Exploring systematicity between phonological and context-cooccurrence representations of the mental lexicon

Running Head: EXPLORING SYSTEMATICITY

Exploring Systematicity between Phonological and Context-Cooccurrence Representations of the Mental Lexicon

Monica Tamariz

Language Evolution and Computation Research Unit

School of Philosophy, Psychology and Language Science

The University of Edinburgh

monica@ling.ed.ac.uk

Abstract

This paper investigates the existence of systematicity between two similarity-based representations of the lexicon, one focusing on word-form and another one based on cooccurrence statistics in speech, which captures aspects of syntax and semantics. An analysis of the three most frequent form-homogeneous word groups in a Spanish speech corpus (cvcv, cvccv and cvcvcv words) supports the existence of systematicity: words that sound similar tend to occur in the same lexical contexts in speech. A lexicon that is highly systematic in this respect, however, may lead to confusion between similar-sounding words that appear in similar contexts. Exploring the impact of different phonological features on systematicity reveal that while some features (such as sharing consonants or the stress pattern) seem to underlie the measured systematicity, others (particularly, sharing the stressed vowel) oppose it, perhaps to help discriminate between words that systematicity may render ambiguous.

The mental lexicon is a complex structure internally organised along relationships of similarity and difference between lexical items. In the words of Saussure ([1916] 1983: 118), "a linguistic system is a series of phonetic differences matched with a series of conceptual differences". Priming studies show that words are organized in terms of their similarities and differences in phonological form (Goldinger, Luce & Pisoni, 1989; Luce, Pisoni & Goldinger, 1990), meaning (Meyer & Schevaneldt, 1971; Shelton & Martin, 1992), syntax (Bock, 1986), orthography (Humphreys, Evett & Quinlan, 1990) and even affective content (Wurm, Vakoch, Aycock, & Childers, 2003). Such comprehensive lexicon is at the core of construction grammar (Croft, 2001; Croft & Cruse, 2004), usage-based approaches (Langacker, 1990; Tomasello, 2003) and statistical language models including connectionist models (Elman, 1991), Data-Oriented Parsing (Bod, Scha & Sima'an, 2003), analogical models (Skousen, Lonsdale & Parkinson, 2002) and cooccurrence-based approaches (Landauer & Dumais, 1997; Lund & Burgess, 1996; Redington, Chater & Finch, 1998).

The lexicon of a language, then, can be represented as the set of difference, or similarity, values between every word pair in a lexicon. Similarity-based models of the lexicon are able to extract taxonomic information (Byrd, Calzolari, Chodorow, Klavans, Neff & Rizk 1987), form noun and verb taxonomies (Amsler & White, 1979), determine the grammatical category of words (Monaghan, Chater and Christiansen, 2005), create semantic networks (Alshawi, 1989), create semantic lexical hierarchies (Beckwith, Fellbaum, Gross, & Miller, 1991), reflect the acquisition of semantic

features (Guthrie, Slator, Wilks, & Bruce, 1990; Pustejovsky, 1991) and construct semantically coherent word-sense clusters (Slator, 1991; Wilks, Fass, Guo, McDonald, Plate, & Slator, 1993). Miikkulainen's (1997) unsupervised model DISLEX consists of orthographic, phonological and semantic feature maps. The geometry of each map and the interconnections between maps are configured by Hebbian learning and self-organization based on the cooccurrence of the lexical symbols and their meanings. Philips' (1999) connectionist mental lexicon, apart from lexical semantics, includes information about grammatical category, frequency and phonology. The Analogical Model of Language (AML) (Skousen, 1995; Skousen, Lonsdale & Parkinson, 2002), proposed as an alternative to connectionist language models, attempts to reflect how speakers determine linguistic behaviours. When speakers need to perform an operation on an unfamiliar word such as derive it or place stress on it, they access their mental lexicon and search for words that are similar to the word in question and then they apply the derivation or stress pattern found in words that are similar to the target word.

Describing the lexicon using similarity at different levels (phonology, semantics, syntax etc.) allows us to explore interactions between these levels. In this paper we investigate two main hypotheses, namely (a) that there is a significant level of systematicity between phonological and semantic-syntactic aspects of the lexicon and (b) that, since systematicity may pose problems for communication by introducing ambiguity, its effects will be countered by other processes: we will look for traces of processes supporting both systematicity and discriminability in the structure of the lexicon.

Systematicity in language

The existence of systematicity between word forms and word use in the linguistic context presupposes a degree of intralinguistic determinism - given the distributional patterns in a word's use, there is a bias for its form to contribute to the overall lexicon systematicity, and vice versa, given a word's form, there is pressure for its use in context to be similar to that of similarly sounding words. Therefore form is not arbitrary, which brings up Saussure's arbitrariness of the sign principle. For Saussure ([1916] 1983) a linguistic sign is a sound pattern linked to a concept. He distinguished between two types of relationships that signs are involved in: signification, or the association between form and concept, and value, determined by the relationships among signs. Saussure proposed arbitrariness at the level of signification, but qualified it at the level of the value, where he sees associative and syntagmatic interdependences between signs "which combine to set a limit to arbitrariness" (ibid.: 131). Jespersen (1922: 397) also proposed a non-arbitrariness of the value of the sign in his defence of sound symbolism, the notion that sounds carry intrinsic meaning. Relatedly, Sapir (1929) and Firth (1935) felt that speech sounds do carry meaning, but they suggested their meaning was not inherent to them. Rather, it was a result of "phonetic habit", a tendency to give similar meanings to words with similar sounds, much in anticipation of Bergen's (2004) conclusion that phonaesthemes (frequent sound-meaning pairings like *gl-* in words relating to vision and light or *sn-* in words relating to mouth and nose in English), while not being constituent units that can participate in compositionality nor behaving exactly like morphemes, do play a role in the structural organization of the lexicon.

The relationship between linguistic structure and meaning is fundamentally systematic, as evidenced by the compositional relationships between syntax and grammatical meaning, with similar syntactic structures expressing similar relationships between concepts, or between morpho-phonology and meaning, with morphemes with similar phonology denoting similar word syntactic properties. It should not come as a surprise, then, that the relationships between word phonology and words syntax and semantics also show a degree of systematicity. This effect is nevertheless expected to be small, as many other conflicting constraints act on words' phonology, syntax and semantics, not least the need to make words that tend to occur in the same contexts in speech sound different from each other so that they can be easily distinguished. A degree of systematicity may be useful in language acquisition and comprehension, by allowing a person hearing a word for the first time to extract meaning information from either phonological or context cues and make inferences based on that information about the other domain.

