size. Mean bigram frequency accounted for less than 0.2% of unique variance in their analyses of a large scale lexical decision study.

4. Note that although it is interesting to ask *why* noun modifiers and adjectives share important phonological properties (or why they are not more different), the answer is not relevant to the child language learner, who is not endowed with the history of the language. From the learner's perspective, the input exhibits correlations among different types of information, including (a) lexical statistics (e.g., similarity relations among words with respect to semantics or phonology), and (b) distributional statistics (e.g., co-occurrences of words in sentences). The conjunction of these types of information underlies the development of grammatical categories such as noun, verb, and modifier (Kelly, 1992). See MacDonald (1997, 1999) for discussion of the origins of some distributional patterns and how learners and comprehenders make use of them.

5. Kim et al. (1991) stated that adding semantic information would be helpful only if each word's semantics were represented by an "orthogonal activation vector," in which case the units would stand in for lexical representations and "in no sense would they be *semantic*" [italics in original]. Kim et al. do not provide evidence supporting this claim, and in fact models using distributed representations had already been shown to capture phenomena usually thought to require lexical entries (see Seidenberg & McClelland, 1989, and appendix for discussion).

6. Experiments 1 and 2–3 exhibit complementary misattributions about our theory. Experiment 1 attributes to us the idea that acceptability of the modifier should be a function of its phonological familiarity vis á vis the rest of the lexicon, whereas our actual claim is about the phonological properties of modifiers, a much narrower class. BP err in the opposite direction in Experiments 2–3, where examples such as RO GARDEN and HOE INSTALLER suggest to them that words should be subject to deformation based on a phonological constraint concerning modifiers, ignoring all other uses of the words in the language.

7. After acknowledging longstanding problems that led to the demise of level-ordering, BP provide a discussion of some more recent linguistic theorizing that is said to be relevant because it maintains the regular-irregular dichotomy. However, such work does not preserve the ordered application of rules at different levels of lexical structure as in the level-ordering theory on which their predictions depend. This characterization of the linguistic literature also omits other work (such as Bybee, 1995 and elsewhere; Hay, 2003; Bresnan, in press; research in various versions of optimality theory, e.g., Boersma, 1998; Zurow, 2000) that is more compatible with our approach. The main references to level-ordering in the recent literature are (a) demonstrations of what is wrong with it empirically (e.g., Lardiere, 1995; Nicoladis, 2003, 2004, 2005) and (b) Pinker and colleagues' use of the concept in arguing for a words and rules approach and against connectionist theories.

8. The same holds for BP's concerns about our use of the term "modifier," which they take pains to clarify several times. Our use of "modifier" was in keeping with the claim that people develop useful generalizations about the class of prenominal modifiers, of which there are several types (adjectives, nouns, others). This usage is common and should have been unremarkable; see, for example, the Oxford English Dictionary (*modifier: A word, phrase, or clause which modifies another; (now) esp. an element within a noun phrase which characterizes more specifically what the head refers to*); and in the technical literature, Sproat (1992, p. 37): "Nominal compounds are instances of modifier-head constructions (Levi, 1978). For example, one can think of *handgun* as referring to a particular kind of gun, as determined by the head *gun*, where the modifier *hand* tells you what kind of gun you have." Quibbles over a term such as "modifier" seem like distractions from more important issues.

References

Allen, J., & Seidenberg, M. S. (1999). The emergence of grammaticality in connectionist networks. In B. MacWhinney (Ed.), *The emergence of language* (pp. 115–151). Mahwah, NJ: Erlbaum.

Andrews, S. (1992). Frequency and neighborhood effects on lexical access: Lexical similarity or orthographic redundancy? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *18*, 234–254.

Baayen, R. H., Feldman, L. B., & Schreuder, R., 2006. Morphological influences on the recognition of monosyllabic monomorphemic words. *Journal of Memory and Language, 55*, 290–313.

Balota, D. A., Cortese, M. J., Sergent-Marshall, S. D., & Yap, J. J. (2004). Visual word recognition of single-syllable words. *Journal of Experimental Psychology: General, 133*, 283–316.

Berent, I., & Pinker, S. (2007). The Dislike of Regular Plurals in Compounds: Phonological or Morphological. *The Mental Lexicon, 2*, 129–181.

Bock, K., & Levelt, W. J. M. (1994). Language production. Grammatical encoding. In M. A. Gernsbacher (Ed.), *Handbook of psycholinguistics* (pp. 945–984). San Diego: Academic Press.

