Exploring the Adaptive Structure of the Mental Lexicon

Mónica Tamariz-Martel Mirêlis

PhD
University of Edinburgh
2004

Declaration

I declare that this thesis has been composed by myself and that the research reported here has been conducted by myself unless otherwise indicated.

Mónica Tamariz-Martel Mirêlis

Edinburgh, 27 October 2004.

Acknowledgements

I wish to thank my supervisors, Simon Kirby and Richard Shillcock, for their support and encouragement throughout the preparation and writing of this thesis.

I would also like to thank Scott McDonald and Mark Ellison for their help and comments, and the following people for their various contributions: Ellen Bard, Dan Dediu, Christine Haunz, Padraic Monaghan, Andrew Thompson, Viktor Tron, my father Juan Tamariz, the Language Evolution and Computation research group and the Centre for Connectionist modelling of Cognitive Processes research group.

For computing support, thanks to Eddie Dubourg, Cedric Macmartin, Mike Bennet and Morag Brown. For financial support, I am grateful to the Engineering and Physical Science Research Council.

For everything in general. big thanks to James, Ana Catriona and Alejandro Átomo.

Abstract

The mental lexicon is a complex structure organised in terms of phonology, semantics and syntax, among other levels. In this thesis I propose that this structure can be explained in terms of the pressures acting on it: every aspect of the organisation of the lexicon is an adaptation ultimately related to the function of language as a tool for human communication, or to the fact that language has to be learned by subsequent generations of people. A collection of methods, most of which are applied to a Spanish speech corpus, reveal structure at different levels of the lexicon.

- The patterns of intra-word distribution of phonological information may be a consequence of pressures for optimal representation of the lexicon in the brain, and of the pressure to facilitate speech segmentation.
- An analysis of perceived phonological similarity between words shows that the sharing of different aspects of phonological similarity is related to different functions. Phonological similarity perception sometimes relates to morphology (the stressed final vowel determines verb tense and person) and at other times shows processing biases (similarity in the word initial and final segments is more readily perceived than in word-internal segments).
- Another similarity analysis focuses on cooccurrence in speech to create a representation of the lexicon where the position of a word is determined by the words that tend to occur in its close vicinity. Variations of context-based lexical space naturally categorise words syntactically and semantically.
- A higher level of lexicon structure is revealed by examining the relationships between the phonological and the cooccurrence similarity spaces. A study in Spanish supports the universality of the small but significant correlation between these two spaces found in English by Shillcock, Kirby, McDonald and Brew (2001). This systematicity across levels of representation adds an extra layer of structure that may help lexical acquisition and recognition. I apply it to a new paradigm to determine the function of parameters of *phonological* similarity based on their relationships with the *syntactic-semantic* level. I find that while some aspects of a language's phonology maintain systematicity, others work against it, perhaps responding to the opposed pressure for word identification.

This thesis is an exploratory approach to the study of the mental lexicon structure that uses existing and new methodology to deepen our understanding of the relationships between language use and language structure.

Contents

Chapter 1. Introduction

The mental lexicon is a complex structure where words are organised in terms of their phonology, syntax, semantics as well as other non-linguistic aspects. In this thesis I take the mental lexicon to embody the human language capacity, a robust system adapted ultimately to the pressures imposed by communication and by learnability. I assume that linguistic aspects of the mental lexicon are reflected in the structure of speech, and, conversely, that the linguistic information contained in speech contributes to the development of new mental lexicons in human children and to the subsequent adjustments to the existing lexicons in adulthood. I use mainly corpus-based methods to analyze patterns of information in speech - which reflect the structure of the mental lexicon - and explain them as adaptations to the pressures that may have brought them about.

This chapter justifies and expands on these assumptions. It first defines and characterizes the object of study - the mental lexicon. It reviews recent literature on language structure, complexity and adaptiveness to motivate the adoption of a complex, adaptive mental lexicon model. Then it reviews the literature on statistical learning, the link between the patterns found in speech and the mental lexicon structure inferred from them. Finally, it sketches the different methodologies employed and it provides an overview of the contents of chapters two to six, stating their aims, motivating the research they present and explaining how they relate to each other in the general organisation of the thesis.

1.1 Defining the mental lexicon

Throughout this thesis I assume that organisation of the lexicon is based on relationships between words. I emphasize that the adaptive structure of the lexicon is a consequence of the pressures acting on it. From this standpoint, for instance, syntax can be viewed as an emergent property, a consequence of the way the lexicon is structured.

The mental lexicon is accessed in every act of linguistic communication. We need to find the word that denotes the meaning we want to express, or the meaning of a word we hear or read. These basic tasks are bound to be greatly facilitated if the mental lexicon is organised in some way. Priming studies show that words are linked to each other along many dimensions. Some dimensions are studied by linguistic disciplines (phonological, semantic, syntactic) and others, by other disciplines (emotional, social, context-interactional). When a word is activated, other words of similar form (Goldinger, Luce & Pisoni, 1989; Luce, Pisoni & Goldinger, 1990), meaning (Meyer & Schevaneldt, 1971; see Neely 1991 for review), syntax (Sereno, 1991), orthography (Segui & Grainger, 1990), emotional content (Wurm, Vakoch, Aycock, & Childers, 2003) etc are also activated, suggesting that the mental lexicon is complex and highly interconnected. Words are defined to a large extent in terms of their fluid relationships of similarity to the rest of the words. At a given point in time we can ask: Are two words pronounced in a similar way? Do they point to similar concepts? Do they tend to occur close to the same words in speech? Are they used in similar social situations? Do they have similar affective connotations? In this thesis I present a model of the mental lexicon based on the relationships between words at different levels.

We can define the mental lexicon as the collection of words one speaker knows and the relationships between them, and the lexicon of a language as the collection of words in a language and the relationships between them. The individual mental lexicons - in Chomskyan terminology, I-language, or internal language (Chomsky, 1986) - are more or less complete instantiations of the lexicon of the language. The lexicon is also manifested as the words in the speech stream during communication, and as writing - Chomsky's E-language, or external language. This definition practically equates the lexicon with the language capacity. Indeed, throughout this thesis I assume a mental lexicon that incorporates at least the major elements of language - phonology, syntax and semantics - and can be used interchangeably with the term 'language'.

Language has been characterized by some authors as a complex adaptive system (CAS), and language change, as the evolution of a CAS (Gell-Mann, 1994; Kirby, 1999). A CAS is essentially a system that adapts through a process of self-organisation and selection. Dooley (1997) gives a nominal definition of CAS[1]: '(It) behaves/evolves according to three key principles: order is emergent as opposed to predetermined, the system's history is irreversible, and the system's future is often unpredictable. The basic building blocks of the CAS are agents, semi-autonomous units that seek to maximize some measure of goodness, or fitness, by evolving over time...' Mufwene (2001) also supports the view that languages are complex adaptive systems, and goes as far as defining them as having life. Being an adaptive system necessarily implies that the mental lexicon constantly evolves. In this

[1] Complex adaptive system: definition in Dooley (1997), based on the works of Gell-Mann (1994), Holland (1995), Jantsch (1980), Maturna and Varela (1992), and Prigogine and Stengers (1984).

thesis I look at a synchronic sample of Spanish speech. The main source of data is a corpus of transcribed speech recorded in Spain in the early 1990's (Marcos Marin, 1992). However, the explanations of the traits of the lexicon always take into account that the mental lexicon is a system evolving under the effect of a set of often conflicting, probably interacting and potentially changing pressures.

1.1.1 Pressures on the structure of the lexicon

The lexicon is under many pressures: words have to be able to be pronounced, transmitted, processed and decoded, and they have to be acquired by new speakers; the representations of the lexicon have to be stored in the brain in such a way that there are connections between the different aspects of a single word as well as over whole categories of words; words and the relationships between them need to allow people to communicate concepts and their relationships. This indicates that the lexicon is content-addressable at every level, allowing us to access words in terms of syntactic category, phonological characteristics and meaning among others.

Therefore, factors such as the nature of the human neural substrate underlying language processing, the characteristics of our articulatory and auditory systems, principles of efficiency of information storage and information transmission, the nature of concept representations, and human parental and social relationships, among others, constrain the structure of the mental lexicon. Some of these factors are universal and some are not. All normal human newborns are capable of learning any human language, so factors originating in the processor (any elements of human hardware related to speech perception, processing and production) must be universal. Other factors such as social and interactional pressures and language-external influences (such as the concepts that speakers can talk about), vary across societies, and therefore are not universal. Cross language analyses of the information in speech can help determine the universality of the pressures.

The mental lexicon must adapt to optimise language processing, and this has to happen efficiently over the brain substrate supporting the lexicon, which has its own properties and limitations.

1.1.1.1 Homeostasis

The structure of the lexicon must be flexible enough to adapt to all the pressures that act on it. On the other hand it must be robust enough to maintain its identity. The pressures mentioned above can be viewed as dimensions in an adaptive landscape. The lexicon is constantly adapting to changes in these pressures in order to find an optimal state in the landscape at each moment. The mechanism by which the lexicon, as a complex system, is able to juggle all those often contrary pressures is called homeostasis, defined in the Merriam Webster Online Dictionary as 'a relatively stable state of equilibrium or a tendency toward such a state between the different but interdependent elements or groups of elements of an organism, population, or group'.

De Rosnay (1997) states that homeostasis is the essential condition for the stability and survival of complex systems. It helps the system withstand the multitude of pressures that act on it. Homeostasis' main effect is a resistance to change, but it accommodates necessary alterations. The system reacts to every change in the environment in order to maintain the internal balances. Aitchison (2001) emphasizes 'the extraordinarily strong tendency of language to maintain and neaten its patterns'. As happens in the 'butterfly effect' in another complex system, the weather, the reactions are unpredictable or even counterintuitive (Forrester, 1975). When one expects an effect as the result of a precise action, an unexpected and often contrary action occurs instead. This is so because of the complexity of the system and of the relationships between its elements. In pharmacology, a new drug that treats one problem usually has many unforeseen collateral effects. This happens because of the intricate interrelations within physiological systems. In the context of trying to establish a plausible origin of language, Keller (1994) offers a generalisation of Mandeville's paradox[2] that highlights the fact that individual actions - prompted by individual selfish ('bad') motives - can have emergent effects that are positive for the society as a whole. Aitchison (2001) provides some examples of attempts by language to restore a structure equilibrium which have 'lead to quite massive, unforeseen disruptive changes, which trigger one another off in a long sequence'. The disruption has always been kept in check by homeostatic pressures, which is proven by the facts that language has never ceased to be learned by humans nor stopped serving its purpose as a system of human communication. In a complex system such as the lexicon, the consequences of one change to one aspect of the representation of one element can potentially reach the whole system, as illustrated by examples of sound shifts such as Grimm's Law (Bammesberger, 1992, cited in Aitchison, 2001). Grimm's Law describes how in the Germanic branch of Proto-Indo-European [bh][dh][gh] became [b][d][g]; [b][d][g] became [p][t][k] and [p][t][k]became [f][th][h]. Another example is the American Northern cities vowel shift, which is still taking place (Labov, 1994, cited in Aitchison, 2001).

[2] Mandeville published in 1705 a poem entitled 'The fable of the bees', whose leitmotiv was that every single individual vice made a beneficial contribution to the well-being of society (Keller 1994).

These long-range effects are also present in distributed connectionist models of the mental lexicon, where the processing of each word is affected by the whole lexicon. In a distributed representation, each node in the network participates in the representation of all the words. Examples of this are Seidenberg and McClelland's (1989) feedforward network model of reading; Plaut, McClelland, Seidenberg and Patterson's (1986) attractor network model of reading; Hinton and Shallice's (1991) model of dyslexia based on a lesioned attractor network; and Gaskell and Marslen-Wilson's (2002) study of the effects of the competition between phonological and semantic aspects of word representations.

A homeostatic lexicon is a delicately balanced system where a change in one of the levels of representation of one element may affect the structure of the whole system. A change in a word's meaning or in its syntactic use, or a growing trend to pronounce a vowel differently will have consequences for all the words in the lexicon. Other factors such as the pressure for being a useful communication tool for humans, or the pressure for being easy to learn by human infants keep the possible changes in check.

1.1.1.2 The principal pressures: communication and learnability

Among the pressures operating on the lexicon described above I emphasize the preservation of (or the quest for) structural characteristics of the lexicon that allow humans to communicate. The main such characteristic is a correspondence between the conceptual and the linguistic domains through symbolic reference, the uniquely human system of reference (Deacon, 1997). Peirce (1897, 1903) proposed three levels of reference: iconic, indexical and symbolic. Icons are signs that resemble the objects they stand for, such as a photograph of a dog representing a dog. Indexes indicate or provide clues; as to what their references are, for instance the symptoms of a disease or a thermometer for the temperature. In symbols, the relationship between sign and meaning is arbitrary; examples of symbols include words (where the form does not resemble neither does it indicate the meaning), colour codes, and in general any form of conventional language.

Other traits of the lexicon that allow humans to communicate efficiently are adaptations to the pressures imposed by language production, perception and processing capacities. The phonological inventory, speech rate, prosody and other traits of speech are adapted to characteristics of the organs of speech and of the auditory system, which constrain the sounds we can produce and perceive. The temporal nature of speech and the potential noise in the environment affect the structure of the phonological information transmitted in utterances (this theme is developed in chapter two). Memory capacity affects, for example, the amount of information that can be stored in the short-term memory in order to process aspects of language. For instance, Baddeley, Thomson and Buchanan (1974) described the word length effect whereby lists of short words are easier to recall than lists of long words. It also affects how many items we can store in the short-term memory in order to process the syntactic relationships between them (see Caplan & Waters, 1999 for recent review).

