PARAMETERS OF WORD-FORM SIMILARITY IN SPANISH

Mónica Tamariz-Martel Mirêlis

monica@ling.ed.ac.uk

TAAL, The University of Edinburgh

ABSTRACT

Shillcock et al. found a significant positive correlation between the representational spaces of word-form and word-meaning in English. If this correlation is useful for word acquisition and comprehension, it could be adaptive to maximise it. We devise a method to measure word-form similarity that maximises the form-meaning correlation in Spanish. We show that the correlation is present in Spanish too and find out more about the organization of words in the morpho-phonological representational space.

1 INTRODUCTION

Our starting point is the finding by Shillcock et al. (2001) that there is a significant positive correlation between the how close words are to each other in terms of their meanings and in terms of their form (sounds) in English. They established that words that have similar meanings are likely to sound similar, and words that have very different meanings are likely to sound very different from each other. They took the 1,700 most frequent monomorphemic monosyllabic words in the spoken part of the BNC corpus and listed all the possible pairwise combination between them. Then they measured the semantic distance and the phonological distance between the two words in each pair and found a significant correlation coefficient Pearson's $r = 0.08$.

Their semantic distances are based on the context that words appear in. Two words are semantically close if they tend to occur surrounded by the same context words in speech, or how interchangeable they could be. Two words are semantically far apart if they tend not to appear surrounded by the same contexts.

We will assume that this correlation is useful, that it has evolved or been preserved in languages because it has a function. We hypothesize that the correlation could help word acquisition in children and new word comprehension in adults in the folowing way: when someone hears a new word, it immediately finds its place in the form representational space, because we perceive phonological similarities with other words, so we know which other words it is "form-similar" to. Because of the existence of this correlation, we know that its meaning is likely to be relatively close to the meaning of the words it is "form-similar" to, so we have a clue as to what it might mean.

Our next assumption is that if the correlation is useful, it might be adaptive to maximise it (given the many other constraints that restrict lexical form and meaning).

The aims of this study are (a) to try to establish whether the form-meaning correlation exists in Spanish and (b) to see what parameters of form similarity drive the correlation.

2 THE HILL-CLIMBING ALGORITHM

We use a hill-climbing algorithm that acts upon the parameters of word-form similarity in order to maximise the correlation between form-similarity and meaning-similarity in a set of word-pairs.

We will carry out the same study in three homogeneous groups of Spanish words extracted from a phonologically transcribed (citation transcription) orthographic Spanish speech corpus (Marcos Marin, 1992). The groups are cv-cv, cvccv and cv-cv-cv words. Table 1 gives some information on these word-groups.

	Nr. words	Nr. Pairs
CV-CV	229	52,226
CVCCV	174	15,051
CV-CV-CV	146	10,585

Table 1. The word-groups.

When we have out pair lists, we measure the semantic distances between the two words in each pair following Shillcock et al. (2001)'s method.

Then we proceed to construct the way we will measure form similarity: First we identify a comprehensive set of word-form similarity parameters (see examples in Table 2, complete set in Appendix 1), that is, factors that make two words sound more similar.

Stress	kAsa – mOto
Consonant1	mEsa – mUCo
Place of art 1	mirO – pasO
Cross vowel	kAmpo - nUnka
Syll. Structure	po·drE - kA·bra

Table 2. Some examples of parameters of form similarity.

So we have a list of word-pairs, with their respective meaning distances, and also for each pair we can list the parameters the two words share (see Table 3).

W1	W2	Meaning dist.	Form dist.		
			stress	*phonemes*	*features*
gAla	dICa	0.4801	+ S	+ V2	+ s1
gAla	pikA	0.3526		+ V2	+ m1
gAla	gERa	0.5505	+ S	+ V1 + V2	+ p2 + s2
gAla	rAza	0.6043	+ S	+ V1 + V2	+ s1
rAza	dICa	0.5418	+ S	+ V2	+ s1 + s2
rAza	rIka	0.6668	+ S	+ C1 + V2	+ s2

Table 3. The form-meaning correlation matrix.

On the right-hand side column of table 3 we have a list of the parameters that the two words in the pair share. E.g. the first pair share the stress on the same syllable, the second vowel and the sonority of the first consonant; the second pair share the second vowel and the manner of articulation of the first consonant etc.

Now we have all the elements to run the hill-climbing algorithm itself, which is an iteration of the following steps:

a) To start with, each parameter is assigned a random score between -1 and 1.
b) Then, using those scores, the form distance score of each word pair is calculated.
c) Those scores are aligned with the "semantic" distances for the same word pairs, and the resulting correlation coefficient (Pearson's r) is calculated.

d) Then, we modify one randomly chosen score by adding or subtracting a small amount to it. After this, steps (b) and (c) above are executed again.
e) If the resulting Pearson's r is higher (meaning a better correlation), the modification is kept. If it is the same or lower, the modification is discarded.
f) Steps (b) to (e) are repeated until the growth curve of the correlation coefficient becomes flat, i.e., when no modification in any of the scores generates a better Pearson's r value.

