



# What makes words sound similar?

Ulrike Hahn\*, Todd M. Bailey

*School of Psychology, Cardiff University, Tower Building, Park Place, Cardiff CF10 3AT, UK*

Received 12 November 2003; revised 16 June 2004; accepted 10 September 2004

## Abstract

Although similarity plays an important role in accounts of language processing, there are surprisingly few direct empirical studies of the phonological similarity between words, and it is therefore not clear whether similarity comparisons between words involve processes similar to those involved in other cognitive domains. In five experiments, participants chose which of two monosyllabic pseudo-words sounded more similar to a target pseudo-word. Our results are generally consistent with the structural alignment theory of comparisons between complex mental representations, suggesting that phonological word similarity parallels similarity involving other kinds of information including visual objects and scenes, events, and word meanings. We use our results to test new metrics of word similarity, and identify predictions for future similarity research both in the domain of word sounds and in other cognitive domains.

© 2004 Elsevier B.V. All rights reserved.

**Keywords:** Word similarity; Phonological similarity; Phonological comparisons; Structural alignment; Phoneme edit distance

Similarity is one of the most central concepts in cognitive psychology, invoked in domains as diverse as memory, problem solving, categorisation, and reasoning (Goldstone, 1993). Similarity also plays an important role in explanations of language processing, including word recognition, speech production, verbal short term memory, and inflectional morphology, as discussed below. It is widely assumed that the similarity in sound between two words depends on comparisons between phonemes in corresponding positions (e.g. Frisch, Large, & Pisoni, 2000; Greenberg & Jenkins, 1964; Luce, 1986; Metsala, 1997; Vitevitch & Luce, 1999). For example, the words *creaks* and *keeps*

\* Corresponding author. Fax: +44 2920874858.

E-mail addresses: [hahnu@cardiff.ac.uk](mailto:hahnu@cardiff.ac.uk) (U. Hahn), [baileym1@cardiff.ac.uk](mailto:baileym1@cardiff.ac.uk) (T.M. Bailey).

/kriks/  
 |||||  
 /k ips/

Fig. 1. Alignment of *creaks* and *keeps*.

(phonemically transcribed as /kriks/ and /kips/ using the International Phonetic Alphabet) could be aligned to yield the phoneme correspondences shown in Fig. 1.

The phonemes in one word might correspond with either matching or mismatching phonemes in the other word, or they might not correspond to anything at all (represented here as a gap). A simple count of the number of correspondences between mismatching phonemes or gaps produces the most common measure of dissimilarity between words, the phoneme edit distance (Greenberg & Jenkins, 1964; Landauer & Streeter, 1973; Luce, 1986). Variants of phoneme edit distance have proven to be useful estimates of phonological dissimilarity in a range of language processing tasks, including perceptual identification, auditory lexical decision and auditory word naming (Luce, 1986), as well as verbal memory (e.g. Frisch et al., 2000). Nevertheless, it is universally acknowledged to be a crude approximation at best. Given the importance of word similarity to language processing, there are surprisingly few studies of what actually determines the similarity between word sounds. Without an understanding of word similarity, however, we cannot fully understand any of the language processing tasks which are influenced by phonological comparisons between words.

Though our understanding of word similarity is limited, the idea that it depends on comparisons between aligned parts resonates with recent research on similarity within the general cognitive literature. The theory of structural alignment maintains that comparisons between complex mental representations (visual scenes, events, word meanings, etc.) involve an alignment process which identifies corresponding components based on their featural and structural properties (e.g. Gentner, 1983, 1989; Gentner & Markman, 1994; Goldstone, 1994b; Markman, 1996, 2001; and in the context of analogy see Holyoak & Thagard, 1989; Hummel & Holyoak, 1997; Keane, Ledgeway, & Duff, 1994). The present work takes as its starting point the theory of structural alignment and the empirical findings on general cognitive similarity processes, and relates these to word similarity. This motivates our empirical investigation of several aspects of word similarity. The results of this investigation are then used to develop better measures of word similarity. This empirical program demonstrates how linking word similarity and general similarity processes can be beneficial to our understanding of both.

## 1. Previous work

Similarity comparisons between the sounds of words affect performance in a wide range of language processing tasks. For example, Greenberg and Jenkins (1964) related wordlikeness judgments for nonwords to their similarity to real words (also see Bailey & Hahn, 2001 and references therein). Baddeley (1966) found that phonological similarity among list items interferes with serial recall of words (also see Li, Schweickert, & Gandour, 2000, for a recent

study, and many other works). Fay and Cutler (1977) identified a class of spontaneous speech error (the malapropism) in which a phonologically similar but semantically inappropriate word is substituted for the intended word. Stemberger (1990) reviewed evidence that the presence of two phonologically similar words in an utterance increases the likelihood of mispronouncing one or both words. Frisch (1996) argued that asymmetries in sound substitutions in speech errors arise from similarity neighborhoods in the mental lexicon. Luce (1986) found that words which sound similar to many other words are more difficult to recognise in noise than words which are not. Bybee and Slobin (1982) showed that the way a word is inflected is affected by the inflection of similar sounding words (also see Prasada & Pinker, 1993).

However, despite a universal consensus that phonological similarity between words is important in language processing, there is little agreement on how best to measure the relevant similarities. Some studies of language processing have relied on unstated intuitions of word similarity (e.g. Baddeley, 1966). Others have adopted a priori measures of similarity such as rhyme (e.g. Prasada & Pinker, 1993) or the number of matching and mismatching phonemes (e.g. Greenberg & Jenkins, 1964). While these works have demonstrated a basic role for word similarity in a range of language processing tasks, they have not provided a detailed understanding of the psychological similarity between words. However, such an understanding is necessary for fine-grained prediction and explanation of linguistic performance. Mueller, Seymour, Kieras, and Meyer (2003), for example, provide evidence to suggest that seemingly conflicting evidence on the nature of verbal short-term memory might be attributable to less than ideal measurement and control of word similarity.

The development of a more detailed understanding of word similarity will necessarily require more direct examinations than mere demonstration that some measure of word similarity correlates with linguistic performance. One direct study of word similarity is that of Vitz and Winkler (1973). Vitz and Winkler compared participants' similarity judgments of word pairs to a measure of phonological distance based on the proportion of mismatching phonemes between the two words. Their measure of dissimilarity was essentially a metric of edit distance, normalized by length. First, they aligned words to maximize the number of matching phonemes in corresponding positions, as in Fig. 1. Then, they counted the number of correspondences between mismatching phonemes or gaps and divided by the total number of phoneme positions. This measure yielded good predictions of similarity judgments for pairs of words varying from one to four syllables in length. Nevertheless, immediate shortcomings can be identified in measures of word dissimilarity based on simple phoneme edit distance, including the Vitz–Winkler measure. For one, it seems likely that some syllable positions might be more important than others (e.g. Nelson & Nelson, 1970). Also, simple edit distance measures make a binary distinction between matching and mismatching phonemes so that *tuck* is as similar to *muck* as it is to *duck*. A more realistic measure of word similarity will probably have to take account of degrees of similarity between phonemes based on shared sub-phonemic features (or see Luce, 1986 for an alternative approach based on empirical phoneme confusability rather than theoretical features).

Few researchers would argue with these shortcomings of phoneme edit distance; Vitz and Winkler, for example, explicitly mention these and other shortcomings of their own

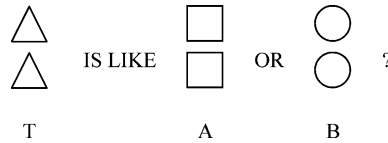


Fig. 2. An illustration of Tversky's (1977) contrast model (Goldstone et al., 1991).

edit distance based measure. Empirically, the shortcomings of phoneme edit distance are demonstrated by the results of Bailey and Hahn (2001) who found that a measure of word similarity which included phoneme similarities doubled the predictive accuracy of a simple phoneme edit distance measure in explaining wordlikeness judgments. This suggests that crude measures of word similarity considerably limit our understanding of a wide range of language processing tasks.

At present, there is insufficient data to guide the development of more adequate measures of word similarity. Furthermore, because word similarity—in contrast to phoneme similarity—has not previously attracted theoretical interest in its own right, psycholinguistics currently lacks a theoretical framework for word similarity from which detailed testable predictions could be derived. Consequently we asked whether general theories of similarity developed in other cognitive domains might provide useful guidance on where to begin.

General research on similarity was long dominated by a feature-based account—Tversky's (1977) contrast model. His equation

$$\text{SIM}(A, B) = \theta f(A \cap B) - \alpha f(A - B) - \beta f(B - A)$$

defines the similarity of representations  $A$  and  $B$  as a function of their shared features, minus those distinctive to  $A$ , minus again those specific to  $B$ . The parameters  $\alpha$ ,  $\beta$ , and  $\theta$  are weighting terms that depend on the task. An illustration of this theory is given by Goldstone, Medin, and Gentner (1991), reproduced in Fig. 2 (above). If we assume these shapes are represented with simple perceptual features, then Tversky's model predicts that subjects would find the triangles in set  $T$  more similar to the squares than the circles in Fig. 2, as they have more features in common (straight lines, corners, etc.).

Tversky's model can just as easily be applied to phonological features as to visual features, and it has indeed been examined in the context of phoneme similarities (Frisch, 1996). More relevant to word similarity, however, is that general research on similarity has recently come to emphasize the role of structure. Goldstone et al. (1991) provide a second example, reproduced in Fig. 3. By simply adding a square to each component from Fig. 2 our perception changes dramatically. Now the two triangles over a square in set  $T$  are seen

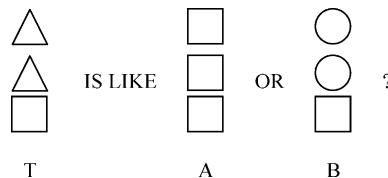


Fig. 3. Violation of the contrast model (Goldstone et al., 1991).

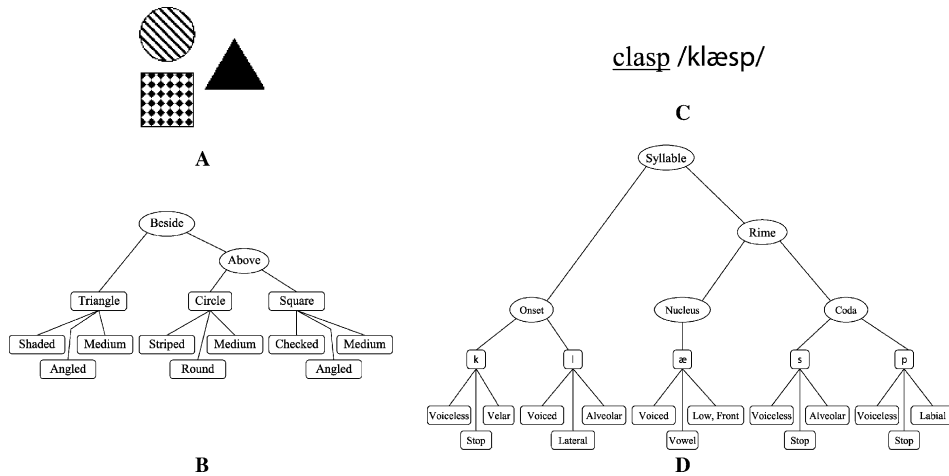


Fig. 4. (A) Simple configuration of objects, (B) its relational structure, (C) an English syllable, and (D) its phonological structure. (A) and (B) are based on Gentner and Markman (1994, Fig. 2).

as more similar to *B* than to *A*. Thus, Goldstone et al. argue that similarity comprises both attributional (featural) and relational (structural) components of hierarchical representations. Similarity comparisons between such representations require *structural alignment* to identify one-to-one mappings between corresponding components.

There are clear parallels between the alignment of hierarchical structures proposed for general similarity comparisons and the correspondences that must be established between phonological representations to compare words. Consider the simple configuration of objects in Fig. 4A (based on Gentner & Markman, 1994, Fig. 2), which can be represented as a hierarchical organization of relations among the individual shapes, which are themselves represented as collections of basic features as shown in Fig. 4B. The relational structure of this configuration of objects is parallel to the phonological structure of words, as illustrated in Fig. 4C and D. A monosyllabic word like *clasp* (/klæsp/) is phonologically composed of an onset (here /kl/), and a rime (/æsp/). The onset is the initial part of a syllable, including any consonants that precede the vowel.<sup>1</sup> The rime is composed of the nucleus (typically a vowel) and the coda which includes any consonants following the nucleus. This hierarchical organization of syllables into onset and rime explains, for example, why spontaneous speech errors are particularly likely to blend the onset of one word (e.g. the /kl/ of *close*) with the rime of another semantically related word (e.g. the /lr/ of *near*) to produce an unintended utterance (*clear*). Blends are less likely to break up the first word at a point which separates phonemes within the onset or within the rime (MacKay, 1972; also see Fowler, Treiman, & Gross, 1993; Fudge, 1987; Treiman, 1983). One hierarchical level down, the phonemes that constitute onset, nucleus and coda are themselves defined in terms of

<sup>1</sup> See Stemberger (1990) and Stemberger and Treiman (1986) for evidence from speech errors which motivates additional hierarchical structure within syllable onsets.

features describing their articulation within the vocal tract. Arguments for decomposing phonemes into phonological features are discussed in more detail below. We focus here on traditional major class features, that is, place of articulation, manner of articulation, and voicing, as illustrated in Fig. 4D.

The initial parallels between the relational structures of visual scenes and the hierarchical structure of words suggests that the structural alignment process which has been invoked to explain similarity in other cognitive domains might equally be applied to the phonological representations of words. The experiments reported in this paper explore these parallels further by testing predictions derived from structural alignment theory with words. The experimental results provide much needed data on what makes words sound similar, data that we then use to develop better measures of word similarity. At the same time, however, they demonstrate that the general cognitive literature on similarity provides a useful theoretical framework for the study of word similarity.