Some authors have emphasized the role of learnability in shaping language. Kirby and Hurford (2002) propose that universal language characteristics are ultimately adaptations to the successful transmission of language from individual to individual and from generation to generation, as is well exemplified in their iterated learning model (ILM), where the key pressure behind the emergence of a linguistic trait (e.g. compositionality) is the cultural transmission bottleneck. A key conclusion of the ILM for compositionality emergence is that if the training set is too small, it does not allow for generalisation, and if it is too large, the pressure to generalize diminishes and holistic (non-compositional) languages are equally adaptive. Compositionality is an adaptation of language whose function is to make language learnable given the 'poverty of the stimulus' – the fact that children are exposed to limited, imperfect linguistic data, and that they receive practically no feedback on their performance. Here the environmental constraint is the nature of the human relations that lead to the poverty of the stimulus, the adaptation that language has evolved to match it is compositionality. Kirby, Smith and Brighton (2004) also emphasize learnability as the main pressure acting on language. Deacon (1997) writes: 'The structure of language is under intense selection pressure because in its reproduction from generation to generation it must pass through a narrow bottleneck: children's minds. (...) So, languages should change through history in ways that tend to conform to children's expectations; those that employ a more kid-friendly logic should come to outnumber and replace those that don't'.

The pressures acting on language (or the lexicon) can be explained at two different levels of adaptation: the phylogenetic and the cultural levels. Phylogenetic explanations focus on the emergence of language within the evolution of the *Homo* species since its appearance between 10^7 and 10^6 years ago. During this time, parts of the anatomy of humans evolved in such a way that language became possible. Language is an adaptation with a function in human societies that make humans fitter. Language is a human phenotypic trait with a genetic basis which is the object of natural selection (Deacon, 1997; Hurford, 1989; Komarova & Nowak, 2003; Nowak & Komarova, 2001; Pinker & Bloom, 1990; see also Wagner, Reggia, Uriagereka, & Wilkinson, 2003 for an exhaustive review of computational simulations of language emergence). If the evolutionary advantage that language conferred on humans is that of communication, then the neuroanatomical traits that allow language to serve as a communication tool are under great pressure to be preserved. During hominid evolution, language adapted to being easily learned, and/or the brain adapted to learn it easily. From the point of view of the structure of language, the traits that make it easily learnable are also under great pressure to be preserved.

Cultural evolution explanations focus on the evolution of language since the appearance of *Homo sapiens sapiens* between 10^5 and 10^4 years ago. In this evolutionarily short time, human anatomy has remained essentially unchanged. Lexicon traits are adaptations whose function ultimately leads to the better transmission of language and to the better communication of concepts. Lexicon traits are coded in a transmissible replicable medium – speech - and are the object of lexicon selection. Croft’s (2000) ‘linguemes’, Kirby's (1999) 'variants', Mufwene’s (2001) ‘linguistic features’, Nettle's (1999) 'linguistic items' or Worden’s (2000) ‘word feature structures’ are pieces of linguistic information that are selected for or against in the framework of linguistic evolution. In these studies, human fitness is only one environmental factor, and not the driver of evolution. The drivers of evolution are the reproducibility of language itself, that is, its ability to be replicated in successive generations of humans; in other words, its adaptation to be learned by human infants.

These two explanations complement each other by focusing on different timescales and levels at which language evolution takes place. In the phylogenetic approach, the emergence of language in humans, there is also a place for a cultural evolution approach. A joint approach sheds light on the co-evolution of language and humans. In the phylogenetic timescale, an evolving processor precipitated language evolution. In the cultural timescale, languages are stable with respect to their developmental environment, the processor, because this environment is stable, and other pressures come to the foreground to guide the cultural evolution of language.

This thesis relates to cultural explanations of language evolution, presenting the ever-changing lexicon as the result of juggling the multiple pressures that act on it. What is crucial in my arguments is that the pressures leave their mark in the structure of the lexicon. My main assumption is that an analysis of the different pressures can help characterize the structure of the lexicon, and conversely, the structure of the lexicon can be explained in terms of the pressures that operate on it. I assume that the lexicon is a complex system with emergent properties that cannot be attributed to any one of its elements, but are only apparent when the system is taken as a whole. I look for those properties at different levels of representation, such as the phonological and the syntactic-semantic level. I also look for emergent properties involving both of those levels simultaneously, and propose that one of the pressures acting on the structure of the mental lexicon is the tendency towards systematicity, or structure-preserving mappings across different levels of representation.

1.1.1.3 Summary

This section has defined the mental lexicon as a complex entity embodying at least the principal elements of language, which is able to change its structure adaptively to accommodate external pressures while preserving a structure that allows it to serve as a communication system between humans and to be learned easily by infants. The following chapters of the thesis will offer explanations of the structure of the mental lexicon that take into account the pressures that operate on it. Summing up,

- The mental lexicon is an embodiment of language where linguistic information is encoded in the complex relationships between words at many levels.
- The lexicon is a complex adaptive structure that constantly responds to the many pressures that operate on it.
- This thesis attempts to explain characteristics of the structure of the lexicon as adaptations to those pressures.

1.1.2 Statistical learning

The methods I employ in the thesis use speech data to infer the structure of the mental lexicon. I am therefore assuming that the brain is able to perform complex calculations on the input it receives from speech in order to develop and subsequently adjust the mental lexicon. This is only possible if it is sensitive to statistical patterns of information in speech, and this sensitivity is explained by the statistical learning literature.

Throughout this thesis I assume that most of the support for language acquisition is not in the human brain but in the structure of language itself (Deacon, 1997). This is an application to language of the more general principle proposed by Anderson (1991) that 'the mind has the structure it has because the world has the structure it has'. The methods described below help reveal the complex patterns of information and of organisation embedded in the structure of the lexicon. The patterns are probabilistic or statistical, such as the calculation of information as entropy or the definition of the position of a word in the semantic space as a pattern of cooccurrences with other words.

During language acquisition these patterns are extracted by infants from the linguistic environment, and assimilated to gradually configure the lexicon structure of their language. This presupposes that human infants are sensitive to statistical patterns in the input speech. This section reviews studies that show mechanisms for the statistical learning of various aspects of language.

Plunkett (1997) offers a review of multidisciplinary approaches to early speech perception, word recognition, word learning and the acquisition of grammatical inflections which, he suggests, demonstrate 'how linguistic development can be driven by the interaction of general learning mechanisms, highly sensitive to particular statistical regularities in the input, with a richly structured environment which provides the necessary ingredients for the emergence of linguistic representations that support mature language processing'. Redington and Chater (1997) review successful computational probabilistic and distributional approaches to speech segmentation and the acquisition of inflectional morphology, syntactic category and lexical semantics and end their review with an optimistic note that a combination of different sources of information might one day attain close to human performance. The volume edited by Bod, Jay and Jannedy (2003) contains probabilistic approaches not only to phonology, morphology, syntax and semantics, but also to sociolinguistics and language change. The following sections briefly review the literature of statistical learning of phonology, syntax and semantics.

1.1.2.1 Phonology

Probabilistic cues in speech help infants to acquire the phonological system of their language. Maye, Werker and Gerken (2002) showed that phonological categories are inferred from statistical modes in use of the phonetic space: they determined that infants are sensitive to patterns in input speech to track the distribution of speech sounds in their mother tongue.

They familiarized 6 and 8 month old infants to unimodal and bimodal distributions of instances of use of a continuum of speech sounds based on a phonetic contrast, and then tested the infants' categorisation of the sounds. The unimodal distribution should indicate that the contrast is linguistically irrelevant and, as predicted, children exposed to it acquired a one-category representation of the sounds. The bimodal distribution should indicate that the contrast is linguistically important, and children exposed to it acquired a two-category representation. Pierrehumbert (2003a) argues that infants learn from superficial statistical properties of speech, but later on, when the lexicon is well developed, the phonological system is refined by internal feedback from type statistics over the lexicon (phonotactic information). This later refinement exploits the confluences across levels of representation that make bootstrapping and generalisation possible.

Peperkamp and Dupoux (2002) studied how infants who still do not have a semantic lexicon might infer the underlying word forms that appear in speech as different phonological variants. They examined word phonological variants containing phonemes and allophones in different languages and proposed a learning mechanism based on an examination of the distribution of either surface segments[3] or surface word forms. They conclude that semantic knowledge is unnecessary to retrieve word forms from a structured set of variant instances of the word in speech.

Maye, Werker and Gerken (2002) suggest that 'in addition to its probable role in speech perception, this sensitivity [to probabilistic patterns in speech] contributes to word segmentation (Saffran, 2001; Saffran, Aslin & Newport,

[3] The word 'segment' is used synonymously with 'phoneme' as a more theory-neutral term.

1996; Saffran, Newport & Aslin, 1996; Christiansen, Allen & Seideberg, 1998) and the acquisition of constraints on speech sound sequences (Jusczyk, Luce and Charles-Luce, 1994 & Zamuner, 2001) and grammatical structure (Gómez & Gerken, 1999; and Saffran, 2001)'.

For recent reviews on statistical phonological learning, see Peperkamp (2003) and Pierrehumbert (2003b).

1.1.2.2 Syntax

Redington and Chater (1997) differentiate between language external and language internal approaches to learning syntactic categories. Semantic bootstrapping (Pinker, 1984) is a language external approach that exploits the correlation between word categories (especially noun and verb) and objects and actions in the environment. Language internal approaches can make use of regularities between phonology and syntactic categories (Kelly, 1992), regularities between intonation and syntactic structure (Morgan & Newport, 1981) and distributional analysis. Cooccurrence statistics is a type of distributional information that can be extracted with computational or connectionist methods. It creates word representations that capture the cooccurrences of target and context words in a corpus within a small window (typically between 2 and 10 words), which reflect syntactic category. Mintz (2003) introduces the idea of 'frames', or frequent combinations of two words with one intervening word, and argues its validity to predict syntactic category. Connectionist models use a form of Hebbian learning to capture the cooccurrence statistics of the corpus in the weights of the network (Rumelhart & McClelland, 1986).

Manning (2003) reviews the trend that linguistic units are continuous and quantitative in contrast with generative grammar's discrete and qualitative units. He motivates a probabilistic approach to syntax acquisition with arguments from language acquisition, language change, and typological and social variation, and exemplifies it with an exploration of verb subcategorizaton.

1.1.2.3 Semantics

Some recent statistical approaches to semantic learning focus particularly on learning symbolic association, the relationship between an object and its name. Smith and Vogt's (2004) cross-situational statistical learning model suggests that learning word form-meaning pairings is achieved through inference over multiple contexts; the form that more consistently co-occurs with one aspect of the context will be assigned to that aspect. Symbolic links are under the pressure of biases such as 'mutual exclusivity' - the tendency towards a one-to-one mapping between forms and meanings (Marksman, 1989; Merriman & Bowman, 1989; see also Smith, 2004) or the 'whole object bias' – the fact that when children learn a new word, they prefer to associate it with an object rather than with a feature of an object or an action (McNamara, 1982). Unpublished research by Houston-Price (2004) supports a distributional mechanism for the acquisition of the mapping between a label and an object. She showed 15 and 18 month infants two visual stimuli. The infants then heard a label, and a few seconds later the target stimulus would move (salience condition) or the image of a face would turn towards it, and the infants heard the label again. The infants never learned the association of the label and the correct stimulus, but did learn the association of the label to the *incorrect* stimulus in the salience condition. This means that they associated the label with the stimulus that was more consistent during the two presentations of the label, namely the *stationary* image. This supports the view that acquisition of naming is mediated by probabilistic distributions of features of the stimuli together with probabilistic distributions of occurrences of the labels.

Word cooccurrence statistics, apart from capturing syntactic category, are also widely used to construct semantic representations (Lund & Burgess, 1996; Schultz, 1993; Landauer & Dumais, 1997; and McDonald, 2000; see also Curran, 2004, for review). Cooccurrence statistics capturing semantics will be further reviewed and developed in chapter four.

Having seen a definition of a complex, adaptive mental lexicon; the corpus-based methodologies employed in the thesis; and a review of statistical learning that psycholinguistically grounds the corpus-based methodologies, I now present an overview of chapters two to six of the thesis.

1.2 Thesis overview

1.2.1 Methods

Underlying the whole thesis is the assumption that the information necessary to configure the mental lexicon in a human brain is found in speech (and text). Language acquisition is possible thanks to the human brain's sensitivity to the linguistic information in speech – humans raised in the absence of linguistic input do not develop language. The lexicon configuration and associations between words at all levels change over a speakers' lifetime owing to exposure to more and more speech (and text). The main point is that relevant analyses of speech should reveal the patterns of information that shape the mental lexicon.

Most of the research reported in this thesis is corpus-based. The main corpus used is the 'Corpus oral de referencia del Español' (Marcos Marin, 1992), a one million word Spanish transcribed speech corpus compiled in Spain in the early 1990's. I also use parts of a Spanish transcribed speech corpus of interactions between a young child and her close relatives, the ‘Maria’ corpus (Lopez Ornat, 1994). Both corpora include Spanish spoken in Spain only, mainly the standard Castilian variety, which is important for consistency. Corpora are assumed to be representative samples of a language. In the case of Spanish only a one million word corpus of transcribed speech was available for research purposes, although larger news text corpora were available. A corpus of transcribed speech, with all its grammatical and speech errors, repetitions etc, is a more accurate representation of the external manifestation of language than a corpus of written text. Consequently, I chose to work with the speech corpus and relied on statistical analyses to reduce the impact of the size of the corpus.

One consequence of using two contemporary corpora is that it offers a synchronic snapshot of Spanish. I have not presented any diachronic analyses, although some of the conclusions reached in the thesis have implications for language change. In the conclusion chapter I briefly sketch a theoretical framework for the evolution of the adaptive lexicon that takes into account the findings of the thesis.