Shillcock, Kirby, McDonald and Brew (2001) reported a small but significant level of systematicity between two similarity-based geometrical representations of a subset of the English lexicon (the 1733 most frequent monosyllabic, monomorphemic English words in the British National Corpus). They estimated the phonological and the semantic distance between all the possible word pairs. For the phonological distance they applied the Wagner-Fisher edit distance algorithm - the number of changes, including deletions and insertions, necessary to turn one word into the other (Wagner & Fisher, 1974) - using values for the distance between segments and assigning penalties for mismatches between segment features such as vowel/consonant, vowel length,

consonant voicing etc., and an extra penalty for deletions and insertions. For the semantic distance they followed Lund & Burgess' (1996) vector-space method and constructed a 500-dimension vector space based on lexical cooccurrences in the 100 million-word British National Corpus. The corpus was lemmatised to reduce vector sparseness and semantic distance was measured as 1- cosine of the angle between two word cooccurrence vectors. They obtained a correlation between phonological and semantic distances of Pearson's $r = 0.061$, which a Monte-Carlo analysis showed to be highly significant ($p < .001$, one-tailed).

Experiment 1: Measuring Systematicity in the Spanish Lexicon

We test the question of whether systematicity in the lexicon similar to that found in English by Shillcock et al.'s (2001) is also found in another language, namely, Spanish.

Materials

Our materials are extracted from a corpus of orthographically transcribed Spanish spontaneous speech (897,395 tokens; 38,847 types) (Marcos Marin, 1992). We use three word sets: all the cvcv, cvccv and cvcvcv phonetically transcribed words of frequency greater than or equal to 20 in the corpus. These were the three most frequent CV word structures in the corpus. The 252 cvcv word types account for 50,639 tokens, the 146 cvccv word types, for 23,423 tokens and the 148 cvcvcv word types, for 11,475 tokens. Together, they make up 9.5% of all the corpus tokens and 1.4% of the word types − Shillcock et al.'s (2001) 1,733 word types account for "almost two-thirds" of the tokens

in the spoken part of the BNC, but only 0.2% of the tokens in the whole BNC, where the lexical statistics were calculated, and 3% of the types in the spoken part of the BNC, but only 0.04% of the whole BNC. The absolute size of the corpora used also affects the number of words used, and we used the largest Spanish transcribed speech corpus available to the best of our knowledge.

Methods

Phonological similarity metric.

We measure phonological similarity between all possible word pairs within each word set by applying norms obtained from an empirical study, based on human similarity judgments, that measured the relative impact of different parameters such as sharing the initial consonant, the vowels, the stress position etc. on perceived word similarity (Tamariz, 2005). (This method was designed to quantify the contribution of individual parameters to overall perceived phonological similarity, which we need in Experiment 2 to determine how different parameters contribute to systematicity).

The norms were calculated separately for three word groups with different CV structure. An online form presented participants with cvcv, cvccv or cvcvcv pseudo-word orthographic triads like the one shown in Fig. 1 randomly ordered for each participant. Participants judged which of the two pseudo-words on the right was more similar to the one on the left. They were instructed to focus on how the stimulus pseudo-words would sound and all stimuli were perfectly orthographically transparent. The stimulus pseudo-words were matched to the word-types of similar CV structure in the corpus in the frequency of the consonants in the different positions and in the number of

phonological neighbours. In each triad, the two pseudo-words on the right were similar to each other, and different to the one on the left, except that each of the former shared one phonological parameter each with the latter. Table 1 shows the list of all phonological parameters probed. All the possible parameter combinations for cvcv, cvccv and cvcvcv words were presented. For each parameter combination, two triads using different pseudo-words were prepared.

The results were analyzed separately for each word group. For each pairwise comparison of parameters, the counts of responses in favour of each parameter (a, b) were used to calculate a weight $w = (a - b) / (a + b)$, expressing the confidence that one parameter was favoured, for example if all respondents preferred the same parameter, its weight is 1; if the responses were half and half, the comparison's weight is 0. The impact value of each parameter on word phonological similarity is the sum of the positive weights for that parameter with respect to all the other parameters. Normalized parameter values are shown in Table 2.

The similarity for a word pair is the sum of the values of the parameters that the two words share. For example, /mésa/ and /móno/ share the initial consonant and the stress on the first syllable, so using the parameter values in Table 2, their similarity value in the 'syntax' condition is 0.074 + 0.133 = 0.207. The similarity measures for all word-pairs in a group are the components of the phonological similarity matrix. Two such matrices are calculated for each word group, one including (syntax condition) and one excluding (no syntax condition) stress-related parameters. Stress is left in the syntax condition because it captures morphosyntactic information, particularly verb inflection, in Spanish

Context-cooccurrence similarity metric.

Context-cooccurrence similarity is used as an estimate of semantic and morphosyntactic similarity. Context-cooccurrence statistics are based on the idea that the meaning of a word is determined by the linguistic contexts in which it occurs. One such model is Landauer and Dumais' (1997) Latent Semantic Analysis (LSA) counted occurrences of target words in whole articles of an encyclopaedia, and constructed a matrix of rows representing word types by columns representing the articles in which the types appear. Each value corresponds to the number of times the word type occurs in the article. After reducing the dimensionality, they obtained a 300-dimension matrix representing a semantic space where the similarity between word types or between articles can be calculated. The LSA approach has been used to account for aspects of semantic similarity (Kintsch, 2001) and to perform complex tasks such as metaphor interpretation (Kintsch & Bowles, 2002), complex problem solving (Quesada, Kintsch & Gomez, 2001), automatic essay grading (Foltz, Laham & Landauer, 1999) and automatic tutoring (Kintsch, Steinhart, Stahl, Matthews & Lamb, 2000; Wiemer-Hastings, Wiemer-Hastings & Graesser, 1999). Simpler, computationally less expensive context space models have been used to categorise words syntactically (Christiansen & Monaghan, 2006; Daelemans, 1999), categorise words semantically (Curran, 2004; Levy, Bullinaria & Patel, 1998; McDonald, 2000) and model semantic and associative priming (Lund & Burgess, 1996; Lund, Burgess & Atchley, 1995; Lund, Burgess & Audet, 1996; McDonald, 2000; McDonald & Lowe, 1998). LSA has been found to be highly correlated with context space models (Yan, Li & Song, 2004).

