Effects of Featural Similarity and Overlap Position on Lexical Confusions and Overt Similarity Judgments

Sarah C. Creel, Delphine Dahan, & Daniel Swingley

Department of Psychology
University of Pennsylvania, Philadelphia, PA, USA

creel@psych.upenn.edu

Abstract

Spoken word recognition involves selecting one word out of many competing candidate word-forms. However, there is disagreement as to what candidates actually compete with each other given a particular signal, though similarity is commonly invoked. We investigate whether the word similarity literature, drawing from explicit similarity judgments of word forms, can speak to this current state of confusion. Many word similarity models emphasize the primacy of featural overlap. With a single set of stimuli, we find that featural overlap and overlap position influence a non-time-pressured lexical learning task as well as explicit similarity judgments. The implications for appropriate metrics of similarity in word recognition and models of word similarity are discussed.

Index terms: spoken word recognition, featural similarity, confusability metric, artificial lexicon

1. Introduction

There is general agreement that the process of spoken word recognition involves selecting one word out of many competing candidate word-forms. There is considerably less agreement as to which candidates actually compete given a particular signal, though competing words are usually described as "similar" to the correct word. The Cohort model of word recognition [1] suggests that words are activated to the degree that they match the acoustic input at a given point in time. This tends to favor candidate words that match initial portions of the input (i.e. "cohorts"). Ambiguity in word recognition is dominated by competition among words that overlap at onset.

A largely separate literature considers word similarity [2,3,4] and finds strong effects of featural similarity. These featural similarity models are often based on explicit similarity ratings, not on tasks that explicitly tap lexical processing. For instance, in an XAB similarity judgment task, Hahn and Bailey [3] find greater sensitivity to featural differences in coda position than in onset position. If coda sensitivity is greater, then words differing by their codas are less similar than words differing by their onsets. This suggests that the greatest ambiguity or confusion would arise among words that overlap in their rime. Here we examine similarity in both word recognition and in similarity judgment, using stimuli of the sort typically used in the literature on explicit similarity judgment.

Word recognition and word similarity studies are often not parallel on a number of dimensions. First, it is rare for identical sets of stimuli to be used between word recognition studies and word similarity or confusability studies. Thus, differences in outcomes between similarity judgment and word recognition tasks (e.g. priming and lexical decision) may derive from the processing demands of the tasks, or merely from differences in the stimulus sets. Another disjunct between these two literatures is that, while similarity judgment tasks are not time-pressured, word recognition tasks regularly rely on time-pressured measures because it is otherwise difficult to achieve any variance in performance among the familiar words that serve as stimuli. Under time pressure, participants may respond (e.g. with a lexical decision) before they have finished processing the entire word. In eyetracking tasks without explicit time pressure, competition effects may derive primarily from fixations driven by initial portions of the speech signal. Either procedure could lead to an emphasis on word-initial information that overestimates its importance in day-to-day word recognition.

The present study is an attempt to evaluate the applicability of feature-based similarity models to the word recognition process. We use a single set of stimuli in both a lexical task and a similarity judgment task. Thus, we can compare similarity ratings and actual word recognition data for the same stimuli. Additionally, the lexical task we use imposes no time pressure. Therefore, any onset effects we see do not simply result inevitably from responses made based on partial (word-initial) information.

Participants learned a carefully constructed set of novel words as labels for unfamiliar pictures [5,6]. We then assessed which words were likely to be mistaken for one another in a picture verification task. Our dependent measure was the false alarm rate (incorrect "yes" responses to a mislabeled picture) to various types of similar words. Such a set of novel words balances the stimulus requirements for a measure of lexical access and a measure of word similarity, which is often appraised with nonsense words to avoid semantic or morphological interference [2]. Following the picture verification task, participants made similarity ratings of word pairs drawn from the set of learned words.

2. Method

2.1. Participants

Participants ($N = 17$) were native English speakers and were undergraduates at the University of Pennsylvania.

2.2. Stimuli

We constructed an artificial lexicon of 32 novel consonant-vowel-consonant (CVC) words (Table 1). Participants learned the words as labels for 32 novel black-and-white shapes, previously used in other word-learning experiments [6,7]. The lexicon was designed to include as many types of similarity relations between CVCs as possible. The words were constructed from a set of 8 onset consonants, 8 vowels, and 8 offset consonants.

Words		Onsets	Codas
dʒup dʒub	tʃup tʃub	*Similar*	*Similar*
gidʒ git	kidʒ kit	*Similar*	*Dissimilar*
bæf bæv	sæf sæv	*Dissimilar*	*Similar*
pɔɪtʃ zɔɪtʃ	pɔɪd zɔɪd	*Dissimilar*	*Dissimilar*

Table 1. Examples of the 32 stimuli used in the current experiment. The "wrong" labels that could occur for "bæf" are indicated with arrows. Four-word groups shared similar or dissimilar onsets and codas as indicated.