Boersma, P. (1998). *Functional Phonology* [LOT International Series 11]. The Hague: Holland Academic Graphics.

Booij, G. E. (1989). Complex verbs and the theory of level ordering. In G. E. Booij & J. van Marle (Eds.), *Yearbook of Morphology 1989* (pp. 21–30). Foris, Dordrecht, The Netherlands: Springer.

Booij, G. E. (1993). Against split morphology. In G. E. Booij & J. van Marle (Eds.), *Yearbook of Morphology 1993* (pp. 27–50). Foris, Dordrecht, The Netherlands: Springer.

Bresnan, J. (in press). Is syntactic knowledge probabilistic? Experiments with the English dative alternation. In S. Featherston & W. Sternefeld (Eds.), *Proceedings of the International Conference on Linguistic Evidence*. Berlin: Mouton de Gruyter.

Bybee, J. (1995). Regular morphology and the lexicon. *Language and Cognitive Processes, 10*, 425–455.

Cassidy, K. W., Kelly, M. H., & Sharoni, L. J. (1999). Inferring gender from name phonology. *Journal of Experimental Psychology: General*, *128*, 362–381.

Ferreira, V. S., Slevc, L. R., & Rogers, E. S. (2005). How do speakers avoid ambiguous linguistic expressions? *Cognition*, *96*, 263–284.

Gordon, P. (1985). Level-ordering in lexical development. *Cognition*, *21*, 73–93.

Harm, M. W., & Seidenberg, M. S. (1999). Phonology, reading acquisition, and dyslexia: Insights from connectionist models. *Psychological Review*, *106*, 491–528.

Harm, M. W., & Seidenberg, M. S. (2004). Computing the meanings of words in reading: Cooperative division of labor between visual and phonological processes. *Psychological Review, 111*, 662–720.

Haskell, T. R., & MacDonald, M. C. (2003). Conflicting cues and competition in subject–verb agreement. *Journal of Memory and Language*, *48*, 760–778.

Haskell, T. R., & MacDonald, M. C. (2005). Constituent structure and linear order in language production: Evidence from subject verb agreement. *Journal of Experimental Psychology: Learning, Memory and Cognition, 35*, 891–904.

Haskell, T. R., MacDonald, M. C., & Seidenberg, M. S. (2003). Language learning and innateness: Some implications of *compounds research*. *Cognitive Psychology*, *47*, 119–163.

Hay, J. (2003). *Causes and consequences of word structure*. New York: Routledge.

Joanisse, M. F., & Seidenberg, M. S. (1999). Impairments in verb morphology after brain injury: A connectionist model. *Proceedings of the National Academy of Sciences, 96*, 7592–7597.

Kelly, M. H. (1992). Using sound to solve syntactic problems: The role of phonology in grammatical category assignments. *Psychological Review, 99,* 349–364.

Kelly, M. H. (2004). Word onset patterns and lexical stress in English. *Journal of Memory and Language*, *50*, 231–244.

Kim, J. J., Pinker, S., Prince A., & Prasada, S. (1991). Why no mere mortal has ever flown out to center field. *Cognitive Science, 15,* 173–218.

Kiparsky, P. (1982). Lexical morphology and phonology. In I. S. Yang (Ed.), *Linguistics in the morning calm* (pp. 3–91). Seoul: Hansin.

Lardiere, D. (1995). L2 acquisition of English synthetic compounding is not constrained by level-ordering (and neither, probably, is L1). *Second language research, 11*, 20–56.

Levi, J. (1978). *The syntax and semantics of complex nominals*. New York: Academic Press.

MacDonald, M. C. (1997). Lexical representations and sentence processing: An introduction. *Language and Cognitive Processes*, *12*, 121–136.

MacDonald, M. C. (1999). Distributional information in language comprehension, production, and acquisition: Three puzzles and a moral. In B. MacWhinney (Ed.), *The emergence of language* (pp. 177–196). Mahwah, NJ: Erlbaum.

MacWhinney, B., & Leinbach, A. J. (1991). Implementations are not conceptualizations: Revising the verb learning model. *Cognition, 40*, 121–157.

Marcus, M., Santorini, B., & Marcinkiewicz, M. A. (1993). Building a large annotated corpus of English: the Penn Treebank. *Computational Linguistics, 19*, 313–330.

McClelland, J. L., & Patterson, K. E. (2002). Rules or connections in past-tense inflections: what does the evidence rule out? *Trends in Cognitive Sciences, 6*, 465–472.