This algorithm gives us at the end a value between -1 and 1 for each of the parameters. A positive value means that sharing that parameter draws two words closer together in the representational space, and a negative value, that it pushes them further apart. The absolute value orf the parameter reflects how much closer or further apart it draws two words.

3 RESULTS

3.1 The correlation in Spanish

We run the hill-climbing algorithm three times for each word group and obtained the same results every time, indicating that we did not reach local maximum results, but rather a global correlation maximum. We obtained these maximum correlation coefficients (Pearson's r) for the three word groups: for cv-cv, $r = 0.13$ ($p<0.05$), for cvccv, $r = 0.24$ and for cv-cv-cv, $r = 0.14$. We have not calculated the significance of the last two results because we use the computationally long and expenxive Monte-Carlo analysis, and the higher Pearson's r values suggest that they must be at least as significant as the first one.

We can say, then, that a significant correlation between the form and meaning representational spaces is not unique to English, and that it occurs at least in one other language, namely Spanish.

3.2 The parameters

An analysis of the parameter values that configure the phonological spaces that generate the best correlation coefficients reveals the correlation is driven by morphology and phonology. For a full list of the final parameter values, see Appendix 2.

3.2.1 Morphological factors

Morphology in Spanish concentrates at the end of words, e.g. verbal, feminine and plural inflections all have the extra morpheme(s) attached to the end. Our word sets do not contain any prefixes, so all information about inflection and derived words is found in the final one or two phonemes. Additionally, many verbal forms in our two bisyllabic word sets are characterized by being stressed on the second syllable (as opposed to the vast majority of nouns, adjectives and other parts of speech, which are stressed on the first syllable).

	=stress1	=str.v1	=stress2	=str.v2
cv-cv	+0.9	+0.1	-0.2	+1
cvccv	+0.9	+0.1	-0.4	+1

Table 4. Values of stress-related parameters in bisyllabic words: same stress and same stressed vowel in the first and second syllables. Note that parameter values are within the range of (-1, +1).

For bisyllabic words, the highest absolute-value parameters at the end of the hill-climbing algorithm runs are related to stress. If we start with all words being equidistant from each other and apply all stress-related parameters (seen in Table 4) to them, we see the following process:

Same stress on the first syllable (+0.9). Words that have the stress on the first syllable (the majority of words in both word sets) are strongly attracted to each other.

Same stressed vowel on the first syllable (+0.1). Words sharing this parameter are only slightly drawn towards each other a little further.

Same stress on the second syllable (-0.25). Words stressed on the second syllable, which had been untouched by the process described in the last paragraph, are pushed even further apart from each other.

Same stressed vowel on the second syllable (+1). Words sharing this parameter, which were very widely scattered, are strongly drawn towards each other, forming distinct clusters.

This process produces a configuration of the form representational space where the vast majority of words are packed together in a large cluster. The rest are packed into five much smaller totally separate clusters, corresponding to words ending in stressed a, e, i, o and u. What is interesting is that these five clusters closely match different verb tense forms in Spanish (see Appendix 3). Since the effect of the highest value parameters is to cluster morphologically similar words together, we can say that our algorithm is driven to a large extent by word morphology, particularly by verb morphology.

This is not surprising because we are correlating the form of words (and morphological form is quite distinctive – all plurals ending in s, most feminines in a, each verb tense in its particular phoneme cluster etc) with the contexts words appear in. Verbs usually are placed in precise places in the sentence, so we can expect that there is a group of words that typically occurs near verbs, but this result shows that the correlation obtained with our method does indeed pick up dimensions of word meaning.

3.2.2 Phonological factors

Even though morphology accounts for a considerable part of the correlation, phonology, in the form of parameters related to phonemes and to sub-phonemic features, also plays a part in the configuration of the form representational space. We find, for example, that the parameters related to sharing consonants or vowels have values that we can relate to an intuitive assessment of phonological similarity:

Same syllable-final consonant in cvc-cv words (+0.80). E.g. *bá**r**ko-kó**r**te, bá**s**ta-pé**s**te, bá**n**ko-tó**n**to*. Spanish speakers can intuitively tell that sharing this consonant adds a lot to making words sound similar.

Sharing several syllable-onset consonants (+0.30 for 2-syllable words, +0.75 for 3-syllable words). E.g. ***k**á**m**a-**k**ó**m**a, **b**ár**k**o-**b**án**k**o, **m**o**n**é**d**a- **m**e**n**ú**d**o*. Here again, it is obvious that words sharing several consonants do sound similar, and the more consonants they share, the more similar they sound.

Sharing several vowels (-0.30, -0.07 for 2-syllable words, +0.60 for 3-syllable words). In the case of vowels, we have to take into account the probability distribution of vowel combinations. There are only 5 vowels in Spanish, and therefore, very few possible vowel combinations for 2-syllable words (5^2=25), but many more for 3-syllable words (5^3=125). 25 combinations are clearly insufficient to correlate to the wide variety of possible meanings or meaning categories that exist in bisyllabic words, so this parameter can't have a bearing in the correlation between meaning and phonology. This is reflected in the low negative values seen above. In 3-syllable words, however, the higher number of different vowel templates is enough for them to encode (to some extent) different meanings or meaning categories. In short, there are many bisyllabic words that share the same two vowels, so this factor can't mean much; however, in 3-syllable words, less words have the same vowel template, which allows the templates to "mean" something.