The two sets of consonants (p, b, dʒ, tʃ, s, z, g, k at onset; p, b, dʒ, tʃ, f, v, d, t at coda) were selected to maximize featural variation in both positions, while insuring a number of segments in each position differing by a single phonological feature (voicing). The vowels used were (u, ʊ, aʊ, i, æ, aɪ, ɔɪ, ʌ). In each position, each segment occurred in exactly 4 words.

Within the set of words to be learned, there were sets of eight that were phonologically related. In each set, four words had one vowel and four had a different vowel. The four words sharing a vowel (see Table 1) were constructed such that, defining word 1 (e.g. /bæf/) as the "target," word 2 shared the first two segments (a cohort competitor, /bæv/) and word 3 shared the last two segments (a rhyme competitor, /sæf/). Word 4 shared only the vowel (/sæv/). For each of the two coda consonants between cohorts (/bæf/, /bæv/), the coda consonants could differ by a single phonological feature (voicing) or multiple features. This was also true of the onset consonants in rhyme items (/bæf/, /sæf/). Thus, each word (/bæf/) had several potential competitors (words sharing some segments). The primary competitors of interest for the current study were the cohort and rhyme. Both overlap by two contiguous segments, and share a C and a V.

Recordings were made in a quiet room with an Edirol UA5 audio capture USB interface to a PC, by a female native English speaker from Pennsylvania who read from randomized word lists. The experimenter selected tokens that were the clearest in segmental content, free of noise artifacts, and uniform in prosodic quality.

2.3. Procedure

Participants completed three parts of the experiment: training, testing, and similarity rating, taking 30-40 min in total. In the first phase, participants learned names for 32 shapes over 512 trials. They were not informed that there would be a test later, but were told that this was the first part of the experiment and were requested to "pay attention as [they were] learning these words." On each trial, a picture appeared and simultaneously its label was played. The picture appeared slightly to the left of center on 50% of trials, and slightly to the right on the other 50%. The participant indicated, using the arrow keys on the number keypad (4 and 6), the side on which the picture had appeared. This left-right manipulation served merely to keep participants attentive. In each of 16 blocks of 32 trials, each picture-label combination appeared once. The order of the trials within a block was random. Across participants, four different random picture-word assignments were used.

In the second phase, participants completed 4 blocks of a picture verification task (128 trials). A trial consisted of the simultaneous presentation of a picture and a spoken CVC label. Participants were asked to respond "yes" when the label was the correct word for that picture, and "no" when the label was incorrect, again using the 4 (no) and 6 (yes) keys. The label was correct 25% of the time. On the incorrect trials, 1/3 presented the picture labeled with its cohort competitor, 1/3 presented the picture with its rhyme competitor, and 1/3 presented the picture with a phonologically unrelated item from the newly-learned lexicon. Trials were counterbalanced so that within each block of 32 trials, each picture only occurred once and each label only occurred once. Blocks occurred in different orders for different participants.

Finally, participants completed 2 blocks of a similarity-rating task (64 trials). Pairs of words were spoken with a 500 ms interstimulus interval. On these trials, a rating scale was present throughout to remind participants to rate each pair as follows: 0 if the two words were identical; 1 if the two words were very similar, ranging to 7 if the words were extremely dissimilar. These trials were identical to the first two blocks of trials the participant received, except that the picture-word pairing was replaced with a word-word pairing. For instance, a trial that during the picture verification task consisted of a picture learned as /bæf/ with the verbal label "sæf" would, in this similarity task, consist of the sequence of words "bæf sæf."

3. Results

The lexical learning data (Figure 1) are analyzed in terms of "yes" responses. Note that for correctly-labeled objects, "yes" is the right response, but for all other trials, it constitutes a false alarm. The rate of false alarms is assumed to reflect the degree to which the learned label and the presented lure label are lexically confusable, in effect mapping out a "tuning function" for word detection. Two participants were excluded for having "yes" rates no higher to correct than to unrelated items. In keeping with previous results (e.g. [6]), cohorts showed the highest false alarm rates, with rhymes lagging behind, and phonologically unrelated items showing a relatively low rate of false alarms (Figure 1). There were also effects of featural overlap in both cohort and rhyme trials.