## 2. Overview of the experiments

Our studies aimed to clarify the nature of word similarity and whether it is analogous to similarity as observed in other cognitive domains, focussing on monosyllables. These studies used a two-choice procedure in which a target monosyllable was paired with two choice monosyllables, and participants chose which of these two candidates sounded more similar to the target. The experiments examined, in the context of word similarity, three key effects from the general similarity literature. These are that (a) similarity typically increases with the number of shared features in corresponding components, (b) similarity is less influenced by unaligned components than aligned components, and (c) similarity increases more when matching features are clustered together than when they are distributed across components. In addition, we examined whether similarity is more sensitive to shared features in some parts of words than in others. In detail, the series of experiments proceeds as follows. Experiment 1A examines whether word similarity is influenced by sub-phonemic features. Experiment 1B tests the influence of alignment on similarity comparisons between words of different length. Experiment 1C then combines the key elements of Experiments 1A and 1B; it examines whether word similarity is influenced by the degree of sub-phonemic featural match between aligned phonemes in words of different lengths. Experiment 1D examines the effects of clustering, and Experiment 1E, finally, draws together aspects of the results of all preceding four experiments by examining whether onsets and codas differ with regards to their influence on the overall similarity between words.

## 3. Experiment 1A

Similarity-based processes in general cognitive domains are widely recognized to be sensitive not just to identical objects, but to degrees of similarity between objects based on the extent to which they share the same features (Tversky, 1977). All else being equal,

similarity between items increases as a function of the number of features they share. On a theoretical level, Experiment 1A tests whether this is paralleled by words. It seeks to confirm that word similarity is affected not just by matching phonemes, but also by the degree of similarity between mismatching phonemes, so that *tuck* is more similar to *duck* than to *muck* because /t/ is more similar to /d/ than to /m/. This similarity between phonemes is represented by shared sub-phonemic features—/t/ and /d/ share the same place and manner of articulation, whereas /t/ and /m/ have no features in common.

There is a broad consensus in phonological theory that speech sounds are represented in terms of articulatory features. Featural representations of speech sounds were developed originally to explain so-called natural classes, that is, sets of phonemes that languages often treat as members of a single category in phonological patterns (see [Kenstowicz, 1994](#), for an introduction). In general, commonalities between phonemes are explained in terms of shared or similar places of articulation, manners of articulation, and voicing.<sup>2</sup> Two phonemes are more or less similar depending on how many of these major class features they share. In addition to linguistic evidence from natural classes in phonological patterns, there is a wealth of psycholinguistic evidence indicating the importance of phonological features to language processing (e.g. [Connine, Blasko, & Titone, 1993](#); [Frisch, 1996](#); [Jusczyk, Goodman, & Baumann, 1999](#); [Shattuck-Hufnagel, & Klatt, 1979](#); [Stemberger, 1991](#); [Wickelgren, 1965, 1966](#)).

Nevertheless, the evidence that phonemic features are relevant specifically to word similarity is less clear cut. Most of the positive evidence comes from studies which are not directly examining word similarity, but are dealing with phenomena such as priming, lexical influence or word association that have natural interpretations in terms of word similarity ([Bailey & Hahn, 2001](#); [Connine et al., 1993](#); [Goldinger, Luce, & Pisoni, 1989](#); [Greenberg & Jenkins, 1964](#); [Marslen-Wilson, Moss, & van Halen, 1996](#)). We have identified three previous studies which directly examine effects of shared sub-phonemic features on word similarity. [Greenberg and Jenkins \(1964\)](#) found that participants judged syllables whose initial consonants agreed in either place of articulation or voicing to be more similar than those whose initial consonants agreed in neither place of articulation, nor voicing. [Sendlmeier \(1987\)](#) examined similarity ratings for pairs of nonwords one to three syllables in length. From multidimensional scaling solutions, Sendlmeier concluded that the phonemic similarity among consonants was important only for monosyllables, and then only in initial position. For other syllable positions and for longer stimuli, other factors (stress location, number of syllables, vowel quality, presence of consonant clusters, etc.) were more salient than consonant similarities. In contrast to these two studies, [Vitz and Winkler \(1973\)](#) found that featural similarities between phonemes did not contribute to an improved measure of word similarity compared to a phoneme edit distance measure that simply distinguished between identical and non-identical phonemes.

The issue of whether or not word similarity is affected by the number of shared sub-phonemic features is sufficiently important to make the current state of empirical

---

<sup>2</sup> While the decomposition of phonemes into major class features is widely accepted, proposals to subdivide the major class features into more fine-grained features have yet to achieve a consensus on exactly which fine-grained features are relevant (cf. [Chomsky & Halle, 1968](#); [Kenstowicz, 1994](#); [Frisch, 1996](#); [Wickelgren, 1965, 1966](#)).



research on this matter seem unsatisfactory. Consequently, our first experiment examines the role of sub-phonemic features in word similarity directly by eliciting forced choices between comparison words differing only in the number of sub-phonemic features they share with a target word. Several considerations motivated our choice of this meta-linguistic task. First, word similarity has been invoked as an explanatory construct in a very wide range of linguistic and psycholinguistic contexts, involving both online and diachronic processes, so that there is no single context that could be viewed as privileged. This suggests the use of a comparatively neutral task as opposed to a particular online processing task such as word recognition under noise or priming. Furthermore explicit judgments (i.e. “Which of the two choice words is more similar to the target?”) seem methodologically desirable because they are likely to maximise the impact of word similarity on participants’ responses. This contrasts with tasks in which similarity forms only an (assumed) implicit component such as priming, confusability, or same-different judgements. An ultimate understanding of similarity requires evidence from both direct examinations and embedded processing; however, direct examinations are likely to provide the greatest resolution, thus providing a natural starting point. This is presumably why the vast majority of general research on similarity has employed either forced choice or ratings procedures (e.g. Goldstone et al., 1991; Goldstone & Medin, 1994; Markman & Gentner, 1993; Tversky, 1977; Tversky & Gati, 1982). Given, furthermore, that we are seeking to determine the applicability of findings from this research to the special case of word similarity, it seemed desirable to follow the experimental procedures employed in these studies as best possible.

A potential problem with explicit judgments of this kind is that participants might develop extraneous high-level strategies of responding. To avoid this, the patterns present in a stimulus set should, if at all possible, remain opaque to the participants. This requires the extensive use of fillers. Our strategy was to use the items of our five studies as each other’s fillers. The five studies were conducted together, with each participant responding to stimuli from all five studies.

Our first experiment asked participants to compare a target word (e.g. /pʌsp/) to two choice candidates (e.g. /bʌsp/, /gʌsp/) which differed from the target by one and two phonological features, respectively, in a single phoneme. We predicted that the candidate that differed from the target by a single feature would be chosen more often than the other one.

### 3.1. Method

*Participants.* Twenty-eight psychology undergraduates at Cardiff University took part in the study for course credit. The same participants supplied data for Experiments 1A–E, with each participant responding to intermixed stimuli from all five studies.

*Stimuli.* We used nonwords for stimuli in order to avoid contamination of similarity judgments by semantic factors, word frequency or age of acquisition. Single-syllable, pronounceable nonwords were generated from a syllable-construction grammar based on word-initial and word-final onsets and codas and the vowels with which they occur in the CELEX database of English words (online <http://www.kun.nl/celex/>). Of these, we



selected nonwords that had one to three initial consonants and one to two final consonants. We excluded items that contained non-native sequences (including /ʃm/, /ʃn/, /kn/, /pw/, /ps/, /pf/, /dz/, /ts/, /sf/, /sv/, /zl/, /zz/, /ʒ/),<sup>3</sup> and those that were likely to be interpreted as multimorphemic (i.e. those ending in /tθ/, /s/, /z/, /t/, /d/). This left some 11,000 nonwords from which our stimuli were chosen. Of these, 447 *echo syllables* began and ended with the same consonant. Echo syllables are ideal for making onset and coda contrasts as comparable as possible. Target words for Experiments 1A, D, and E were chosen from these 447 (the set of echo syllables was too limited for Experiments 1B and C, so we chose stimuli for these experiments from the wider set of 11,000 nonwords). Each experiment involved triads composed of a target nonword and two choice candidates that were compared to the target. Choice candidates differed from the targets by one or two phonemes depending on the hypothesis being tested in each experiment. To obtain stimuli that would be broadly representative, we assembled the full set of suitable triads for each experiment and then chose a random sample from the full set for use in that experiment. In identifying suitable triads we carefully controlled the specific features being compared, because different features do not necessarily carry equal weight in their contributions to similarity (Greenberg & Jenkins, 1964). Therefore, our stimuli always compared a match on one feature (e.g. /d/ and /s/, which have an alveolar place of articulation) with a match on a second feature in addition to the same first feature match (e.g. /d/ and /z/, which are also both voiced).

We made digital recordings (16 bit, 20 kHz) of a male speaker of British English pronouncing the stimulus words in a professional recording studio, and scaled the recorded waveforms to the same RMS amplitude in order to achieve roughly comparable loudness levels for all stimuli.

For Experiment 1A, there were 20 target-candidate triads of nonwords that differed in the initial consonant (the onset stimuli), and 20 triads that differed in the final consonant (the coda stimuli). To make onset and coda stimuli as comparable as possible, each target ended with the same consonant with which it began (e.g. /pʌsp/). A choice candidate was derived from each target by changing a single feature (e.g. /bʌsp/). A second choice candidate was derived from the first by changing an additional second feature of the same phoneme (e.g. /gʌsp/). The stimuli for all five experiments are listed in Appendix A.

*Procedure.* Participants were tested one at a time in a sound-insulated room. They sat in front of a computer screen and listened to the auditory stimuli on headphones. A script for SuperLab Pro 2.0 (<http://www.cedrus.com>) controlled the experiment on a PC. On each trial, participants first heard a target word followed by one of the two corresponding choice candidates. After a gap of 500 ms, participants heard the target again, followed by the other choice candidate. There was a 250 ms. gap between each target and the following candidate. Coinciding with the beginning of each word the computer screen identified the auditory stimulus as 'Target', '1', or '2'.

<sup>3</sup> The phoneme /ʒ/ was avoided because it does not occur in single-syllable English content words. The sequence /klw/, which occurs in the French loanword *cloisonné*, should be classified as a non-native sequence, but was overlooked and appears in one stimulus triad of Experiment 1A. It may be worth noting that such loanwords are generally more common in Britain, where the research was conducted, than in, say, many parts of North America.

Participants judged whether the first or second target-candidate pair was more similar and pressed either '1' or '2' on the computer keyboard to indicate their preference. All participants responded to the same 180 target-candidate triads (40 triads each for Experiments 1A–D plus 20 triads for Experiment 1E). The order of presentation of the two target-candidate pairs for each triad was counterbalanced within each experiment across participants. In addition to the 180 test trials, two foil trials with obvious answers were included to identify inattentive participants (e.g. the target /blek/ was compared to choice candidates /blek/ and /gla'b/). Finally, there were four rest trials that instructed participants to take a short break and then to press any key when they were ready to resume the study. The presentation order of the 186 trials was randomised for each participant.

### 3.2. Results

The data from one participant with impaired hearing was excluded. Seven participants responded incorrectly to one of the foil trials and were excluded from further analysis,

Table 1

Percentage of responses choosing candidate *A* rather than *B* as being more similar to the target word

Experiment	Pos'n	Sample Stimuli	Schematic Stimuli			
			Target	A	B	A%
1A. Featural differences	Onset	/fɔʔf/-/pɔʔf/-/tɔʔf/	C...C	C <sup>1</sup> ...C	C <sup>1,2</sup> ...C	62*
	Coda	/mum/-/mub/-/mug/	C...C	C...C <sup>1</sup>	C...C <sup>1,2</sup>	70*
1B. Insert or replace C	Onset	/dæf/-/dræf/-/ræf/	C...	CĈ...	Ĉ...	51
	Coda	/kɪθ/-/kɪfθ/-/kɪf/	...C	...ĈC	...Ĉ	64*
1C. Singleton-cluster comparison	Onset	/koʃ/-/gloʃ/-/bloʃ/	C...	C <sup>1</sup> Ĉ...	C <sup>1,2</sup> Ĉ...	55
	Coda	/ðæp/-/ðælb/-/ðælv/	...C	...ĈC <sup>1</sup>	...ĈC <sup>1,2</sup>	63*
1D. Clustered feature matches	Onset	/gɔg/-/gɔm/-/bɔŋ/	C...C	C...C <sup>1,2</sup>	C <sup>1</sup> ...C <sup>2</sup>	51
	Coda	/ba'b/-/fa'b/-/pa'v/	C...C	C <sup>1,2</sup> ...C	C <sup>1</sup> ...C <sup>2</sup>	73*
1E. Onset vs. coda salience		/ʃæʃ/-/fæʃ/-/ʃæf/	C...C	C <sup>1</sup> ...C	C...C <sup>1</sup>	58 <sup>†</sup>

C and Ĉ represent different consonants. C<sup>1</sup> and C<sup>2</sup> each differ from C by a single feature; C<sup>1,2</sup> differs from C by two features, and from C<sup>1</sup> and C<sup>2</sup> by one feature each. \**p* < .05. <sup>†</sup>*p* < .10.

leaving data from 20 attentive participants. Responses for each target-candidate triad in all five experiments are given in Appendix A. Summary results are shown in Table 1. The data were subjected to repeated measures analysis of variance (ANOVA) that took both item and subject variance into account as recommended by Raaijmakers, Schrijnemakers, and Gremmen (1999).<sup>4</sup>  $F'$  statistics tested whether one choice candidate was chosen significantly more often than the other.<sup>5</sup>

The question for Experiment 1A was whether similarity judgments were sensitive to the number of featural differences between choice words differing from the target by a single phoneme. Did participants choose target-candidate pairs that differed by a single feature more often than they chose pairs that differed by two features? In both onsets and codas, target-candidate pairs with one feature difference were judged to be more similar than pairs with two feature differences ( $M=62$  and 38% in onsets;  $M=70$  and 30% in codas). These differences were statistically significant,  $F'(1,20)=8$ ,  $MS_{\text{err}}=0.69$ ,  $p=.01$  for onsets,  $F'(1,16)=40$ ,  $MS_{\text{err}}=0.39$ ,  $p<.001$  for codas (two-tailed tests are reported throughout).

### 3.3. Discussion

Similarity judgments were clearly sensitive to the degree of sub-phonemic featural match between choice words and target. This suggests that word similarity depends on sub-phonemic features, and indicates that our experimental task is sensitive to this dependency. The effect of featural matches was observed in both onsets and codas, though the difference was somewhat stronger in codas.