We follow Lund & Burgess' (1996) method, which is computationally less expensive and is more appropriate to out speech corpus, which is not naturally divided into semantic units comparable to the encyclopaedia articles used in LSA. In our method, each target word is geometrically located by a vector whose components represent how often the target appear in the vicinity of each of a set of high-frequency context words in the Spanish speech corpus. The vicinity is defined by a 'window' of five words before or after the target word. Similarity between each word pair is calculated as the cosine of the angle between the two corresponding vectors; the similarity measures for all word-pairs in a group are the components of the context-cooccurrence similarity matrix.

Two such matrices are calculated for each word group using two different sets of context words: for the 'syntax' condition, the 394 words of frequency greater or equal to 200 in the corpus; for the 'no syntax' condition, the 320 content words remaining after removing function words - determiners, prepositions, conjunctions and auxiliary verbs *ser, estar* (be) and *haber* (have). We do not consider cooccurrence with function words in the no-syntax condition as it reflects much of a word's syntax (Finch & Chater, 1992; Mintz, 2003; Redington Chater & Finch, 1998).

Systematicity metric.

Systematicity between the phonological and context-cooccurrence matrices is measured with Fisher Divergence, a symmetric variant of Fisher information developed by Ellison and Kirby (2006) to measure the distance between languages with the aim of building a taxonomical classification of languages. We discarded Pearson's r because similarity values do not meet the required assumptions of data normality and independence. Fisher Divergence is designed to correlate distance or similarity matrices and therefore does

not require independent or normally distributed data, and it takes unitless probability distributions as input. Additionally, it measures the confusion probability for each word-pair, which can be interpreted as the probability that a word is mistaken for the other. This method returns a unitless value representing the divergence between the two matrices. The significance of this value is calculated using the Mantel test (Mantel, 1967; Legendre & Legendre, 1998), a type of Monte-Carlo analysis designed to calculate the significance of the systematicity between the two distance or similarity matrices. We calculate the correlation between 10,000 random permutations of the rows and columns of the phonological similarity matrix and the veridical context-cooccurrence similarity matrix (note that permutating the rows and columns has the same effect as scrambling the word pairs before calculating the pairwise phonological similarities). The Mantel test usually employs Pearson's *r* or Spearman's rank as correlation measures, but we use Fisher Divergence instead for the reasons given above, noting that the choice of correlation test does not affect the validity of Mantel's test.

Results and Discussion

Table 3 shows the systematicities measured between the phonological and context-cooccurrence similarity matrices described above for three different word sets (cvcv, cvccv and cvcvcv) in two conditions ('syntax' and 'no syntax') and their significance among 10,000 randomizations of the pairwise distances.

Systematicity is significant in longer words and in the 'syntax' condition, failing to reach statistical significance in cvcv and cvccv words in the 'no syntax' condition. These results extend those obtained by Shillcock et al. (2001) to Spanish data,

supporting the hypothesis that this systematicity is not restricted to English. The different results in the 'syntax' and 'no syntax' conditions confirm the expected boosting effect of syntactic cues on systematicity. Failure to reach systematicity in the cvcv group may be partially explained by the fact that the space of possible cvcv words is very densely populated in Spanish, which does not leave much room for structure to emerge; *p*-values below 0.1, however, may indicate that word phonology is not totally uncorrelated with word meaning.

Having provided new support for the existence of a significant degree of systematicity at least in the 'syntax' condition, our phonological similarity metric, together with the fact that we have tested three independent, form-homogeneous word groups, allows us to further investigate the differential contribution of parameters of word similarity (see Table 1) to the systematicity. If the main cause of systematicity is morphology, we should expect that word-ends should show higher levels of systematicity than word beginnings, as all morphology concentrates at the end of the Spanish words in our word groups (4% of words were noun or adjective pairs like 'baja/bajo' and 28% of words were verbs, whose last phoneme in cvcv and cvccv words or last three phonemes in cvcvcv words encode morphology). We designed a second study to explore whether all features of phonological similarity relate systematically to context-cooccurrence statistics to the same extent.

Experiment 2. The Phonological Correlates of Systematicity in the Lexicon

Systematicity, as we have measured it, implies that words that occur in similar contexts in speech tend to sound similar in the lexicon; this introduces an ambiguity that goes against the principle of least effort (Zipf, 1949) for hearers in their task of uniquely mapping a word form to its meaning. In our second study we investigate the hypothesis that a pressure opposed to systematicity and favouring the discriminability of words also impacts the structure of the lexicon, and predict a lexicon configuration that reflects a trade-off between the two pressures.

We focus on how different aspects of phonological word similarity contribute to systematicity with word context-cooccurrence statistics and address the following questions: Which parameters of phonological similarity do words tend to share (and tend not to share) when they share context-cooccurrence statistics? Is the empirically obtained set of phonological parameter values particularly good for the correlation? We examine the role of vowels, consonants and stress position within the word. We will test two hypotheses: (1) that the empirically obtained parameter configuration obtains a better systematicity than most randomly generated configurations because of the pressure towards systematicity in the lexicon, and (2) that some parameters of word phonology specifically respond to the pressure for systematicity between phonology and context-cooccurrence statistics, while other parameters may respond to different pressures.

Materials

Three parallel studies use the same three independent word groups as Experiment 1, which are tested in the ‘syntax’ and ‘no syntax’ conditions explained above.

Methods

For each set in each condition we perform a random search algorithm to calculate the systematicity between 2000 randomly generated phonological similarity spaces and the veridical context-cooccurrence similarity space. We calculate the impact of each parameter as the beta coefficients in the linear regression of the randomly generated phonological similarity parameter values with respect to the systematicity values obtained with them.