These effects were confirmed with a 2-factor ANOVA with Featural Similarity (similar [voicing difference] vs. dissimilar [3 featural differences]) and Overlap Position (initial vs. final) as factors. There were main effects of Featural Similarity ($F_1(1,14) = 18.87$, $p = .0008$; $F_2(1,60) = 67.9$, $p < .0001$), with higher rates of "yes" when target and lure-label differed only by a single feature, and of Overlap Position ($F_1(1,14) = 32.48$, $p < .0001$; $F_2(1,60) = 49.75$, $p < .0001$) such that initial overlap (cohort similarity) led to higher rates of "yes" responses. There was no interaction ($F_1(1,14) = 1.78$, $p = .2$; $F_2(1,60) = 1.99$, $p = .16$).

Similarity data (Figure 2) were analyzed after a simple linear transformation (1-rating/7) so that the most similar items received the highest ratings, and the ratings ranged from 0 to 1. In the similarity data, we compared perceived similarity for pairs of words. Cohort words were judged more similar to one another than rhyme words were to one another, while one-feature-different ("similar") words were rated much more similar than multiple-feature-different ("dissimilar") words. It is also worth noting that the distribution of "yes" responses in the picture verification task is more continuously graded from condition to condition than the similarity data, which are essentially at ceiling and floor for identical and unrelated word pairs, and in a compact middle ground for cohort and rhyme pairs.

We confirmed this appraisal of the similarity ratings with a 2-factor ANOVA, again using Featural Similarity and Overlap Position as factors. There was an effect of Featural Similarity ($F_1(1,14) = 51.34$, $p < .0001$; $F_2(1,60) = 62.17$, $p < .0001$), with high-similarity words judged more similar than low-similarity words (.59 vs. 44), and a effect of Overlap Position ($F_1(1,14) = 5.8$, $p = .03$; $F_2(1,60) = 11.14$, $p = .002$), with cohorts being judged more similar than rhymes (.55 vs. .48). There was no interaction (Fs < 1).

4. Discussion

Using a single set of stimuli, we found intriguing effects of featural overlap and position of overlap in both a word recognition task and explicit similarity ratings. The most significant findings with respect to our goal of defining word similarity *as it functions in word recognition* are the effects of these factors in the lexical learning task. First, we found an effect of featural similarity: more featurally-similar words were more likely to be mistaken for one another. This supports feature-based, not just segment-based (e.g. the Shortcut Rule, [8,4]), computations of word similarity in word recognition models. We also found effects of position of overlap: words that overlapped initially were more likely to be mistaken for one another than words that differed initially and overlapped later.

The position-of-overlap effect, while typical of word recognition generally, cannot be dismissed as an artifact of attention only to early portions of the signal. Learners were under no explicit time pressure. Importantly, the rate of false alarms varied by degree of featural similarity in the final segment: if cohort-trial responses were based simply on the initial segment(s) of words, no effects of final-segment similarity would have emerged in the cohort trials. Thus, this initial overlap effect must be inherent to lexical processing and not an artifact of early responding.

The similarity judgment results are also tantalizing. Consonant with much of the work using word similarity judgments (e.g. [9,2,3]), we find strong effects of featural similarity. Surprisingly, and at odds with the similarity-judgment literature (and our own intuitions), we also saw effects of position of overlap. The results of [3] are particularly difficult to reconcile at first glance. The differences may be due to the fact that participants had learned meanings for all words tested, or the particular similarity judgment tasks used. Nonetheless, in the larger context of feature-based models of word similarity judgments, it is clear that featural mismatch alone is necessary but not sufficient to explain our lexical recognition data.

