Configuring the Phonological Organization of the Mental Lexicon Using Syntactic and Semantic Information

Configuring the phonological organization of the mental lexicon using syntactic and semantic information

Monica Tamariz (monica@ling.ed.ac.uk)
Department of Linguistics, AFB, 40 George Square
Edinburgh EH8 9LL, UK

Abstract

This paper tests the hypothesis that the lexicon structure responds to two opposed pressures. First, the pressure for isomorphic representations means that words that occur in similar contexts tend to sound similar. Second, the pressure for disambiguation needs a way to distinguish the similar-sounding words that occur in similar contexts. This corpus-based study finds that while some aspects of the phonological organization of the lexicon respond to the first pressure, others respond to the second. The results presented here support the idea of a complex lexicon able to find solutions and adapt its structure to disparate, often conflicting pressures.

Introduction

The phonological structure of the monolingual mental lexicon has been studied with different methodologies based on lexical recognition (Cutler, Dahan & Van Donselaar, 1997), production (Van Son & Pols, 2003) or syntactic structure (Kelly 1996, Christiansen & Monaghan, in press). This paper presents a theory-independent corpus-based method that explores aspects of the phonological organization of the mental lexicon. This method assumes and is based on an isomorphism between a phonological and a cooccurrence-based representation of the mental lexicon, the latter capturing syntactic and semantic factors. This means that syntax and semantics, through their systematic relationship with phonology, play a part in the evaluation of the parameters that configure phonological organization.

Background: Isomorphism in the lexicon. In an analogical model of the lexicon words are defined in terms of their relationships to the rest of the words in the lexicon at different levels. For example, at the phonological level, words that sound similar are closer together than words that sound differently; in a cooccurrence based representation, words that tend to occur in similar contexts in speech are closer together than words that do not share context-words. Shillcock, Kirby, McDonald and Brew (2001) considered the 1733 most frequent monosyllabic, monomorphemic English words in the British National Corpus and calculated first the phonological distances between all the possible word-pairs. They first produced values for the distance between segments - they assigned penalties for mismatches between segment features such as vowel/consonant, vowel length, consonant voicing etc. For the calculation of each word pairwise 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 the mismatch penalties described above for the changes, and an extra penalty for deletions and insertions. For the cooccurrence distance they constructed a cooccurrence-based 500-dimension vector space based on the 100 million-word British National Corpus. They lemmatized the corpus to reduce vector sparseness and measured the cooccurrence distance as 1 - cosine of the angle between two word cooccurrence vectors. Finally, they obtained a measure of isomorphism as a correlation between the phonological and the cooccurrence distances of Pearson's $r = 0.061$, which a Monte-Carlo analysis showed to be highly significant ($p<0.001$, one-tailed).

Tamariz (2005) replicated these results for Spanish using partially overlapping methodology and measuring similarity values instead of distances. In this case, two subsets of the Spanish lexicon were used: all cvcv and all cvccv words of frequency larger or equal to 20 in a one-million word corpus of phonetically transcribed Spanish speech (Marcos Marin, 1992). For the calculation of the phonological space, identity in a number of parameters of word-form similarity was considered for each word-pair. In these homogeneous word groups, the parameters considered were sharing each of the consonants; sharing all the consonants; sharing each of the vowels; sharing all vowels; having the stress on the same syllable; and having the same stressed vowel in the same position. Each of these parameters was assigned a value according to its faring in a psycholinguistic study of their relative impact of word similarity judgments. For cooccurrence similarity, a vector space was constructed based on the corpus mentioned above. The isomorphism was measured with Fisher Divergence, an information-theoretical metric of the divergence between two similarity matrices. Isomorphism was calculated in two conditions: in the 'syntax' condition, the context words that provide the

dimensions of the cooccurrence vector space included high-frequency content and function words, and the metric of phonological similarity included stress parameters that capture morphosyntax. In the 'no syntax' condition, cooccurrence vector space included content words only and the phonological similarity metric excludes stress parameters. Monte-Carlo analyses showed significant isomorphism in most cases for cvcv and cvccv word groups ('syntax' condition: $p<0.05$ and $p<0.001$; 'no syntax' condition: $p<0.05$ and $p=0.09$, respectively).