The one non corpus-based study is the empirical exploration of the phonological space presented in chapter three. Here I used a psycholinguistic-inspired forced-choice paradigm to collect participants’ impressions on phonological similarity between words. The data collection was done on the Internet in order to reach Spanish speakers living in Spain. The results are then used in chapter five to construct a quantitative representation of the phonological level of the lexicon based on word similarity, and also in chapter six to ground corpus-based findings about phonological similarity.

The corpus data has been subject to several very different analyses, in many cases based only on the words of two consonant-vowel structures: cvcv, such as *mesa, pelo, mira,* and cvccv, such as *banco, tengo* or *marca*. In chapter two I use a method drawn from Information theory first developed to improve encryption techniques. I extract all the words of structure e.g. cvcv and consider the entropy - an elaboration of the probability distributions - of each phone for each segment position. The result quantifies the distribution of phonological information over words of that structure.

Chapter four exploits a very different type of information contained in the corpus. It focuses on word distributional information, which quantifies the degree to which words tend to occur near each other in speech. This is assumed in the literature to capture syntactic and semantic information, and I use two versions of the corpus (the surface forms and the lemmatised forms) and two measures of cooccurrence (one including content and function words in the calculations, and other including content words only) to establish the effect of those parameters in the kind of information captured by the cooccurrence representation of speech.

Chapters five and six bring together the phonological information obtained in chapter three and the semantic and syntactic information extracted in chapter four and tests the existence of systematicity between them. These two chapters are based on the calculation of Fisher divergence, extensively described in chapter five, developed to calculate the correlation between similarity matrices, and a Monte-Carlo analysis to measure the significance of such a correlation.

Chapter six uses methods from Artificial Intelligence to explore the effects of the parameters of phonological similarity between words on the correlation found in chapter five. A random search of the parameter space returns information on the behaviour of the phonological parameters with respect to the phonology-semantics systematicity. A hill-climbing search finds the parameter configuration that yields the optimal phonology-semantics systematicity. The novelty in this methodology lies in the use of *phonology-semantics* systematicity to characterize the function of different aspects of the *phonological* lexicon.

Having briefly presented the methodologies employed in the thesis, I now outline the structure of the thesis.

1.2.2 Thesis organisation

The lexicon is shaped, among others, by communication pressures. Words need to be stored in the representational space; transmitted over a potentially noisy channel, and decoded by the listener, and all of this has to happen in an efficient way. In chapter two I employ Information theory tools to measure the efficiency of lexical information storage, transmission and decoding. Comparing the within-word information and redundancy distributions of different representations of a set of words can help determine how well adapted each representation is to the requirements of its representational space.

I propose a lexicon architecture where words are represented over different levels - phonological, syntactic, semantic. The lexicon is defined by relationships of similarity between words at each level. The position of each word in such a structure is given by its similarity to every other word at each level. Chapters three and four are devoted to quantitatively define the phonological and the syntax-semantic lexicon levels, respectively.

Chapter three presents an empirical exploration of the phonological similarity space: a psycholinguistic paradigm obtains relative values for several parameters of phonological similarity between two words such as ‘sharing the first consonant’ or ‘having the same stressed vowel in final position’. These results fill a gap in the literature on phonological similarity in general, and of Spanish in particular. I offer explanations of the resulting parameter values, linking them to the psycholinguistic literature. These phonological similarity parameter values, derived from human judgements, are used later in the thesis (chapter five) to calculate the phonological similarity between words in two subsets of the lexicon. The pairwise similarity measures obtained are a quantitative expression of the configuration of the phonological level of the lexicon.

Chapter four deals with the semantic level of the lexicon, also configured as the set of all pairwise similarity values between words. In this case the similarity between two words is based on whether they tend to occur close to the same words in speech. I review cooccurrence-based similarity metrics and then construct a syntactic-semantic lexicon similarity representation. I explore what types of information we can extract from cooccurrence-based word representations of the Spanish lexicon, such as information on syntactic category, gender, and meaning. Since syntactic gender is not present in English, the analysis of gender classification is a novel application of cooccurrence-based word representations.

The structure of the lexicon is constrained by the nature of its neural substrate, so characterizing the organisation of aspects of language can help infer aspects of the nature of the underlying brain substrate, and this knowledge can help design language processing computer architectures capable of similar functions. Conversely, the existing knowledge in neuroanatomy and neurophysiology can help narrow the choice of the possible brain architectures that would support our lexicon. In this respect, when the brain needs to deal with complex inputs, it tends to use representations that preserve the structure of stimuli over different parts of the brain. The best studied brain systems that are closest to language in anatomical and functional terms - visual, auditory, somatosensory and motor - use structure-preserving (systematic) representations at their various processing stages, so I propose that they are a good candidate to be the ones used in language processing.

With this in mind, I assume that structure-preserving systematicity is a general property of the neural substrate, and that the relationships between the conceptual and the linguistic levels are crucial in the structure of the lexicon. Therefore, we should see traces of that systematicity between the semantic level and other linguistic levels such as phonology and syntax. This brings us to the main hypothesis in chapter five, namely that there is a pressure towards a structure-preserving mapping between the phonological and the syntax-semantic lexical representations. This hypothesis has been tested for English. Shillcock, Kirby, McDonald and Brew (2001, *submitted*) showed that words that occur in similar contexts are more phonologically similar than expected by chance. In chapter five I replicate their study with a slightly different methodology to test whether the same is true for Spanish, which would support the universality of systematicity between word phonology and syntax-semantics. If this correlation is universal, then it would have to be explained in terms of the evolution of the lexicon against the background of universal pressures. Chapter five continues with an attempt to remove syntactic information from the data and the methods in order to establish the influence of a correlation between the form and the meaning levels of the lexicon. Finally, I replicate Shillcock et al.'s (2001) methodology again to test whether, as they found in English, the correlation is driven by 'communicatively salient' words. The results obtained with my limited data suggest that this is also the case in Spanish.

The phonology-semantics correlation is not without problems. It seems to challenge the Saussurean principle of arbitrariness of the sign, which states that the form of a word is arbitrary and independent of its meaning; according to Saussure, a dog could equally be called ‘caterpillar’, but a structure-preserving mapping between the space of form representations and context representations would not allow this. I address this philosophical issue and suggest that a dog could indeed suddenly be called ‘caterpillar’, but then the rest of word forms in the lexicon would also need to change to accommodate this change. While for any one word the form is independent of the meaning, the relationships between forms are not independent from the relationships between meanings.

The correlation found in chapter five is the basis of further explorations in chapter six, where the measure of systematicity is used to quantify and characterize the phonological lexicon. I use the correlation between phonology and *semantics* to build a quantitative *phonological* space for Spanish. This methodology reveals the effects of robust pressures working not only towards but also against systematicity between phonology and syntax-semantics.

To sum up,

- This thesis explores the structure of language embodied in the mental lexicon, a complex system of words organized along many dimensions and defined in terms of relationships of similarity between each other.

- The exploratory methods in this thesis seek to define the structure of different levels of the mental lexicon by quantifying relationships within and between words. I explain these levels of organisation in terms of the pressures acting on the lexicon at the phonological and syntax-semantic levels, and also of a pressure for systematicity across those different levels of representation.
- The novel applications of the methods are intended to explore new ways to tackle the structure of the mental lexicon and to pave the way for further research with larger corpora and different languages as well as with experimental paradigms that can offer new insights into the forces that shape language.

Chapter 2. The distribution of phonological information within Spanish words

This chapter[1] describes the information profile, a measure of the distribution of phonological information across segments in a word set. It calculates the information profiles of a representation of the mental lexicon and of the words uttered in speech. Assuming that the lexicon structure is the result of the pressures that act on it, I propose that the intra-word distribution of phonological information can be explained in terms of the pressures acting on the mental lexicon representation in the brain and on the phonological representation of speech.

2.1 Introduction

In this chapter I measure the distribution of phonological information in Spanish words, and try to explain it as adaptations to aspects of the two main pressures we saw in chapter one: serving as a tool for human communication and being easy to learn by humans.

A word has to meet several phonological requirements in order to be part of a language. It has to be distinguishable from the other words, and it has to use the segments[2] and conform to the phonotactics of the language. In this chapter I suggest it also tends to fit in with the rest of the words in terms of its phonological information structure. I assume the principle behind the rational analysis proposed by Anderson (1991) that the output of cognitive processes is an optimal response to the information-processing demands of the environment, and hypothesize that the information structure of the

[1] Parts of this chapter were contained in Tamariz and Shillcock (2001); a copy of this paper is included as Appendix H.

[2] The word 'segmen't is used synonymously with 'phoneme' as a more theory-neutral term.

mental lexicon reflects the processing demands of the lexicon substrate in the brain. Additionally, I assume that the information structure of words as sequences of sounds in speech reflects the information processing and communication demands of the potentially noisy medium over which they are transmitted -words are produced in a way that minimizes the loss of information due to the noise in the communication channel and to potential misperception by the listener.

In view of these assumptions I propose that the distribution of phonological information of word representations at different levels of abstraction can help reveal the optimal processing solutions arrived at by the brain. I examine the information structure of the lexicon (the collection of word types we use when we speak) and of the words used in speech (word tokens) expecting to see adaptations to the demands of the different representational spaces in which the two word systems occur. In addition to this, I analyze the information structure of words from a corpus of child-directed speech, assuming the demands of the small, growing lexicon of a young child are different from those of a larger, more stable adult lexicon, and that adults have evolved a way of speaking to children that optimally meets those demands (see Elliot, 1981 on child-directed speech).

I use the concepts of entropy and redundancy from communication theory introduced by Shannon (1948) to construct the information profile of word systems. These profiles reflect the distribution of information across the segment positions of words, and I argue that they can be used to measure how suitably and efficiently each word representation (the lexicon and speech) is adapted to the demands of its representational space.

In 1975, Grice formulated the ‘cooperation principle’ in a series of maxims or rules stating what a good conversation contribution should be like, but that are applicable to all forms of human communication:

- Maxim of quantity (make your contribution as informative as is required, but not more informative than is required).

- Maxim of quality (try to make a contribution that is true).
- Maxim of relevance (be relevant).
- Maxim of manner (be perspicuous: avoid obscurity and ambiguity; be brief and orderly).

The maxims of quantity and manner are particularly relevant to this chapter. The phonological lexicon should contain enough information to allow clear and unambiguous communication, but at the same time it has to be stored in the limited representational space in the brain. Speech should be economical without compromising clarity.

In this corpus-based study, I assume the set of word types in the corpus represent the phonological forms of words in the mental lexicon, subject mainly to the pressure to optimize the usage of the available representational space. If this was the only pressure acting on the phonological lexicon, all segments would be as likely to occur in all positions, in fact all the possible segment combinations would exist as words. However, in reality other pressures interact with optimal storage – phonotactics limit the possible segment combinations, the pressure for prompt lexical recognition concentrates the information towards the beginning of words, and efficient processing constraints press for a monotonic structure where information is incremental and where later information does not invalidate or contradict earlier information.

The child-directed types correspond to the developing mental lexicon. We should expect to see an evolution towards the mature mental lexicon structure in samples of speech directed to an increasingly older child. The developing mental lexicon is laying out the scaffolding of its structure, and can be expected to prioritize clarity over optimisation of storage. The main pressures here are to adapt to the processor sensitivities, reflected in phonological and lexical acquisition.

Secondly, I assume that the word tokens, including frequency information, represent the phonological form of speech. Speech is subjected to the pressures of communication. The main pressure on communication is that of efficient transmission of the acoustic signal from the speaker to the hearer in the face of potential noise. Pressures to overcome potential noise should be the same in adults as in children, so in the child-directed token set, as in the case of adult tokens, we can expect that an important constraint in terms of information content should be efficient transmission. The phonological form of speech is also subject to other processing constraints. One of the crucial problems facing speech recognition is the segmentation of the speech stream into words. I expect to find an information profile of the word phonology that facilitates segmentation, both in child-directed and in normal speech.

I first look at the corpora from which the data are extracted and explain how to build the information profiles using the concept of entropy. Then I look at the information profiles of the different word systems to see how they reflect the efficiency of the representations in the face of the demands of their respective representational spaces.

2.2 Data

Our main source of data is the 'Corpus oral de referencia del español' an orthographical Spanish speech corpus (Marcos Marín, 1992) containing one million words (with a vocabulary of 41,000 word types). This corpus is made up of transcribed recordings ranging from everyday conversations to radio broadcasts and technical and scientific addresses. In the second part of the study I also use a corpus of Spanish speech addressed to a child extracted from the 83,000 word corpus "Maria" (Lopez Ornat, 1994), described in more detail in § 2.6.1 below.

Note that this study includes unfinished words, mistakes and, crucially, all derived and inflected words, assuming in principle the Full Listing hypothesis (Jackendoff, 1975; Butterworth, 1983) whereby all word forms are stored in the mental lexicon and productive morphological rules are only used to produce or understand novel words.

2.2.1 Transcription

Both corpora, ‘Corpus oral de referencia del español’ and ‘Maria’, are orthographically transcribed. In the absence of a large speech corpus phonetically transcribed by experts, these are the best large speech data sets available for Spanish. The corpora were automatically transcribed using 50 phonemes and allophones: vowels: /a/, /e/, /i/, /o/, /u/, /á/, /é/, /í/, /ó/, /ú/. Consonants: /p/, /b/, /t/, /d/, /k/, /g/, /m/, /n/, /ɲ/, /ɾ/, /r/, /f/, /ɵ/, /s/, /ɟ/, /χ/, /l/, /ʎ/, /tʃ/; semivowels: /i/, /u/; semiconsonant /j/, voiced approximants /β/, /ð/, /ɣ/, voiceless approximants /β/, /ð/, /ɣ/, labiodental /m/, dental /n/ and /l/, palatalised /n/ and /l/, velarized /n/, /z/, dental voiced /s/, dental /s/, fricative /ɾ/, voiced /ɵ/ and a silenced consonant /Ø/. The transcription included phoneme interactions such as assimilation, following the rules set out in Rios Mestre (1999). Diphthongs were treated as two separate segments, as is usual in Spanish phonological research. The corpora were divided into chunks separated by pauses - change of speaker, punctuation mark and pause marked in the corpus. The resulting text was automatically transcribed word by word and then phoneme interactions were introduced at word boundaries within the chunks, following the same rules as for the intra-word transcription.