The random search algorithm comprises: (1) Generation of a set of random parameter values. (Random values were generated by a *perl* program independently for each parameter and the set was converted into a normal distribution, because Fisher Divergence is sensitive to the absolute value of the components in the matrices compared.) (2) Computation of the values in the phonological similarity matrix in a word set using the random parameter values. (3) Calculation of the systematicity between the random phonological similarity matrix and the veridical context-cooccurrence similarities matrix using Fisher Divergence. Steps 1 to 3 are repeated 2,000 times, and for each repetition the random parameter values are recorded, as is the Fisher Divergence obtained with them.

This results in a hyperspace whose dimensions are the parameters of phonological similarity. Each set of random parameter values represents a point in a phonological hyperspace which has an associated systematicity value (its Fisher Divergence).

The impact of each parameter on the systematicity is measured with multiple linear regression analysis. This tells us the extent to which each parameter predicts Fisher

Divergence. (Note that because high Fisher Divergence indicates low systematicity, we use the negative of the beta coefficients as the metric of each parameter's impact on systematicity.)

Additionally, we compare the systematicity obtained with the veridical, empirical parameter values with those obtained with random parameter values.

The results from Experiment 1 are compared with the results of the random search for each word group. A match between a parameter's impact value and empirical values points to a link between perceived word phonological similarity and word semantic and syntactic similarity.

Results

First, a multiple regression analysis explores whether systematicity is a function of phonological similarity parameters: R^2 values, as shown in Table 4, reflect the combined impact of all phonological similarity parameters on systematicity in the three word groups and the two conditions (all $p < .001$). The results indicate that overall, variance in systematicity is a function of the parameters of phonological similarity employed. Nonlinearities were explored: exponential functions were best fitted to consonant parameters and sigmoid functions to vowel parameters. The difference between nonlinear and linear function R^2 were negligible, so only the latter are considered here.

Second, we quantified the impact of each phonological similarity parameter on systematicity: Fig. 2 shows the beta coefficients (with the opposite sign, because high Disher Divergence represents low systematicity) for each parameter against

systematicity (removing outliers did not alter the results). For ease of identification, the bars in the graph are coded with different colours for consonant-related, vowel-related and stress-related parameters. Sharing more than one segment has a more positive (or less negative) impact on systematicity than sharing single segments – sharing more than one consonant (tc, tc13, tc12, tc23, 3c) has a more positive impact than sharing one consonant (c1, c2, c3); similarly, sharing more than one vowel (tv, tv13, tv12, tv23, 3v) has a more positive impact than sharing one consonant (v1, v2, v3); moreover, while sharing consonants and stress tend to have a positive impact on systematicity, vowel parameters and the stressed vowel on the penultimate syllable have a negative impact. These results are cross-validated across data-sets, as the parameter impact values are highly coherent across the three word groups: counterpart parameter impact values measured in different word groups correlate significantly for all word-group pairings in both conditions (Table 5; all $p < .01$). This indicates the robustness of the methodology and shows that the same phonological parameters have equivalent impact on systematicity in three independent subsets of the Spanish lexicon.

Third, we examined how well the empirically obtained values for the parameters of word similarity (Table 1) are adapted to the pressure for systematicity. Table 6 shows the systematicity values obtained with the empirical parameters (the same values shown in Table 3), and their significance, this time measured as the rank position of the veridical Fisher Divergence among the 2,000 Fisher Divergence values calculated with random phonological parameter sets. The veridical, empirical parameter configurations are outliers in the distribution of random configurations in all but one condition, which further supports the significance of the systematicity measured in Experiment 1. This

represents evidence for the hypothesis that word phonological parameters that support systematicity are more salient when judging similarity than parameters that do not support systematicity.

Discussion

We argued earlier that systematicity introduces ambiguity in communication, and we hypothesized a pressure against systematicity and for discriminability operating on lexical structure. The beta coefficients of the phonological similarity parameters obtained in the three independent word groups suggest that there are two classes of parameters of phonological word similarity with respect to systematicity in Spanish:

1. Systematic parameters: Individual and groups of consonants, stress position and the identity of the final stressed vowel all impact systematicity positively, indicating that words sharing these phonological traits also tend to have similar context-cooccurrence statistics. These parameters tend to be less salient in a word similarity detection task (the test where the empirical parameter values originated) than predicted by the systematicity-driven random search. They are also either closely linked to narrow niches of syntactic function (e.g. the final stressed vowel encoding verb tense and person in the 'syntax' condition) or offer many combinatorial possibilities (e.g. the consonants in a word), and these two factors could help drive systematicity between phonology and word cooccurrence: the links with syntactic function obviously so; the high combinatorial power better allowing systematic relationships between the phonological space and the multidimensional cooccurrence space. This leads us to conclude that the morphological information encoded in the final stressed vowel for

verbs, but not that encoded in the final vowel for nouns and adjectives, plays a role in the systematicity measured in our datasets.

2. Discriminating parameters. Vowel parameters and the identity of the penultimate-syllable stressed vowel tend to impact systematicity negatively, which means that words sharing these phonological traits tend to have different cooccurrence-based distributional statistics. These parameters are more perceptually salient than predicted by the systematicity-driven random search and allow few combinatorial possibilities - there are only 5 vowels in Spanish, but 18 consonants. Unlike systematic parameters, these discriminating parameters are not related to morphosyntactic function.

The behaviour of systematic and discriminating phonological parameters can be explained in functionalist terms (Newmeyer, 2004) as adaptations. While information processing principles would favour systematic mappings, in a highly systematic lexicon words that tend to occur in similar contexts in speech would also tend to sound similar. From this conflict emerges the pressure for a salient phonological difference between words in an otherwise systematic lexicon. While systematic parameters could be responding to the pressure for systematicity, we may argue that discriminating parameters have taken on the role of dispelling the ambiguities brought about by systematicity.