An adaptive lexicon. The general principle underlying the methodology presented in this paper is that quantitative aspects of the mental lexicon structure can be explained in terms of the pressures that brought them about. The isomorphism measured by Shillcock et al. (2001) and by Tamariz (2005) may be responding to a pressure for structure-preserving representations originating in the nervous system, with clear examples in the visual, auditory, somatosensory and motor systems, where the structure of a stimulus is systematically preserved in its cortical representation. In the case of a complex stimulus, its different aspects are processed by different brain mechanisms, and all partial representations are isomorphic with each other (as well as with the stimulus). This brings about important processing advantages, such as allowing synthetic and analytical processes.

However, a highly isomorphic lexicon also presents one distinct disadvantage: words with similar meanings, used in similar contexts, would tend to sound similar. This poses a problem for communication: two words that sound the similar are usually distinguished by the context, but if their contexts are also similar (caused by the pressure for isomorphism), they will be easily confused.

In the rest of this paper I test the hypothesis that the lexicon organization reflects the pressure for isomorphism, but also the effects of an opposed pressure that works to differentiate the forms of words that tend to occur in similar contexts.

The impact of phonological parameters on phonological-cooccurrence isomorphism

The point of departure for this experiment is the test to determine the existence of isomorphism employed by Tamariz (2005) described above. The method employed is a random search of a parameter space, which indicates the behavior of dependent variables with respect to an independent variable. In this case, we are interested in the behavior of parameters of phonological similarity with respect to the isomorphism between the phonological and the cooccurrence spaces.

The algorithm works as follows: In each word group and each condition, I generate a random configuration of values of parameters of phonological similarity, which is used to calculate the phonological similarity in all the word pairs of a sample of the lexicon. I correlate (using Fisher divergence) these pairwise similarity values with the cooccurrence similarity values for the same word group. The novelty of this approach is that the *phonology-cooccurrence* isomorphism value is used to evaluate the *phonological* parameters - a high correlation indicates that the random parameter values tend to contribute to the isomorphism, and a low correlation indicates that the random parameter values tend to go against it.

Data. Random searches were performed in the 'syntax' and 'no syntax' conditions in three independent phonological spaces: those formed by cvcv, cvccv and cvcvcv words of frequency larger or equal to 20 from the corpus (252 words of structure cvcv, 146 cvccv words and 148 cvcvcv words).

Procedure. The random search is based on measuring the correlation (isomorphism) between many randomly generated phonological similarity spaces and the veridical cooccurrence similarity space. An analysis of the covariance of the random phonological parameter values with the correlation values will reveal which parameters are driving the correlation. The algorithm includes the following steps: (1) Generation of a set of random parameter values. (2) Computation of the phonological similarity values for all the word pairs using the random parameter values. (3) Calculation of the Fisher divergence (a measure of isomorphism) between the obtained pairwise phonological similarity values and the veridical cooccurrence similarities for the same word pairs. (4) Record the random parameter values and the Fisher divergence obtained with them. (5) 2,000 repetitions of steps 1 to 4.

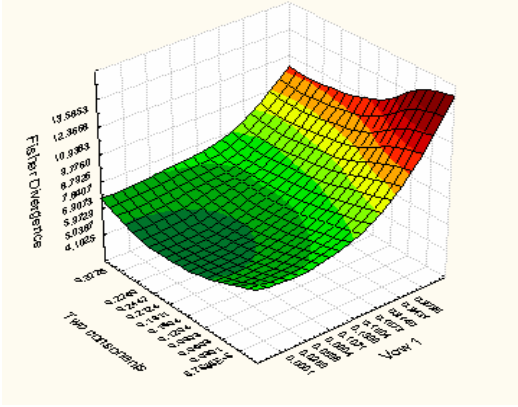

Figure 1: Illustration of the parameter space (cvcv words, 'syntax' condition) obtained with a random search. The surface is created by 2000 3-D points. The horizontal position of each point is given by the values of the phonological parameters 'first vowel' and 'two consonants'. The vertical position is the isomorphism value (Fisher divergence) of that point.

