Phonemic Repertoire and Similarity within the Vocabulary

Anne Cutler[1], Dennis Norris[2] & Núria Sebastián-Gallés[3]

[1]Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands
[2]MRC Cognitive and Brain Sciences Unit, Cambridge, United Kingdom
[3]Departament de Psicologia Bàsica, Universitat de Barcelona, Spain
anne.cutler@mpi.nl

Abstract

Language-specific differences in the size and distribution of the phonemic repertoire can have implications for the task facing listeners in recognising spoken words. A language with more phonemes will allow shorter words and reduced embedding of short words within longer ones, decreasing the potential for spurious lexical competitors to be activated by speech signals. We demonstrate that this is the case via comparative analyses of the vocabularies of English and Spanish. A language which uses suprasegmental as well as segmental contrasts, however, can substantially reduce the extent of spurious embedding.

1. Introduction

The units of speech are phonemes, and every language has its own set. But there is less variation than there might have been in the size of these sets. Phonemic inventories cluster around a mean of about 30 phonemes, and the most commonly occurring inventory size is 25 [1].

Differences in phoneme repertoire distribution have significant implications for listener processing of spoken language. For instance, there is evidence that listener awareness of the distributional makeup affects phonetic processing. Costa, Cutler and Sebastián-Gallés [2] examined phoneme detection in the relatively balanced Dutch phonemic inventory - 16 vowels and 19 consonants – versus the unbalanced Spanish inventory – five vowels, 20 consonants. They exploited the known effect of that a target sound is detected faster if other irrelevant phonetic variation is removed. In Dutch, strong effects of contextual uncertainty were observed for detection of both vowels and consonants, but in Spanish, consonant uncertainty had a stronger effect on vowel detection than vowel uncertainty had on consonant detection. This suggests that the Spanish listeners were well aware that the *potential* for variability was greater in one direction than the other (though *actual* variability in the experiments did not in fact differ).

More far-reaching effects may occur in the domain of word recognition. For listeners, one of the most important aspects of similarity within the vocabulary is embedding of words within other words. Whenever listeners hear the English word *barber*, they also hear the English word *bar*; likewise, hearing the Spanish word *bárbaro* entails hearing the Spanish word *bar*. Although virtually all languages make do with a relatively limited stock of a few dozen phonemes, every language creates from this limited material a huge vocabulary, running into the tens, indeed hundreds of thousands of words. This inevitably means that the words of a language often resemble other words, occur fortuitously embedded within other words, and so on, all of which potentially causes problems for the listener attempting to understand speech. Much research over the past decades has shown that spoken-word recognition involves multiple concurrent activation of word candidates, and competition between them [3]. The more competing words are activated, the slower recognition of the intended word can be [4]. Therefore, the extent of embedding within the lexicon can play a significant role in listening.

In the present study we compared two vocabularies: one typical, the other less typical. Spanish has 25 phonemes: five vowels, and four times as many consonants. In this respect Spanish displays approximately the modal phoneme repertoire. English, on the other hand, has a crowded phonemic repertoire with 24 consonants and (according to dialect) at least 20 vowels. Thus we contrasted a language with a modal phonemic repertoire of 25 phonemes, of which only a fifth are vowels, with a language with nearly twice as many phonemes overall, with roughly 40% of those being vowels, to ask what implications these differences might have for the word-recognition task facing listeners of the two languages.

We further took into account the role played by differences in the patterning of lexical stress. Both English and Spanish have lexical stress; however, they differ in the realisation of stress variation. In English, vowels in unstressed syllables may be the central vowel schwa or may reduce to a more central form, but this may never happen with the vowels in stressed syllables. Spanish, however, does not extend stress variation to segmental structure in this way; the five Spanish vowels exhibit the same spectral quality in unstressed as in stressed syllables.

This too is a difference with consequences for our understanding of how listeners process spoken words. Spoken-word recognition experiments have shown that Spanish listeners use stress to distinguish words [5], and moreover, the effects of a stress mismatch and a segmental mismatch on word recognition are parallel, suggesting that listeners are as adept in exploiting suprasegmental as segmental information. There is evidence that listeners in English can use suprasegmental information, too, but they use it with less than maximal efficiency [6]. This difference between English and Spanish may arise because English listeners rarely need to depend on suprasegmental inter-word distinctions alone, whereas Spanish listeners often have to. We tested this suggestion by comparing the savings achieved when stress is taken into account in recognising words of the two languages.