Our results confirm the findings of Greenberg and Jenkins (1964) and Sendlmeier (1987) that features in syllable onsets are relevant to explicit judgments of word similarity. In contrast to Sendlmeier, we found effects of features at the ends of words that were at least as strong as those at the beginning. Because our experiment 1A focused exclusively on one question, it will have had considerably more power than Sendlmeier's more general study to test this specific hypothesis. Our experiment also measured the effect of our featural manipulation directly, whereas Sendlmeier's conclusions were based on the results of multidimensional scaling. These factors may have masked effects of coda features in Sendlmeier's study.

Our results differ from those of Vitz and Winkler (1973) who concluded that phoneme similarities based on sub-phonemic features did not contribute to word similarity. Several methodological factors are likely to have contributed to this

<sup>4</sup> It would be inappropriate to simply compute ANOVAs based on average responses across the 20 target-candidate triads for each comparison. We are interested in drawing conclusions about similarity among monosyllables in general based on the sample we actually tested. Averaging across target-candidate sets would implicitly treat items as a fixed effect rather than a random effect and any conclusions drawn from such an analysis would apply only to the particular stimuli actually tested. Instead, items need to be treated as a random effect in order to generalise the results beyond the stimuli used in our study.

<sup>5</sup> Because these analyses take both item and subject variance into account in testing the hypotheses of interest, there is nothing to be gained by performing separate analyses by items and by subjects. Indeed, such tests would likely yield misleading results. See Raaijmakers, Schrijnemakers, and Gremmen (1999).

discrepancy. Our study focused specifically on the question of featural differences between consonants, and all our word pairs were minimally different. In contrast, Vitz and Winkler conducted a general survey of word similarity that contrasted vowels as well as consonants, and their word pairs varied from highly similar to totally dissimilar. These differences in design result in greater statistical power for our study to test this specific hypothesis.

The finding that judgments were sensitive to feature differences is generally consistent with many previous findings on the impact of phonological features in a wide range of language processing tasks, and in particular with the indirect studies of the impact of features on word similarity conducted by Bailey and Hahn (2001), Connine et al. (1993) and Marslen-Wilson et al. (1996). This consistency suggests that the word similarity judgment task taps into similar, if not identical, phonological representations and processes as those relevant to these other tasks.

The conclusion that word similarity depends not only on the number of matching phonemes between two same-length words, but also on shared sub-phonemic features, is consistent with similarity processes observed in non-linguistic domains. The next three experiments provide further tests of the parallel between linguistic and non-linguistic similarity processes.

#### 4. Experiment 1B

Experiment 1B tested whether word similarity was more sensitive to alignable phoneme substitutions or nonalignable insertions. Arguably the central result of research on similarity within the structural alignment framework has been the finding that similarity comparisons focus primarily on matches and mismatches between aligned components (e.g. Goldstone, 1994b, 1996; Goldstone & Medin, 1994; Markman, 1996; Markman & Gentner, 1993). Goldstone (1994b), for example, distinguished between so-called “matches in place” (MIPS), which are featural matches between aligned components, and “matches out of place” (MOPS), which are featural matches between unaligned components. MIPS were found to increase similarity more than MOPS. Mismatches, likewise, can be either alignable or unalignable.

In one study that is particularly relevant to the present experiment, Markman and Gentner (1996) asked people to judge which of two choice pictures was more similar to a target picture, as illustrated in Fig. 5. The target depicts an archer aiming at a bull’s eye mounted on a wall. The choice pictures are similar to the target except for the presence of a bird sitting on the wall, which is not present in the target picture. In one choice picture, shown on the right, the bull’s eye is gone and the bird becomes its substitute. This is an alignable difference between target and choice pictures. In the other choice picture, the bull’s eye remains and the bird is an extraneous, inserted element with no correspondent in the target picture—an unalignable difference. On 88% of trials, participants indicated that the insertion choice was more similar to the target than the substitution choice. This result and numerous others support the conclusion that picture similarity is based primarily on correspondences between aligned elements. Insertions, which give rise to nonalignable

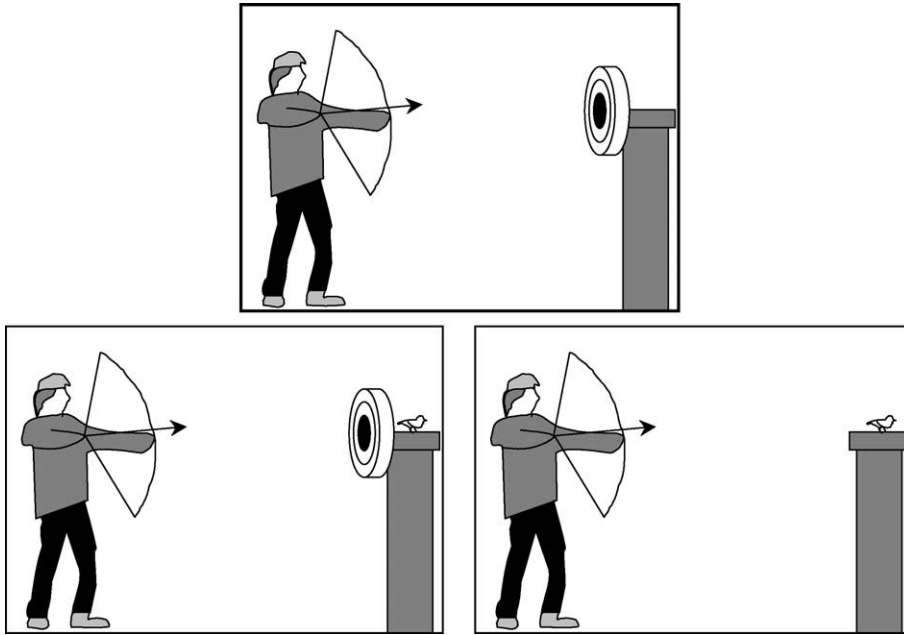


Fig. 5. Insertion and substitution in a scene depicting predicate-argument relations between objects (after Markman & Gentner, 1996, Fig. 4, p. 242).

elements with no direct correspondent, have relatively little impact on similarity judgments.

Does this finding extend to words? If so, then inserting a consonant into a word should produce a highly similar word (e.g. *bat-brat*), whereas replacing a consonant (*bat-rat*) should produce a word that is less similar to the original. Experiment 1B was designed to test this prediction.

#### 4.1. Method

Twenty target words were compared to choice candidates in which one candidate replaced the initial consonant of the target with a different consonant,  $\hat{C}$ , and the other candidate inserted the same consonant,  $\hat{C}$ , immediately after the initial consonant, creating a consonant cluster (e.g. /kug/-/jug/, /kug/-/kjug/). Similar manipulations of coda consonants in another 20 target words either replaced the final consonant or inserted a consonant immediately before it (see Appendix A).

#### 4.2. Results and discussion

Results are summarized in Table 1. For syllable onset comparisons, there was virtually no difference between target-candidate pairs differing by an insertion and those differing

by a substitution ( $M=51$  and 49%, resp.),  $F' < 1$ . However, when contrasts were in syllable codas, pairs differing by an insertion were judged to be more similar than pairs differing by a substitution ( $M=64$  and 36%),  $F'(1,16)=17$ ,  $MS_{err}=0.45$ ,  $p=.001$ . We observed the effect predicted by structural alignment, but only in syllable codas.

The structural alignment interpretation of the coda effect is that comparison of a word pair like *rat-ran* puts the consonants /t/ and /n/ into alignment with each other, an aligned difference that has a substantial effect on overall similarity. In contrast, a pair like *rat-rant* aligns the final consonants of both words (as well as the initial consonants and the vowels) and leaves the /n/ as an unaligned difference that has a minimal impact on the overall similarity between the two words. If this interpretation is correct, it indicates that comparisons of words, like comparisons between other complex mental representations, (1) involve an alignment process that identifies correspondences between constituent parts, and that (2) overall similarity assessment is affected by the nature of the established alignments. This interpretation requires the alignment of a final singleton consonant with the final consonant in a cluster, at least when those consonants are identical as in the current experiment. This result is incompatible with any scheme that aligns coda phonemes in strict left-to-right order (e.g. Luce, 1986; Luce & Pisoni, 1998).

Regardless of the question of alignment, our data indicate that a phoneme insertion (at least in the coda) makes for a smaller difference between words than a phoneme substitution. This finding is at odds with the common practice of treating insertions, deletions, and substitutions the same in computing edit distances (e.g. Bailey & Hahn, 2001; Vitevitch & Luce, 1999; Vitz & Winkler, 1973). Our result suggests that substitutions should be assigned a greater cost than insertions or deletions in measures of word similarity based on edit distance.

Our data are also incompatible with another common practice that identifies neighborhoods of similar same-length words and assumes (usually implicitly) that different-length words are dissimilar. Neighborhoods of similar words are often used to distinguish between words that could be facilitated or inhibited by many similar-sounding words and those that are not affected by many similar-sounding words. In particular, similarity neighborhoods are often defined in terms of same-length words differing by a single phoneme substitution. This definition excludes words that differ by the addition or deletion of a single phoneme. Our result indicates that these excluded words may be substantially more similar than the words included in the usual so-called neighborhood.

Finally, the predicted difference between alignable and nonalignable differences was not observed in syllable onsets. The difference between onsets and codas might indicate that onset comparisons are less salient than coda comparisons, and that our task lacked the statistical power to detect the smaller effect in onsets. This explanation would be compatible with either an alignment-based or non-alignment based view of word similarity. An alternative explanation could be that onset consonants align in a strict right-to-left sequence, as assumed by Luce (Luce, 1986; Luce & Pisoni, 1998). A right-to-left alignment of onset consonants would align the /b/ of *bat* with the /r/ of *brat*, for example. Such an alignment produces one aligned difference (/b/-/r/) and one unaligned difference (the /b/ of *brat*). If similarity comparisons are dominated by aligned differences, as predicted by results in non-linguistic similarity comparisons, then this

alignment of *bat-brat* would be only marginally less similar than the comparison of *bat-rat*, involving just an aligned difference. There is some evidence against this interpretation, however. Stemberger and Treiman (1986) argue on the basis of speech errors that the consonants in onset clusters occupy distinct C1 and C2 slots, and that a singleton onset consonant occupies a C1 slot. If structural alignment is sensitive to different types of consonant slots, Stemberger and Treiman's analysis would predict that the initial consonants of *bat-brat* would align with each other, so that the comparison would involve just an unaligned difference. Thus, Stemberger and Treiman's analysis points toward the conclusion that the difference we observed between onsets and codas is due to a difference in salience between these parts of the syllable in comparisons between words.

## 5. Experiment 1C

Under the structural alignment interpretation, Experiment 1B suggested that a singleton final consonant can align with a matching final consonant in a cluster. The results for onset comparisons were less clear, but were consistent with alignment between a singleton initial consonant and a matching initial consonant in a cluster. Experiment 1C combines the phoneme insertion aspect of Experiment 1B with the sub-phonemic feature manipulation of Experiment 1A, to test whether featural differences are relevant for comparisons between words of different lengths. If the degree of featural match between two phonemes is important to word similarity then we must know which phonemes to compare when words are of different lengths. The results of Experiment 1B are suggestive, but to what extent are non-identical components aligned? Consider the insertion choice picture of Fig. 5, that is, the picture in the lower left that has a bird in addition to the bull's eye. This picture is clearly very similar to the standard picture in the top of Fig. 5. Intuitively, if the bull's eye in the insertion choice picture were of a different shape, the picture would be less similar to the standard. If the bull's eye were of a different shape and was also painted in different colours, then the resulting picture would be even less similar to the standard. Thus, the similarity of the pictures would vary as a function of the degree of match between their aligned components, as opposed to depending entirely on identity matches between aligned parts. This intuition has been widely confirmed within the structural alignment literature using visual scenes (e.g. Goldstone & Medin, 1994; Markman & Gentner, 1996).

Does the same pattern hold for similarities between words? We have found no experimental evidence one way or the other in published studies of word similarity. In direct analogy to the results obtained with visual scenes, we would expect people to be sensitive to degrees of featural similarity between aligned phonemes. For example, the word pair *pet-pend* (differing by a single feature in the final consonants) should be more similar than *pet-pens* (differing by two features). This prediction incorporates, and therefore tests, the claim of structural alignment that component parts are aligned, and that featural similarity among aligned parts contributes to the overall comparison.



### 5.1. Method

We generated target words whose onset (or coda) consisted of a singleton consonant (e.g. /miv/) and two choice candidates containing consonant clusters (e.g. /briv/, /griv/). One candidate was derived from the target by replacing the singleton onset (or coda) consonant with a consonant cluster. The initial (final) consonant of the cluster differed from the original singleton by just one feature. The second candidate was derived from the first by changing an additional second feature of the initial (final) consonant. There were 20 onset triads and 20 coda triads.

### 5.2. Results and discussion

As indicated in Table 1, there was no significant effect for onsets, though pairs differing by a single feature in the initial consonant were judged to be somewhat more similar than pairs differing by two features ( $M=55$  and 45%, resp.),  $F'(1,16)=2.2$ ,  $MSerr=0.45$ ,  $p=.15$ . For codas, pairs differing by a single feature in the final consonant were judged to be significantly more similar than pairs differing by two features ( $M=63$  and 37%, resp.),  $F'(1,16)=6.7$ ,  $MSerr=0.48$ ,  $p=.002$ . Again, the effect predicted by structural alignment was observed, but only in syllable codas.

Word similarity was sensitive to featural differences between consonants in word final position even though the words under comparison were of different lengths. This finding is consistent with a structural alignment interpretation, incorporating an alignment process combined with comparisons among aligned components that are sensitive to their featural similarities. As in Experiment 1B, the structural alignment interpretation requires the alignment of a final singleton consonant with the final consonant in a cluster. In contrast to Experiment 1B, in which the aligned consonants were identical, the present study requires non-identical consonants to be aligned. The results, like those of Experiment 1B, are not compatible with a strict left-to-right alignment of coda consonants (e.g. Luce & Pisoni, 1998). Our data do not show whether word-final consonants are always aligned with each other based on their structural position, or whether a final singleton can align with whichever consonant in a cluster has more featural similarities. For example, would the word pair *hen-hemp* align final /n/ with /m/ (one featural difference, but non-final) or with /p/ (three featural differences, but final)? The theory of structural alignment is compatible with either alignment, depending on whether the alignment process is dominated by structural or featural properties of words (Markman & Gentner, 1993).