2.2.2 Word sets

For a clearer picture of the profiles, particularly towards the end of the word, I work with four sets of equal length words: all the 4, 5, 6 and 7 segment long words from the corpus. Word recognition typically occurs before the end of the word is uttered (Marslen-Wilson & Tyler, 1980), and information about word-length is usually available once the nucleus is being processed (Grosjean, 1985), so I assume an idealised processing where recognition processes are restricting their activities to the subset of words in the lexicon that match the word being uttered in terms of approximate overall length. The particular word lengths were chosen because the structure of shorter words is simpler - although I did not use the 1, 2 and 3 segment long words because they do not allow for so much incremental interpretation, meaning that for the set of same length words, the first segment carries most of the information of the word. Also, working with equal-length word groups means less variation in the internal structure of each word-length group, which could potentially obscure the word-internal information distribution. Additionally, the selected word lengths are equidistant from the modes of the word-length distributions of the types and the tokens (see Figure 2.1). The sum of these four word lengths accounts for 37% of the tokens and 45% of the types. So any conclusions of this study are restricted to words of intermediate length (4-7 segments).

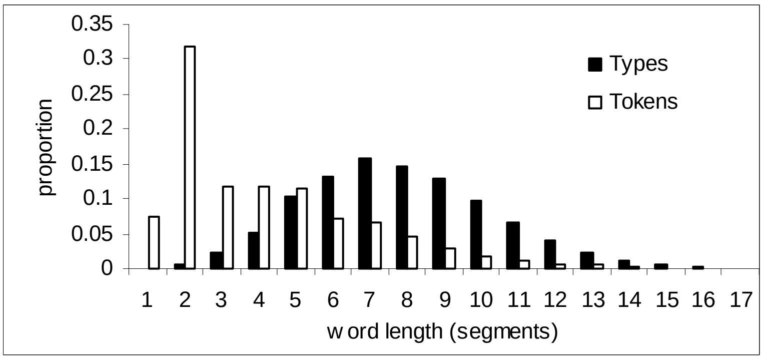


Figure 2.1. Word-length distribution in the 42,000 tokens and one million types from the speech corpus.

2.3 Methodology

2.3.1 Entropy

Goldsmith (2000) suggests that information theory concepts such as probability and entropy are the natural quantitative measures of many of the concepts used by linguists in general and by phonologists in particular.

I will use entropy as introduced by Shannon (1948), and further developed as a tool to estimate the efficiency of communication and information (Shannon, 1951). Recent developments in new information technologies have highlighted the need for an efficient way to store and transmit information, and entropy has been used extensively in encryption studies (e.g. Cachin, 1997; Van Droogenbroek & Delvaux, 2002), and also in speech recognition studies (e.g. Yannakoudakis & Hutton, 1992; Shen, Hung & Lee, 1998), speech production studies (Van Son & Pols, 2003) and in natural language processing (e.g. Berger, Della Pietra & Della Pietra, 1996).

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. For example, the finite scheme formed by the possible outcomes when throwing a dice has maximum entropy: each of the six sides of the dice has 1/6 probability of occurring and it is very difficult to predict what the outcome will be (high entropy). A loaded dice, on the other hand, has an unequal probability distribution, and the outcome is less uncertain (low entropy), with, say, number three having a ½ probability of occurring. Entropy is a statistical measure of irregularity and it has been defined as the amount of surprise experienced when encountering a new element.

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

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

(Note that I use base 2 logarithms throughout this chapter.)

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 the relative entropy allows us to compare entropies from systems with a different number of events.

$H_{max} = \log n$

(where n is the number of possible outcomes);

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

Redundancy 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. If entropy reflects irregularity, redundancy measures regularity in a system. Redundancy is defined as:

$R = 1 - H_{rel}$

2. 3. 2 Calculation of the Information Profile

The entropy and redundancy of letters in a text has been used in corpus studies before, for instance in the calculation of the entropy of letters in a dictionary of over 93 thousand words (Yannakoudakis & Angelidakis, 1988). This study examines the variation of entropy within words. I use the information profile of a set of words as calculated by Shillcock, Hicks, Cairns, Chater and Levy (1995) and Tamariz and Shillcock (2001) - a plot of the relative entropy of each segment position of a set of words.

The information profile is a plot of the entropy calculated for each segment position of a set of words of equal length, using the set of Spanish segments as the finite scheme - for each position, the possible 'outcomes' are the Spanish segments (/a/, /b/, /d/, /e/, /f/ etc).

The calculation of the entropy for each position is as follows. First, count the occurrences of each segment in the position. The probability of each segment is the number of actual occurrences divided by the total number of all segments in that position. The entropy for that position is given by the sum of the products of the probability of each segment multiplied by its logarithm. This entropy divided by the maximum entropy for that position equals the relative entropy.

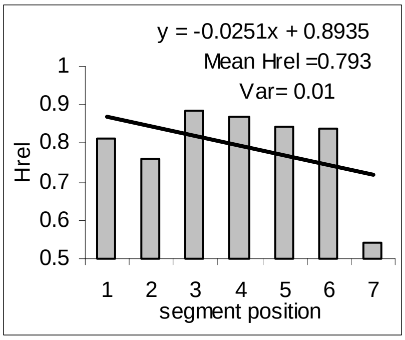


Figure 2.2. Information profile of words of length 7. The linear trend line equation is shown above with a slope (m) value of -0.0251. The mean relative entropy (H_{rel}) value and the variance are also shown.

The plot of the relative entropy for all word positions is the information profile (e.g. the one shown in Figure 2.2). The shape of the information profile can be represented by the slope[3] of the linear trend line of the entropy values. Equation (1) gives the linear trend line:

(1) $y = mx + n$

In equation (1), (m) is the slope, and it indicates the overall shape of the information profile, particularly the difference in entropy levels at the beginning and the end of the word. We have to take into account that for equally shaped information profiles, the slope gets flatter as word-length increases. Figure 2.2 shows a typical information profile, with entropy rising after the first two positions and dropping sharply in the final position. The slope (m) shows a negative value, indicating that the linear trend line drops towards the end of the words. A positive value would indicate a line rising from left to right, and a zero, a perfectly horizontal trend line. Because all the slope values obtained are negative, the figures below will show (-m) for

[3] Another possible representation of the information profile is the variance of segmental entropy values. The slope retains information about overall shape of the profile, but it is not normalised for word length. The variance normalizes for word-length, but it loses the information about the overall shape of the profile. The results of the comparisons presented in this chapter are the same whether we use the variance or the slope of the linear trend line, with higher variances correlated with steeper slopes.

clarity. I will also examine the mean relative entropy across the profile, simply the average of the segmental relative entropy values.

The information profile as defined above means that, as we perceive the sequence of speech segments, we have information about how predictable or unpredictable each segment is.

I now report some applications of the information profile and then go on to compare the information profile of words in speech and in the mental lexicon.

2.4 Applications of the information profile

2.4.1 The levelling effect of accurate word representations

Noting that in the absence of other constraints, the phonological information profile of words would tend to be flat, with information distributed evenly across word segments, Tamariz and Shillcock (2001) proposed the principle that processes that make the representation of words more robust yield flatter information profiles, and compared the slopes of information profiles generated by the same speech corpus used in the present chapter.

2.4.1.1 Fast-speech

Tamariz and Shillcock (2001) compared two transcriptions of the words in a corpus of Spanish transcribed speech (Marcos Marin, 1992) - a citation transcription (the idealised pronunciation found, for example, in dictionary entries) and a transcription including fast speech processes (described in § 2.2.1 above, including assimilation, which occurs when the articulation of one consonant affects the way an adjacent consonant is pronounced, e.g. the fact that a *n* is pronounced as *ng* when it comes before a velar consonant such as *k* or *g*). The fast speech transcription consistently yielded flatter information profiles in four samples of a corpus of Spanish speech (the words of length 4, 5, 6 and 7). Additionally, the fast speech transcription generated lower entropy levels (higher redundancy). Speech communication is under pressure to overcome environmental noise. The higher predictability introduced by consonant assimilation in the transcription may be a response to that pressure. It could help deal with the loss of information produced by random noise and thus enhance communication.

Fast-speech processes are an adaptation to the ‘least effort’ pressure. This pressure is in conflict with that of intelligibility. The production mechanisms attempt to alter the pronunciation in order to do as little work as possible, but the pressure of intelligibility only allows those changes that do not hinder comprehension. The whole system has evolved under the two pressures and seems to have reached a state where ‘lazy’ actions are also good for communication.

2.4.1.2 Inflection and derivation

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). Following this hypothesis, the full forms of words need to be stored in the mental lexicon, as proposed by Jackendoff (1975) and 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. I compared the information profiles of the speech types with those of matching words (4, 5, 6 and 7-segment words) from a dictionary wordlist (the 28,000 headwords from the Harrap Compact Spanish Dictionary, 1999). The dictionary profiles yield steeper slopes (one-tailed paired t-test, t=3.86, df=3, p<0.05), and lower levels of entropy (one-tailed paired t-test, t=3.85, df=3, p<0.05) than the speech lexicon. The dictionary lexicon contains almost no inflected or derived words. It has steeper information profiles, indicating that inflections and derivations alone are not responsible for lower entropy towards the end of words, and its entropy levels are lower, indicating it is a less complex system.

The comparisons of fast-speech versus citation transcription and of the speech lexicon versus a dictionary lexicon support the hypothesis that the information profile slope steepness is a good predictor of quality of representation.

2.4.2 Other word segmentations

Single segments are not the only possible units to calculate entropy. I tested the effect of using the finite scheme of bigrams[4] instead of that of segments in the comparison of citation and assimilation transcriptions, and found similar results to those with the segment finite scheme. For the speech tokens, assimilation transcription profiles are significantly flatter than the citation profiles (one-tailed paired t-test, t=5.01, df=3, p<0.01). The average relative entropy of bigram profiles is lower than that of the segment profiles, but it reacts similarly to transcription, with lower entropy for assimilation transcription (one-tailed paired t-test, t=14.65, df=3, p<0.001).

These results indicate that the distribution of information over the full word length can be measured equally well using bigrams or segments. The flattening effect of the assimilation transcription also holds when measured with bigrams. One of the principles of this study is to use the least computationally expensive methodology that is sensitive to the information required, so the similar results obtained with the large bigram finite scheme endorse the use of the smaller segment finite scheme in the calculation of the information profiles.

2.5 Word representations in speech and in the mental lexicon

By adopting a vision of language as embodied in the mental lexicon, this thesis is focusing on words as the basic units of language. I now examine the intra-word distribution of phonological information in two sets of words extracted from the same speech corpus, namely the list of the unique words used by the speakers (types) and the collection of all the word tokens uttered.

[4] Bigrams can be equated to the transitions between one segment and the next (e.g. in the word *admitir* we find the bigrams *ad, dm, mi, it, ti, ir*).

This section examines these two aspects of the same corpus to see how well they are adapted to the different pressures acting on them.

2.5.1 The mental lexicon

I take the set of all word types in the corpus to be a representation of a full-listing mental lexicon (Butterworth, 1983). I assume that phonology plays an important role in the organization of the mental lexicon.

I now appeal to a pressure towards an optimal processing strategy for the storage of the mental lexicon to explain aspects of the information profile of the word types. Shillcock et al. (1995) propose the general principle of maximum storage efficiency, whereby information should be spread as evenly as possible over the representational space in the brain. Entropy is a measure of information content. If the demands of efficient storage were the only factor at play, all segments in the lexicon would have the same probability of occurring anywhere in the word, and then the relative entropy would equal one (maximum uncertainty). However, in reality, constraints such as morphological rules, phonotactic limitations and even sound symbolism (the observation that certain sounds appear to convey certain hues of meaning– see, e.g. Hinton, Nichols & Ohala, 1994) introduce redundancy in the system, preventing storage from being maximally efficient. An analysis of the information profile or word types can reveal the effect of those constraints on the within-word information structure of the mental lexicon.

2.5.2 Speech transmission and speech segmentation

As outlined in the introduction, I assume that the tokens from the corpus represent speech. Word tokens are the word types plus information about their frequency in speech.

This assumption with all its consequences (that analyses of the tokens will reveal the pressures acting on speech) is a testable one. I assume that if we assigned random or scrambled frequencies to the word types, the resulting information profiles would be different from those obtained with the real frequencies, and crucially, they would be different for different word groups.

Speech occurs over a noisy channel. Noise, in the information theory sense, is a random disturbance of the channel that introduces uncertainty in the correspondence between the produced and the received signal. In a noisy channel, a more redundant (lower entropy) message will be more likely to be reconstructed without errors by the receiver. High frequency, on the other hand, reduces uncertainty. In the calculation of the entropy of word tokens, all token occurrences are taken into account, meaning that more frequent words contribute more to the information profile. This means that because high-frequency words are uttered many times, it is very likely that they are partially obscured by a random noise occurrence. But because high-frequency words are taken into account many times and therefore contribute more to the information profile (a general property of the lexicon, tacitly known by speakers of a language), they are easier to reconstruct without error.