2. Method

Analyses of the vocabulary structure of Spanish were based on a newly available phonetically transcribed version of

LEXESP [7]. Analyses of English vocabulary structure were based on phonetic transcriptions in the English corpora in CELEX [8].

For the purposes of all analyses, homophones (e.g. *see*/*sea*) were represented as a single lexical form. CELEX contains many entries that correspond to two printed words (e.g. *see to*); these were excluded from the analysis, as were all entries with a written frequency of zero. This produced a lexicon of 60,000 words. The full LEXESP database contains 120,000 words. For both vocabularies, frequency-of-occurrence statistics were also available. To generate a lexicon comparable in size to the CELEX lexicon, we excluded all words with a frequency of 1, which produced a LEXESP sublexicon of 73,000 words. The availability of frequency statistics also enabled us to translate the distributional statistics for the vocabulary to estimates of likely occurrence of word tokens in samples of natural speech. For the word-length analyses described below, we separately present type statistics and estimated token statistics. CELEX provides both written and spoken frequencies, but because the transcriptions in LEXESP were derived entirely from a written corpus, we used the written frequencies from CELEX also.

We first analysed word length. The hypothesis at issue here was that average word length would be inversely correlated with phonemic repertoire size, i.e. Spanish words would on average be longer than English words. We next analysed frequency of lexical embedding; here we hypothesised that embedding frequency would also vary inversely with phoneme repertoire size, i.e. there would be (even) more embeddings in Spanish than in English. Finally, we analysed the effects of taking stress into account in word recognition: here we hypothesised that doing so would result in greater savings in Spanish than in English, because in the latter language most word pairs differing in stress also differ in segmental structure.

3. Results

3.1. Word length

(a) *Phonemes*. All languages prefer (relatively) shorter words to longer ones, and shorter words tend to have higher frequency than longer words (Zipf's Law). Yet a language with relatively more phonemes will clearly have more scope for making short words than a language with relatively fewer phonemes. The possible population of CVC words in a language with 20 consonants and five vowels is 20x5x20=2000 (assuming no constraints at all on where in a syllable phonemes may occur); with 24 consonants and 20 vowels there are 11520 possibilities.

Word types in the CELEX English vocabulary have a mean word length of 6.94 phonemes; word types in the LEXESP Spanish vocabulary have a mean length of 8.3 phonemes. Estimated token lengths when frequency was taken into account were, of course, shorter, but the difference remained: a mean of 3.54 phonemes for English and 4.62 for Spanish.

Figures 1a to 1d present the distributions of word length in phonemes across the two vocabularies, and the estimated token occurrence in speech samples from each vocabulary. It can be seen that the vocabularies differ quite a lot, the peaks of the distribution being markedly lower for English than for Spanish. The token numbers in speech differ less; in each case the mode of the distribution is at two phonemes. This is because about half of the words in the speech of both languages are function words, and both languages' function words are very short. Nevertheless, there are still likely to be many more long words in samples of Spanish than of English speech.