As in Experiment 1B, the predicted effect of aligned initial consonants was not observed in the present experiment. Again it could either be that onset comparisons are less salient than coda comparisons, or that onset consonants are aligned in a strict right-to-left sequence. This second possibility would predict no difference whatsoever between pairs like *tip-drip* and *tip-grip*, because in each case /t/ would be aligned with /r/ to produce an alignable difference, and the initial consonant of the cluster (either /d/ or /g/) would be an additional unaligned difference. Although we cannot rule out this possibility entirely, we note that the magnitude of the difference between codas and onsets in the present experiment (63 vs. 55% choice A responses) is identical to the difference in Experiment 1A (70 vs. 63%). It is quite possible that onset comparisons were sensitive to

featural differences in this experiment as in Experiment 1A, but that there was a general decline in people's ability to identify more similar word pairs in the present experiment compared with that one. Such a decline would be consistent with a very general magnitude estimation phenomenon.

It is a universal of magnitude estimation in psychophysical contexts that sensitivity varies with the absolute magnitudes under consideration. Our ability to distinguish whether or not two objects have equal weight, for example, is dependent on how heavy these objects are: differences of a few grams that are discernible when comparing two sheets of paper cannot be detected when comparing the weight of two filled suitcases. Generally, the smallest noticeable difference between two magnitudes is a constant fraction of these magnitudes (Weber, 1835/1978). This effect follows from certain assumptions about choice behaviour, as discussed by Luce (1959). In particular, under conditions of imperfect discrimination, it may be reasonable to suppose that people's judgments when asked to identify, say, the larger of two psychophysical values (e.g. the heavier of two suitcases) have a probability distribution described by Luce's choice rule:

$$\text{probability of response } A = \frac{\text{Magnitude } A}{\text{Magnitude } A + \text{Magnitude } B}.$$

It is easy to see that response probabilities according to this equation will be very sensitive to a unit difference between  $A$  and  $B$  when  $A$  and  $B$  are both small (close to zero), and less sensitive to a unit difference as  $A$  and  $B$  become larger.

It seems natural to interpret the forced choice task of our experiments as requiring participants to make judgments about the magnitudes of phonological distances between pairs of words. In Experiment 1A, the distance between words in a pair is minimal, and it is relatively easy for participants to judge which of two pairs is more similar. In Experiment 1C, the distance between words in a pair is increased by the insertion of an extraneous (unaligned) consonant, and it is more difficult for participants to judge which of two pairs is more similar. The decline in sensitivity to a given difference in phonological distance (one versus two features) follows the classic pattern for magnitude estimation.

## 6. Experiment 1D

Under the structural alignment interpretation, Experiments 1A and 1C demonstrated that judgments of word similarity are sensitive to the degree of featural match between aligned components of words, at least within codas. Experiment 1D examined whether the same number of featural matches gives rise to different degrees of similarity depending on whether the matches are clustered together in a single phoneme or whether they are distributed across multiple phonemes.

Goldstone and Medin (1994) found that the similarity between two visual scenes was greater when matching features were concentrated than when they were distributed more widely among parts of the scenes. Consequently, the importance of a given "match in place" (MIP) depended on how other MIPS were distributed. When comparing two scenes each of which comprised a pair of butterflies, for example, similarity was greatest if four

MIPS were clustered into the right-hand butterfly of each butterfly pair. Similarity was reduced when three MIPS were located in the right-hand butterfly and one in the left-hand butterfly of each pair, and it was reduced further if the distribution was two to two across the butterflies of each pair. Across a range of studies, the similarity between two ensembles depended not just on the number of corresponding matches between the ensembles, but also on the degree to which these matches were clustered into a particular component.

In the context of words, an analogue is the question of whether a target word such as *dad*, is more similar to *pad* or to *bat*. Overall, both *pad* and *bat* match in exactly the same features, but the distribution of these matches is different: while *pad* matches in three out of three features in two of its phoneme positions (1-3-3), *bat* matches three out of three features in only one position (2-3-2). The greater clustering of matches in *pad* should give it a higher similarity to the target if Goldstone and Medin's (1994) findings carry over to words.

### 6.1. Method

Experiment 1D asked participants to compare a target word to two choice candidates that matched the target in all but two phonological features (e.g. /fuf/-/zuf/, /fuf/-/suv/). Each target word ended with the same consonant with which it began. The distribution of feature differences across onset and coda was such that they either maximised the matching features in one position (/fuf/-/zuf/) or distributed feature matches evenly so that the candidates matched in two out of three features in both onset and coda positions (/fuf/-/suv/). There were 20 onset triads that included a target-candidate pair with feature matches clustered in onsets, and 20 coda triads that included a target-candidate pair with feature matches clustered in codas.

### 6.2. Results and discussion

As indicated in Table 1, word pairs with featural matches clustered in the onset were no more similar than pairs with matches distributed between onset and coda, ( $M=51$  vs. 49%),  $F' < 1$ . In contrast, word pairs with matches clustered in the coda were more similar than pairs with matches distributed between onset and coda, ( $M=73$  vs. 27%),  $F'(1,17)=20$ ,  $MSerr=0.41$ ,  $p < .001$ .

This pattern of results confirms the prediction that clustered matches would increase similarity more than distributed matches, but suggests that this effect is modulated by syllable position. The results are consistent with clustered matches having an effect in both onsets and codas if the cluster effect interacts with a tendency for coda matches to be more salient. To illustrate, suppose we compared the target word 'deed' to choice words 'deep' and 'beet'. Across phoneme positions, the distribution of feature matches with the target word is (3-3-1) for 'deep', the clustered-match choice, and (2-3-2) for 'beet', the distributed-match choice. The total number of feature matches is the same for both candidates, but the clustering of matches in 'deep' should make it more similar to the target than 'beet'. However, if matched coda features have greater salience than matched onset features, then the feature matches would appear as if they were something like (3-3-1)

or (2-3-2), and similarity comparisons would depend less on the clustering of onset matches and more on the number of coda matches. Therefore, although ‘deep’ has greater clustering of feature matches, ‘beet’ has more coda matches, and the effect of clustering and the effect of coda salience for matches work against each other. If there was just an effect of clustered matches, we should have observed a preference for cluster candidates like ‘deep’. If there was just an effect of coda salience we should have observed a preference for distributed candidates like ‘beet’. Consequently, our null result for stimuli with clustered matches occurring in onsets suggests that the two effects cancelled each other out.

In contrast, when matches are clustered in codas, clustering and coda salience work together instead of against each other. Consider the target word ‘dad’ and the choice words ‘pad’ and ‘bat’, with the distributions of feature matches (1-3-3) and (2-3-2), respectively. Coda salience would make these feature matches appear something like (1-3-3) and (2-3-2). Here, the choice word ‘pad’ has greater clustering and it also has more coda matches, so both effects favour the clustered choice over the distributed choice ‘bat’.

For both onsets and codas, then, matches clustered within a phoneme seem to increase similarity more than the same matches distributed more widely. Why should this be so? There are at least two interpretations of this effect in the present experiment. The first interpretation is that similarity increases more than linearly as the number of shared features in corresponding phonemes increases.<sup>6</sup> The first shared feature increases similarity a little bit, the second feature shared by the same phonemes increases similarity by a greater amount, and a third shared feature produces an even greater increase in similarity. Thus, the importance of each shared feature may be compounded by the presence of other shared features. The second interpretation of the observed clustering effects is based on the next level in the hierarchy of phonological units, the phoneme. In our stimuli, the clustered match items always shared a consonant as well as the vowel with the target word, whereas the distributed match items shared only a vowel with the target. Thus it is possible that our results reflect the number of exact matches at the phoneme level instead of showing a general effect of clustered matches at the feature level, *per se*.

Although the results of the present study could reflect either an effect of clustered matches, an effect of whole phoneme matches, or perhaps a combination of both, previous work with visual figures suggests that clustering plays a role in similarity that is independent of matches at higher levels in hierarchical representations. Goldstone and Medin (1994), for example, found an effect of clustering that is not due to exact matches in higher level components. Their butterfly couplets included configurations that had no exact butterfly matches, but that differed only in the degree of clustering of feature matches. Goldstone and Medin found that similarity increased as the clustering of feature matches increased, whether or not the number of exact butterfly matches also increased. Similarly, Goldstone (1993) reports a number of studies in which clustered items in visual displays capture participants’ attention, and Goldstone concludes that “concentrated information influences judgments more than distributed information” in many areas (p. 577). Finally, Tversky and Gati (1982) also found that similarity was increased more by

---

<sup>6</sup> A similar effect was observed by Marslen-Wilson, Moss, and van Halen (1996), in a study of semantic priming with mispronounced prime words.

matches accumulated along a single psychological dimension than by the same total number of matches distributed across more than one dimension. If word similarity is analogous to similarity in other domains, then these studies all predict a preference for clustered match words even when no phonemes match exactly. So, for example, word pairs like *dad-fat*, with feature matches distributed as (0-3-2), should be more similar than *dad-pass*, with matches distributed as (1-3-1).

Whether the cluster effect we observed in this study is ultimately due to clustering of feature matches, to matches between whole phonemes, or to a combination of both, the present experiment demonstrates another relational aspect of word similarity, showing that it depends not only on numbers of matches or mismatches, but also on the distribution of feature matches relative to each other.

## 7. Experiment 1E

Each of the experiments above has suggested that the impact of a given match or mismatch on perceived similarity was affected by whether it was found in the onset or the coda. Matches in codas seemed to be more salient, and thus carry greater weight in similarity assessment than the same matches in onsets. Our final experiment tests directly for differential salience of onsets and codas.

General considerations of perception and memory alone do not make clear predictions: codas might benefit from recency, but onsets will benefit from primacy effects. Likewise, there is conflicting evidence from studies of language processing. For example, on one hand, sensitivity to word beginnings may develop before sensitivity to the ends of words (Duncan, Seymour, & Hill, 1997; Jusczyk et al., 1999). On the other hand, shared phonemes at the ends of words induce facilitatory phonological priming in word-nonword lexical decision, but shared initial phonemes do not (Słowiacek & Pisoni, 1986). Likewise, it is easier to quickly repeat a pair of words that shares coda consonants than a pair that shares onset consonants (Sevold & Dell, 1994).

There has been almost no work that has examined potential differences between onsets and codas in their influence on word similarity directly, and so far the evidence found is conflicting. As already mentioned above, Sendlmeier (1987) reported that shared features of consonants had some effect for monosyllables, but only in onsets and not in codas. Similarly, Nelson and Nelson (1970) found that three-letter words that shared only their initial sounds (e.g. *pat-pin*) were more similar than words that shared only their final sounds (e.g. *pat-fit*). However, when Nelson and Nelson examined words that had the same vowel, those whose final consonants were also the same (e.g. *lag-bag*) were more similar than words with the same initial sound (*lag-lad*). This latter finding suggests that coda matches are more important than onset matches, which is consistent with our results in Experiments 1A–D. Nelson and Nelson's results, in particular, suggest that the relative importance of onset and coda matches may vary depending on whether the words being compared have the same vowel sound.

The present experiment tests whether Nelson and Nelson's (1970) finding of greater coda influence in the case of CVC minimal pairs generalises to nonwords with a rich variety of simple and complex onsets and codas. We focus on the high end of the word

similarity spectrum, and compare words with the same vowel, because it is very similar lexical items that exert the greatest influence in language processing (cf. e.g. Bailey & Hahn, 2001).

### 7.1. Method

Experiment 1E asked participants to compare a target word to choice candidates that differed from the target by the same single feature in either the initial or final sound (e.g. /fæf/-/fæf/, /fæf/-/fæf/). Each target word ended with the same consonant with which it began. These target-candidate pairs are not just minimal pairs (differing from the target by a single phoneme)—they are minimally different in that they share all features but one with the target. If similarity between minimally different words is more sensitive to coda matches than onset matches, then /fæf/ should be more similar to /fæf/ than /fæf/ is. There were 20 target-candidate triads.

### 7.2. Results and discussion

As indicated in Table 1, minimally different word pairs with matching final consonants were somewhat more similar than pairs that had matching initial consonants ( $M=58$  vs. 42%, resp.). This difference was marginally significant in a two-tailed test,  $F'(1,24)=2.9$ ,  $MSerr=0.85$ ,  $p=.077$ . The trend was in the same direction as that observed by Nelson and Nelson (1970) for minimal pairs. Finally, a post hoc omnibus test of the overall contrast between onsets and codas across Experiments 1A–D shows a significant difference between onsets and codas,  $F'(1.04, 66.0)=13.1$ ,  $p=.001$ .

When a single feature was changed in either the initial or final consonant, leaving a matching phoneme in the other position, matches in codas tended to produce greater similarity than matches in onsets. This result is in line with the findings in Experiments 1A–D, all of which suggested greater coda salience. Our results are also consistent with the findings of Nelson and Nelson (1970) for words that differed in only one phoneme position. We conclude that comparisons between codas influence similarity more than do comparisons between onsets, at least for words of comparatively high degrees of similarity.

Why might such a difference between onsets and codas occur? One possibility is that coda differences are actually easier to detect than onset differences. A detectability advantage for codas would enable them to more reliably influence similarity. However, evidence reviewed by Sussman, Bessell, Dalston, and Majors (1997) suggests that, if anything, onsets are articulated more precisely than codas. This should make it easier to detect differences between onsets than codas, but this is not what our results suggest. Further evidence to this effect stems from Benki's (2003b) finding that CV syllables are more intelligible under masking noise than are VC syllables (see Benki, 2003b for further discussion of perceptual differences between onsets and codas). An alternative explanation is that coda differences, though not easier to detect, are given greater weight. In other words, similarity judgments themselves focus on the ends of words rather than the beginnings. A third possibility is that the differential salience observed has its explanation in hierarchical structure: similarity may be higher for items that have larger, higher-level

matching constituents, and lower for items whose matches are confined to smaller, lower levels of structure. This would explain the effects found in Experiment 1E because, in our study, word pairs that matched in the final phoneme matched not only in the coda, but also in the entire syllable rime. In contrast, the pairs with the same initial phoneme matched only up to the level of the onset, that is, one hierarchical level less, as seen in Fig. 2.<sup>7</sup>

These last two explanations are distinct, but mutually compatible—and both might be relevant to the current results. Though we cannot at present provide a definitive explanation for the greater salience of codas in our studies the issue can be resolved empirically. The explanation in terms of level of highest hierarchical match would predict different results for stimuli in which the syllable nucleus did not match, as this rules out a match in rime without affecting the experimental manipulation performed in the present experiment. Matching in terms of hierarchical structure might also explain Nelson and Nelson's (1970) observed interaction between the relative salience of onsets and codas and the overall similarity of the words under comparison. Their results suggest that matching rimes might count more than matching onsets, while matching onsets count more than matching codas.

