Real World Constraints on the Mental Lexicon: Assimilation, the Speech Lexicon and the Information Structure of Spanish Words

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

Richard C. Shillcock (rcs@cogsci.ed.ac.uk)
Department of Cognitive Science, 2 Buccleuch Place
Edinburgh EH8 9LW, UK

Abstract

This paper focuses on the optimum use of representational space by words in speech and in the mental lexicon. In order to do this we draw the concept of entropy from information theory and use it to plot the information contour of words. We compare different representations of Spanish speech: a citation vs. a fast-speech transcription of a speech corpus and a dictionary lexicon vs. a speech lexicon. We also compare the information profiles yielded by the speech corpus vs. that of the speech lexicon in order to contrast the representation of words over two representational spaces: time and storage space in the brain. Finally we discuss the implications for the mental lexicon and interpret the analyses we present as evidence for a version of Butterworth's (1983) Full Listing Hypothesis.

Introduction

In this paper we focus on the optimum use of representational space by words over time (the sequence of sounds in speech) and over space (the storage site of the mental lexicon in the brain). We draw the concept of entropy from information theory and propose that it can be used to study the information structure of the set of words uttered in speech and of those stored in the mental lexicon in the face of the constraints of communication and of storage, respectively, in a potentially noisy medium.

We have two representational spaces for words: time and storage space. Further, we will consider the phonology and morphology of word systems. Our data sets are phonetic representations of words, and recent research demonstrates that information on the probabilistic distribution of phonemes in words is used in language processing (see Frisch, Large & Pisoni, 2000 for review). Morphology is involved in this research because we will be comparing groups of words with different inflectional and derivational features. We will initially assume the Full Listing Hypothesis (Butterworth, 1983): every word-form, including inflected and derived forms, is explicitly listed in the mental lexicon.

Shillcock, Hicks, Cairns, Chater and Levy (1995) suggest the general principle of the presentation of information in the brain that information should be spread as evenly as possible over time or over the representational space. Therefore, if the entropy of the mental lexicon is to be maximized so that the storage over a limited space is most efficient, then all the phonemes will tend to occur as evenly as possible in each segment position of the word. The phonology of each individual word, because it will have an effect on the entropy of the system, affects whether it is likely to become part of the mental lexicon.

Shillcock et al. stated that "the optimum contour across the phonological information in a spoken word is flat; fast-speech processes cause the information contour to become more level". We generalize this notion and propose the Levelling Effect of Realistic Representations (LERR): *processes that make the representation of words more accurate will flatten the information profiles*.

In order to test this, we will use Spanish word systems to calculate the slope and overall level of entropy of a citation (idealized pronunciation of the word in isolation) transcription and of a fast-speech (more realistic) transcription and of a dictionary lexicon and the speech lexicon. Our prediction is that the second system in each comparison should yield flatter information contours. We also compare a representation of words over time and another one over storage space - a speech corpus and the speech lexicon.

Entropy

We will use the concept of entropy in the context of information theory (Shannon, 1948), also employed in speech recognition studies (e.g. Yannakoudakis & Hutton, 1992). Entropy H is defined for a finite scheme

(i.e., a set of events such that one and only one must occur in each instance, together with the probability of them occurring) as a reasonable measure of the uncertainty or the information that each instance carries. E.g. the finite scheme formed by the possible outcomes when throwing a dice has maximum entropy: each side of the dice has 1/6 probability of occurring and it is very difficult to predict what the outcome will be. A loaded dice, on the other hand, has an unequal probability distribution, and the outcome is less uncertain. In this research, the possible events are the phonemes and allophones, and for each word only one of them can occur at each segment position.

For probabilities (p_1, p_2, p_3...p_n):

$H = - \Sigma (p_i \cdot \log p_i)$

The relative entropy H_{rel} is the measured entropy divided by the maximum entropy H_{max}, which is the entropy when the probabilities of each event occurring are equal and the uncertainty is maximized. Using H_{rel} allows us to compare entropies from systems with a different number of events (in this case, a system with 30 phonemes with another one with 50).

$H_{max} = \log n$

$H_{rel} = H / H_{max}$

Redundancy R is a measure of the constraints on the choices. When redundancy is high, the system is highly organized, and more predictable, i.e. some choices are more likely than others, as in the case of the loaded dice.

$R = 1 - H_{rel}$

In order to obtain the information profiles of words (see Figure 1), the entropy was calculated separately for each segment position in a set of left-justified words of equal length, i.e., for the first phoneme in the words, the second phoneme etc.