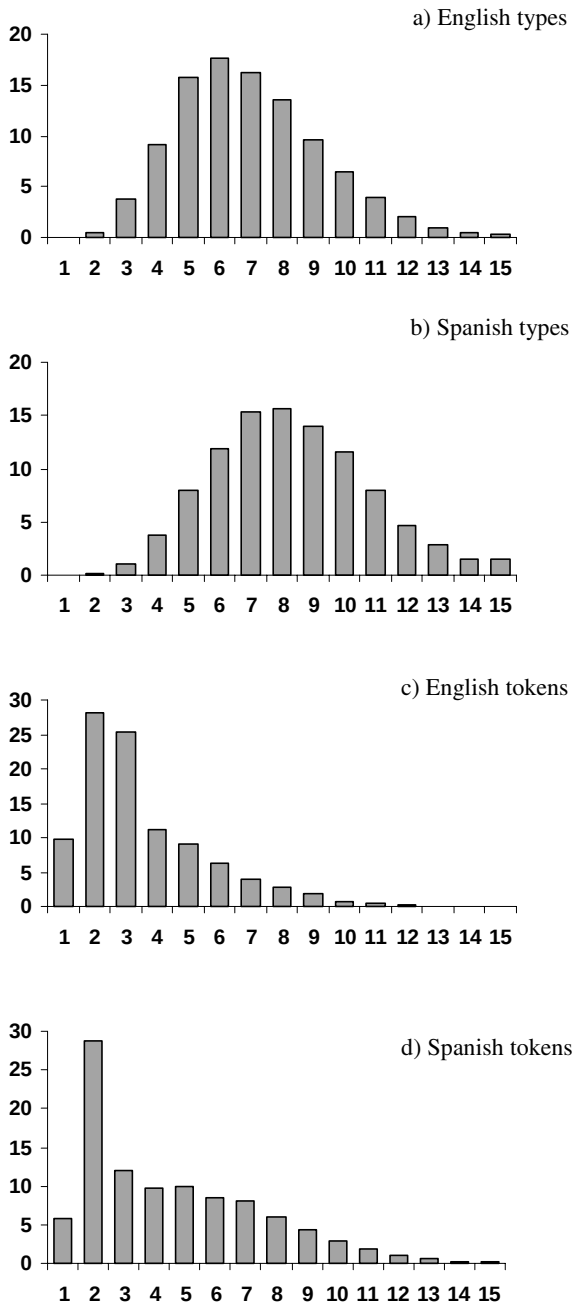

Figure 1: Distribution of word types in the vocabulary of English (1a) and Spanish (1b) and likely distribution of word tokens in speech samples from each vocabulary (1c, 1d), as a function of word length in phonemes, from 1 to 15 or more.

(b) *Syllables*. Many occurrence restrictions constrain phoneme patterning within syllables – neither English /h/ nor Spanish /f/ may occur in syllable-final position, for instance. Further, languages differ in the kinds of syllables they allow and prefer. Spanish prefers open syllables to closed and disprefers consonant clusters; English allows a large range of consonant clusters in both onset and coda of syllables. This implies that the differences apparent in Figure 1 will also imply a difference in number of syllables, with Spanish words tending to have more syllables than English. This is indeed so: we calculated a mean length of 3.48 syllables for Spanish and 2.72 for English. Even in the token counts, with the large preponderance of monosyllabic function words in both languages, the difference remains: the means there reduce to 2.02 syllables for Spanish but 1.43 for English.

3.2. Lexical similarity

McQueen, Cutler, Norris and Briscoe [9] calculated statistics on embedding within the English vocabulary, basing their analyses on a 26,000-word lexicon. They found that embedding was rife within the English word stock; 84% of polysyllabic words contained shorter words embedded within them.

Table 1: Proportions of Polysyllabic MWs in CELEX with EWs of Different Lengths, Stress Ignored

MW Length	*EW:No. Sylls*	*Location of Onset of EW in MW*				
		1st	*2nd*	*3rd*	*4th*	*5th*
Two	1	0.506	0.431			
Three	1	0.418	0.278	0.329		
	2	0.207	0.216			
Four	1	0.385	0.206	0.327	0.221	
	2	0.126	0.068	0.169		
	3	0.134	0.121			
Five	1	0.448	0.185	0.242	0.340	0.181
	2	0.100	0.056	0.066	0.122	
	3	0.029	0.034	0.088		
	4	0.111	0.136			

Their analyses considered phonemic similarity only, with no account of lexical stress. We recalculated their analyses for English from CELEX, and calculated the same statistics for Spanish from LEXESP. These initial statistics for the two languages are presented in tables 1 and 2, in the format used by McQueen et al. [9]. In the tables, MW signifies matrix word, EW signifies embedded word.