The present experiment is the only one in our series that was not informed by previous work on similarity in other domains. The question suggested itself directly for words, but it might nevertheless apply to similarity in other areas. Differential weighting, for example due to attentional focus, is a familiar concept within the general literature on similarity. However, it has to our knowledge only been applied to individual features or dimensions, not to positional locations within an object or scene. It seems possible, however, that visual scenes that have matching objects in central positions might be more similar than comparable scenes that have matching objects in peripheral locations, even under circumstances where all differences are readily detectable by the observer. The greater attention afforded to the eye region in the context of face recognition points in this direction (e.g. Davies, Ellis, & Shepherd, 1977; Haig, 1985). Finally, if the depth of the hierarchy in which a match occurs is the factor driving our results, then analogous effects might be observable with a simple configuration of objects such as that in Fig. 4A above: The same featural change should decrease similarity more when affecting the square than when affecting the triangle as the square participates in an additional level of relational match. In short, the difference we observe between onsets and codas raises several important questions about similarity comparisons, not just between words, but between other complex mental representations as well.

## 8. Computational models of word similarity

In systematic manipulations of word similarity, Experiments 1A–E identified four key effects that are not predicted by phoneme edit distance. Word similarity is (1) sensitive to the number of sub-phonemic feature matches, (2) affected by the extent to which feature

---

<sup>7</sup> As one reviewer pointed out the observed preference would also arise if coarticulation systematically had a greater effect for codas than for onsets. In other words, if the shared vowel varied more across the coda contrasts, than across the onset contrasts, then coda differences would have a larger impact overall.



matches are clustered or distributed across different phonemes, (3) more sensitive to phoneme substitutions than insertions, and (4) more sensitive to coda than onset comparisons. Except in the case of Experiment 1D, phoneme edit distance predicts no differences in similarity for any of our comparisons. This illustrates the fact that phoneme edit distance makes few distinctions among highly similar words—precisely the words that have the greatest interactions in language processing tasks. In the case of Experiment 1D, phoneme edit distance predicts the same results for onset and coda comparisons, contrary to our experimental results.

In this section of the paper, we incrementally develop a more elaborate model of word similarity, beginning with phoneme edit distance and assessing the impact of various extensions suggested by our data. It would be trivial and uninteresting to propose one model to explain part of our data and a different model for another part (e.g. separate models for onsets and codas). We therefore focus on developing a single model that provides an adequate account of our entire data set. The models we consider all assume a phoneme alignment process that minimizes the number of unaligned phonemes and the number of mismatches between aligned phonemes. In this sense, these models are all instantiations of structural alignment.

We tested five edit distance models of word similarity in total, which are summarized below. Further details are given in Appendix B. For each model, we computed the predicted proportion of choice *A* responses for all 180 syllable comparisons, using Luce's choice rule (Luce, 1959) to obtain predicted response proportions from edit distance values. Thus, for comparisons between target *T* and choice words *A* and *B*, the predicted proportion of responses indicating that *A* is more similar to *T* is

$$\text{probability of response } A = \frac{\text{EditDistance}_{TB}}{\text{EditDistance}_{TA} + \text{EditDistance}_{TB}}. \quad (1)$$

Predicted response probabilities for each model were fit to the data using nonlinear regression to minimize squared deviations from the actual response probabilities.

### 8.1. Extending phoneme edit distance

The first four edit distance models are basic phoneme edit distance plus three simple extensions motivated by our data. The basic phoneme edit distance (PED) model counts the number of phoneme substitutions, insertions, or deletions required to transform one word into another. As described above, it is computed by aligning the phonemes of two words so as to minimize the number of mismatches. PED adds up the cost of all mismatches (insertions, deletions, and substitutions), assigning a uniform cost of 1 to each mismatch.

Because Experiment 1B identified differences in the effect of consonant substitutions and insertions on word similarity, the PED + InsDel model extends basic edit distance by assigning different costs for insertions (and deletions) compared to substitutions. This is accomplished by letting the cost of an insertion or deletion be a free parameter in

the model, allowing it to be either greater or less than the cost of a substitution, which remains fixed at 1.

Because Experiments 1A–E consistently found differences between onsets and codas, the PED+InsDel+Coda model includes a further extension that assigns differential weight to mismatches in onsets versus codas. This is accomplished by multiplying the cost of a mismatch in the coda by a coda weighting factor (CodaWeight), which is a free parameter. The cost of an onset mismatch is multiplied by  $1 - \text{CodaWeight}$ .

Finally, because Experiments 1A and 1C showed that similarity judgments were sensitive to featural differences, the FED+InsDel+Coda model further enhances the PED+InsDel+Coda model by assigning the cost of a phoneme substitution according to the number of featural differences involved (place, manner, and voicing). This model does not involve any additional free parameters.

The results predicted by the models above for each of our nine stimulus sets are shown in Fig. 6A and B. Fig. 6A shows mean actual and predicted fractions of judgments indicating that choice A is more similar than B to the target word, averaged across the 20 comparisons within each stimulus set (see Table 1). For convenience, the squared deviations of predicted from actual response probabilities are shown in Fig. 6B, grouped by model rather than by stimulus set.

In order to assess the overall ability of each model to account for our pattern of empirical results, we also compared the average predicted response rate with the observed response rate for each of our nine stimulus sets, and computed an  $R^2$  statistic measuring the extent to which the model explained the overall variance from chance responding. These statistics are shown in Table 2, along with the number of free parameters for each model. Because the  $R^2$  statistic was intended to assess the ability of the model to explain the effects of the qualitatively different stimulus manipulations, this statistic was computed on cell means, not on response rates for individual stimulus items. Given that our stimuli were a random sample of the triads relevant to each comparison, these results estimate

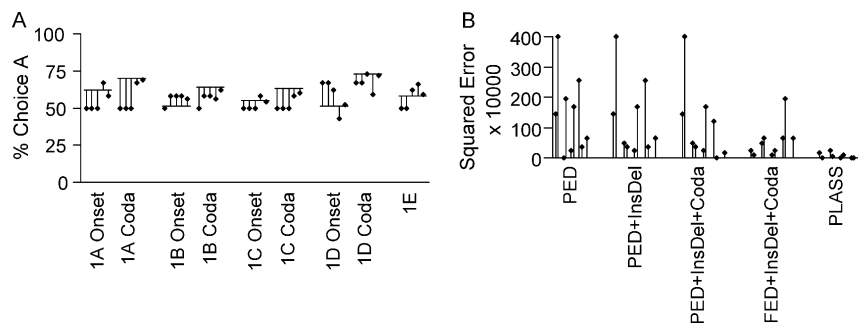


Fig. 6. Summary of model fits. (A) shows actual and predicted fraction of choice A responses grouped by the nine stimulus sets. Horizontal lines show actual responses, dots show the responses predicted by five models, in order: PED, PED+InsDel, PED+InsDel+Coda, FED+InsDel+Coda, and PLASS. (B) shows squared error grouped by the five models. Within each group are data for the nine stimulus sets corresponding, in order, to onset and coda comparisons for Experiment 1A and similarly for 1B–D, plus comparisons for 1E.

Table 2

Comparison of metrics, showing the number of free parameters (*Df*), and the percentage of variance from chance explained among the average responses for the nine stimulus sets ( $R^2$ )

Metric	<i>Df</i>	$R^2$ (%)
PED	0	16
PED + InsDel	1	24
PED + InsDel + Coda	2	38
FED + InsDel + Coda	2	70
PLASS	5	97

the extent to which each model captures the variation between these nine types of comparisons in the wider population of relevant monosyllables.

*Phoneme edit distance.* It is clear from Fig. 6A and B that PED does not accurately predict the pattern of average results. This is particularly true for the coda comparisons of Experiment 1A and the onset comparisons of Experiment 1D. It is true to a lesser extent of the coda comparisons of Experiments 1B and 1C. Overall, PED explained  $R^2=16\%$  of the variance from chance responding for our nine different types of stimuli.

*Insertions versus substitutions.* The best-fit parameter value assigned greater weight to substitutions than insertions or deletions by a ratio of 4:3. The InsDel parameter was motivated by Experiment 1B, and it is only in the responses to this set of stimuli that the predictions of PED + InsDel differ from those of PED (see Fig. 6). Overall, PED + InsDel explained  $R^2=24\%$  of the variance.

*Onset versus coda.* The best-fit parameter value assigned greater weight to mismatches in codas than onsets by a ratio of 2:1. Although differences between onsets and codas were observed in all of our experiments, the Coda parameter is only effective in adjusting the predicted responses for Experiments 1D and 1E (see Fig. 6B). It has no effect on the predicted results of Experiments 1A–C because when edit distances are entered into Luce's choice rule, if both word pairs involve only onsets or both involve only codas, the syllable part weighting parameters in the numerator and denominator terms simply cancel out. Overall, PED + InsDel + Coda explained  $R^2=38\%$  of the variance. We return to the differential salience of onsets and codas in the final PLASS model below.

*Feature matches.* The best-fit value of the CodaWeight parameter was about the same as in the PED + InsDel + Coda model, with a 2:1 ratio between the weight given to matches in codas compared to onsets. The best-fit value of the InsDel parameter put the cost of an insertion or deletion at about the same level as 1.6 feature substitutions. The incorporation of sub-phonemic features clearly gives better predictions for the results of Experiment 1A and 1C. This model gives fairly accurate predictions for all stimulus sets except the coda comparisons of Experiment 1D. Nevertheless, FED + InsDel + Coda fails to predict the differences between onset and coda comparisons in Experiments 1A–C because the syllable part weighting parameter cancels out as discussed above. Overall, FED + InsDel + Coda explained  $R^2=70\%$  of the variance.

### 8.2. Syllable part power law

The various PED and FED models do not predict the differences between onset and coda manipulations that we observed in Experiments 1A–C, because even when they incorporate a multiplicative syllable part weighting factor (e.g. PED + InsDel + Coda), it simply cancels out when applied to the stimulus sets for these studies. Instead of using different multipliers depending on the syllable part involved, the power law of aligned syllable structures (PLASS) uses different exponents, so the cost of a phoneme substitution, insertion or deletion is raised to a different power depending on the part of the syllable it is in. This power law transformation allows different response profiles for onsets and codas in Experiments 1A–C.

In the PLASS model, one free parameter determines the cost of feature substitutions, and a second free parameter determines the cost of a phoneme insertion or deletion relative to the cost of feature substitutions. The costs of all differences within the same syllable part are added up and then raised to the power  $P_{\text{Onset}}$  or  $P_{\text{Coda}}$  depending on the syllable part involved. These values are then added across the whole word and raised to the power  $P_{\text{Word}}$ :

$$\text{Edit Distance} = \left[ \left( \sum_{\text{Onset Diffs}} \text{Cost of Diff} \right)^{P_{\text{Onset}}} + \left( \sum_{\text{Coda Diffs}} \text{Cost of Diff} \right)^{P_{\text{Coda}}} \right]^{P_{\text{Word}}}.$$

The PLASS model could be generalized in an obvious way to take additional levels of syllable structure into account, such as nucleus and rime constituents, but the simpler formulation given here is sufficient for our data.

The best-fit parameter values raised coda differences to a larger power than onset differences by a ratio of 5:2. The best-fit cost of an insertion or deletion was about the same as 1.3 featural substitutions. Predicted choice A response fractions are shown Fig. 6A as the last dot of each group. Squared error is shown in Fig. 6B. This model very accurately predicts the results for all nine stimulus sets. Overall, PLASS explains 97% of the variance from chance responding.

### 8.3. Discussion

The standard phoneme edit distance model does not provide an adequate account of our similarity judgments. It explained only a small fraction of the variance in responses among our nine stimulus sets. This result demonstrates quite dramatically that, although it provides useful coarse distinctions between similar and dissimilar words, phoneme edit distance does not explain finer degrees of similarity among highly similar words. These degrees of similarity were readily apparent in our study and are likely to be salient enough to make a difference in many psycholinguistic experiments where word similarity is relevant.

Obvious elaborations of phoneme edit distance can discriminate between phoneme insertions or deletions and substitutions, and can differentiate to some extent between

different parts of the syllable. However, the PED+InsDel+Coda model is still based on phoneme matching, and does not take account of featural differences between aligned phonemes. Few readers will be surprised that models which ignore featural differences fail to predict responses for comparisons that specifically manipulate featural differences. Incorporating featural differences, the FED+InsDel+Coda model provides a much improved fit to the data, and still has just two free parameters. However, despite its parameterized weighting for onsets versus codas, it does not capture distinctions between onsets and codas at all in three of our experiments (1A–C), and captures them only poorly in the other two experiments.

In contrast, the PLASS model fits our data very well, including the differences observed between onsets and codas in all five experiments. Of particular interest is the data from Experiment 1D, which is fit well by PLASS but not, for example, by FED+InsDel+Coda. In contrast to the other models we considered, which combine the costs attributed to individual mismatches in a linear way to compute overall word similarity, PLASS combines individual costs in different parts of a syllable nonlinearly. Because of this, interactions between the parameters of PLASS are more complex. Crucially, they interact in the right way to fit the similarity judgments for Experiment 1D as well as the other experiments. As we argued above, the similarity judgments for this experiment result from interactions between syllable part salience (which is greater for codas than onsets) and distributional similarity enhancement (greater for clustered than distributed feature matches). A model with the wrong sort of interactions between its parameters is unlikely to simultaneously fit Experiment 1D and the other experiments as well, because effects of syllable part salience and distributional similarity enhancement are constrained by the data from all five experiments. In PLASS, distributional similarity enhancement results from raising the cost of featural changes to a power less than 1 (either  $P_{\text{Onset}}$  or  $P_{\text{Coda}}$ ). In effect, this means that each additional feature change in the same syllable position counts less than the previous one or, equivalently, each additional match counts more. Thus, the importance of each shared feature is compounded by the presence of other shared features in the same syllable position. This non-linear effect of clustered feature matches or mismatches interacts with syllable position salience (a greater value for  $P_{\text{Onset}}$  or  $P_{\text{Coda}}$ ) to provide a good fit for the whole data set.