The process of speech perception includes the segmentation of the continuous sound stream into words. Segmentation is, therefore, a critical issue for speech representation (tokens), but not for the representation of words in the mental lexicon (types). Different approaches emphasize different factors that help achieve speech segmentation. In English, metrical structure seems to be a good predictor of word boundaries, which tend to be found immediately before strong syllables (Cutler 1990, Cutler & Butterfield, 1992). Cutler and Carter (1987) found that over 90% of all English content words begin with a strong syllable, but this is not the case in Spanish, so the equivalent of metrical structure (stress) might not be so relevant to segmentation in that language. Norris, McQueen, Cutler and Butterfield (1997) suggested that a Possible Word Constraint (PWC) could ease word recognition by limiting the number of lexical candidates activated by a given input. This constraint requires that, whenever possible, the input should be segmented into a string of feasible words. Any segmentation resulting in impossible words (e.g. a single consonant) is not allowed. Norris et al. 1997 used a word spotting task to demonstrate that adults find words such as *apple* more easily in *vuffapple* than in *fapple*, because *vuff* is a possible word, whereas *f* is not. The PWC approach requires previous knowledge of the possible lexical units, which infants do not possess during the initial stages of language acquisition (although Johnson, Jusczyk, Cutler and Norris's 2003 results indicate that 12-month-old infants already use the possible word constraint in segmentation of fluent speech). McQueen (1998) adds to the PWC the phonotactics and other statistical regularities that constrain what can be a possible word and where words can start and end. Saffran, Newport and Aslin (1996) suggest that distributional cues are crucial in the initial lexical segmentation of (adult) language learners. They found that the transitional probabilities between syllables in a language were enough for learners of an artificial language to hypothesize word boundaries (even though prosodic cues would enhance performance). Cairns, Shillcock, Chater and Levy (1997) used neural networks and conventional statistics to demonstrate that segmental distributional information in English is an important cue to segmentation and it could allow infants to bootstrap into increasingly complex strategies to end with an adult segmentation competence.

Most of the variation between information profiles in the present study is found at the beginning and, particularly, at the end of words, where word boundary transitions occur. This is also relevant to the principle that phonological reduction usually takes place at the end of the word or the syllable, occasionally leading to material being dropped, together with a compacting of the beginning of the word. The information profile of words thus reflects these differential probabilities that help speech segmentation by showing increased redundancy at the end of words – during word processing, the appearance of redundant segments (e.g. one of the reduced set of segments that usually occur at the end of words) provides a cue to the word ending.

In summary, phonological statistical regularities such as distributional cues seem to play an important role in speech stream segmentation. This should be reflected in a more marked drop of entropy in the final position of the information profiles of the tokens as compared to the types.

2.5.3 Comparison of the information profiles

A comparison of the information profiles generated by the types and the tokens, in the light of the pressures acting on the mental lexicon and on the communication channel that is speech, indicates how well adapted each word set is to the constraints of its representational substrate. I expect the information profile of the types to be flatter and to have higher levels of entropy, reflecting a pressure for a more efficient use of the representational space; on the other hand, I expect the profiles of the tokens to be more redundant, reflecting the complexity introduced by the different word frequencies, and steeper, reflecting the pressure to facilitate the segmentation of speech into words.

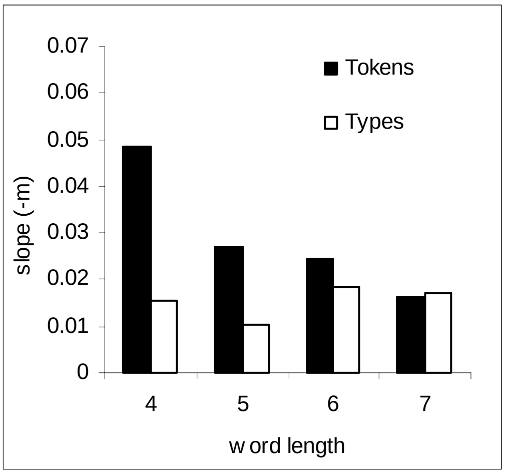


Figure 2.3. Slopes of the information profiles of the corpus (tokens) and the lexicon (types) across the four word lengths.

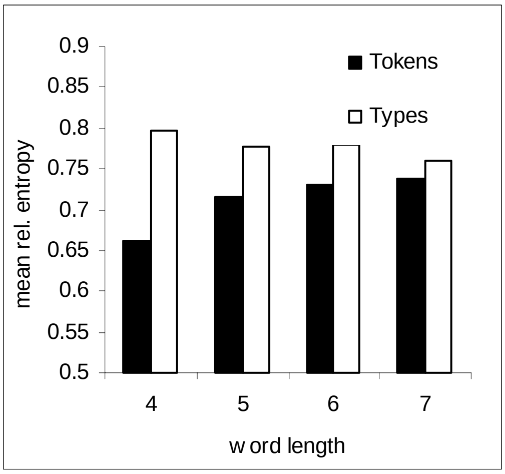


Figure 2.4. Mean relative entropy of the information profiles of the corpus (tokens) and the lexicon (types) across the four word lengths.

Figure 2.3 shows that the word types generate mostly flatter information profiles, although the results are only marginally significant (one-tailed paired t-test, t=2.57, df=3, p=0.08). As we saw in Figure 2.2, the main contributor to the descending slope of information profiles is the last segment. If we compute the effect of this segment alone by subtracting the slope of all-but-the-last-segment from the slope of all the segments, we obtain significant differences between the tokens and the types (one-tailed paired t-test, t=5.28, df=3, p<0.01). Similarly, if we measure the effect of the last segment using the level of relative entropy of the last segment as a percentage of the mean relative entropy of the other segments, the comparison between tokens and types is significant across the four word lengths (one-tailed paired t-test, t=6.74, df=3, p<0.01).

We see in Figure 2.4 that, as predicted, the mean relative entropy values are significantly higher for the types (one-tailed paired t-test, t=3.66, df=3, p<0.05).

Slopes are flatter in the types, indicating a more even distribution of entropy across all word segments. Given the phonotactic constraints of the language, which must account for at least some of the redundancy and differential entropy across segments, it seems that segments are very evenly spread in the lexicon, particularly in shorter words, allowing (or caused by the need of) an efficient representation. Word-final segments, particularly, are significantly more evenly spread in the types than in the tokens.

The tokens present steeper slopes, reflecting lower entropy levels in the last word segments, as predicted. Given the flatter profiles found in the types, this trend must be due to the fact that more frequent words show a more marked low entropy final segment, or more predictable (redundant) word-end patterns.

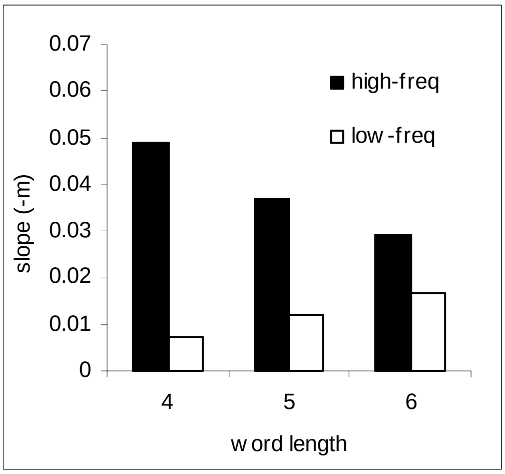


Figure 2.5. Slopes of the information profiles of the 100 most frequent word types and of 100 types of frequency=4 across three word lengths.

Figure 2.5 shows the slopes of the information profiles of the most frequent 100 types in the corpus (frequency range [50,000-877]) and 100 types of frequency 4 from word length groups 4, 5 and 6 (there were not enough words of length 7 in the corpus to include them in this comparison). The slope values indicate that the high-frequency words, particularly shorter ones, generate steeper profiles (one-tailed paired t-test=4.48, df-2, $p<0.05$). Shorter words tend to have higher frequencies than longer ones (see Figure 2.1 above), suggesting that high-frequency words make better use of the statistical regularities of the lexicon to become more easily recognizable as independent units. The results shown in Figure 2.5 also suggest that frequent words have more informative beginnings and more redundant endings. This could help understand the progressive phonological reduction and eventual dropping of the endings of high-frequency words, and also the higher communicative effectiveness of frequent words – the information concentrated at the beginning of the word allows early recognition.

Relative entropy values are higher in the types than in the tokens (Figure 2.4), reflecting the higher complexity introduced by the frequencies (most words being considered more than once).

However, it is also apparent that while the high level of relative entropy is constant in the lexicon (types), in the corpus (tokens), it is lower for shorter words. This can be due partly to the fact that shorter words are more frequent – and the frequency distribution of longer words is closer to the distribution of types (frequency = 1 for all of them). Additionally, the high level of entropy in the lexicon (even distribution of phoneme frequencies across the word) in the short words suggests that many of the possible phoneme combinations do exist as words. However, the lower level in the corpus reflects the fact that some of these words are being used much more frequently than others.

The information profiles of the lexicon reflects that it is adapted to an efficient storage solution, and that of word tokens reflects that they are well adapted to being segmented from the speech stream. However, a child's lexicon and ability to segment speech may have different requirements, which should be reflected in the information profile.

2.6 Child-directed speech

I assume that speech addressed to a child is adapted to help develop an optimal strategy for lexical acquisition. Because children need to be able to identify and process new words, the structure of their smaller, ever-expanding mental lexicon should reflect a greater emphasis on lexical acquisition demands and less on lexical efficient storage demands than the adult mental lexicon. On the other hand, the potential environmental noise in the channel transmitting the message and thus the constraint for producing intelligible speech is the same for adults and for children, therefore the sets of adult and child-directed tokens should show the same adaptation to intelligibility constraints.

Young children are acquiring basic phonologically-encoded features of the language, such as word segmentation, so it can be expected that they need word boundaries to be more marked. The child's main language input comes from their caregivers, who seem to adapt their speech to their young audience in different respects. Speech addressed to children is characterized by special prosody (Kemler Nelson, Hirsh-Pasek, Jusczyk & Wright-Cassidy, 1989), high pitch and other distinctive acoustic measures (Fernald & Kuhl, 1987; Slaney & McRoberts, 2003), short Mean Utterance Length and simple syntactic structure (Tse, Kwong, Chang & Li, 2002), clearer phonological segmental information (Kuhl, Andruski, Chistovich et al. 1997), the inclusion of special vocabulary and special grammatical uses ('mummy is here' instead of 'I am here'), and a higher neighbourhood density, which implies that children first learn words with more frequent sounds and sound combinations (Coady & Aslin, 2003). I argue that child-directed speech also tends to be different from normal adult speech in the informational contour of the words used, and child-directed speech contours should emphasize word boundaries. Many studies suggest that children are sensitive to the phonological statistical information in a language from an early age, and they seem to use it in the segmentation of continuous speech into words. Christophe, Dupoux, Bertoncini and Mehler's (1994) experiments carried out with three-day-old infants in French suggest that they can discriminate between items that contain a word boundary and items that do not. This result indicates that newborns could be sensitive to cues that correlate with word boundaries, and that they could use these cues during lexical acquisition. Mattys and Jusczyk (2001) report that 6-12 month old infants are already sensitive to the probabilistic phonotactics of the language that is spoken around them. Jusczyk, Luce and Charles-Luce (1994) report that 9-months infants prefer to listen to lists of monosyllables containing phoneme sequences that are frequent in their language than to lists containing infrequent (although legal) sequences. There are other reports of sensitivity of 10 month old children to cues to word boundaries such as statistical regularities (Jusczyk, 1999), and of the sensitivity of 9 month olds to how phonotactic sequences typically align with word boundaries (Mattys, Jusczyk, Luce & Morgan, 1999). Nine-month olds also prefer legal over illegal word boundary clusters (clusters of sounds which are allowed or not allowed to occur at the beginning of a word in a specified language) within their own language (Friederici & Wessels, 1993). These pieces of research were carried out using English and French, but I assume that the same strategy is employed in Spanish, which is, like French, a syllable-timed language. Also, the fact that these effects are observed as such early ages suggests that they are not language-specific. Cairns, Shillcock, Chater and Levy (1997) demonstrated that the distribution of phonetic segments in English is an important cue to segmentation. Statistical information is also the basis of word segmentation by a connectionist network trained with child-directed speech (Christiansen, Allen & Seidenberg, 1998).

In summary, very young infants are sensitive to statistical cues to segmentation in the spoken language they hear. These cues could also help lexical acquisition. More predictable word-endings in the lexicon could help the child segment words from speech and recognize them as lexical units. This should be reflected as a drop in entropy at the end of the word type information profile.

2.6.1 Data

The data in this section come from "Maria" corpus (Lopez Ornat, 1994), an 83,000 word corpus of speech interaction between an only child (between the ages of 1.25 and 4) and her parents and, to a much lesser extent, other relatives. For the present study only the speech addressed to the child was taken into account. After removing the child's utterances and all corpus annotations 41,138 word tokens were left. The lexicon used in the corpus of child-directed speech has 3,895 word types, which represents a rather restricted vocabulary compared with the numbers of word types found in similar size samples from the adult speech corpus and from the text corpus. The average type:token ratio per 10,000 is 0.21 (range 0.20, 0.22) for 5 samples of the adult corpus and 0.16 (range 0.14, 0.17) for 4 samples of the child corpus. Incidentally, the four chronologically ordered 10,000-word samples from the child corpus produced increasingly higher type:token ratio values, reflecting the increasingly varied vocabulary employed when addressing a growing child. Another difference between the two corpora was the word length distribution. Child-directed speech was made up of shorter words, on average: the mode for the normal (adult) speech corpus was 7 segments compared with 5 for the child-directed speech corpus, and there were fewer long words in the child-directed speech (Figure 2.6).