Mirković, J., MacDonald, M. C., & Seidenberg, M. S. (2005). Where does gender come from? Evidence from a complex inflectional system. *Language and Cognitive Processes, 20*, 139–168.

Nicoladis, E. (2003). Compounding is not contingent on level-ordering in acquisition. *Cognitive Development, 18*, 319–338.

Nicoladis, E. (2004). Level-ordering does not constrain children's ungrammatical compounds. *Brain and Language, 90*, 487–494.

Nicoladis, E. (2005). When level-ordering is not used in the formation of English compounds. *First Language, 25*, 331–346.

Pinker, S. (1994). *The language instinct*. New York: William Morrow & Co.

Pinker, S. (1999). *Words and rules: The ingredients of language*. New York: Basic Books.

Pinker, S., & Prince, A. (1988). On language and connectionism: Analysis of a parallel distributed processing model of language acquisition. *Cognition*, *28*, 73–193.

Pinker, S., & Ullman, M. (2002). The past and future of the past tense. *Trends in Cognitive Sciences, 6*, 456–463.

Postma, A. (2000). Detection of errors during speech production: A review of speech monitoring models. *Cognition, 77*, 97–131.

Prasada, S., & Pinker, S. (1993). Generalisation of regular and irregular morphological patterns. *Language and Cognitive Processes*, *8*, 1–56.

Rayner, K., Foorman, B. R., Perfetti, E., Pesetsky, D., & Seidenberg, M. S. (2001). How psychological science informs the teaching of reading. *Psychological Science in the Public Interest*, *2*, 31–74.

Rumelhart, D. E., & McClelland, J. L. (1986). On learning the past tenses of English verbs. In J. L. McClelland, D. E. Rumelhart, & the PDP research group (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition. Volume II.* (pp. 216–271). Cambridge, MA: MIT Press.

Sandak, R., Mencl, W. E., Frost, S. J., Rueckl, J. G., Katz, L., Moore, D., Mason, S. M., Fulbright, R. K., Constable, R. T., & Pugh, K. R. (2004). The neurobiology of adaptive learning in reading: A contrast of different training conditions. *Cognitive, Affective, and Behavioral Neuroscience, 4*, 67–88.

Seidenberg, M. S., & MacDonald, M. C. (1999). A probabilistic constraints approach to language acquisition and processing. *Cognitive Science*, *23*, 569–588.

Seidenberg, M. S., & McClelland, J. L. (1989). A distributed developmental model of word recognition and naming. *Psychological Review, 96*, 523–568.

Seidenberg, M. S., & Gonnerman, L. M. (2000). Explaining derivational morphology as the convergence of codes. *Trends in Cognitive Sciences*, *4*, 353–361.

Seidenberg, M. S., & Plaut, D. C. (2006). Progress in understanding word reading: Data fitting versus theory building. In S. Andrews (Ed.), *From inkmarks to ideas: Current issues in lexical processing* (pp. 25–49). Hove, UK: Psychology Press.

Seidenberg, M. S. & Tanenhaus, M. K. (1979). Orthographic effects on rhyme-monitoring, *Journal of Experimental Psychology: Human Learning and Memory, 5*, 546–554.

Siegel, D. 1974, *Topics in English Morphology.* Doctoral dissertation, MIT.

Sproat, R. (1992). *Morphology and computation*. Cambridge, MA: MIT Press.

Thornton, R., & MacDonald, M. C. (2003). Plausibility and grammatical agreement. *Journal of Memory and Language, 48*, 740–759.

Treiman, R., Mullennix, J., Bijeljac-Babic, R., & Richmond-Welty, E. D. (1995). The special role of rimes in the description, use, and acquisition of English orthography. *Journal of Memory and Language, 124*, 107–136.

Zimmer, B. (2007). Astronaut drives 900 miles wearing… *Language Log*. Retrieved February 15, 2007, from http://itre.cis.upenn.edu/~myl/languagelog/archives/004200.html.
Zurow, K. (2000). *Exceptions and regularities in phonology*. Doctoral dissertation, UCLA department of linguistics, Los Angeles.