Our comparisons between models illustrate the considerable power this set of empirical data provides for discriminating between different models of word similarity. This power derives from interactions of the various factors our experiments examined individually, each of which constrains any potential model from a different direction. In short, it is factorially more difficult for a model to accommodate the data as a set than to provide individual accounts of each experiment separately. This point is illustrated in Fig. 6 by the comparison between PED+InsDel+Coda and FED+InsDel+Coda. Including featural similarity in FED+InsDel+Coda improves performance overall, particularly on Experiments 1A and 1C, but there is a trade-off with Experiment 1E and the coda conditions of 1B and 1D, all of which see reduced performance in FED+InsDel+Coda compared to PED+InsDel+Coda (the same model except for sensitivity to features).

It is also important to realize that models are constrained just as much by ‘null’ results, where similarity judgments are equally divided between the two choices, as by

results that show a significant preference one way or the other. For example, it is clear from Fig. 6 that FED+InsDel+Coda fits the data of Experiment 1D very poorly, including both the onset manipulation in which participants showed no clear preference for either choice, and the coda manipulation in which participants did show a consistent preference. Although significant inferential tests are informative, it is really the contrast between when effects occur and when they do not that is important. The ability to accurately predict the size of effects across the board is the goal of detailed models, whether individual effects are statistically significant or not.

The recent literature has seen several proposals for new measures of word similarity. None of these measures could accommodate the full range of our present results. Bailey and Hahn (2001) tested three new measures in their ability to predict lexical influences on sequence typicality, all of which provided improvements over basic edit distance. Mueller et al. (2003) also offered a new measure (PSIMETRICA), though they did not directly test its performance relative to basic edit distance. PSIMETRICA could not predict the lesser cost of insertions relative to substitutions, because unaligned phonemes are simply given a mismatch score corresponding to an average substitution mismatch value, nor would it allow for the effects of clustering. Of Bailey and Hahn's three metrics, only the first captures both of these results. However, neither PSIMETRICA nor any of the measures considered by Bailey and Hahn (2001) provide the right treatment of onset-coda weighting to capture our results.

Finally, we should note that although the two models that incorporate featural similarity (FED+InsDel+Coda and PLASS) explain most of the variance across our nine stimulus sets, the individual comparison triads within each stimulus set exhibit further variability (see Appendix A), most of which is not explained by these models. Only further empirical and theoretical work can determine the extent to which this inter-item variability reflects systematic processes versus noise, and the extent to which models would need to include more detailed sub-phonemic information (see e.g. Marlsen-Wilson & Warren, 1994; McMurray, Tanenhaus, & Aslin, 2002; McQueen, Norris, & Cutler, 1999; Smits, Warner, McQueen, & Cutler, 2003).

## 9. General discussion

This series of experiments provides basic new data on what makes words sound similar, and in particular, how the sound similarity between words is related to their component parts. Our findings have relevance for models of performance in a wide variety of language processing tasks. Specifically, word similarity was shown to be sensitive to the degree of mismatch in terms of phonological features—similarity declined as the number of mismatching sub-phonemic features increased (Experiments 1A and 1C). This effect was observed both in words of the same length and in words of different lengths. It was also demonstrated that the location of matches and mismatches is crucial. Mismatches between mutually corresponding parts of words affected similarity more than did extraneous parts occurring in one word only (Experiment 1B), which meant that word similarity was less affected by phoneme insertions than by substitutions. At the same time, clustered feature matches resulted in higher similarity

than the same matches distributed across multiple phonemes (Experiment 1D). Finally, the same degree of featural match had greater impact in coda positions than in onsets (Experiments 1A–E). Although the widely used phoneme edit distance measure of word similarity did not provide an adequate explanation of our findings, we identified two improved models that incorporated featural similarities for phonemes, reduced costs for phoneme insertions and deletions compared to substitutions, and greater salience for differences in codas compared to onsets.

These experiments were guided by studies of similarity processes outside the domain of language, and our results largely confirmed predictions derived from key effects identified in those studies. In particular, comparisons between words appear to involve an alignment process that places phonemes into correspondence (Experiments 1B–C); word similarity was more sensitive to aligned than unaligned differences (Experiment 1B), and word similarity was more sensitive to clustered than distributed feature matches (Experiment 1D). Our results suggest that findings with geometric shapes, visual scenes and conceptual descriptions carry over to phonological similarity between words. That the application of general theories of similarity to words should prove so fruitful is remarkable given that language is often viewed as special. For example, Fodor (1983) suggests that the processing of phonological information is dealt with by an informationally encapsulated, dedicated module, whereas semantically rich materials such as schematic shapes, visual scenes, and word meanings—things that have been the focus of general research on similarity—fall firmly within the central systems, which Fodor believes to be intractable. From this perspective it is surprising and noteworthy that theories of similarity developed within the context of high level cognition should carry over so well to similarity in the way words sound.

Moreover, the potential for productive interchange between the general similarity literature and further investigations on word similarity is by no means exhausted by the experiments presented here. Crucial issues that must be resolved by further research are the extent to which our results carry over to multi-syllabic words and to other tasks, specifically to online processing tasks. We address each of these issues in turn. With respect to multi-syllabic words, segmentation and syllable stress complicate the design of well-controlled stimulus materials, but there is nothing inherent in our experiments that would preclude replication with multi-syllabic words. We expect additional influences on syllable similarity from syllable stress; apart from this, however, structural alignment theory would predict that our results with monosyllables should carry over directly to the comparison between aligned syllables of multi-syllabic words. Furthermore, we expect that the predictions of structural alignment theory will serve well, not just at the level of phonemes and their alignment within monosyllabic words as examined here, but also at the level of syllables and their alignment within multisyllabic words. Issues such as the comparative influence of clustered versus distributed matches or of substitutions versus insertions arise again when it comes to determining which syllables in two polysyllabic words are aligned and how the various syllable matches and mismatches affect similarity. For example, in parallel to our findings with clustered versus distributed phoneme matches, we would expect two tri-syllabic words which matched exactly in two syllables to be perceived as more similar



than two words which match exactly in only one syllable, even if the two words have the same number of featural and phonemic matches.

The second issue of online processing tasks is an issue that word similarity shares with similarity in other domains. How results from explicit similarity judgments carry over to other tasks is equally an issue for the general literature on similarity and for the present investigation of word similarity. In contrast to much of the work with visual scenes or word meanings, research on word similarity has the benefit that in addition to possessing comparatively well-developed, independently motivated theories of phonological representation, such representations seem comparatively constrained. That these representations seemed to matter in our experiments suggests that the word similarity judgment task draws on the same phonological representations and processes as those relevant to these other tasks. It would thus be more than surprising if our results did not generalize at all to other language processing tasks.

However, given the ubiquity of task and context dependent modulation of similarity within the general cognitive literature (see e.g. Goldstone, 1994a) it would make word similarity decidedly different if no contextual effects emerged with words (see also Wickelgren, 1965, 1966). Because context sensitivity has been a central property of similarity in other cognitive domains, the general literature on similarity provides a wealth of research that might guide the search for possible task dependent shifts in word similarity and explain such shifts where they emerge.

Specifically, the general similarity literature provides several distinct mechanisms by which perceived similarity might change as a function of tasks. We introduce these as a way of outlining how different tasks such as word recognition, inflection, or memory might influence the relevant similarities between words. Ultimately, of course, only empirical investigations of word similarity across different contexts will be able to resolve to what extent word similarities remain constant or change systematically with tasks. One aspect which is likely to be influential given the sequential nature of language is timing. Lamberts (1995, 1998) has captured time-course effects in the categorization of pictorial stimuli within a similarity-based model. Here, the underlying representation of the stimulus materials and consequently their perceived similarities are built up over time because individual component features are sampled by the perceptual system at differential rates. Furthermore, time course effects are also a by-product of the structural alignment process itself. Goldstone and Medin (1994) reported that the relative influence of matches-in-place and matches-out-of-place changed as a function of time. Early in the comparison process, locally consistent matches are more important than globally consistent matches; however, this relationship reverses over time. As a result, matches-out-of-place—which constitute matches between objects whose correspondence is not globally consistent—diminish in their influence on similarity. Given that language processing tasks differ considerably in the time scale on which they operate, time-dependent differences in similarity seem a distinct possibility and both the time course of representation formation and the time course of alignment would seem plausible factors in the context of language as well. For example, there is ample evidence in the context of word recognition, that speech input is evaluated against the lexicon as it unfolds (e.g. Marslen-Wilson, 1987). Obviously the relevant similarities with regards to competing lexical items will be those that pertain to

the word as it has so far been extracted, in direct analogy to Lamberts' investigations with visual objects. If identification of the sequence 'xylo...' as xylophone is possible long before the final syllable is reached, then that syllable will not partake in the relevant lexical competitions of similar sounding words. More generally, this raises the question of how task dependent our finding of greater coda sensitivity will be. One study that bears on these issues is the investigation of the time course of spoken word recognition by Allopena, Magnuson, and Tanenhaus (1998). These authors used eye movements as an index of activation for competitors that shared an onset, shared a coda or were phonologically unrelated. Of these competitors, the onset competitor reached a higher overall level of activation, but its activation also faded more quickly. The activation of the coda competitors was more enduring and eventually exceeded that of the onset. This pattern closely matched that predicted by TRACE (McClelland & Elman, 1986). The reason the overall peak of the onset match is higher is because the target word has already accumulated considerable evidence by the time the coda information comes in. This greater peak fits with findings of greater onset influence than coda influence in recognition under noise (e.g. Benki, 2003a). At the same time, Allopena et al.'s results are compatible with our present results, because coda matches become more important than onset matches over time.

Beyond time dependent effects, the general literature on similarity also contains other examples of systematic task-dependent changes. Differences in the salience of particular features and thus similarity were found by Nosofsky (1986) who succeeded in relating stimulus identification, recognition and classification of schematic visual stimuli within a single similarity-based model. While the information contained in the underlying stimulus representation remains constant across all three tasks, each task gives different weight to the various stimulus components in a way that facilitates the task at hand. The resultant differences in similarity across tasks are captured through differential attention weights which modify a single underlying spatial representation of similarity. Differences in feature weighting were also put forward by Tversky (1977) to explain asymmetries found in directional as opposed to non-directional similarity judgments (i.e. "how similar is a to b" versus "how similar are a and b"). Tversky further used differential weighting of common as opposed to distinctive features to explain differences in perceived similarity which emerged between tasks encouraging an emphasis on similarity and tasks encouraging an emphasis on difference. It is an intriguing possibility that analogous differences between tasks emphasizing commonality as opposed to difference might emerge in psycholinguistics. The dominant context in which word similarity is invoked within psycholinguistics is the influence of lexical neighbors, i.e. similar sounding words within our mental lexicon, on a particular process. Processes affected by lexical neighbors fall neatly into two categories depending on the specific effect that lexical neighbors have. In some, such as the naming of real words, lexical neighbors compete with the word being processed and have an inhibitory influence—the more lexical neighbors a particular target has the more processing is slowed (Luce, 1986). In other processes, such as judgments of sequence typicality, lexical neighbors have a facilitatory influence—the more neighbors, the greater the perceived typicality (Bailey & Hahn, 2001; Frisch, Large, & Pisoni, 2000). Tasks in which the influence of neighbors is competitive might naturally emphasize difference

whereas tasks in which neighbors make joint contributions might naturally emphasize commonalities. This might possibly give rise to corresponding shifts in word similarity between these two types of task that are analogous to Tversky's findings.

In conclusion, there is a wealth of findings and models relevant to task-dependencies in the general similarity literature. Viewing word similarity in this context suggests that task dependent shifts in word similarity are most likely where tasks differ substantially in time course or in the relative emphasis they give to commonality or difference.

This last issue of commonality versus difference, finally, raises an important fundamental question for any theory of similarity whatever the domain—a question with which we would like to conclude our discussion of issues for further research. Traditionally, some theories of similarity have been phrased in terms of commonality, others—such as psycholinguistic edit distance measures or spatial models of similarity—have been phrased in terms of difference. Yet other models, such as Tversky's (1977) contrast model and recent structural alignment accounts, combine elements of both: for Tversky, similarity is a function of both matching *and* mismatching features; structural alignment theories distinguish aligned and unaligned matches and mismatches. Furthermore, different strands of the structural alignment literature vary in the emphasis they place on commonalities or differences. Goldstone and Medin (e.g. 1994) speak of “matches in place” versus “matches out of place”, in other words aligned and unaligned matches, an approach that emphasizes commonality. By contrast, Markman and Gentner (e.g. 1996) speak of aligned versus unaligned differences, an approach that emphasizes difference. The two modes of presentation within structural alignment theory are not isomorphic and do not make identical predictions as can be seen from Table 3, basically because an unaligned difference can be either a “match out of place” or no match at all.

Table 3  
Counts of phoneme matches and differences for six syllables compared to pa

Comparison	Matches		Differences	
	Aligned	Unaligned	Aligned	Unaligned
<i>pa</i>	2			
<i>pat</i>	2			1
<i>pap</i>	2	1		1
<i>ta</i>	1		1	
<i>tat</i>	1		1	1
<i>tap</i>	1	1	1	1

Boxes identify pairs of syllables which are equally similar to pa in terms of the number of matches or the number of differences.

As is also apparent from Table 3, the critical items on which the two variants disagree arise naturally with sequential materials such as words. In the discussion of our experiments above we have variously spoken of matches or mismatches depending on which perspective was most natural. This reflects not only differences in ease of exposition, however, but an underlying theoretical tension within current theories of similarity that can only be resolved by further research. It can only be determined empirically whether the predictions of MIPS and MOPS or that of aligned vs. unaligned differences are confirmed, or whether, in fact, both must be combined. Word similarity seems an ideal testing ground for this endeavor.