All tables give the proportion of matrix words of different lengths (MW Length) with an embedded word of a given length (EW: Number of Syllables) beginning at a given syllable position (Nth Syllable). For example, in Table 1 *canvas* contributes to the 0.506 of two-syllable MWs with a one-syllable EW (*can)* beginning at the first syllable of the MW. We computed the statistics to MW length 6, but to simplify the tables, the (very small) numbers for these long MWs have been omitted. It can clearly be seen that for virtually all word lengths and positions, Spanish has considerably more embeddings than English. As we shall see, however, the effects of taking lexical stress into account change this picture dramatically.

Table 2: Proportions of Polysyllabic MWs in LEXESP with EWs of Different Lengths, Stress Ignored

MW Length	*EW:No. Sylls*	*Location of Onset of EW in MW*				
		1st	*2nd*	*3rd*	*4th*	*5th*
Two	1	0.616	0.510			
Three	1	0.736	0.577	0.531		
	2	0.316	0.226			
Four	1	0.799	0.632	0.589	0.513	
	2	0.271	0.238	0.213		
	3	0.323	0.114			
Five	1	0.833	0.690	0.626	0.564	0.181
	2	0.293	0.239	0.226	0.208	
	3	0.182	0.097	0.104		
	4	0.273	0.076			

Table 3: Proportions of Polysyllabic MWs in CELEX with EWs of Different Lengths, Stress Considered

MW Length	*EW:No. Sylls.*	*Location of Onset of EW in MW*				
		1st	*2nd*	*3rd*	*4th*	*5th*
Two	1	0.439	0.122			
Three	1	0.249	0.182	0.062		
	2	0.188	0.129			
Four	1	0.107	0.119	0.154	0.012	
	2	0.066	0.055	0.074		
	3	0.130	0.113			
Five	1	0.014	0.078	0.119	0.156	0.004
	2	0.014	0.038	0.049	0.053	
	3	0.023	0.033	0.075		
	4	0.106	0.134			

Table 4: Proportions of Polysyllabic MWs in LEXESP with EWs of Different Lengths, Stress Considered

MW Length	*EW:No. Sylls*	*Location of Onset of EW in MW*				
		1st	*2nd*	*3rd*	*4th*	*5th*
Two	1	0.444	0.204			
Three	1	0.290	0.433	0.216		
	2	0.097	0.193			
Four	1	0.346	0.218	0.461	0.213	
	2	0.025	0.057	0.172		
	3	0.093	0.109			
Five	1	0.398	0.196	0.237	0.433	0.229
	2	0.021	0.020	0.053	0.170	
	3	0.035	0.020	0.102		
	4	0.073	0.075			

3.3. Lexical stress and lexical similarity

Tables 3 and 4 show the embedding statistics when the transcriptions include lexical stress. In this case in English, for example, *see* would be embedded at the beginning of *secret* but not of *seniority*, and *tea* would be embedded at the end of *settee* but not of *hasty*.

Tables 3 and 4 show much less difference than Tables 1 and 2. The major differences in Tables 3 and 4 concern monosyllabic words. For words of two and three syllables, the number of monosyllabic words embedded at the first syllable of the matrix word is similar for Spanish and English, but for all other word lengths and positions, Spanish has more monosyllabic embeddings. However, this is entirely due to the fact that LEXESP codes monosyllabic function words as having no stress. Most Spanish polysyllabic words have penultimate stress. Unstressed monosyllables can therefore generally be embedded at any position other than the penultimate syllable. If we recompute the statistics with all monosyllabic words marked as stressed (Table 5), then all of the monosyllabic embeddings located in unstressed syllables are eliminated, and the majority of monosyllabic embeddings are then located on the penultimate syllable. In other words, if function words could be prevented from matching unstressed syllables in polysyllabic words (by being filtered out on syntactic grounds, for example), then the number of effective embeddings in Spanish would decrease considerably.