The result of the random search is a multidimensional hyperspace, the dimensions being the parameters of phonological similarity. Each set of random parameter values represents a point in the hyperspace. Each point has an associated isomorphism, value, given by the Fisher divergence. Figure 1 illustrates this hyperspace, showing just two of the dimensions involved in its configuration (*tc* and *v1*).

Analysis of the results. I consider three aspects of each phonological parameter: its linear covariance and the regression model that best fits its behavior with respect to the phonological-cooccurrence isomorphism, and its relationship with psychologically measured parameter values.

First, the linear covariance of each parameter with the Fisher divergence measures the effect of each parameter on isomorphism. A low Fisher divergence value indicates a high isomorphism, so I use the negative of Pearson's *r* as the measure of the linear covariance of the parameter with isomorphism. A positive covariance indicates that high values of the parameter in question improve the isomorphism measured; a negative covariance indicates that low value of the parameter improves the isomorphism; a covariance near zero indicates that the parameter does not greatly affect the isomorphism.

Second, I examine how different regression equations model the relationship between the phonological parameters and isomorphism. I run ten nonlinear standard regression functions on the data (logarithmic, inverse, quadratic, cubic, compound, power, sigmoid, growth, exponential and logistic) and obtain an r^2 value for each of them (r^2 =regression sum of squares/total sum of squares), a measure of how well each model fits the data. Similar performance across the independent word groups cvcv, cvccv and cvcvcv will cross validate the models.

Third, I compare the linear impact value of each parameter with empirically obtained values for the same parameters. The psychologically informed values come from a study reported in Tamariz (2005) which directly compared the salience of each parameter of phonological similarity in a word similarity judgment task. That study only included cvcv and cvccv words, so cvcvcv is not analyzed in this respect.

Results

Linear covariance. The linear covariance of the parameter values with isomorphism was obtained for the three word-groups in the two conditions. With minor variations, the linear impact values shown in Figure 2 apply to all groups and conditions.

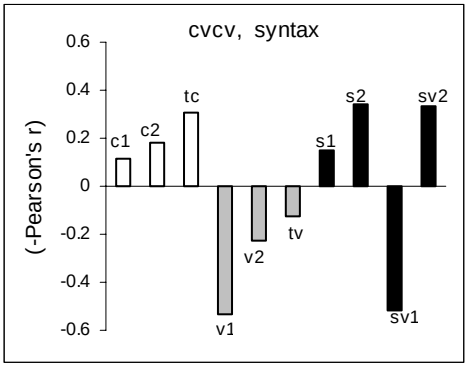

Figure 2. Linear impact parameter values for the cvcv word group in the 'syntax' condition (white = consonants; grey = vowels; black = stress).

The bars in Figure 2 show the covariance of each phonological parameter with the isomorphism between the phonology and the cooccurrence spaces. The results for all word groups and conditions were highly coherent (R^2 values ranging from .84 to .95), which indicates that each parameter is subserving the same function in three independent subsets of the lexicon. This has a double effect: while providing extra evidence for the existence of an isomorphism between the two levels of the lexicon at hand, the coherence of the results from independent data indicates the reliability of the methodology.