The results make a clear difference between the role of consonants and vowels with respect to systematicity in Spanish. In the results above (see Fig. 2) most vowels show a negative impact on systematicity and most consonants, a positive impact (the only consistent exception to the latter being the consonant cluster (second and third consonants) in cvccv words, which are strongly phonotactically constrained and

therefore have a low combinatory power). Several studies suggest vowels and consonants are processed separately, suggesting they might underlie different functions in language perception and production, and therefore play different roles in the structure of the lexicon. Cole, Yan, Mak, Fanty and Bailey (1996) carried out experiments with English speech where either consonants or vowels had been rendered incomprehensible and found that vowels are clearly more important for recognition than obstruent consonants. Boatman, Hall, Goldstein, Lesser and Gordon's (1997) experiments with implanted subdural electrodes showed that electrical interference at different brain sites could impair either consonant discrimination or vowel and tone discrimination. A study of two Italian-speaking aphasics with selective impaired processing of vowels and consonants, respectively, suggests that vowels and consonants are processed by different neural mechanisms (Caramazza, Chialant, Capazzo & Miceli, 2000). Monaghan and Shillcock's (2003) connectionist model of Caramazza et al.'s effect showed that separable processing of vowels and consonants can be an emergent effect of a divided processor operating on feature-based representations. In a study in Spanish, Perea and Lupker (2004) found that nonwords created by transposing two consonants of a target word primed the target word (e.g. *caniso* primed *casino*), but transposition of two vowels did not lead to priming (e.g. *anamil* did not prime *animal*). Perea and Lupker propose that these differences could arise at the sub-lexical phonological level, and mention that the transposition of two consonants preserves more of the sound of the original than the transposition of two vowels. Lian and Karslen (2004) tested the recall of consonant-vowel-consonant nonword lists in Norwegian. Consonant frame lists (*kal, kol, kul*) were recalled and recognised better than rime lists (*kal, mal, sal*), showing an

advantage of vowel variation over consonant variation in this kind of tasks. Consonant frame lists could be found in the isomorphic consonant-based dimension (a *k_l* cluster). It is then easy to memorise which of the few possible vowels (Norwegian has 11 vowels) were present. Together, these results suggest that vowels and consonants are processed separately and might contribute to lexicon structure in different ways.

Some studies support the hypothesis that some of our proposed 'discriminating' parameters, namely vowels, may be particularly important in Spanish word recognition. Ikeno et al. (2003) explain that when foreigners from different language backgrounds speak English, their foreign accent reflects their native language characteristics. For instance, Flege, Bohn and Jang (1997) report that Koreans - whose native language distinguishes between long and short vowels - exaggerate the long-short vowel distinction in English. Ikeno et al. (2003) report that Spanish speakers tend to use more full vowels and less *schwas* than native English speakers when speaking English, probably because reduction to schwa is does not occur in Spanish.

A number of studies further suggest that stress information is processed independently of segmental information. Cutler (1986) shows that, in English, stress distinctions between pairs such as trusty-trustee do not affect the outcome of lexical decision tasks; French speakers' judgement about nonword similarity is not affected by stress differences (Dupoux, Pallier, Sebastian-Galles, & Mehler, 1997). The effect in English is explained by the fact that word stress strongly correlates with segmental information – vowel quality – with most stressed vowels pronounced fully and most unstressed vowels reduced to schwa; therefore, stress information is redundant and speakers can rely on segmental information only. In French, all words are stressed on

the last syllable, so stress does not help differentiate between words and it is not attended to in similarity judgments. In Spanish, unlike in French, stress can be in any of the last three syllables of a word and, unlike in English, stress information cannot be predicted from segmental information. In Spanish and similar languages, prosody may help reduce the number of competitors in word recognition, i.e. the number of candidates activated given an acoustic input (see review in Cutler, Dahan & van Donselaar, 1997). Pallier, Cutler and Sebastian-Gallés (1997) compared the abilities of Spanish and Dutch speakers to separately process segmental and stress information with a classification task of cvcv words. Their results suggest that in these languages, segmental information cannot be processed independently of stress information. In Dutch, stress contrasts are usually accompanied by syllable weight contrasts, with stress falling on the strong syllable, but in Spanish, stress is independent of weight, with many cvcv words made up of two equal weight syllables. As expected, Pallier et al. (1997) found that segmental judgements are more affected by stress in Spanish than in Dutch.

All this constitutes evidence that, in Spanish, systematic parameters have links with syntactic function; that systematic parameters have higher combinatorial power than discriminating parameters; that different neural mechanisms may underlie processing of consonants (systematic) and vowels (discriminating); and finally, that discriminating parameters vowel identity and stress may be important for word recognition. This evidence supports the division of function, again in Spanish, between systematic parameters (help maintain systematicity, which in turns helps generalisation and inference) and discriminating parameters (help word recognition in a systematic lexicon) suggested by the results of the present study.

This exploratory study poses many exciting questions that could be answered by examining different languages, such as: Are there universal biases towards certain phonological parameters responding preferentially to systematicity and to word discriminability? Does the number of consonants and vowels in a language interact with these biases? What characteristics do systematic and discriminating parameters show cross-linguistically?

In conclusion, starting with the assumption that lexical items are represented at least in two ways – according to how they sound and according to their context-cooccurrence statistics in speech, the present studies support the existence of a systematic mapping between these two representations in Spanish, extending previous results in English, and suggest ways in which different aspects of the phonological representation have adapted to a trade-off between the pressure for systematicity in the lexicon and the opposite pressure for word discriminability.

Acknowledgments

The research reported in this paper was financially supported by EPSRC studentship award nr 00304518, ESRC Postdoctoral Fellowship R39681 and a Leverhulme Trust Early Career Fellowship.

References

Alshawi, H. (1989). Analysing the Dictionary Definitions. In B. Boguraev & T. Briscoe (Eds.), *Computational Lexicography for Natural Language Processing*. London: Longman.

Amsler R.A. and White, J. (1979). Development of a computational methodology for deriving natural language semantic structures via analysis of machine-readable dictionaries. *National Science Foundation technical report, MCS77-01315*.

Beckwith, R., Fellbaum, C., Gross, D. and Miller, G.A. (1991). WordNet: A Lexical Database Organized on Psycholinguistic Principles. In U. Zernik (Ed.), *Lexical Acquisition: Exploiting On-line Resources to Build a Lexicon.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Bergen, B. (2004). The psychological reality of phonasethemes. *Language*, 80(2): 290-311.

Boatman, D., Hall, C., Goldstein, M.H., Lesser, R. and Gordon, B. (1997). Neuroperceptual differences in consonant and vowel discrimination: As revealed by direct cortical electrical interference. *Cortex, 33(1)*, 83-98.

Bock, J.K. (1986). Syntactic persistence in language production. *Cognitive Psychology*, *18(3)*, 355-387.