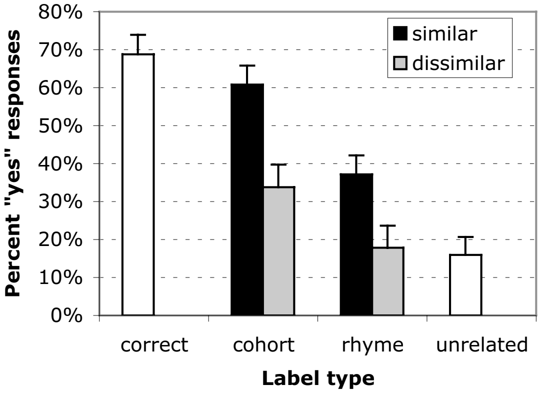

Figure 1 *"Yes" responses to pictures labeled with correct or incorrect CVCs.*

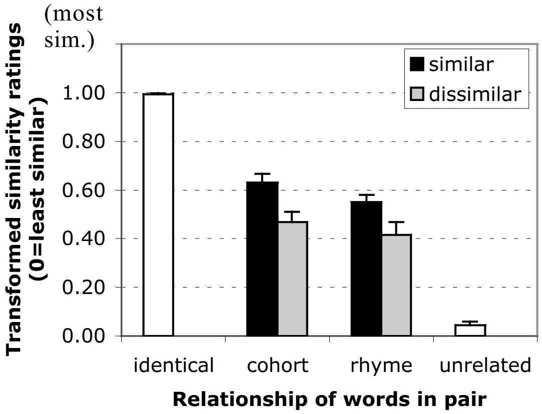

Figure 2 *Transformed similarity ratings for pairs of CVCs.*

Our work has interesting implications in both the word recognition and word similarity domains. In word recognition, our lexical learning data speak to an unresolved issue of lexical "neighborhood" relationships (what words form the competitor set of a spoken word). On the one hand, De Cara and Goswami [10] show that many words have large numbers of rhyme competitors, especially words with many neighbors. They argue that, given this information, many effects of neighborhood density may be carried by these rhyme neighbors. On the other hand, Vitevitch [11] has found evidence that cohort neighbors have a predominant influence on shadowing and lexical decision times. The current study suggests that they may both be right: while cohort competition is robust, high featural similarity may make for strong rhyme competitors. Lower featural similarity rhyme words, due to their lack of confusability, might be allowed to proliferate unchecked in the lexicon, contributing to [10]'s results.

The implications for word similarity judgments primarily concern the role of order information. In our task, order information played a role in similarity judgments, counter to featural similarity models that do not explicitly account for position of overlap. It may be that the learning task that preceded the similarity judgment task influenced similarity ratings. Continuing this line of reasoning, if lexical experience with a word can change its perceived similarity to other words, perhaps by reallocation of attention to various dimensions (e.g. [12], then explicit similarity judgments of nonwords may not be able to render the transient lexical similarity effects that populate the word recognition literature (or in our work, the lasting similarity effects; see also [6,7]).

5. Conclusion

Models of similarity judgments suggest that featural overlap predicts performance. In our lexical learning task, not only featural overlap but also position of overlap is a crucial factor in explaining word confusions. This is true despite the fact that our task imposes no time pressure. Moreover, using the same set of stimuli, we find that position of overlap may influence similarity judgments. This work demonstrates that the asymmetry between similarity judgments on the whole and word recognition performance merits continuing investigation.

6. Acknowledgements

Thanks to Sarah L. Johnstone for recording the stimuli. This work was supported by grant NSF 0433567 to DD and DS.

7. References

[1] Marslen-Wilson, W., "Functional parallelism in spoken word-recognition", Cognition, 25(1-2):71-102, 1987.

[2] Frisch, S. A., Large, N. R., and Pisoni, D. B., "Perception of wordlikeness: effects of segment probability and length on the processing of nonwords", Journal of Memory and Language, 42:481–496, 2000.

[3] Hahn, U., and Bailey, T. M., "What makes words sound similar?", Cognition, 97(3):227-267, 2005.

[4] Vitz, P. C., and Winkler, B. S., "Predicting the judged 'similarity of sound' of English words", Journal of Verbal Learning and Verbal Behavior, 12(4):373-388, 1973.

[5] Magnuson, J. S., Tanenhaus, M. K., Aslin, R. N., and Dahan, D., "The time course of spoken word learning and recognition: studies with artificial lexicons", Journal of Experimental Psychology: General, 132:202-227, 2003.

[6] Creel, S. C., Aslin, R. N., and Tanenhaus, M. K., "Acquiring an artificial lexicon: Segment type and order information in early lexical entries", Journal of Memory and Language, 54:1-19, 2006.

[7] Creel, S. C., Tanenhaus, M. T., and Aslin, R. N., "Consequences of lexical stress on learning an artificial lexicon", Journal of Experimental Psychology: Learning, Memory, and Cognition, 32:15-32, 2006.

[8] Luce, P. A., and Pisoni, D. B., "Recognizing spoken words: the neighborhood activation model", Ear and Hearing, 19:1–36, 1998.

[9] Bailey, T. M., & Hahn, U., "Phoneme similarity and confusability", Journal of Memory and Language, 52:339–362, 2005.

[10]De Cara, B., and Goswami, U., "Similarity relations among spoken words: the special status of rimes in English", Behavior Research Methods, Instruments, and Computers, 34(3):416-423, 2000.

[11] Vitevitch, M. S., "Influence of onset density on spoken word recognition", Journal of Experimental Psychology: Human Perception and Performance, 28(2):270-278, 2002.

[12] Nosofsky, R. M., "Attention, similarity, and the identification-categorization relationship", Journal of Experimental Psychology: General, 115:39-57, 1986.