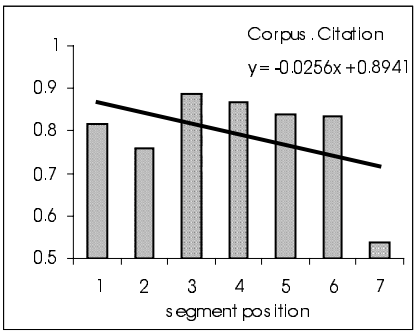

Figure 1: Information profile of 7-segment words from the citation transcription of the speech corpus.

The information profile of the word was measured as the linear trendline of these individual segment entropies. The slope (m) (multiplied by (-1)) of these trendlines and the mean relative entropy for each word length are shown in the figures below. E.g. In Figure 1, (-m)=0.0256. The flatness of the slope refers literally to how horizontal the trendline is.

Transcriptions

We have restricted ourselves to phonemic representations of word and will not report data concerning the distributions of phonemic features. We have used citation transcription rules (the idealised pronunciation of the isolated word) and fast-speech rules (an attempt to represent normal speech more realistically). Both citation and fast-speech rules were applied uniformly to the whole data sets. For the citation transcription we used 29 phonemes including 5 stressed vowels; for the fast-speech transcription we used 50 phonemes and allophones:

Citation transcription: Vowels: /a/, /e/, /i/, /o/, /u/, /á/, /é/, /í/, /ó/, /ú/. Consonants: /p/, /b/, /t/, /d/, /k/, /g/, /m/, /n/, /ɲ/, /ɾ/, /r/, /f/, /θ/, /s/, /ʝ/, /χ/, /l/, /ʎ/, /tʃ/.

Fast-speech transcription: The above plus semivowel /i̯/, /u̯/, voiced approximants /β̞/, /ð̞/, /ɣ̞/, voiceless approximants /β̞̥/, /ð̞̥/, /ɣ̞̥/, labiodental /m/, dental /n/ and /l/, palatalised /n/ and /l/, velarized /n/, /z/, dental voiced /s/, dental /s̪/, fricative /ɾ̞/, voiced /θ/ and a silenced consonant /Ø/. The transcription was made following the rules for consonant interactions, such as feature assimilation, set out by Rios Mestre (1999, chapter 5). Diphthongs were treated as two separate segments, as is usual in Spanish. Rules to mark stressed vowels were applied to all but monosyllabic words without an orthographic accent. For the corpus, the whole text was used, including repetitions and false starts of words. After deleting all the tags, the corpus was divided into chunks separated by pauses (change of speaker, comma, full stop, or pause marked in the transcription). The resulting text was transcribed automatically word by word (orthographic forms being replaced by phonetic forms) and then word boundary effects were introduced within the chunks, following the same rules as for the intra-word transcription.

Data

We used these three sets of data:

The speech corpus: a 707,000 word Spanish speech corpus, including repetitions and unfinished words. This corpus was developed by Marcos Marín of the Universidad Autonoma de Madrid in 1992 and contains transcribed speech from a wide range of registers and fields, from everyday conversation to academic talks and political speeches.

The dictionary lexicon: a 28,000 word Spanish word lexicon (the Spanish headword list of the Harrap Compact Spanish Dictionary, excluding abbreviations). This list does not include inflections, but approximately 40% of the words are derived words (we take the infinitive of verbs and the simple form of the noun as

the basic forms). This word system could represent a mental lexicon where that only word stems are listed and where inflected words are assembled during speech production.

The speech lexicon: the 42,000 word types found in the corpus. Some 80% of these types were derived and inflected words. We take this word system to be the most realistic representation of the mental lexicon, assuming Butterworth (1983)'s Full Listing Hypothesis, where all the wordforms are individually represented in the mental lexicon.

The dictionary lexicon and the speech lexicon share only ~30% of the words. The remaining ~70% of the words in the dictionary lexicon are mostly low frequency words which do not appear in our sample of speech. The new ~70% in the speech lexicon are verbal inflections (~35%), plurals and feminine inflections (~25%), some derived words absent from the dictionary lexicon (~4%), unfinished or mispronounced words (~4%) and proper nouns (~2%).