## 10. Conclusion

Similarity between words is invoked in the explanation of a wide range of psycholinguistic tasks. A detailed understanding of performance on these tasks consequently requires an adequate understanding of what makes words sound similar. Crude measures of similarity will allow only crude predictions about the influence of similar sounding words such as predictions that rest on a coarse distinction between “high”- and “low”-similarity words. Accurate predictions about the processing of particular individual words are possible only once it is properly understood which factors determine word similarity and how the influence of these factors is combined.

This paper contributes towards this goal in two ways. First, our five experiments have provided new data to guide the design of more detailed models of word similarity, and we have improved the basic edit distance measure in light of our results. Secondly, and possibly more importantly in light of the fact that there are other aspects of word similarity that will require investigation, this paper shows that the general cognitive literature on similarity provides a productive theoretical framework for the study of word similarity. It does so both by demonstrating experimentally that central predictions of general theories of similarity are confirmed for words, and by outlining the way in which applying these theories to words raises a wealth of interesting issues for further research.

## Acknowledgements

The order of authors is arbitrary. Much of the research reported in this paper was conducted while the first author was at the University of Warwick and the second author at the University of Oxford. It was supported in part by a grant from the Biotechnology and Biological Sciences Research Council, UK to Kim Plunkett, Department of Experimental Psychology, Oxford, and by a British Academy grant to the first author. We are grateful to John Culling, Andy Brand and Christopher McGowan for their help in recording the materials for these experiments, and to several anonymous reviewers for helpful comments. We are also indebted to Daniel Lock who conducted an initial exploration of the issues in the context of a final year project at Warwick University.

## Appendix A. Stimuli and results by target-candidate set

Tables show percentage of responses choosing candidate *A* rather than *B* as being more similar to the target syllable. Items are arranged in decreasing order of response *A* percentage. *T*, *A*, and *B* signify Target, Candidate *A*, and Candidate *B*, resp.

### A. Featural differences

Onset				Coda			
T (C...C)	A (C <sup>1</sup> ...C)	B (C <sup>1,2</sup> ...C)	A	T (C...C)	A (C...C <sup>1</sup> )	B (C...C <sup>1,2</sup> )	A
plɪmp	flɪmp	slɪmp	90%	plɪp	plɪb	plɪg	90%
kʷɪlk	twɪlk	swɪlk	80%	brub	brug	bruk	85%
čoč	foč	θoč	75%	fɾʌf	fɾʌv	fɾʌm	85%
fɔʃf	pɔʃf	tɔʃf	75%	mum	mub	mug	85%
ʃɪnj	zɪnj	vɪnj	75%	θruθ	θruð	θruv	85%
klæsk	glæsk	blæsk	75%	klwɑ:k	klwɑ:p	klwɑ:f	80%
krusk	grusk	drusk	75%	bleb	blep	blef	80%
pʌsp	bʌsp	gʌsp	75%	čʌč	čʌʃ	čʌθ	80%
va <sup>w</sup> v	ba <sup>w</sup> v	ga <sup>w</sup> v	75%	jeʃ	ječ	jeʃ	70%
vɒv	fɒv	pɒv	65%	flelf	flelv	flelm	70%
brib	drib	trib	60%	feʃ	feč	fej	70%
fɾæf	pɾæf	tɾæf	60%	fæf	fæp	fæk	65%
mɪ <sup>ɹ</sup> m	bɪ <sup>ɹ</sup> m	dɪ <sup>ɹ</sup> m	60%	grʌg	grʌk	grʌp	65%
feʃʃ	fɛʃʃ	veʃʃ	55%	gwɛg	gwɛk	gwɛp	60%
čæč	ʃæč	fæč	50%	plup	plub	pluv	60%
kɾʌk	pɾʌk	bɾʌk	45%	vjuv	vjub	vjug	55%
mɪlm	bɪlm	pɪlm	40%	ʃiʃ	ʃiɜ	ʃiʃ	55%
pɪsp	kɪsp	gɪsp	40%	pep	pef	peθ	55%
pɾɒmp	bɾɒmp	dɾɒmp	40%	lel	len	lem	50%
lɪ <sup>ɹ</sup> l	zɪ <sup>ɹ</sup> l	vɪ <sup>ɹ</sup> l	25%	ʃʌʃ	ʃʌč	ʃʌʃ	50%

### B. Insert or replace a consonant

Onset				Coda			
T (...C)	A (CĈ...)	B (Ĉ...)	A	T (...C)	A (...ĈC)	B (...Ĉ)	A
fa:ɜ	fla:ɜ	la:ɜ	85%	zič	zinč	zim	90%
kug	kjug	jug	80%	smɔ:č	smɔ:nč	smɔ:n	90%
foð	froð	roð	80%	stɪp	stɪmp	stɪm	80%
denʃ	dwɛnʃ	wɛnʃ	75%	slʌk	slʌlk	slʌl	75%
kɒč	klɒč	lɒč	60%	ɜɪθ	ɜɪfθ	ɜɪf	75%
peʃč	pleʃč	leʃč	60%	vjeθ	vjeʃθ	vjeʃ	70%
dʌʃʃ	drʌʃʃ	rʌʃʃ	55%	pɾɪʃ	pɾɪʃʃ	pɾɪʃ	70%
ʃa <sup>w</sup> ʃ	ʃɾa <sup>w</sup> ʃ	ɾa <sup>w</sup> ʃ	55%	spʌʃ	spʌʃʃ	spʌʃ	70%

bib	blib	lib	50%	bæv	bælv	bæl	70%
geŋθ	greŋθ	reŋθ	50%	stap	stalp	stal	65%
tɪlm	trɪlm	rɪlm	50%	skwɪm	skwɪlm	skwɪl	60%
dɒsp	drɒsp	rɒsp	45%	spuf	spumf	spum	60%
tɪθ	trɪθ	rɪθ	45%	jʌʃ	ʃʌŋʃ	ʃʌn	55%
fɛʒ	fleʒ	leʒ	45%	θɔ:θ	θɔ:mθ	θɔ:m	55%
dæf	dræf	ræf	40%	kɪθ	kɪfθ	kɪf	55%
bɒŋʃ	brɒŋʃ	rɒŋʃ	35%	splɪn	splɪln	splɪl	55%
vɑ:f	vwa:f	wɑ:f	35%	blep	blemp	blɛm	50%
pæʃ	plæʃ	læʃ	35%	trɛp	trɛlp	trɛl	50%
θæv	θræv	ræv	30%	blev	blɛlv	blɛl	40%
vɛf	vjɛf	jɛf	20%	plɪn	plɪln	plɪl	40%

## C. Singleton-cluster featural differences

Onset				Coda			
T (C...)	A (C <sup>1</sup> Ĉ...)	B (C <sup>1,2</sup> Ĉ...)	A	T (...C)	A (...ĈC <sup>1</sup> )	B (...ĈC <sup>1,2</sup> )	A
penθ	brɛnθ	grɛnθ	85%	mæf	mælv	mælb	95%
fɪʒ	vjuʒ	mjuʒ	75%	vʌʃ	vʌmf	vʌmp	85%
koʃ	gloʃ	bloʃ	70%	glæf	glælv	glælb	80%
penč	brɛnč	grɛnč	70%	sɪʃ	sɪmf	sɪmp	75%
keč	greč	breč	60%	plʌb	plʌsp	plʌsk	70%
lub	njub	mjub	60%	wʌk	wʌmp	wʌmf	70%
mɑ:sp	blɑ:sp	glɑ:sp	60%	drɛŋ	drɛlm	drɛlv	70%
mɪv	brɪv	grɪv	60%	dɪθ	dɪmf	dɪmp	70%
vɪʃ	brɪʃ	grɪʃ	60%	skʌb	skʌsp	skʌsk	65%
tʊ³m	dju³m	bju³m	55%	krʌk	krʌmp	krʌmf	65%
væm	blæm	glæm	55%	plɒg	plɒsk	plɒsp	60%
mawj	blawj	glawj	50%	splʌk	splʌmp	splʌmf	60%
zun	vjun	bjun	50%	pjæp	pjælb	pjælv	60%
zʊk	vjuk	mjuk	50%	ðæp	ðælb	ðælv	60%
vɑ:sk	blɑ:sk	glɑ:sk	45%	spɹɛf	spɹɛlv	spɹɛlm	50%
zuʒ	djuʒ	gjuʒ	45%	drɪg	drɪsk	drɪsp	50%
duč	bjuč	mjuč	40%	kwɛŋ	kwɛlm	kwɛlv	50%
kɛlv	grɛlv	brɛlv	40%	fʌθ	fʌmf	fʌmp	50%
nʊʃ	djuʃ	gjuʃ	40%	ʒɪg	ʒɪsk	ʒɪsp	45%
kɛpθ	glɛpθ	blɛpθ	30%	vwa:b	vwa:sp	vwa:sk	30%

## D. Clustered vs. distributed matches

Onset				Coda			
T (C...C)	A (C...C <sup>1,2</sup> )	B (C <sup>1</sup> ...C <sup>2</sup> )	A	T (C...C)	A (C <sup>1,2</sup> ...C)	B (C <sup>1</sup> ...C <sup>2</sup> )	A
viv	vin	zim	95%	blælb	flælb	plælv	90%
gɒg	gɒm	bɒŋ	75%	ʃiʃ	ziʃ	siʒ	90%
bra:b	bra:n	dra:m	70%	flɪf	blɪf	plɪv	85%
pelp	pelv	belf	70%	geg	mæg	bɛŋ	85%
plep	plɛg	klɛb	60%	ðið	fið	viθ	80%
ða <sup>i</sup> ð	ða <sup>i</sup> f	va <sup>i</sup> θ	60%	grɪg	trɪg	dri:k	80%
vev	vep	bef	60%	ma:m	da:m	na:b	80%
fʌf	fʌb	pʌv	55%	nɛn	vɛn	zɛm	80%
prep	prev	bref	55%	plop	glop	klob	80%
pra <sup>i</sup> p	pra <sup>i</sup> v	bra <sup>i</sup> f	55%	ba <sup>i</sup> b	fa <sup>i</sup> b	pa <sup>i</sup> v	75%
pra:p	pra:v	bra:f	50%	fuf	zuf	suv	75%
klek	kleb	glep	50%	brælb	frælb	prælv	70%
brob	brof	prov	45%	fɛf	bɛf	pɛv	70%
fof	foð	voθ	40%	bɔ:b	fɔ:b	vɔ:p	65%
plæp	plæv	flæb	40%	brʌb	trʌb	drʌp	60%
blob	blof	plov	35%	frɒf	brɒf	prɒv	60%
va:v	va:p	fa:b	35%	kɹɛsk	bɹɛsk	ɡɹɛsp	60%
flɪf	flɪb	plɪv	30%	θeθ	veθ	ðɛf	60%
ʒæʒ	ʒæʃ	zæč	25%	krʌsk	brʌsk	ɡrʌsp	55%
čɛč	čɛʒ	ʃɛʒ	20%	θa <sup>w</sup> θ	va <sup>w</sup> θ	fa <sup>w</sup> ð	50%

## E. Onset vs. coda differences

T (C...C)	A (C <sup>1</sup> ...C)	B (C...C <sup>1</sup> )	A
blʌlb	plʌlb	blʌlp	85%
fla:f	pla:f	fla:p	80%
glug	blug	glub	80%
čɛč	ʒɛč	čɛʒ	75%
klɛsk	plesk	klɛsp	75%
ʃæʃ	fæʃ	ʃæf	75%
fra <sup>i</sup> f	θra <sup>i</sup> f	fra <sup>i</sup> θ	70%
čɛnč	ʒɛnč	čɛnʒ	65%
fɒf	ʃɒf	fɒʃ	60%
kluk	gluk	klug	60%
ðoð	voð	ðov	55%
θriθ	friθ	θrif	55%
brɔ:b	prɔ:b	brɔ:p	50%
grog	krog	grok	50%
pla <sup>i</sup> p	kla <sup>i</sup> p	pla <sup>i</sup> k	50%



prʌsp	krʌsp	prʌsk	50%
mɜ:m	nɜ:m	mɜ:n	45%
plɔ:p	klɔ:p	plɔ:k	35%
nɒn	mɒn	nɒm	30%
bob	gob	bog	25%

---

## Appendix B. Model details

The various PED and FED models (PED, PED+InsDel, PED+InsDel+Coda, and FED+InsDel+Coda) are special cases of a general edit distance model. In comparing two words, each difference (phoneme substitution, insertion, or deletion) is assigned a cost:

$$\text{Cost of Diff} = \begin{cases} \text{SubCost} & \text{For a phoneme substitution} \\ \text{InsDelCost} & \text{For an insertion or deletion} \end{cases}$$

For the various PED models, SubCost is simply set to 1 for all phoneme substitutions. For FED+InsDel+Coda, the cost of a substitution (SubCost) is determined by counting the number of major class features by which two aligned phonemes differ. The cost of an insertion or deletion (InsDelCost) is set to 1 in the simplest model (PED itself). In the other models it is a free parameter that can take on any non-negative value.

After the cost for each difference is determined individually, the costs are added together. In the general model, costs for onset and coda differences are added up separately, and multiplied by a syllable part weighting factor. The weighted syllable part costs are then added together to compute the total edit distance:

$$\begin{aligned} \text{Edit Distance} = & (1 - \text{CodaWeight}) \left( \sum_{\text{Onset Diffs}} \text{Cost of Diff} \right) \\ & + \text{CodaWeight} \left( \sum_{\text{Coda Diffs}} \text{Cost of Diff} \right). \end{aligned}$$

This edit distance is used with Luce's choice rule (Eq. (1) in the main text) to compute predicted response probabilities. In the PED and PED+InsDel models, the coda weighting parameter (CodaWeight) is set to 0.5, so that onsets and codas are weighted equally. In the PED+InsDel+Coda and FED+InsDel+Coda models, CodaWeight is a free parameter that can take on any value between 0 and 1.