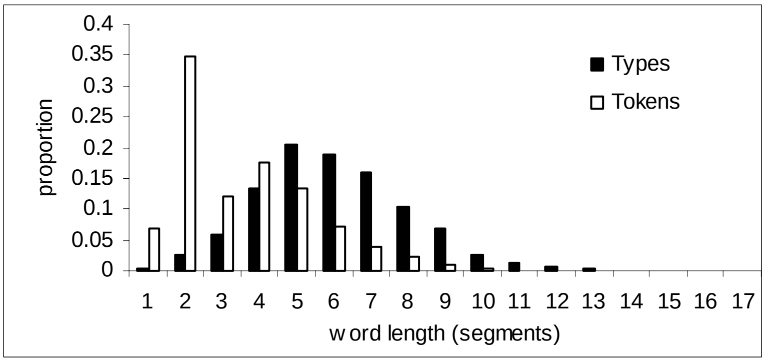


Figure 2.6. Distribution of word-lengths in the 3,895 types and the 41,000 tokens from the corpus of child-directed speech (Cf. Figure 2.1 for the same distribution of adult types and tokens).

The word frequency distribution of the child-directed corpus also has a significantly different standard deviation from that of the adult corpus. I tested this with a Monte-Carlo test: the SD of the 41,138-word child-directed frequency distribution is 73.4. I then calculated the SD of 100 random samples of the same size extracted from the adult corpus, and found that the child-directed SD was significantly higher than any of them, being an outlier ($p<0.01$) of the distribution of adult-corpus sample SD's (thus the observed SD could not have occurred by chance). This is explained by the fact that the child-directed speech contains significantly less very low frequency words than the samples of the adult corpus.

Another feature of this corpus of child-directed speech is the higher presence of nouns and adjectives with diminutive suffixes *-ito, -ita, -itos* and *-itas*. Diminutives are typical of positive affect speech in Spanish, including interaction with children (see Melzi & King, 2003, for a recent review of the use of diminutives in general and in Spanish in particular). Kempe and Brooks (2001) experiments in Russian – where diminutives are also a pervasive feature of child-directed speech – suggest that the function of diminutives is to help acquire grammatical gender. There are almost five times more diminutive types and eight times more diminutive tokens in the child-directed than in the adult speech corpus (1.37% of word types vs. 0.17% in the adult speech, and 5.65% of the word tokens vs. 1.17% in the adult speech).

I assume that the phonological structure of child-directed speech triggers and directs the progressive organisation of a new mental lexicon structure in the child's brain. I argue below that the characteristics of child-directed speech reflect the pressures acting on the developing mental lexicon, which are different and occasionally opposed to those acting on the adult mental lexicon. A comparison of the information profile obtained with the types and tokens from a child-directed and an adult corpus may reflect the effects of those different pressures.

I expect that, unlike adult types, the child-directed types do not yield 'optimally efficient' flat profiles, since children have a lot of representational space available. The child-directed tokens, represented speech, will behave like the adult tokens, since the pressures of communication over a noisy channel are the same for both adults and children.

2.6.2 Comparison of the information profiles

Figures 2.7 and 2.8 show the slopes and level of relative entropy of the information profiles of the tokens and types of the child-directed speech (Cf. Figures 2.3 and 2.4 for the same comparisons on the adult corpus).

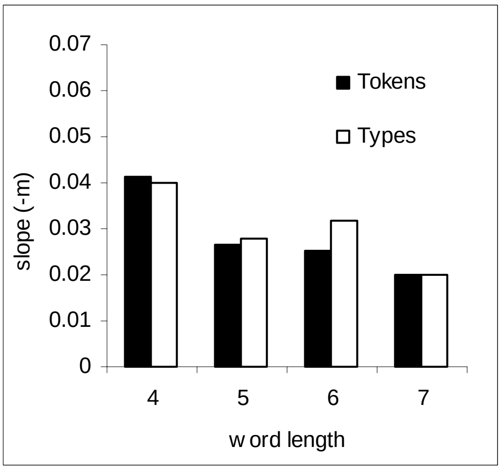


Figure 2.7. Slopes of the information profiles of the tokens and types from the corpus of child-directed speech across the four word lengths.

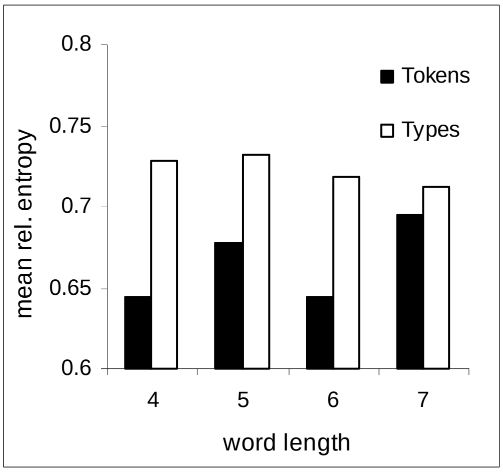


Figure 2.8. Mean relative entropy of the information profiles of the tokens and types from the corpus of child-directed speech across the four word lengths.

Figure 2.7 shows the very similar slopes generated by the tokens and types of child-directed speech. The relative entropy levels (Figure 2.8), however, are significantly different (one-tailed paired t-test, t=4.99, df=3, p= 0.01)

Figures 2.9 and 2.10 compare the slopes and mean relative entropy values of the child-directed information profiles with those of the adult corpus (data presented in § 2.5.3) , averaged over the four word lengths.

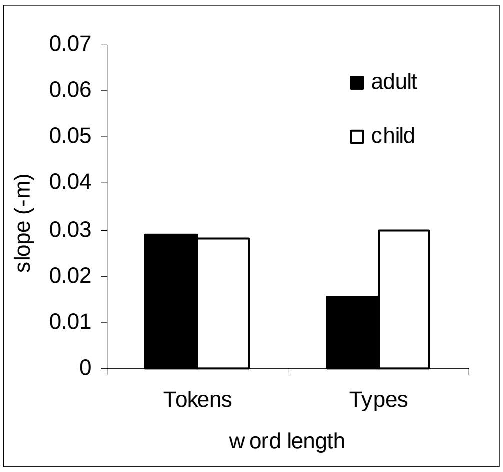


Figure 2.9. Average of the slopes of the information profiles of the tokens ant types of adult speech and child-directed speech across the four word lengths.

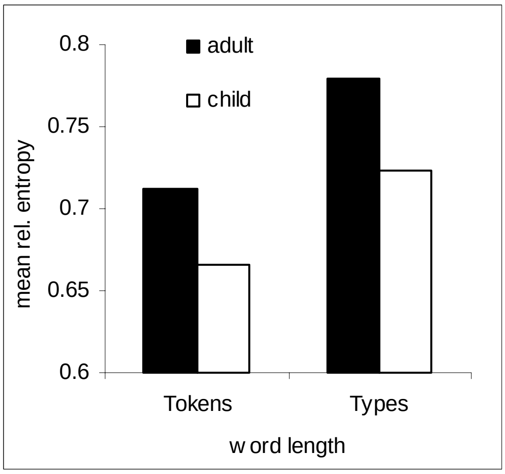


Figure 2.10. Average of the mean relative entropy values of the information profiles of the types and tokens of adult speech and child-directed speech across the four word lengths.

Figure 2.9 shows a summary of the slopes and the relative entropy values for the adult and the child-directed tokens and types (Figures 2.3, 2.4, 2.7 and 2.8). There is no significant difference between the slopes generated by the adult and child-directed corpora. The difference in the slopes of the types, however, is significant across the four word lengths (one-tailed paired t-test, t=4.23, df=3, p<0.05). Figure 2.10 shows significantly higher relative entropy in the adult data (one-tailed paired t-test, t=4.133, df=3, p<0.05 for the tokens and t=12.7, df=3, p<0.01 for the types).

As expected, the slopes are similar in child-directed and adult tokens in the face of similar environmental noise levels and segmentation requirements.

The fact that the child-directed types do not show the flatter slopes found in the adult types is due to the different slopes generated by frequent and infrequent types in both corpora. I calculated the information profiles of the 50 most frequent types (frequency range [1,800-164]) and 50 types of frequency 2 in the 4, 5 and 6-segment words from the corpus of child-directed speech (Figure 2.11).

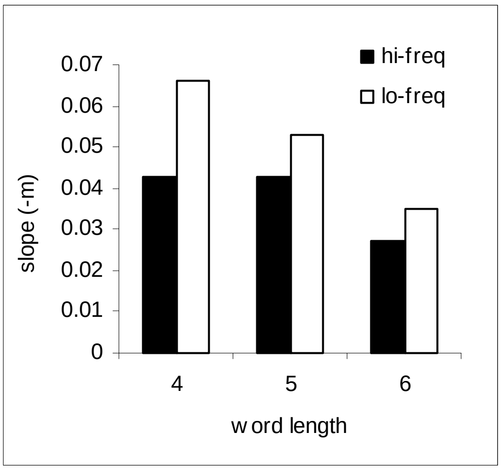


Figure 2.11. Slopes of the information profiles of the 50 most frequent types and of 50 types of frequency=2 across 3 word lengths from the corpus of child-directed speech. (Cf. Figure 2.5 for the similar data from the adult corpus).

The slopes of the high and low-frequency words are significantly different (one-tailed paired t-test, t=4.36, df=2, p<0.05). This result can be compared to the calculations done on the adult corpus shown in Figure 2.5 above, even though the number of words and the frequencies are lower for the child-directed corpus, given the fewer types and tokens in this corpus.

A comparison of Figures 2.11 (for child-directed types) and 2.5 (for adult speech types) reveals that the high-frequency types generate similar slopes in both the adult and the child-directed corpora (between 0.030 and 0.049). However, infrequent types are flatter in the adult (between 0.008 and 0.017) and steeper in the child-directed speech corpus (between 0.066 and 0.035). The type profiles are calculated using one count of every word, as opposed to the tokens, where each word is counted as many times as it appears in the corpus. Therefore, infrequent words have a bigger impact on the types than on the tokens, and this is reflected in the slopes. Additionally, the child-directed corpus contains relatively fewer high-frequency words than the adult corpus - the frequency distribution of the tokens is flatter in the speech-directed corpus (SD = 73.4) than in the adult-directed corpus (SD= 49.5 - average of the SD's of the 100 random samples from the adult corpus of the same size as the child corpus used in the Monte-Carlo test in § 2.6.1). The higher number of low-frequency types in the child-directed corpus explains the steeper slope of the child-directed types.

These steeper profiles of the child-directed types are due to the lower entropy level in the last segment. The more redundant, predictable word-endings relative to normal adult speech may be reflecting the caregivers' speech helping children identify word boundaries. This outcome is achieved at least partly through the higher presence of nouns and adjectives with a diminutive suffix, a strong characteristic of Spanish child-directed speech.

I argued earlier that the flatter slope of low-frequency tokens in adult speech reflected the fact that these words have unusual word endings. I propose that the steeper slope of low-frequency types in child-directed speech reflects the fact that word endings are very predictable. This would help segmentation and thus recognition and acquisition of novel words by children who have never heard them before. Note that most low-frequency words in child-directed speech will be high-frequency in the adult speech, and adult low-frequency words will probably not be present at all in the child-directed speech. Adults will use previous knowledge of a low-frequency word in order to segment it in speech, whereas children need strong clues to the word boundaries of words that are new to them.

The results in Figure 2.10 also confirm the prediction that the level of relative entropy should be lower (higher redundancy) in the child-directed than in the adult types. This difference in relative entropy in the child-directed and adult types is also reflected in the tokens. In line with normal speech, the frequencies in the tokens introduce redundancy and lower the relative entropy of the types.

> **Summary of sections 2.5 and 2.6.** In these two sections I have calculated and compared the information profiles generated by words from an adult-speech and from a child-directed speech corpus.
>
> - For the types – representing the mental lexicon – we have seen that, while adult speech profiles were very flat, child-directed speech profiles are not, suggesting that the efficiency of storage pressure is less strong in the smaller infant mental lexicon than in the adult mental lexicon.
> - As for the tokens – representing words in speech - both adult speech and child-directed speech information profiles were steep, suggesting that the pressures for efficient communication in a potentially noisy medium and for speech segmentation are similar for both children and adults.

- Additionally, we have seen that the even distribution of phonological information in the types (flat profiles for types) is driven by low frequency words in the adult corpus, while frequent words show steep profiles that can be more easily exploited in speech segmentation. In the child corpus, the low-frequency words present very steep profiles, perhaps to help segment new words as they are introduced in the child’s vocabulary.

2.7 The role of features: manner and place of articulation

This section tests a different way of calculating the information profiles. For the calculation of entropy, instead of using the finite set of segments, I now use features such as manner and place of articulation.

Manner of articulation speech features are best transmitted by the auditory channel, whereas place of articulation are best transmitted by the visual channel (Robert-Ribes, Schwartz, Lallouache & Escudier, 1998): in a noisy environment, seeing the speaker's face improves message intelligibility (Girin, Schwartz & Feng, 2001). Conflicting information from the two channels generate fused responses reflected, for instance, in the McGurk effect (McGurk & MacDonald, 1976): when presented simultaneously with the sound ‘ba’ and an image of a face pronouncing ‘ga’, people perceive ‘da’.

I compare the results of the last two sections with similar information profiles calculated with the finite sets of 17 ‘manner of articulation (plus vowel)’ and 19 ‘place of articulation (plus vowel)’ features (see Appendix A for full lists of features).