Table 5: Proportions of Polysyllabic MWs in LEXESP with EWs of Different Lengths, Monosyllabic Words as Stressed

MW Length	*EW:No. Sylls*	*Location of Onset of EW in MW*				
		1st	*2nd*	*3rd*	*4th*	*5th*
Two	1	0.386	0.112			
Three	1	0.037	0.388	0.066		
	2	0.097	0.193			
Four	1	0.002	0.053	0.414	0.032	
	2	0.025	0.057	0.172		
	3	0.093	0.109			
Five	1	0.002	0.010	0.063	0.385	0.026
	2	0.021	0.020	0.053	0.170	
	3	0.035	0.020	0.102		
	4	0.073	0.075			

Comparison of the embedding statistics taking stress into account with the same counts when all stress marking is ignored show that in English the number of embeddings decreases, especially for longer words. Most of the decrease is in terms of monosyllabic words embedded at word onset. However, in Spanish, the savings effect is much more marked, especially comparing the numbers in Table 5 with those in Table 2. Also, the increase in embeddings is spread more evenly, although it is most apparent for one- and two-syllable embedded words. Using stress to constrain the activation of embedded words in Spanish therefore has the potential to ameliorate much of the problem caused by lexical embeddings.

We estimated the comparative benefit of taking account of stress in Spanish and English by computing the number of words embedded in each matrix word of length 2-6 syllables weighted by the frequency of the matrix word. For English, there are 0.94 words embedded in each token of a matrix word when stress is ignored, and this is reduced to 0.59 when stress is taken into account. For Spanish the number of embeddings is 2.32 ignoring stress, 1.19 when content but not function words are marked for stress, and 0.73 when content words and all monosyllabic words are marked for stress. Thus stress information can reduce the number of embeddings by about one-third in English, but by up to two-thirds in Spanish.

4. Conclusions

Our analyses have supported all the hypotheses under test, underpinning our suggestion that the makeup of a language's phonemic repertoire can itself have significant implications for the task facing listeners processing speech signals. Languages with fewer phonemes in their repertoire will tend to have longer words, creating more opportunities for embedding and hence temporary competition. Our analyses show that Spanish indeed has longer words, and, by phonemic counts, much more embedding than English. However, suprasegmental distinctions can reduce this embedding problem to a significant extent.

Spanish listeners, in particular, are greatly assisted by their language's use of suprasegmental distinctions between lexical items. More exactly: Spanish permits embeddings that can be eliminated on the basis of stress, because Spanish uses stress to distinguish words. In comparison, stress plays an almost insignificant role in signaling lexical contrasts in English, and the English vocabulary does not require the listener to display sensitivity to lexical stress in order to eliminate embeddings.

5. Acknowledgements

We thank Antonio Bonafonte and the TALP Research Center of the Universitat Politécnica de Catalunya for facilitating the phonetic transcriptions of the LEXESP database.

6. References

[1] Maddieson, I., *Patterns of Sounds*, Cambridge University Press, Cambridge, 1984.

[2] Costa, A., Cutler, A., and Sebastián-Gallés, N., "Effects of phoneme repertoire on phoneme decision", *Perception & Psychophysics*, 60: 1022-1031, 1998.

[3] Frauenfelder, U.H., and Floccia, C., "The recognition of spoken words", in A. Friederici (Ed.), *Language Comprehension: A Biological Perspective,* pp. 1-40, Springer, Heidelberg, 1998.

[4] Norris, D., McQueen, J.M., and Cutler, A., "Competition and segmentation in spoken word recognition", *Journal of Experimental Psychology: Learning, Memory and Cognition*, 21: 1209-1228, 1995.

[5] Soto-Faraco, S., Sebastián-Gallés, N. and Cutler, A., "Segmental and suprasegmental mismatch in lexical access", *Journal of Memory and Language*, 45: 412-432, 2001.

[6] Cooper, N., Cutler, A., and Wales, R., "Constraints of lexical stress on lexical access in English: Evidence from native and non-native listeners", *Language and Speech,* 45: 207-228, 2002.

[7] Sebastián-Gallés, N., Martí, M.A., Cuetos, F., and Carreiras, M., *LEXESP: Léxico informatizado del español.* Barcelona: Edicions de la Universitat de Barcelona, 2000.

[8] Baayen, R.H., Piepenbrock, R., and Van Rijn, H., *The CELEX lexical database* (CD-ROM)., Linguistic Data Consortium, Philadelphia, PA, 1993.

[9] McQueen, J.M., Cutler, A., Briscoe, T., and Norris, D.G., "Models of continuous speech recognition and the contents of the vocabulary", *Language and Cognitive Processes*, 10: 309-331, 1995.