Bod, R. Scha and K. Sima'an (Eds.), (2003). *Data-Oriented Parsing*. Chicago: University of Chicago Press.

Byrd R.J., Calzolari, N., Chodorow, M.S., Klavans, J.L., Neff, M.S. and Rizk, O.A. (1987). Tools and Methods for Computational Lexicology. *Computational Linguistics, 13(3-4)*, 219-240.

Caramazza A., Chialant D., Capasso R., Miceli G. (2000). Separable processing of consonants and vowels. *Nature, 403(6768)*, 428-430.

Christiansen, M.H. and Monaghan, P. (2006). Discovering verbs through multiple-cue integration. In K. Hirsh-Pasek & R.M. Golinkoff (Eds.), *Action Meets Word: How Children Learn Verbs*. Oxford: Oxford University Press.

Cole, R., Yan, Y., Mak, B., Fanty, M. and Bailey, T. (1996). The contribution of consonants versus vowels to word recognition in fluent speech. International Conference on Acoustics, Speech and Signal Processing, Atlanta, GA.

Croft, W. (2001). Radical Construction Grammar. Syntactic theory in typological perspective. Oxford: Oxford University Press.

Croft, W. and Cruse, D.A. (2004). *Cognitive linguistics.* Cambridge: Cambridge University Press.

Curran. J.R. (2004). *From distributional to semantic similarity*, PhD thesis. University of Edinburgh.

Cutler, A. (1986). Forbear is a homophone: lexical prosody does not constrain lexical access. *Language and Speech, 29*, 201-220.

Cutler, A., Dahan, D. and van Donselaar, W.A. (1997). Prosody in the comprehension of spoken language: A literature review. *Language and Speech, 40 (2)*, 141-202.

Daelemans, W. (1999). Machine learning approaches. In H. van Halteren (Ed.), *Syntactic wordclass tagging*. Dordrecht: Kluwer Academic Publishers.

Dupoux, E., Pallier, C. Sebastian-Galles, N. and Mehler, J. (1997). A destressing "deafness" in French? *Journal of Memory and Language, 36*. 406-421.

Ellison, M.E. and Kirby, S. (2006) Measuring language divergence by intra-lexical comparison. *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the AC*, 273-280.

Elman, J. L. (1991). Distributed representations, simple recurrent networks, and grammatical structure. *Machine Learning, 7*, 195-224.

Finch, S.P. and Chater, N. (1992). Bootstrapping syntactic categories. *Proceedings of the Fourteenth Annual Conference of the Cognitive Science Society*, 820-825.

Firth, J.R. (1935). The Use and Distribution of Certain English Sounds. *English Studies, 17*, 8-18.

Flege, J.E., Bohn, O.S. and Jang, S. (1997). Effects of experience on non-native speakers' production and perception of English vowels. *Journal of Phonetics, 25*, 427-470.

Foltz, P.W., Laham, D. and Landauer, T.K. (1999). The Intelligent Essay Assessor: Applications to educational technology. Interactive Multimedia Education. *Jounal of Computer-Enhanced Learning, 1(2).*

Goldinger, S.D., Luce, P.A. and Pisoni, D.B. (1989). Priming lexical neighbors of spoken words – effects of competition and inhibition. *Journal of Memory and Language, 28 (5)*, 501-518.

Guthrie, L., Slator, B., Wilks, Y. and Bruce, R. (1990). Is there content in Empty Heads? *Proceedings of the 13th International Conference of Computational Linguistics, 3*, 138-143.

Humphreys, G.W., Evett, L.J. and Quinlan, P.T. (1990). Orthographic processing in visual word identification. *Cognitive Psychology, 22(4)*, 517-560.

Ikeno, A., Pellom, B., Cer, D., Thornton, A., Brenier, J.M., Jurafsky, D., Ward, W. and Byrne, W. (2003). Issues in recognition of Spanish-accented spontaneous English, *Proceedings of the ISCA/IEEE Workshop on Spontaneous Speech Processing and Recognition,* Tokyo, Japan.

Jespersen, O. (1922). *Language: Its Nature, Development and* Origin. London: Allen and Unwin.

Kintsch, W. (2001). Predication. *Cognitive Science, 25*, 173-202.

Kintsch, E., Steinhart, D., Stahl, G., Matthews, C., Lamb, R. and the LSA Group. (2000). Developing summarization skills through the use of LSA-based feedback. *Interactive Learning Environments, 8(2)*, 87-109.

Kintsch, W. and Bowles, A. (2002). Metaphor comprehension:What makes a metaphor difficult to understand? *Metaphor and Symbol, 17*, 249-262.

Landauer, T. K. and Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review, 104*, 211-240.

Langacker, R.W. (1990). *Concept, Image, and Symbol: The Cognitive Basis of Grammar*. Berlin and New York: Mouton de Gruyter.

Legendre, P. and Legendre, L. (1998). *Numerical ecology* (2nd English Ed.). Elsevier.

Levy, J.P., Bullinaria, J.A. and Patel, M. (1998). Explorations in the derivation of semantic representations from word co-occurrence statistics. *South Pacific Journal of Psychology, 10(1)*, 99-111.

Lian, A. and Karlsen, P.J. (2004). Advantages and disadvantages of phonological similarity in serial recall and serial recognition of nonwords. *Memory and Cognition, 32(2)*, 223-234.

Luce, P.A., Pisoni, D.B., and Goldinger, S.D. (1990). Similarity neighborhoods of spoken words. In G. Altmann (Ed.), *Cognitive models of speech perception: Psycholinguistic and computational perspectives* (pp. 122-147). Cambridge, MA: MIT Press.

Lund, K. and Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavioral Research: Methods, Instruments and Computers, 28(2)*, 203-208.

Lund, K., Burgess, C. and Atchley, R. (1995). Semantic and associative priming in high-dimensional semantic space. *Proceedings of the 17th Annual Conference of the Cognitive Science Society*, 660-665.

Lund, K., Burgess, C. and Audet, C. (1996). Dissociating semantic and associative word relationships using high-dimensional semantic space. *Proceedings of the 18th Annual Conference of the Cognitive Science Society*, 603-608.

Mantel. N. (1967). The detection of disease clustering and a generalized regression approach. *Cancer Research, 27*, 209-220.