Author's address

Mark S. Seidenberg
Department of Psychology
University of Wisconsin-Madison
Madison, WI 53706, USA
Phone: 608 263–2553
FAX: 608 262–4029

seidenberg@wisc.edu

Appendix

Architecture

The connectionist network used in the simulations consisted of three layers. The input layer consisted of 27 units representing segmental features, and one unit representing stress. In the models incorporating number semantics, an additional two units represented this information. The input layer was connected to a hidden layer consisting of seven units. In turn, this hidden layer had recurrent connections to itself, as well as connections to an output layer consisting of two units. In addition, the network contained a bias unit which was connected to the hidden layer and the output layer, and whose activation was always set to one. The logistic sigmoid activation function was used for all units.

Representation

Each word was presented to the model as a sequence of phonemes. At time t=0, the units representing the segmental features of the first phoneme in the word were set to a value of 1, and all other units in the input layer were set to 0. At time t=1, this process was repeated for the second phoneme, and so on, until all phonemes had been presented. For vowels receiving primary stress, the stress unit was set to 1 concurrently with the appropriate segmental units.

For the simulations including semantics, the two bits representing number semantics coded singular and plural meaning. Singular meaning was represented by turning the first unit on and keeping the second unit off, i.e., a [1 0] pattern. Plural meaning was represented by the complementary pattern, i.e., [0 1]. Thus, the model was provided with two cues to grammatical number (phonology and semantics). The relative contribution of the two cues was modulated by adding "noise" to the number semantics bits; less noise causes the model to rely more on number semantics, more noise causes the model to rely more on phonology. The noise was applied by selecting a random number from a Gaussian distribution with a mean of zero and a standard deviation of either .33 (modest noise) or .5 (high noise). We then took the absolute

value of that number, and adjusted each semantic number bit by that amount. For example, for a singular noun with a noise value of .2, the number semantics bits would be set to [.8 .2]. Note that because the model always had some amount of noise on the semantic units, model was never exposed to discrete singular/plural distinctions of [1 0] and [0 1] in these units.
Two units in the output layer were used to represent singular and plural in the same way as the number semantics bits in the input layer. The two-unit output layer in the noun model contrasts with the single unit output in the original Haskell et al. (2003) model. However, because the activation of the two units in the noun model summed to 1 and the activation of the single unit in the original model ranged from 0–1, this architectural difference has no effect on the computation of activation levels.

For purposes of computing error, the correct classification was compared with the activation of the output units at two time steps after the final phoneme of the word was presented to the model, as it took two time steps for the contribution of this phoneme to reach the output layer.

Materials and Training Procedure

The training set for the model consisted of 100 singular and 100 plural nouns. These nouns were randomly selected from the Treebank version of the Brown corpus (Marcus, Santorini, & Marcinkiewicz, 1993), subject to the constraint that all words had to appear at least three times in the corpus (to ensure that they weren't typographical errors or neologisms). All words were presented to the network equally often during training. Pronunciations for the training items were obtained from the CMU Pronouncing Dictionary (Carnegie Mellon University, Pittsburgh, PA). The test set was comprised of the nonwords from BP's Experiment 1.

Prior to training, all weights in the model were set to randomly chosen values between −0.1 and 0.1. Weights were updated after each cycle through the training set using the back-propagation-through-time algorithm. The learning rate was set to 0.005, and the momentum parameter was set to 0.5. Weight values were decayed by a factor of 0.001 after each pass through the training set.

During training, the performance of the model was assessed after every pass through the training set, and training was halted when additional training did not change the number of incorrect classifications for 100 consecutive epochs (i.e., when performance reached asymptote), or when 1000 epochs had elapsed.

The model was trained three times. Each time, the initial weight values were set to different random values, and a different training set was generated, according to the procedure described above. The reported results represent the average output of the model across these three runs.

A note about the representations used in these simulations: Joanisse and Seidenberg (1999) also developed a model that investigated the use of the conjunction of semantic and phonological information in generating inflected forms. That research was criticized (Pinker & Ullman, 2002) because single, localist units were used to represent base words. Such units could be taken as lexical entries in a words-and-rules type of theory. Similarly, in the present model, individual units were used to code the singular vs. plural distinction, which could be interpreted as grammatical features rather than "semantics". We note here that no significant aspect of the models' performance turns on these choices; in each case the information could as well have been represented by a distributed representation that encoded word meanings rather than words or specific features. Such models have already been shown to exhibit behavior that is standardly taken to require lexical entries (e.g., word frequency effects). Harm and Seidenberg (2004) provide an example of this approach, including simple applications to the generation of inflected forms. The essence of these models is not found in these implementational details; rather, it is the idea that simple networks provide a mechanism for combine multiple types of probabilistic constraints, semantics and phonology most prominently.