An analysis of the results of the linear covariance of parameters with isomorphism reveals several trends in all word groups. Figure 2 shows that consonant parameters (in white) have positive impact on isomorphism, while vowel parameters (in grey) have a negative impact. This is the case in all groups and conditions, except for a negative *c2* (the syllable-final consonant) and *c3* (the second-syllable initial consonant) in cvccv words; sharing all vowels in cvccv and cvcvcv words, and, the last vowel in cvcvcv words. The impact of the final vowel is much lower in the 'no syntax' condition than in the 'syntax' condition in all word groups. This may be explained by the fact that the final vowel carries in many instances morphosyntactic information: when correlated with syntax-laden cooccurrence representations, the phonological representations are more influenced by the weight of the last vowel.

Another similarity between all word groups is that sharing all consonants or vowels (*tc, tv*) tends to have greater impact on isomorphism (higher parameter values) than sharing single consonants or vowels (*c1, c2, c3, v1, v2, v3*). The only exception is sharing the final vowel (*v3*) in the 'syntax' condition in cvcvcv words, with a higher impact than sharing any combination of vowels; again, this may be explained by the morphosyntactic information encoded by the final vowel.

Another common feature of the 'syntax' condition across word groups is the high impact value of stress parameters in the last and one-but-last syllable. In the three word groups, sharing the stress on the same syllable brings two words close together in the phonological similarity space. Because the parameter impact value is so high, we know that words sharing the stress on the same syllable must be close together in the cooccurrence similarity space too. Sharing the same stressed vowel has very different effects depending on the syllable. The same stressed vowel on the final syllable (*sv2*) makes words very phonologically similar. The stressed final vowel encodes important verb morphosyntactic information in Spanish. The fact that the present methodology so clearly picks up the importance of parameter *sv2* in the phonological similarity space when correlated with a syntax-laden cooccurrence space both endorses the methodology and confirms the syntax-phonology correlation proposed by the phonological typicality literature (see review in Christiansen & Monaghan, in press).

Sharing the stressed vowel on the penultimate syllable has a very negative impact on isomorphism. Over 80% of Spanish bi- and trisyllablic words are stressed on this syllable, so the negative impact value indicates that sharing the same stressed vowel in the penultimate syllable (phonologically similar words) corresponds to dissimilarity in the cooccurrence space. This is going against the pressure for isomorphic representations, but may help an opposed pressure: the pressure for words to be easily distinguished from each other, particularly words that occur in similar contexts.

Nonlinear regression models. An examination of the r^2 obtained with the ten nonlinear regression equations (measuring how well models fit the data) reveals some connections with the linear impact parameter values: (1) The linear function is never the best predictor of the isomorphism given the parameter data, but it obtains its best r^2 values for negative-impact parameters. (2) The exponential, growth, logistic and compound functions return very similar r^2 values. (3) The best single predictor of negative-impact parameters (such as same vowel and same stressed vowel in the penultimate syllable) is the sigmoid or S-curve function. (4) The worst single predictor of negative-impact parameters is the inverse function. (5) The best compound predictor of the sign of the parameter impact is the sign of [exponential r^2 minus the S-curve r^2]. In positive impact parameters, exponential r^2 > S-curve r^2. In negative impact parameters, exponential r^2 < S-curve r^2.

Comparison with empirical values. The impact parameter values and the empirical values are significant in the cvccv group ($p<0.05$ in the 'no syntax' condition, $p<0.01$ in the 'syntax' condition), but not in the cvcv one. Figure 3 shows cvcv words in the 'syntax' condition.

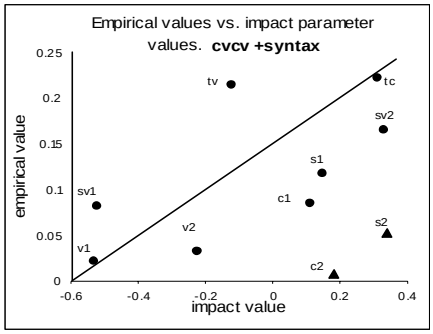

Figure 3. Scatter plot of the parameter impact values against the empirical values.