Marcos Marín, F. (1992). *Corpus oral de referencia del español*, Madrid: UAM.

McDonald, S. and Lowe, W. (1998). Modelling functional priming and the associative boost. *Proceedings of the 20th Annual Conference of the Cognitive Science Society*, 675-680

McDonald, S. (2000). *Environmental determinants of lexical processing effort*. PhD thesis. University of Edinburgh.

Meyer, D.E. and Schevaneldt, R.W. (1971). Facilitation in recognizing pairs of words – Evidence of a dependence between retrieval operations. *Journal of Experimental Psychology, 90(2)*. 227-234.

Miikkulainen, R. (1997). Dyslexic and category-specific aphasic impairments in a self-organizing feature map model of the lexicon. *Brain and language, 59 (2)*, 334-366.

Mintz, T.H. (2003). Frequent frames as a cue for grammatical categories in child directed speech. *Cognition, 90(1)*, 91-117.

Monaghan, P., Chater, N., and Christiansen, M.H. (2005). The differential contribution of phonological and distributional cues in grammatical categorisation. *Cognition, 96,* 143-182.

Monaghan, P. and Shillcock, R.C. (2003). Connectionist modelling of the separable processing of consonants and vowels. *Brain and Language, 86(1)*, 83-98.

Newmeyer. F.J. (2004). Cognitive and functional factors in the evolution of grammar. *European Review, 12,* 245-264.

Pallier, C., Cutler, A. and Sebastian-Gallés, N. (1997). Prosodic structure and phonetic processing: A cross-linguistic study. *Proceedings 5th European Conference on Speech Communication and Technology, 4*, 2131-2134.

Perea, M. and Lupker S.J. (2004). Can CANISO activate CASINO? Transposed-letter similarity effects with nonadjacent letter positions. *Journal of memory and language, 51(2)*, 231-246.

Phillips, B.S. (1999). The mental lexicon: Evidence from lexical diffusion. *Brain and language, (1-2)*, 104-109.

Pustejovsky, J. (1991). The generative lexicon. *Computational linguistics, 17(4)*, 409-441.

Quesada, J.F, Kintsch, W. and Gomez, E. (2001). A computational theory of complex problem solving using the vector space model (part I): Latent Semantic Analysis, through the path of thousands of ants. In J.J. Cañas (Ed.), *Proceedings of the 2001 Cognitive Research with Microworlds Meeting*, 117-131.

Redington, M., Chater, N. and Finch, S. (1998). Distributional information: A powerful cue for acquiring syntactic categories. *Cognitive Science, 22(4)*, 425-469.

Sapir, E. (1929). A Study in Phonetic Symbolism. *Journal of Experimental Psychology, 12*, 225-239

Saussure, F. [1916] 1983. *Course in General Linguistics*. London: Duckworth.

Shelton, J.R. and Martin, R.C. (1992). How semantic is automatic semantic priming? *Journal of Experimental Psychology: Learning Memory and Cognition, 18*, 1191-1210.

Shillcock, R.C., Kirby, McDonald, S. and Brew, C. (2001). Filled pauses and their status in the mental lexicon. *Proceedings of the 2001 Conference of Disfluency in Spontaneous Speech*, 53-56.

Skousen, R. (1995). Analogy: A non-rule alternative to neural networks. *Rivista di linguistica, 7*, 213-231.

Skousen, R., Lonsdale, D. and Parkinson, D.B. (Eds.) (2002). *Analogical Modeling: An exemplar-based approach to language*. Amsterdam: John Benjamins.

Slator, B.M. (1991). Using Context for Sense Preference. In U. Zernik (Ed.), *Lexical Acquisition: Exploiting On-line Resources to Build a Lexicon.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Tamariz, M. (2005). *Exploring the adaptive structure of the mental lexicon*. PhD thesis. The University of Edinburgh.

Tomasello, M. (2003). *Constructing a Language: A Usage-Based Theory of Language Acquisition*. Cambridge MA: Harvard University Press.

Wagner, R.A. and Fisher, M.J. (1974). The string-to-string correction problem. *Journal of the Association for Computing Machinery, 21(1)*, 168-173.

Wiemer-Hastings, P., Wiemer-Hastings, K. and Graesser, A.C. (1999). Improving an Intelligent Tutor's Comprehension of Students with Latent Semantic Analysis. In S.P. Lajoie & M. Vivet (Eds.), *Artifical Intelligence in Education*. 535-542.

Wilks, Y., Fass, D., Guo, C., McDonald, J., Plate, T. and Slator, B. (1993). Providing Machine Tractable Dictionary Tools. In J. Pustejovsky (Ed.), *Semantics and the Lexicon*. Cambridge, MA: MIT Press.

Wurm, L.H., Vakoch, D.A., Aycock, J. and Childers, R.R. (2003). Semantic effects in lexical access: Evidence from single-word naming. *Cognition and Emotion, 17(4)*, 547-565.

Yan, X., Li, X. and Song, D. (2004). A Correlation Analysis on LSA and HAL Semantic Space Models. In *Proceedings of International Symposium on Computational and Information Sciences (CIS'2004)*, LNCS 3314. 710-717.

Zipf, G.K. (1949) *Human behavior and the principle of least effort*. Cambridge, (Mass.): Addison-Wesley.

Table 1

Parameters of Phonological Similarity.

Class	*Parameters*	*Explanation*
Single segment	c1, c2, c3	Same initial, 2nd, 3rd consonant
	v1, v2, v3	Same 1st, 2nd, 3rd vowel
Multiple segment	c1c2, c1c3, c2c3, c1c2c3	Same consonant combinations
	v1v2, v1v3, v2v3, v1v2v3	Same vowel combinations
Syllable structure	str	Same syllabic structure (cvc-cv or cv-ccv) (in cvccv words only)
Stress	s1, s2, s3	Same stress (on 1st, 2nd, 3rd syllable)
	sv1, sv2, sv3	Same stressed vowel (in the 1st, 2nd, 3rd syllable).