From these data, we used 4, 5, 6 and 7-segment transcriptions. Words were separated by length in order to see a clearer picture of the information profiles, especially as far as the word-ending contribution is concerned. Considering that the information profiles of Spanish words follows the same pattern as those of English words as seen in Shillcock et al. (1995), we can extend research in English to Spanish words. In English, word recognition typically occurs before the end of the word is uttered (Marslen-Wilson & Tyler, 1980), and information about word-length is typically available once the nucleus is being processed (Grosjean, 1985). It is, therefore, legitimate to assume that recognition processes are restricting their activities to the subset of words in the lexicon that match the word being uttered both in terms of initial segments and approximate overall length. The particular word lengths were chosen because the structure of shorter words is simpler, and the effects are less likely to be obscured by greater variation in the internal structure of each word-length group. These word lengths are equidistant from the modes of the word-length distribution of the three data sets (lexicon: mode = 8, speech lexicon: mode = 7 and speech corpus: modes = 2, 4 – the mode of the normal distribution is 4, but the proportion of 2-segment words is even higher, accounting for 32% of all tokens). The sum of these four word lengths accounts for 41% of the dictionary lexicon, 45% of the speech lexicon and 37% of the speech corpus.

The effect of the transcription

Shillcock et al. (1995) showed that fast-speech processes cause the information contour to become more level for English, German, Welsh, Irish and Portuguese. Here we compare the slope of the information profiles of 4-7 segment words from the corpus transcribed with citation rules and with fast-speech rules.

As predicted by the LERR principle, Figure 2 confirms that this is also the case for Spanish. The information profile is consistently flatter for the more realistic fast-speech transcriptions in all word lengths. Note that in the figure, a higher value of (–m) indicates a steeper profile.

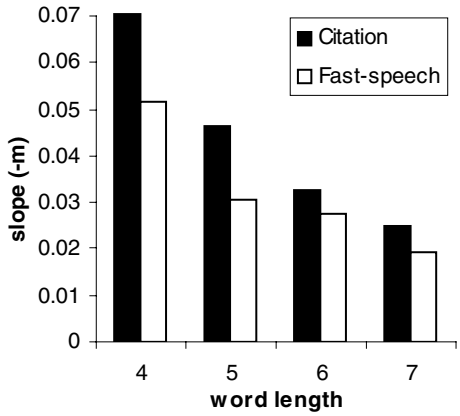

Figure 2: Slopes of the information profiles of the citation and the fast-speech transcriptions applied to the corpus, over the four word lengths.

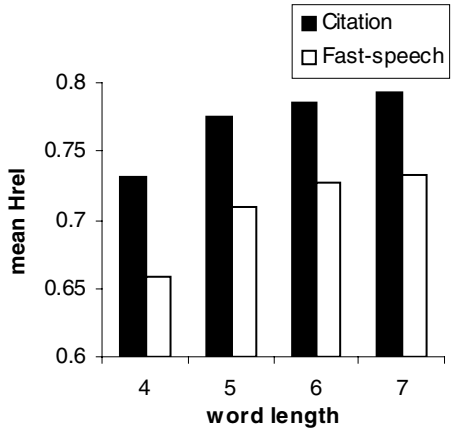

Figure 3: Mean relative entropy of the citation and fast-speech transcriptions over the four word lengths.

Figure 3 shows how the overall entropy is lower for the fast-speech transcription: when we introduce the allophones and the assimilation rules, the system becomes more redundant and thus, more predictable.

The Speech Lexicon

Some current models of lexical access propose two parallel word recognition routes, a whole-word route and a morpheme-based one (e.g. Wurm (1997) for English; Colé, Segui & Taft (1997) for French; Laine, Vainio & Hyona (1999) for Finnish). Following this hypothesis, the full forms of words need to be stored in the mental lexicon (cf. Butterworth, 1983). It is relevant, then, to study the behaviour of the set of all word types, including derived and inflected words, that appear in speech: the speech lexicon.

We have seen that fast-speech transcriptions yield flatter information contours than citation transcriptions, so we will use the fast-speech transcriptions of the speech lexicon, the lexicon and the corpus.

Comparing the slopes of the information profiles of the speech lexicon on the one hand and the dictionary lexicon and the corpus on the other hand will help characterize the active mental lexicon.

Speech lexicon vs. dictionary lexicon

The speech lexicon contains inflected and derived forms, and does not contain the more obscure words that can be found in the dictionary. The LERR principle that data that are closer to real speech should produce flatter information contours is confirmed in Figure 4, where we see that the values of the slope of the information profile of the speech lexicon are lower than those of the dictionary lexicon.