The PLASS model includes five parameters that determine the cost of a feature change ( $F$ ), the cost of a phoneme insertion or deletion ( $I$ ), and the exponents  $P_{\text{Onset}}$ ,  $P_{\text{Coda}}$ , and  $P_{\text{Word}}$ . In practice, it is better to use a more orthogonal set of parameters that allow regressions to converge efficiently. For convenience, in addition to  $F$ , we define four

additional basic parameters

$$I' = \frac{I}{F}, \quad P' = \frac{P_{\text{Onset}} + P_{\text{Coda}}}{2}, \quad \Delta = \frac{P_{\text{Onset}} - P_{\text{Coda}}}{2}, \quad \text{and} \quad P'' = P_{\text{Word}} \cdot P'.$$

In terms of these parameters, the cost of a phoneme substitution is  $F$  times the number of featural differences, the cost of an insertion or deletion is  $I = F \cdot I'$ , and the PLASS regression model is:

$$\text{Edit Distance} = \left[ \left( \sum_{\text{Onset Diffs}} \text{Cost of Diff} \right)^{P' - \Delta} + \left( \sum_{\text{Coda Diffs}} \text{Cost of Diff} \right)^{P' + \Delta} \right]^{P''/P'}.$$

The best-fit values of the basic parameters were  $F = 1.71$ ,  $I' = 1.31$ ,  $P' = 0.54$ ,  $\Delta = 0.22$ , and  $P'' = 0.83$ . These give derived parameter values of  $I = 2.25$ ,  $P_{\text{Onset}} = 0.31$ ,  $P_{\text{Coda}} = 0.76$ , and  $P_{\text{Word}} = 1.55$ .

A possible alternative to the power law transformation to produce differences between onsets and codas in Experiments 1A–C would be an exponential transformation (e.g.  $e^{-Ax}$ ) applied to something like the FED + InsDel + Coda model (Bailey & Hahn, 2001; Hahn & Nakisa, 2000). However, this transformation does not produce magnitude estimation effects. Instead, it would predict that participants' sensitivity to a one-feature change versus a two-feature change would be the same regardless of the overall similarity of the words under comparison. Our intuition is that people would be most sensitive to the contrast between one and two featural differences when there are no other differences between the words under comparison, and that people may not detect the contrast at all when the words involved are quite distinct (e.g. *FLASK-PIT* vs. *FLASK-BIT*). We would expect sensitivity to fall off as a function of overall difference, along the lines of Weber's law, and the results of Experiments 1A and 1C are consistent with this. The power law transformation predicts these results (cf. Luce, 1959, p. 45); the exponential transformation does not.

## References

- Allopenna, P. D., Magnuson, J. S., & Tanenhaus, M. K. (1998). Tracking the time course of spoken word recognition using eye movements: Evidence for continuous mapping models. *Journal of Memory and Language*, 38, 419–439.
- Baddeley, A. D. (1966). Short term memory for word sequences as a function of acoustic, semantic and formal similarity. *Quarterly Journal of Experimental Psychology*, 18, 362–365.
- Bailey, T. M., & Hahn, U. (2001). Determinants of wordlikeness: Phonotactics or lexical neighborhoods? *Journal of Memory and Language*, 44, 568–591.
- Benki, J. R. (2003a). Quantitative evaluation of lexical status, word frequency, and neighborhood density as context effects in spoken word recognition. *Journal of the Acoustical Society of America*, 113, 1689–1705.
- Benki, J. R. (2003b). Analysis of English nonsense syllable recognition in noise. *Phonetica*, 60, 129–157.
- Bybee, J. L., & Slobin, D. I. (1982). Rules and schemas in the development and use of the English past tense. *Language*, 58, 265–289.

- Chomsky, N., & Halle, M. (1968). *The sound pattern of English*. New York: Harper and Row.
- Connine, C. M., Blasko, D. G., & Titone, D. (1993). Do the beginnings of spoken words have a special status in auditory word recognition? *Journal of Memory and Language*, 32, 193–210.
- Davies, G., Ellis, H., & Shepherd, J. (1977). Cue saliency in faces as assessed by the 'photofit' technique. *Perception*, 6, 263–269.
- Duncan, L. G., Seymour, P. H. K., & Hill, S. (1997). How important are rhyme and analogy in beginning reading? *Cognition*, 63, 171–208.
- Fay, D., & Cutler, A. (1977). Malapropisms and the structure of the mental lexicon. *Linguistic Inquiry*, 8, 505–520.
- Fodor, J. A. (1983). *The modularity of mind*. Cambridge, MA: MIT Press.
- Fowler, C. A., Treiman, R., & Gross, J. (1993). The structure of English syllables and polysyllables. *Journal of Memory and Language*, 32, 115–140.
- Frisch, S. (1996). *Similarity and frequency in phonology*. PhD thesis, Department of Linguistics, Northwestern University, Evanston Illinois, U.S. Available online as Rutgers Optimality Archive 198-0597: <http://roa.rutgers.edu/>
- Frisch, S., Large, N. R., & Pisoni, D. B. (2000). Perception of wordlikeness: Effects of segment probability and length on processing nonwords. *Journal of Memory and Language*, 42, 481–496.
- Fudge, E. (1987). Branching structure within the syllable. *Journal of Linguistics*, 23, 359–377.
- Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive Science*, 7, 155–170.
- Gentner, D. (1989). The mechanisms of analogical learning. In S. Vosniadou, & A. Ortony (Eds.), *Similarity and analogical reasoning* (pp. 199–241). New York: Cambridge University Press, 199–241.
- Gentner, D., & Markman, A. B. (1994). Structural alignment in comparison: No difference without similarity. *Psychological Science*, 5, 152–158.
- Goldinger, S. D., Luce, P. A., & Pisoni, D. B. (1989). Priming lexical neighbors of spoken words: Effects of competition and inhibition. *Journal of Memory and Language*, 28, 501–518.
- Goldstone, R. L. (1993). Feature distribution and biased estimation of visual displays. *Journal of Experimental Psychology: Human Perception and Performance*, 19, 564–579.
- Goldstone, R. L. (1994a). The role of similarity in categorization: Providing a groundwork. *Cognition*, 52, 125–157.
- Goldstone, R. L. (1994b). Similarity, interactive activation, and mapping. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 20, 3–28.
- Goldstone, R. L. (1996). Alignment-based nonmonotonicities in similarity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 22, 988–1001.
- Goldstone, R. L., & Medin, D. (1994). Similarity, interactive activation, and mapping: An overview. In K. J. Holyoak, & J. A. Barnden (Eds.), *Analogical connections*. Norwood, NJ: Abex.
- Goldstone, R. L., Medin, D. L., & Gentner, D. (1991). Relational similarity and the nonindependence of features in similarity judgements. *Cognitive Psychology*, 23, 222–262.
- Greenberg, J. H., & Jenkins, J. J. (1964). Studies in the Psychological Correlates of the Sound System of American English. *Word*, 20, 157–177.
- Hahn, U., & Nakisa, R. C. (2000). German inflection: Single route or dual route? *Cognitive Psychology*, 41, 313–360.
- Haig, N. (1985). How faces differ—A new comparative technique. *Perception*, 14, 601–615.
- Holyoak, K. J., & Thagard, P. (1989). Analogical mapping by constraint satisfaction. *Cognitive Science*, 13, 295–355.
- Hummel, J. E., & Holyoak, K. J. (1997). Distributed representations of structure: A theory of analogical access and mapping. *Psychological Review*, 104, 427–466.
- Jusczyk, P. W., Goodman, M. B., & Baumann, A. (1999). Nine-month-olds' attention to sound similarities in syllables. *Journal of Memory and Language*, 40, 62–82.
- Keane, M. T., Ledgeway, T., & Duff, S. (1994). Constraints on analogical mapping: A comparison of three models. *Cognitive Science*, 18, 387–438.
- Kenstowicz, M. (1994). *Phonology in generative grammar*. Cambridge, MA: Blackwell.
- Lamberts, K. (1995). Categorization under time pressure. *Journal of Experimental Psychology: General*, 124, 161–180.

- Lamberts, K. (1998). The time course of categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 24, 695–711.
- Landauer, T. K., & Streeter, L. A. (1973). Structural differences between common and rare words: Failure of equivalence assumptions for theories of word recognition. *Journal of Verbal Learning and Verbal Behavior*, 12, 119–131.
- Li, X. J., Schweickert, R., & Gandour, J. (2000). The phonological similarity effect in immediate recall: Positions of shared phonemes. *Memory and Cognition*, 28(7), 1116–1125.
- Luce, P. A. (1986). *Neighborhoods of words in the mental lexicon*. PhD Dissertation, Department of Psychology, Indiana University, Bloomington, Indiana.
- Luce, P. A., & Pisoni, D. B. (1998). Recognizing spoken words: The neighborhood activation model. *Ear and Hearing*, 19, 1–36.
- Luce, R. D. (1959). *Individual choice behavior: A theoretical analysis*. New York: Wiley.
- MacKay, D. G. (1972). The structure of words and syllables: Evidence from errors in speech. *Cognitive Psychology*, 3, 210–227.
- Markman, A. B. (1996). Structural alignment in similarity and difference judgment. *Psychonomic Bulletin and Review*, 3, 227–230.
- Markman, A. B. (2001). Structural alignment. In U. Hahn, & M. C. A. Ramscar (Eds.), *Similarity and categorization*. Oxford: Oxford University.
- Markman, A. B., & Gentner, D. (1993). Splitting the differences—A structural alignment view of similarity. *Journal of Memory and Language*, 32, 517–535.
- Markman, A. B., & Gentner, D. (1996). Commonalities and differences in similarity comparisons. *Memory and Cognition*, 24, 235–249.
- Marslen-Wilson, W. (1987). Functional parallelism in spoken word recognition. *Cognition*, 25, 71–102.
- Marslen-Wilson, W., Moss, H. E., & Van Halen, S. (1996). Perceptual distance and competition in lexical access. *Journal of Experimental Psychology: Human Perception and Performance*, 22, 1376–1392.
- Marslen-Wilson, W., & Warren, P. (1994). Levels of perceptual representation and process in lexical access: Words, phonemes and features. *Psychological Review*, 101, 653–675.
- McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology*, 18, 1–86.
- McMurray, B., Tanenhaus, M. K., & Aslin, R. N. (2002). Gradient effects of within-category phonetic variation on lexical access. *Cognition*, 86, B33–B42.
- McQueen, J. M., Norris, D., & Cutler, A. (1999). Lexical influences in phonetic decision making: Evidence from subcategorical mismatches. *Journal of Experimental Psychology: Human Perception and Performance*, 25, 1363–1389.
- Metsala, J. L. (1997). An examination of word frequency and neighborhood density in the development of spoken-word recognition. *Memory and Cognition*, 25, 47–56.
- Mueller, S. T., Seymour, T. L., Kieras, D. E., & Meyer, D. E. (2003). Theoretical implications of articulatory duration, phonological similarity, and phonological complexity in verbal working memory. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 29, 1353–1380.
- Nelson, D. L., & Nelson, L. D. (1970). Rated acoustic (articulatory) similarity for word pairs varying in number and ordinal position of common letters. *Psychonomic Science*, 19, 81–82.
- Nosofsky, R. M. (1986). Attention, similarity, and the identification–categorization relationship. *Journal of Experimental Psychology: General*, 115, 39–57.
- Prasada, S., & Pinker, S. (1993). Generalization of regular and irregular morphological patterns. *Language and Cognitive Processes*, 8, 1–56.
- Raaijmakers, J. G. W., Schrijnemakers, J. M. C., & Gremmen, F. (1999). How to deal with “The language-as-fixed-effect fallacy”: Common misconceptions and alternative solutions. *Journal of Memory and Language*, 41, 416–426.
- Sendmeier, W. F. (1987). Auditive judgments of word similarity. *Zeitschrift für Phonetik Sprachwissenschaft und Kommunikationsforschung*, 40, 538–546.
- Sevald, C. A., & Dell, G. S. (1994). The sequential cuing effect in speech production. *Cognition*, 53, 91–127.
- Shattuck-Hufnagel, S., & Klatt, D. (1979). The limited use of distinctive features and markedness in speech production: Evidence from speech error data. *Journal of Verbal Learning and Verbal Behavior*, 18, 41–55.

- Slowiaczek, L. M., & Pisoni, D. B. (1986). Effects of phonological similarity on priming in auditory lexical decision. *Memory and Cognition*, 14, 230–237.
- Smits, R., Warner, N., McQueen, J. M., & Cutler, A. (2003). Unfolding of phonetic information over time: A database of Dutch phoneme perception. *Journal of the Acoustical Society of America*, 113, 563–574.
- Stemberger, J. (1991). Apparent anti-frequency effects in language production: The addition bias and phonological underspecification. *Journal of Memory and Language*, 20, 41–55.
- Stemberger, J. P. (1990). Wordshape errors in language production. *Cognition*, 35, 123–157.
- Stemberger, J. P., & Treiman, R. (1986). The internal structure of word—initial consonant clusters. *Journal of Memory and Language*, 25, 163–180.
- Sussman, H. M., Bussell, N., Dalston, E., & Majors, T. (1997). An investigation of stop place of articulation as a function of syllable position: A locus equation perspective. *Journal of the Acoustical Society of America*, 101, 2826–2838.
- Treiman, R. (1983). The structure of spoken syllables. *Cognition*, 15, 49–74.
- Tversky, A. (1977). Features of similarity. *Psychological Review*, 84, 327–352.
- Tversky, A., & Gati, I. (1982). Similarity, separability, and the triangle inequality. *Psychological Review*, 89, 123–154.
- Vitevitch, M. S., & Luce, P. A. (1999). Probabilistic phonotactics and neighborhood activation in spoken word recognition. *Journal of Memory and Language*, 40, 374–408.
- Vitz, P. C., & Winkler, B. S. (1973). Predicting the judged “similarity of sound” of English words. *Journal of Verbal Learning and Verbal Behaviour*, 12, 373–388.
- Weber, E. H. (1835/1978). *Concerning touch* (H. E. Ross, Trans.). New York: Academic Press.
- Wicklegren, W. A. (1965). Distinctive features and errors in short-term memory for English vowels. *Journal of the Acoustical Society of America*, 38, 583–588.
- Wicklegren, W. A. (1966). Distinctive features and errors in short-term memory for English consonants. *Journal of the Acoustical Society of America*, 39, 388–398.