Table 2

Empirically Obtained Values of the Parameters of Phonological Similarity for the Three Word Groups cvcv, cvccv and cvcvcv in the Two Conditions 'Syntax' and 'No Syntax'.

cvcvcv.stx		cvcvcv.nostx		cvccv.stx		cvccv.nostx		cvcv.stx		cvcv.nostx	
c1	0.025	c1	0.047	c1	0.053	c1	0.081	c1	0.074	c1	0.178
c2	0.017	c2	0.036	c2	0.023	c2	0.028	c2	0.007	c2	0.009
c3	0.032	c3	0.067	c3	0	c3	0	c1c2	0.021	c1c2	0.388
c1c2	0.049	c1c2	0.099	c1c3	0.083	c1c3	0.105	v1	0.032	v1	0.021
c1c3	0.064	c1c3	0.107	c2c3	0.07	c2c3	0.094	v2	0.195	v2	0.072
c2c3	0.056	c2c3	0.11	c1c2c3	0.151	c1c2c3	0.32	v1v2	0.188	v1v2	0.332
c1c2c3	0.087	c1c2c3	0.167	v1	0.053	v1	0.082	s1	0.133		
v1	0.005	v1	0.01	v2	0.069	v2	0.043	s2	0.06		
v2	0	v2	0	v1v2	0.132	v1v2	0.246	sv1	0.073		
v3	0.018	v3	0.03	s1	0.095			sv2	0.217		
v1v2	0.023	v1v2	0.041	s2	0.078						
v1v3	0.036	v1v3	0.064	sv1	0.031						
v2v3	0.047	v2v3	0.089	sv2	0.135						
v1v2v3	0.067	v1v2v3	0.133	str	0.027						
s1	0.075										
s2	0.067										
s3	0.077										
sv1	0.073										
sv2	0.079										
sv3	0.102										

Table 3

Fisher Divergence Values Obtained in Experiment 1 and their Significances.

Group	*Syntax*		*No Syntax*	
	FD	*sig (p)*	*FD*	*sig (p)*
cvcv	5.03	< 0.05	7.79	= 0.06
cvccv	2.18	< 0.001	3.69	= 0.09
cvcvcv	2.36	< 0.001	3.84	< 0.01

Table 4.

Multiple Linear Regression Adjusted R^2.

Group	'syntax'	'no syntax'
cvcv	.890	.804
cvccv	.874	.770
cvcvcv	.919	.899

Table 5

Consistency of Counterpart Parameter Values across Word-Groups.

R^2	'syntax'		R^2	'no syntax'	
	cvcv (10)	cvccv (14)		cvcv (6)	cvccv (10)
cvccv (14)	0.86		cvccv (10)	0.84	
cvcvcv (20)	0.90	0.94	cvcvcv (14)	0.95	0.90

Table 6

Fisher Divergence Values Obtained in Experiment 2 and their Significances.

Group	*Syntax*		*No Syntax*	
	FD	*sig (p)*	*FD*	*sig (p)*
cvcv	5.03	0.06	7.79	0.01
cvccv	2.18	0.001	3.69	0.006
cvcvcv	2.36	0.001	3.84	0.001

Figure captions

Figure 1. An example pseudo-word triad comparing the effect of sharing parameters 'third consonant' (shared by 1 and 2) and 'stressed vowel' (shared by 1 and 3) on perceived similarity.

Figure 2. Beta coefficients of the parameters of phonological similarity. Two conditions, 'syntax' and 'no syntax' are shown for cvcv, cvccv and cvcvcv words. White bars for consonant-related parameters; grey bars for vowel-related parameters; black bars for stress-related parameters; striped bar for structure-related parameter (in cvccv, syntax condition only). Unless otherwise stated, $p < .01$. For parameter code names, see Table 1 above.

Figure 1.

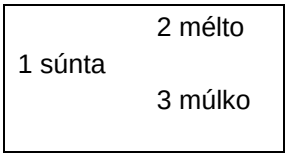
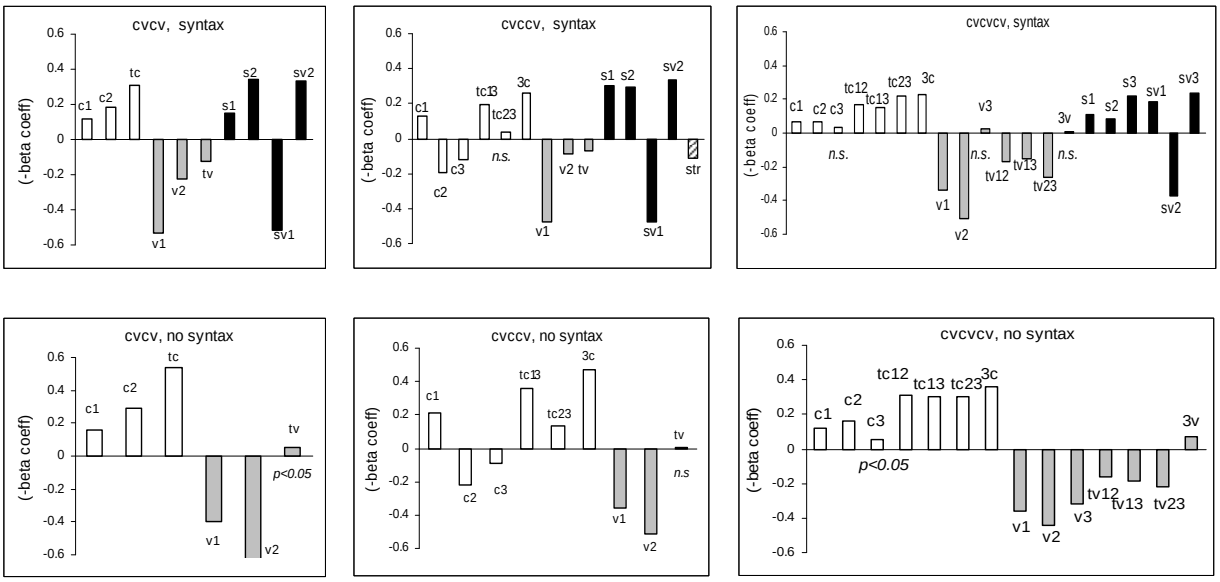


Figure 2.


cvcv, syntax
cvccv, syntax
cvcvcv, syntax
cvcv, no syntax
cvccv, no syntax
cvcvcv, no syntax
(-beta coeff)
p<0.05
n.s.
n.s