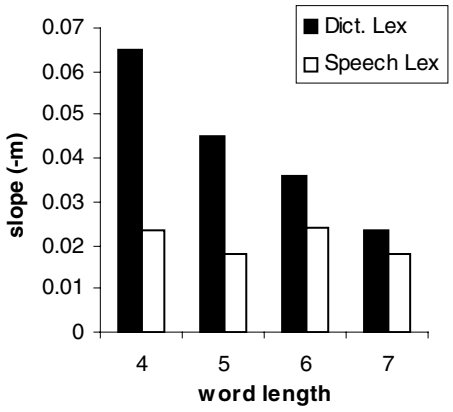

Figure 4: Slopes of the information profiles of the dictionary lexicon and the speech lexicon over the four word lengths.

Figure 5 shows that the overall entropy level is higher for the speech lexicon. This means that the speech lexicon is less redundant than the dictionary lexicon. The representational space is now a limited amount of memory storage space in the brain, and for maximal efficiency redundancy has to be reduced as much as possible. The results from both the slopes and the entropy levels support the Full Listing Hypothesis that all wordforms, particularly inflected forms, are listed in the mental lexicon – the system that includes all wordforms (the speech lexicon) could be stored more efficiently over a limited representational space.

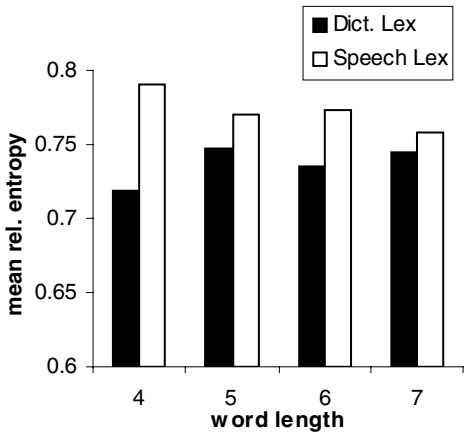

Figure 5: Mean relative entropy of the dictionary lexicon and the speech lexicon over the four word lengths.

Speech lexicon vs. corpus

The fact that entropy and redundancy statistics obtained from a lexicon are different from those obtained from a corpus has been noted by Yannakoudakis and Angelidakis (1988). Here we are comparing the word tokens with the word types in a speech corpus. Figures 6 and 7 show that the speech lexicon has consistently flatter slopes and higher entropy levels than the corpus.

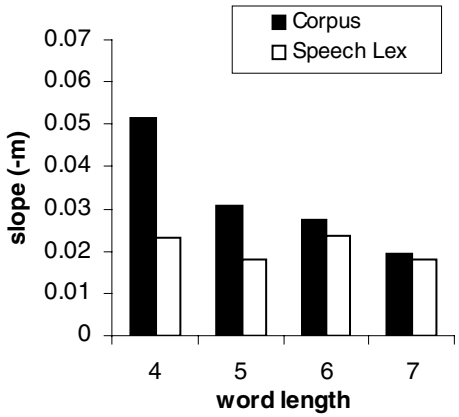

Figure 6: Slopes of the information profiles of the corpus and the speech lexicon across the four word lengths.

We are comparing two representational spaces: words in the brain are constrained by a limited space and words uttered over time are constrained by the efficiency of communication. We saw in the last section that the flat slopes and high entropy levels of the speech lexicon information profiles are best suited to enhance storage efficiency. Slopes in the corpus are relatively flat, but still steeper than those of the speech lexicon. This may reflect the fact that there are other factors affecting the information contour of words in speech, such the need to encode cues to lexical segmentation (signals that indicate where words begin and end). These other factors may be interacting with the optimization of communication.

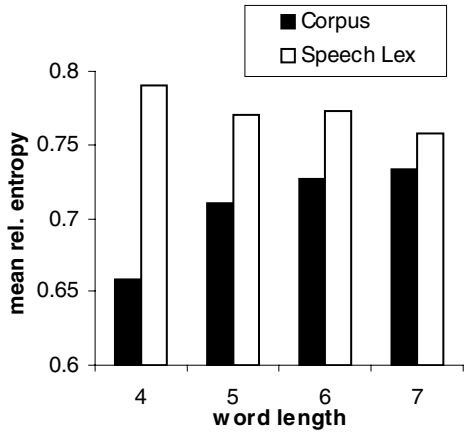

Figure 7: Mean relative entropy of the corpus and the speech lexicon across the four word lengths.

The corpus presents lower entropy levels than the speech lexicon. Speech over time is not constrained by space limitations, but rather by the need to communicate efficiently. The higher redundancy means that this system reduces the uncertainty and is indeed better for communication.