Most of the impact values correlate well with the empirical ones. In general, the corpus-based methodology yields higher consonant values (consonant parameters placed in the bottom-right half of the graphs), while the empirical method yields higher values for the vowels (vowel parameters in the top-left half of the graphs). This indicates that while consonants are more important in an isomorphic lexicon, people focus more on vowels when asked to tell how similar words sound.

Discussion

The analysis of the results suggest that there are two classes of parameters of phonological similarity with respect to the phonological-cooccurrence isomorphism.

Class one parameters: Individual and groups of consonants, the stressed syllable and the identity of the final stressed vowel all have positive impact values on isomorphism and are best modeled by an exponential function $Y = b0 * (e^{(b1*t)})$. Strong isomorphism (low Fisher divergence values) is brought about by high parameter values, indicating that words sharing these phonological traits tend to be close together in the cooccurrence-based space.

The pressure for isomorphism predicts a lexicon where class one parameters have high values. The fact that the empirical values are higher than predicted suggests that pressures other than that for isomorphism are affecting the lexicon.

Class two parameters. Individual vowels and the identity of the penultimate-syllable stressed vowel have negative impact values on isomorphism, possibly playing a role in word differentiation and disambiguation. They are best modeled by a sigmoid curve function $Y = e^{(b0 + (\frac{b1}{t}))}$ and also reasonably well modeled by a linear function Y = b0 + b1t. Strong isomorphism (low Fisher divergence values) is brought about by low parameter values, indicating

that words sharing these phonological traits tend to be far apart in the cooccurrence-based space.

The pressure for word differentiation predicts a lexicon where class two parameters have high values. The fact that the empirical values are lower than predicted suggests that the pressure for word differentiation is not the only one affecting the lexicon.

I propose that each of these two parameter classes is responding selectively to one of the two opposed pressures at hand. In Spanish, Class one parameters such as sharing consonants, stress and the final stressed vowel organize the lexicon in a systematic way leading to isomorphism between the cooccurrence-based and the phonological lexical representations. Class two parameters such as sharing vowels, particularly the stressed vowel in the penultimate syllable, differentiate words that would otherwise be easily confused.

Characteristics of Class one and Class two parameters may help explain why they have adopted these opposed roles in the organization of the lexicon.

Class one parameters are either closely linked to narrow niches of syntactic function (e.g. the final stressed vowel encoding verb tense and person) or offer many combinatorial possibilities (e.g. consonants: there are 18 consonants in Spanish compared with only five vowels). These two characteristics are desirable in parameters that drive isomorphism between phonology and word cooccurrence: the links with syntactic function obviously so; the high combinatorial power better allowing systematic relationships of phonological space with the multidimensional cooccurrence space.

Class two parameters allow fewer combinatorial possibilities (e.g. vowels) and may be related to word differentiation, the pressure opposed to isomorphism in the configuration of the lexicon structure. The fact that, in cvccv words, *c2* and *c3* are Class-two parameters supports the connection with combinatorial power: only seven consonants can occupy the syllable-coda position (*c2*) in Spanish, and the following consonant (*c3*) is constrained by *c2*. One way of determining the importance of the differential combinatorial power of vowels and consonants would be a cross-linguistic comparison of the result of this kind of study in languages with many and with few contrastive vowels.

Another differentiating factor between Class one and Class two parameters, particularly between consonants and vowels, is their neural substrate. Boatman, Hall, Goldstein, Lesser, and Gordon's (1997) experiments with patients with implanted subdural electrodes showed that electrical interference at different brain sites could impair consonant discrimination or vowel and tone discrimination. A study by Caramazza, Chialant, Capazzo & Miceli (2000) 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. In another study in Spanish, Perea and Lupker (2004) found that nonwords created by transposing two consonants of a target words primed the target word (e.g. *caniso* primed *casino*). However, when two vowels were transposing no priming occurred (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. These results, plus the supporting facts of the appearance of vowels as phonological units earlier in life than consonants (Bertoncini, Bijeljac-Babic, Jusczyk, Kennedy & Mehler, 1988) and the earlier spelling of vowels than of consonants in Spanish (Ferreiro & Teberosky, 1982), suggest that, at least in languages like Spanish and Italian, vowels and consonants are processed separately.