Sub-phonemic features: manner and place of articulation, and sonority (<0.10 for cv-cv words, up to ±0.40 for cvc-cv words and up to +0.90 for cv-cv-cv words). In this study of

three word groups it seems that word length is related to the importance of sub-phonemic features for the configuration of the form representational space. In one end, the space for cv-cv words is configured mainly by the consonantal phonemes they share. In the other end, the space for cv-cv-cv words takes much more into account sharing features such as manner and place of articulation and sonority.

4 CONCLUSION

We have proven that the correlation of the configuration of the form and meaning representational spaces exists in Spanish too, supporting the idea that it must be a generalised effect not based on the structure of individual languages, but driven by the brain information storage and retrieval mechanisms.

This correlation is driven, on the form side, by a large extent by morphology but, more interestingly, also by purely phonological factors such as sharing phonemes and sub-phonemic features.

BIBLIOGRAPHY

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

APPENDIX 1:

The complete list of parameters of form similarity.

Same consonant in the same / in a different position
Same vowel in the same / in a different position
Same two (or three) consonants at the same time
Same two (or three) vowels at the same time
Stress on the same syllable
Same stressed vowel
Same manner/place of articulation or same sonority in the same / in a different position

APPENDIX 2:

The parameter values that obtain the best correlation coefficients between form and meaning, for our three word sets. In bold, parameters with an absolute value > 20.

cv-cv

sav2	**0.99578**	xp	0.114493	xm	0.009213
sa1	**0.899314**	sc1	0.102358	sp1	-0.01075
tc	**0.334753**	sm2	0.101748	sm1	-0.0609
sv2	0.184097	sav1	0.077467	ss1	-0.10852
xv	0.137558	sc2	0.039749	tv	-0.13607
sv1	0.126543	sp2	0.031368	sa2	-0.1926
xc	0.120981	ss2	0.019259		

cvccv

sav2	**0.989675**	xp2	0.107022	xp1	-0.03769
sa1	**0.9564**	sp1	0.105998	sc1	-0.05852
sc2	**0.784097**	sm2	0.097245	tv	-0.07879
ss2	**0.491904**	xm1	0.095193	xv	-0.13441
sm3	**0.311284**	sm1	0.088476	sp3	-0.14259
tc2	**0.288583**	sav1	0.084966	sv1	-0.17859
sp2	**0.209201**	ss1	0.029269	**sa2**	**-0.36136**
xc2	**0.207347**	sv2	0.002534	**ss3**	**-0.39799**
xm2	0.141231	xc1	-0.02322	**sstr**	**-0.46451**
tc1	0.131868	sc3	-0.02706		

cv-cv-cv

sm2	**0.922875**	**xc2**	**0.266334**	tc3	-0.12722
thc	**0.785435**	**sp3**	**0.229083**	xv2	-0.17752
sm1	**0.760642**	**v2**	**0.213269**	ss3	-0.19908
c3	**0.675889**	sm3	0.194869	**sp2**	**-0.23612**
tv3	**0.470671**	c2	0.177803	**sp1**	**-0.25792**
tv1	**0.455313**	v1	0.143569	**tv2**	**-0.35082**
xc3	**0.387481**	tc1	0.044267	**sav2**	**-0.54606**
sa1	**0.367025**	xc1	0.027292	**xv3**	**-0.65489**
c1	**0.352679**	ss1	-0.03702	**sav1**	**-0.99138**
thv	**0.315508**	ss2	-0.03848	**sa3**	**-0.99452**
sa2	**0.292292**	xv1	-0.11297		
v3	**0.273391**	tc2	-0.12039		

APPENDIX 3:

The 31 cv-cv words stressed on the last syllable in our 324-word list. [*74% verbs, 22% nouns, 3% adverbs.*]

pasO	v	past	*it happened*
LegO	v	past	*he arrived*
kedO	v	past	*he stayed, remained*
tokO	v	past	*he touched*
LebO	v	past	*he carried*
LamO	v	past	*he called*
deXO	v	past	*he let, left*
ganO	v	Past	*he won*
kayO	v	past	*it/he fell*
mirO	v	past	*he looked at*
sakO	v	past	*he took out*

XosE	n		*Jose (man's name)*
CalE	n		*Chalet*
kafE	n		*Coffee*
dirE	v	fut	*I will say*
pasE	v	past	*I passed*
LamE	v	past	*I called*
LegE	v	past	*I arrived*
kedE	v	past	*I stayed, remained*

papA	n		*Daddy*
mamA	n		*Mummy*
kizA	adv		*Perhaps*
serA	v	fut	*it will be*
berA	v	fut	*he will see*
dirA	v	fut	*he will say*
darA	v	fut	*he will give*

koXI	v	past	*I took*
metI	v	past	*I put into*
salI	v	past	*I went out*

menU	n		*Menu*
perU	pn		*Peru*