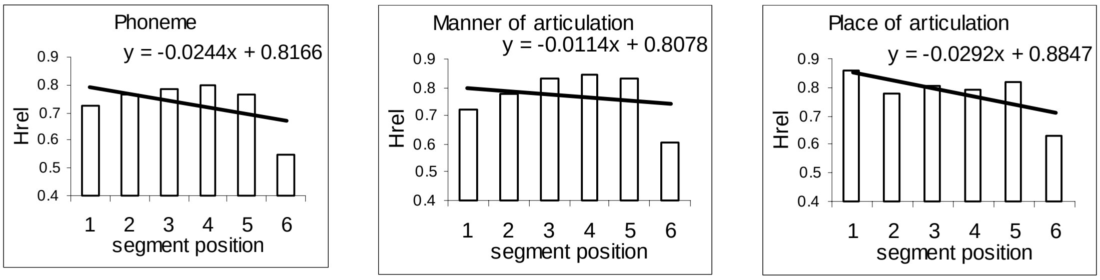


Fig 2.12. Information profiles of 6-segment tokens from the speech corpus, calculated using the finite scheme of phonemes, of manner of articulation features and of place of articulation features.

Manner and place of articulation information are distributed differently in the word. Figure 2.12 shows the profiles by phoneme and by manner and place of articulation of 6-segment long words from the corpus. Place of articulation is most informative in the word-initial position, where manner of articulation is relatively redundant in that position. The highest redundancy in the last segment is best captured by the phoneme analysis.

Figures 2.13 and 2.14 present a comparison of the information profile slopes and mean relative entropy generated by the segment finite set (the results already presented in § 2.5 and § 2.6) and by the manner and place of articulation finite sets.

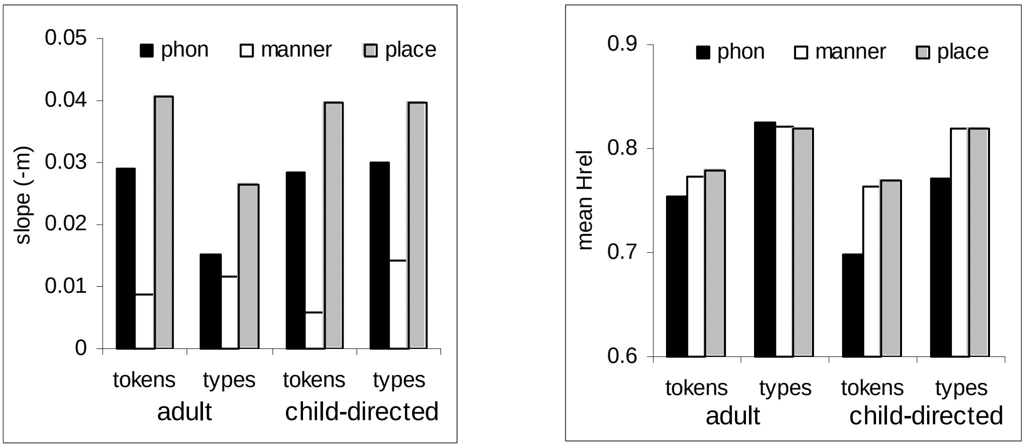


Figure 2.13. Values of the slopes (left) and of the mean H_{rel} (right) averaged over the four word lengths of the analysis by phoneme, by manner of articulation and by place of articulation using the assimilation transcriptions of the speech (adult) corpus and the child-directed corpus.

Figure 2.13 shows that manner of articulation yields the flattest contours, particularly for the tokens, both in the adult and the child-directed corpora. In terms of average relative entropy, both manner and place of articulation behave similarly, with higher values than the phonemes, except in the adult types (lexicon), where the most efficient encoding seems to be attained with phonemes.

Manner of articulation information is more evenly spread across words than place of articulation, and in speech (tokens) this suggests that it is more immune to noise and could have an important role in auditory speech production and recognition. This means that manner of articulation encodes speech robustly in the absence of visual contact. In the lexicon (types), the even spread of manner of articulation information suggests that it produces a more efficient encoding and thus could have an enhanced role in the organization of the phonological mental lexicon. This is also supported by the fact that manner of articulation slopes are steeper in the types than in the tokens – manner of articulation might be encoding internal word structure in the types, but not in the tokens.

Place of articulation, providing visual information, shows much steeper slopes than manner of articulation, even steeper than phoneme slopes. This suggests that place of articulation is not an efficient dimension to organise the mental lexicon storage; however, the sharp difference of entropy between word-beginnings and endings is a good clue to speech segmentation (see Figure 2.12).

Summing up, while manner of articulation seems to encode auditory information more robustly, place of articulation encoding may be responding to the pressure to facilitate speech segmentation.

2.8 Conclusion

This chapter has examined the information profiles of words found in spoken language, a measure of how well different word systems are adapted to the informational requirements of their representational spaces. The information profile is calculated with a computationally inexpensive methodology that still finds reflections of the pressures acting on the distribution of phonological information within Spanish words. The consistency of the profiles’ behaviour over four independent word groups (words of length 4, 5, 6 and 7) supports the robustness of this method. The information profile calculated with different finite sets (segments and features) show comparable results, each reflecting different aspects of the phonological information structure of words.

The profile found in the adult lexicon supports the claim that it reflects phonological distributional features that allow an optimal strategy of storage in the brain. However, the features of the lexicon of child-directed speech do not respond in the same way. Caregivers’ speech is adapted to meet other critical demands that interfere with an efficient storage strategy at this early age.

The profiles of two different token sets (adult-directed and child-directed speech) show that they are equally well adapted to good communication over a potentially noisy medium.

The vocabulary employed with children has a more marked drop in entropy levels at the end of words, which could enhance word-boundary recognition and help with lexical acquisition. In adults, segmentation cues are clearer in the corpus, helping with speech stream segmentation, one of the crucial problems of language recognition.

Calculation of the information profiles generated by manner and place of articulation features suggests that while manner encodes a robust auditory representation of speech, place may serve as a cue for speech segmentation.

In conclusion, I have shown that information profiles of spoken words and of the lexicon are a useful tool to measure distributional aspects of large samples of language, and can be used to test and potentially falsify particular aspects of psycholinguistic theories about speech production and recognition, the mental lexicon, and lexical acquisition.

Chapter 3. The structure of the phonological mental lexicon

This chapter describes a mental lexicon geometry defined by quantifiable relationships at several levels between words. It concentrates on the phonological level of the mental lexicon, particularly on representations based on similarity. It presents a psycholinguistic empirical exploration of a phonological parameter space that can be used as a tool to define the phonological lexicon structure.

3.1 Similarity-based mental lexicon structure

The last two chapters have emphasized the complexity of the mental lexicon. Chapter one described the mental lexicon structure as the result of juggling the many different pressures acting on it. Chapter two described emergent characteristics of the intra-word phonological level, and how this level responded to pressures such as intelligibility, storage and processing constraints. This section reviews approaches to the lexicon based on relationships between words, where each word's phonology, syntax and semantics is defined in terms of its similarity to the rest of the words in the lexicon.

Chapter two considered the phonological lexicon as a set of words that responds *as a system* to communication and acquisition constraints. It examined the distribution of entropy, a measure of the information content, in the phonological representation of words. Entropy is defined in terms of probabilities in a set of elements. It makes no sense to talk about the entropy of one word, but in the statistical framework of linguistic communication, each word occurs in speech with a certain probability. An unconstrained system evolves towards a state of maximum entropy, where all elements occur with the same probability. Deviations from this state of maximum entropy are a reflection of the pressures that operate on the system. In chapter two I found traces of the effect of communication, segmentation and acquisition needs on the phonological lexicon structure.

This chapter focuses again on the lexicon as a system, but at the word level instead of the segment level. I look at explanations of lexicon organisation where each word is defined by its relationships with the rest of the words, for instance the word 'cat' is defined phonologically by its similarity to other words like 'mat' and 'cab', and by its differences from words like 'lease' or 'friendliness'; syntactically, it is defined by its similarity to other nouns like 'chair' and 'glove' and by its difference from words from other categories such as 'the' and 'go'; semantically, it is defined by its similarity to 'dog' and 'purr' and its differences from 'cloud' and 'write'. These examples illustrate the emergence of categories in a lexicon structure based on similarity: words belonging to the same category will be close together along the dimension measuring that category.

The aim of this chapter and the next is to obtain two similarity-based representations of the lexicon: one at the word-form level and another at the cooccurrence-based level. These will be brought together in chapter five, where I test the existence of systematic relationships between them. In the present chapter I review studies suggesting that the lexicon is structured in terms of similarity between words at many levels, and then concentrate on metrics of phonological similarity. Finally I present a psycholinguistic study that measures the relative impact of phonological parameters such as 'sharing the initial consonant' or 'sharing the stress on the final syllable' on perceived word similarity. The resulting parameter values will be used in chapter five's metric of phonological similarity to produce quantitative representations of samples of the Spanish phonological lexicon.

3.1.1 Lexicon levels

Many studies have shown evidence suggesting that similarity plays a role the structure of the mental lexicon. Words can be categorised in terms of their phonology, semantics and syntax, among other levels. One widely used paradigm that reveals these similarity relationships between words is priming. In priming experiments, participants are exposed to a prime word for a short time, and are subsequently shown a target word. Prime and target are related semantically, phonologically or at another level, according to what lexical level the experiment addresses. An analysis of the effect of exposure to the prime on processing of the target reveals aspects of the lexicon organisation and representation at the relevant level. Primes can have facilitatory or inhibitory effects on target processing. Facilitation usually occurs with rapid presentation and it does not rely heavily on attention or processing effort. Inhibition occurs later during lexical processing and may involve more attention or strategic processing (Faust & Gernsbacher, 1996; Neely, 1991). Facilitatory priming reveals a more direct reflection of the mental lexicon structure, since it is not affected by conscious or controlled processing.

The priming effect can be explained by a process of spreading activation (Collins & Loftus, 1975): when a word representation is activated, the activation spreads to word representations that are closely related to it. For example, hearing the word 'cat' activates semantically related words such as 'purr' and 'dog'. When one of these related words is subsequently presented, the participant reacts to it sooner because it was already partially activated. This assumes that the representations of related words are more closely connected to each other than to unrelated words. Priming is proportional to relatedness, so the strength of the priming effect between a target and its prime is a measure of how closely related they are. The relatedness between two words can be defined as the similarity between them at the relevant level.

A whole body of literature on semantic priming suggests a lexicon organised in terms of semantic similarity relationships between words (Meyer & Schevaneldt, 1971; see Neely 1991 for review). The lexical decision task has been used extensively to demonstrate semantic priming: participants have to decide whither a string of characters is a real word or not. The typical effect is that when participants are shown for instance the word ‘cat’, the time they take to recognize it as a word is shorter if they were shown the prime ‘purr’ than if the prime was ‘puff’. Manipulation of the semantic relationships between primes and target words has helped study the structure of the semantic mental lexicon.

Other studies focus on phonological similarity: Frisch, Pierrehumbert and Broe (2004) studied the interactions of different phonotactic constraints in Arabic and found that the more similar two homorganic (same place of articulation) consonants are, the less they tend to cooccur within the same root. They propose a model of the phonological lexicon where constraints are graded rather than absolute, and interact with each other in complex ways. Saffran (2003) and Pierrehumbert, (2001b, 2003b) emphasize statistical learning of phonology, suggesting that the organisation of the phonological lexicon is learned from the statistical properties of the linguistic input.

Morphosyntactic priming results are reported in several papers: Sereno (1991) found that prime words facilitated targets from the same syntactic class in a lexical decision task (but not in a naming task). Sanchez-Casas, Igoa and Garcia-Albea's (2003) priming and lexical decision experiments in Spanish also suggest that morphology is represented in the mental lexicon and it may play a central role in word identification and recognition. Other paradigms also reveal the morphological organisation of the lexicon: Bozsahin (2002) proposes a Combinatory Categorial Grammar-based interface between inflectional morphology, syntax and semantics that exploits systematic relationships between the three lexical levels, and Saffran

(2003) suggests that morphology is also learned from statistical patterns in speech.

The lexicon also seems to be organised in terms of the orthographic form of words. Segui and Grainger (1990) used a priming paradigm to reveal that words were activated by primes with similar orthography, and the number and frequency of the neighbours affected the degree of this activation. Andrews (1997) notes some differences between orthographic priming effects in English and in other languages with more transparent orthography such as Spanish and English, such as less influence of high-frequency orthographic neighbours on lexical retrieval.

Similarity of nonlinguistic lexical aspects also has been found to influence processing and to affect the lexicon structure. Kjellmer (2000) found that foreign words are more likely to enter the lexicon of a language if there was no native equivalent, but also due to social aspects such as fashion and prompting by the media. Type of social interactions has been shown to influence the evolution of the lexicon (Baldwin, 2000; Vogt & Coumans, 2003). Finally, the lexicon also seems to be organised in terms of similarity of words' emotional connotation. Wurm, Vakoch, Aycock and Childers (2003) isolated the effect of very specific emotional lexical connotations such as 'danger' and 'usefulness' on word naming times, and in a perceptual matching and classification task, Mullennix, Bihon, Bricklemyer, Gaston and Keener (2002) showed the effect of emotional tone of voice.