The 'phonological similarity effect' (PSE) (Conrad & Hull, 1964) says that when people are asked to recall a list of words, they perform worse if the words sound similar to each other. In a recent paper, Lian and Karslen (2004) tested the PSE of cvc nonwords with Norwegian participants, analyzing the impact of three parameters of phonological similarity - sharing middle vowels (*mal, sar, tas*), the consonant frames (*kal, kol, kul*) or the rhyme (*kal, mal, sal*) - with two tasks: recall and recognition of the words in the list. Their results bear on the differential processing of consonants and vowels. They found a reversal of PSE in several conditions. Sharing mid-vowels did not produce PSE, and sharing the consonants and sharing the rhyme actually reversed the PSE, that is to say, lists of words sharing the consonant frames and the rhyme were generally recalled and recognized better than distinct word lists. Most relevant to the present discussion is the fact that consonant frame lists (*kal, kol, kul*) were recalled and recognized better than rhyme lists (*kal, mal, sal*), showing an advantage of vowel variation over consonant variation in this kind of tasks. Consonant frame lists could be easily placed in a consonant-based phonological dimension of the lexicon (in the *k_l* position). It is then easy to memorize which of the few possible vowel (Norwegian has 11 vowels) were present.

A number of studies suggest that stress information is processed independently of segmental information. Cutler's (1986) results show that, in English, stress distinctions between pairs such as trusty-trustee do not affect the outcome of lexical decision tasks; French speakers' judgment about nonword similarity is not affected by stress differences either (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 speakers do not pay attention to it when judging similarity. In Spanish, stress information cannot be predicted from segmental information; therefore it should have an impact on similarity judgments.

These studies show evidence that Class one parameters have links with syntactic function and a higher combinatorial power than class two parameters and that different neural mechanisms may underlie processing of consonants (Class one) and vowels (Class two). This, together with the results in this paper, supports the division of function between class one parameters (maintain isomorphism, which in turns helps generalization and inference) and class two parameters (help word recognition in an isomorphic lexicon).

Conclusion

I have presented a new paradigm to study the phonological organization of the lexicon which takes into account other levels of lexical organization such as syntax and semantics, until recently considered to be independent from phonology. The present results support the existence of isomorphism in the lexicon, but also show evidence of the effects of an opposed force that tends to disambiguate word forms that, in a fully isomorphic lexicon, would be easily confused. More generally, this paper supports the idea that the lexicon is an adaptation to the set of pressures that act upon it.

Acknowledgments

This research has benefited from the support of EPSRC studentship award nr. 00304518.

References

Bertoncini, J., Bijeljac-Babic, R., Jusczyk, P.W., Kennedy, L., & Mehler, J. 1988. An investigation of young infants' perceptual representations of speech sounds. *Journal of Experimental Psychology: General*, 117: 21-33.

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

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

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

Conrad, R. & Hull, A.J. 1964. Information, acoustic confusion, and memory span. *British Journal of Psychology*, 55: 429-432.

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

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

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

Ferreiro, E. & Teberosky, A. 1982. *Literacy before schooling*. Portsmouth, NH: Heinemann.

Kelly, M.H. 1996. The role of phonology in grammatical category assignment. In J. Morgan & K. Demuth (eds.) *From signal to syntax*. Hillsdale, NJ: Lawrence Erlbaum Associates, 249-262.

Lian, A. & 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.

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

Perea, M. & 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.

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

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

Van Son, R.J.J.H. & Pols, L.C.W. 2003. Information structure and efficiency in speech production, *Proceedings of EUROSPEECH2003*, Geneva, Switzerland.

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