Discussion

The present study points in the direction of the LERR principle that the more realistic data - the fast-speech transcription and the speech lexicon - produce flatter information profiles.

The flatter profile of the fast-speech transcription can be partly explained in terms of the Markedness Ordering Principle (Shillcock et al., 1995) that when consonant interactions introduce phonological ambiguity, the ambiguity introduced is always in the direction of a less frequent phoneme. As for the comparison between lexicons, let us remember that the 70% of words in the speech lexicon that do not appear in the dictionary lexicon are mostly inflected words, and the 70% of words in the dictionary lexicon not present in the speech lexicon are mainly low-frequency words. The flatter profile of the speech lexicon is due to the fact that the inflected words (which are derived from one third of the dictionary lexicon words) yield a flatter profile than the low-frequency dominated group. This suggests that inflected words are included in the mental lexicon, and so it supports the Full Listing Hypothesis.

Additionally, the overall level of entropy and redundancy gives us an insight into the degree of complexity of a system. Highly organized systems will show low entropy and high redundancy. Fast-speech rules make the system more redundant than the citation rules. This higher predictability helps to deal with the loss of information produced by noise and thus enhance communication. The speech lexicon is less redundant than the dictionary lexicon. Here again, the higher entropy must be attributable to the fact that the phonemes in inflected forms are more evenly distributed over the phonological space than the more obscure words present in the dictionary lexicon.

The comparison between the corpus and the speech lexicon shows the features of the representation that has evolved to enhance communication and storage, respectively. Both systems are “realistic”, and indeed both show relatively flat information contours, but more so the speech lexicon, suggesting that communication has other constraints that interact with this measure, such as word-boundary recognition. This is true particularly for shorter words. The fact that the corpus is markedly more redundant than the speech lexicon is only to be expected, since it reflects the added complexity of different word-frequencies.

In conclusion, we have shown that it is possible to use psychological theories of the mental lexicon and spoken word recognition to make testable predictions concerning distributional information in large samples of language, and, conversely, that data from information distribution may potentially falsify particular aspects of those psychological theories. Our current conclusions from the analyses of Spanish favour versions of Butterworth's original Full Listing Hypothesis, in which all the wordforms encountered in speech are individually stored.

Acknowledgments

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

References

Butterworth, B. (1983). Lexical representation. In B. Butterworth (Ed.), *Development, writing and other language processes, Vol. 2*, London: Academic Press.

Colé, P., Segui, J., & Taft, M. (1997, Words and morphemes as units for lexical access, *Journal of memory and language*, *37 (3)*, 312-330.

Frisch, S. A., Large, N. R. & Pisoni, D. B. (2000). Perception of wordlikeness: Effects of segment probability and length on the processing of nonwords, *Journal of Memory and Language*, *42 (3)*, 481-496.

Grosjean, F. (1985). The recognition of words after their acoustic offset: Evidence and implications. *Perception and Psychophysics*, *38*, 299-310.

Laine M., Vainio, S., & Hyona, J. (1999). Lexical access routes to nouns in a morphologically rich language, *Journal of memory and language*, *40 (1)*, 109-135.

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

Marslen-Wilson, W. D., & Tyler, L. K. (1980). The temporal structure of spoken language understanding, *Cognition*, *8*, 1-71.

Ríos Mestre, A. (1999). La transcripción fonética automática del diccionario electrónico de formas simples flexivas del español: estudio fonológico en el léxico, *Estudios de Lingüística Española*, *4*.

Shannon, C. E. (1948). A mathematical theory of communication, *Bell System Technology Journal*, *27* (July), 379-423 and (October), 623-656.

Shillcock, R.C., Hicks, J., Cairns, P., Chater, N., & Levy, J. P. (1995). Phonological reduction, assimilation, intra-word information structure, and the evolution of the lexicon of English: Why fast speech isn't confusing. *Proceedings of the Seventeenth Annual Conference of the Cognitive Science Society* (pp. 233-238), Hillsdale, NJ: Lawrence Erlbaum Associates.

Yannakoudakis, E. J. & Hutton, P. J. (1992). An assessment of N-phoneme statistics in phoneme guessing algorithms which aim to incorporate phonotactic constraints, *Speech communcation*, *11*, 581-602.

Yannakoudakis, E. J. & Angelidakis, G. (1988). An insight into the entropy and redundancy of the English dictionary, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *10 (6)*, 960-970.

Wurm L. H. (1997). Auditory processing of prefixed English words is both continuous and decompositional, *Journal of memory and language*, *37 (3)*, 438-461.