Having seen how words are organised in the mental lexicon in terms of their semantics, phonology, syntax, orthography and some non-linguistic aspects, the next section briefly reviews theories of category construction

3.1.2 Categories within lexicon levels

The organisation of the mental lexicon is reflected in the existence of word categories at different levels of linguistic description. Category construction has been dealt with by different theories throughout history, as Murphy

(2002) explains: according to the classical theory of concepts (Frege, 1862-1960), concepts are represented in the mind as definitions or lists of necessary conditions to belong to that category. Classic categories in cognitive psychology such as the syntactic category 'noun' or the phonological category 'bilabial' were defined a priori by rules. Later, Rosch's (1978) typicality effect studies showed that people agree about whether an example is a good member of a category to a surprisingly large extent. This prompted two opposing theories, both involving similarity comparisons: in prototype theory, concepts are represented in long-term memory as the best or most prototypical instance, and categorization is achieved by comparing the observed item to stored prototypes and matching it with the prototype it is most similar to. In exemplar theory, many or all instances (exemplars) of a category are stored, and categorization is achieved by comparing the observed item to all the stored exemplars and determining the number of exemplars it is similar to and the extent of this similarity. The exemplar theory of categorisation has been applied to speech perception and production by Jonhson (1997), Lacerda (1995) and Pierrehumbert (2001), and to phonological acquisition by Maye, Werker and Gerken (2000), who propose that the acquisition of linguistic categories such as phonemes is brought about by the memory traces of perception of many exemplars of the categories in speech.

The next section reviews similarity-based models of the mental lexicon.

3.1.3 Similarity-based mental lexicon models

In this section I briefly review models where the structure of the mental lexicon is determined by relationships between words.

Guthrie, Pustejovsky, Wilks and Slator (1996) review analyses performed on machine readable dictionaries that, apart from extracting explicit information such as definitions, exploit implicitly available phonological, semantic and syntactic information. They focus on cooccurrence approaches that extract part of speech (Byrd et al., 1987), form noun and verb taxonomies (Amsler & White, 1979), 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 et al., 1993). All these kinds of information can be included in an analogue of the mental lexicon, a 'lexical database' (e.g. Nakamura & Nagao, 1988), that can be used in natural language processing tasks such as sense disambiguation.

Connectionist models of the lexicon include those of Miikkulainen (1997) and Philips (1999). The former presents an unsupervised connectionist model called DISLEX, consisting 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) proposes a connectionist mental lexicon that, apart from lexical semantics, includes information about grammatical category, frequency and phonology.

The Analogical Model of Language (AML) (Skousen, 1995) was proposed as an alternative to connectionist language models. AML 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. Then they apply the derivation or stress pattern of the similar words to the target word. AML has been used to predict stress placement in Spanish (Barkanyi, 2000; Eddington, 2000), the choice of linking elements in Dutch noun-noun compounds (Krott, Schreuder & Baayen, 2002), and Spanish diminutive formation (Eddington, 2002).

Having reviewed similarity-based approaches to the mental lexicon, the next section outlines the similarity-based mental lexicon model adopted in the rest of this thesis.

3.1.4 A similarity-based model of the mental lexicon

I assume a mental lexicon configuration consisting of different levels of organisation, each of which is defined by relationships of similarity between each word and every other word with respect to that level. In such a configuration, categories emerge as groups of words that are close together in that level of description (see Figure 3.1).

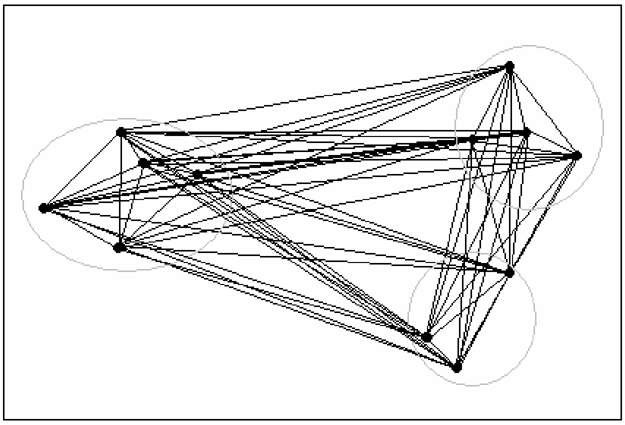


Figure 3.1. One level of description of the words in the lexicon. The black dots are the words; the lines between them represent relationships of similarity: the shorter the line, the more similar the two words it joins. The overall configuration of the lexicon is defined by the similarity relationships between words. Categories emerge from the resulting geometry (the clusters in the grey ovals). Note that in reality the lexicon would not be representable in a two-dimension plane.

For example, the phonological level of the mental lexicon is defined by the phonological similarity between each word and every other word. In such a structure, the identity of the words themselves becomes unimportant. Each word's position is defined by its similarity values to every other word (this could be visualised by rotating the whole structure represented in Figure 3.1; the actual positions of the words is irrelevant, what counts is their relative position to each other). Categories can be identified at different levels (e.g. phonological, semantic) and along different dimensions within levels (e.g. words stressed on the final syllable). I assume that those groupings can be explained in terms of pressures acting on the lexicon structure.

I assume that speech contains the information patterns necessary to organize categories at the different linguistic levels of the lexicon (phonological, semantic, syntactic). This is linked to the important role of statistical learning in using the information in speech during the development of the lexicon in language acquisition: the stochastic patterns in speech incrementally define the relationships between words. I therefore assume that relevant analyses of speech should reveal the patterns of information that shape the mental lexicon.

Using a corpus as an approximation to speech, this chapter and the next explore the information revealed by the analysis of two types of speech information patterns - phonological and cooccurrence-based (including syntactic and semantic information), respectively.

This section has looked at models of a structured mental lexicon organised in terms of similarity between words at different levels. The next section focuses on the phonological level of the lexicon, reviewing metrics of phonological similarity and finally presenting an empirical study to measure the impact of different parameters on perceived phonological similarity between words.

3.2 Phonological similarity

The phonological level of the lexicon is composed of discrete units: the segments of the language. Words are temporal combinations of those segments. The similarity between two words in the phonological space depends on the configuration of the space in that language - two words may be perceived as phonologically similar for example if they share the initial segment; if they rhyme; if they both contain segments with the same place or manner of articulation; if they are stressed on the same syllable.

The function of detecting phonological similarities (or differences) between words is to classify and distinguish lexical items. Some phonological features of a word may contribute more than others to lexical classification. The aspects of words where similarity is more easily detected in a particular language must correspond to the more salient parameters of the phonological word representation in the mental lexicon. Or, in other words, phonological aspects of the mental lexicon organization that contribute more to classification will be the aspects that are easier to detect by the processor. Finding which parameters of word form have a greater impact on similarity, and how they relate to each other can help us understand the functions of those parameters in the organization of the mental lexicon of a language. Such an analysis can contribute to understand the processor's biases to pay attention to specific lexical aspects, and the adaptations of the lexicon to those biases.

This section reviews metrics of phonological similarity and presents an empirical approach to measure the relative importance of word-form parameters for the detection of word-form similarity in Spanish words. This study examines two sets of bisyllabic word structures (cvcv and cvccv) and attempts to establish the impact of the different segments, of stress and of syllabic structure on perceived word-form similarity.

3.2.1 Metrics of phonological similarity

Different lines of research have devised metrics of phonological similarity, from purely psycholinguistic studies to language engineering.

Many methods to measure phonological similarity consider the segmental level. Focusing on speech production, speech error analyses such as slip-of-the-tongue studies provide information on the importance of different segments for overall word-form similarity. By comparing which phonemes are replaced by which in speech errors, Stemberger (1991) composed a confusion matrix that quantifies the degree of confusability between each phoneme pair. The more confusable two phonemes are, the more similar they are assumed to be in a language's phonological representation map. Stemberger's confusion matrix has been used to test the accuracy and psychological plausibility of other similarity metrics. For instance, Frisch (1996) used it to support his choice of a phoneme similarity metric based on phoneme representations derived from Broe's (1993) structured specification theory.

The priming paradigm has been used in psycholinguistics to study word recognition (see an overview in Zwitserlood, 1996). A target word is preceded by a prime word that shares one parameter with the target. If the prime affects the processing of the target, then the parameter they share must be involved in lexical access. In priming studies of phonological similarity, when the prime and target shared the initial segments the results are conflicting (see review in Radeau, Morais & Seguí, 1995), but sharing the final segments, particularly if they rhyme, has been shown to facilitate target processing (see Dumay et al., 2001, for review).

Phonological similarity has also been measured in relation to the phonological similarity effect described by Conrad and Hull (1964), who found that when people are asked to recall a list of words, they perform worse if the words sound similar to each other. (Although Lian & Karslen, 2004, recently found that the effect depends on the type of phonological similarity considered, as reviewed in § 6.2.3.3 in chapter six). This effect is also found when words are read instead of heard, which is best explained by Baddeley and Hitch (1974) model of working memory that includes a component that recodes visual (orthographic) information into a phonological representation. In order to study the phonological similarity effect, researchers needed sets of phonologically similar and dissimilar stimuli. One method used to quantify phonological *dis*similarity is Psimetrica (Phonological SImilarity METRIC Analysis), developed by Mueller, Seymour, Krawitz, Kieras and Meyer (2003) to test models of verbal working memory – yielding results in support of Baddeley's model. For each word pair, Psimetrica returns a multi-dimensional vector that includes information about dissimilarity along a number of parameters such as rhyme, stress pattern or syllable onset match. This technique first defines each word in terms of a number of parameters or dimensions, it then aligns the two words and quantifies the level of matching for each dimension and finally, the results are averaged over all the word pairs to yield the mean phonological dissimilarity profile of the word set.

Several methods for measuring word similarity across languages were developed with the purpose of automatic cognate identification. Cognates are words from different languages that share the same etymological origin, such as 'pronounce' in English and 'pronunciar' in Spanish, both derived from the Latin verb ‘pronuntiare’. These methods look for orthographically or phonetically similar words across different languages. This task involves searching and matching, including finding the word alignment that yields the best possible similarity score. Some of these methods measure the similarity of *orthographic* forms, such as the Longest Common Subsequence Ratio or LCSR (Melamed, 1999), which divides the length of the common subsequence (common characters in the same order) by the length of the longest of the two strings; and Dice’s coefficient, used by Brew and McKelvie (1996) which equals the number of shared bigrams multiplied by two divided by the sum of bigrams from the two strings. Other methods measure the similarity of *phonological* forms, such as ALINE (Kondrak, 2000), that uses a list of parameters based on phonological features ranked by salience and then finds the optimal alignment of strings. The best parameter values for finding cognates are found by a hill-climbing search that optimises the values for the task at hand (in this case, cognate matching).

McMahon and McMahon (2003) propose that quantitative methods drawn from the field of genetics should be applied to language classification into families. They used measures of phonological similarity between cognates to generate an unrooted phylogeny tree for Indo-European languages. Another quantitative approach is that of Kirby and Ellison (in preparation), who carried out a study of language phylogeny based on similarity within and between languages. They created vector representations of the phonological lexicons of 95 different languages (using edit-distances to compare words within each language – and *not* cognates across languages). They then compared the 95 languages using the divergence of their distributions of confusion probabilities. Finally, using a neighbour joining algorithm, they constructed a language phylogenetic tree that reflected a plausible evolution of the Indo-European language family.

Phonological similarity is also used in a spoken document retrieval method (Crestani, 2003) that combined phonological and semantic similarity of the term used in a search with the terms contained in the documents to be searched. Crestani used a metric of phonological similarity between two words devised by Ng (1999) that uses the values in a phone confusion matrix (how liable is each phoneme to be misperceived or used instead of another one).

The last few paragraphs present many studies that have measured phonological similarity, some focusing on individual segments and some on whole words, for a variety of purposes, briefly summarised in Table 3.1.

Parameters	**Paradigm**	**Results**
Shared phonemes	Speech errors	Phoneme confusion matrix
Shared segmental positions	Priming	Determines impact of parameters on lexical processing
Various at different levels (rhyme, stress, shared sequences)	Quantitative methods	Quantifies impact of parameters on phonological similarity

Table 3.1. Summary of metrics of phonological similarity.

The next section presents a metric of similarity between whole word-forms based on identity at the segmental level that measures the relative importance of the position, the stress pattern and the syllabic structure. This metric is different from the ones described above in several respects. First, unlike the phoneme similarity studies, I measure whole word similarity rather than single segments. Second, using a psycholinguistic methodology means that, as in the case of priming studies, I am not measuring pure phonological similarity, but rather word-form similarity, since other factors such as morphology may affect the results. Third, unlike cognate identification and document matching, this empirical metric is not looking for certain types of similarity with a specific purpose in mind. Rather, I offer parameter combinations in a forced choice task and analyze people's responses. Finally, this method does not take into account the identity of the phonemes compared, as spoken document retrieval systems do. Instead, I consider the positions in the words together with information on whether they are consonants or vowels and whether they are stressed or not. I then measure the impact of these parameters on the estimation of the overall perceived similarity between word-forms.

The next section describes the study and discusses the results in the light of current psycholinguistic theories.

3.2.2 Word-form similarity perception in Spanish: an empirical approach

The focus of this thesis is the structure of the mental lexicon, assuming a lexicon organised in terms of relationships of similarity between words at different levels. This section concentrates on building up a quantitative model of the phonological mental lexicon; chapter four presents a quantitative approach to the syntax-semantic lexicon. Chapter five brings the two together and looks for systematic relationships between the two levels.

This section presents an empirical metric of phonological similarity aimed at determining the relative impact of different parameters on the perception of phonological word similarity. Using a forced-choice paradigm on pseudo-words, this Internet-based study tested the impact of sharing single and multiple segments (e.g. sharing the initial consonant; sharing all the vowels) and stress on two word structures: cvcv and cvccv. The resulting parameter values are used later to configure the quantitative model of the phonological lexicon.

3.2.2.1 Participants

All participants had Spanish as their mother tongue and lived in Spain. 55 participants (30 male, 25 female) between their teens and their sixties, from ten Spanish regions participated in this on-line study (See Table 3.2 for full demographic data). They were recruited through an e-mail message sent to a linguistics web forum, to friends and also to university students requesting them to take part in an experiment and forward the message on to their acquaintances. Participants were directed to a web form containing the instructions and the materials. At the end of the form there was a small questionnaire where they were asked about their region of origin, age group, sex and about the main strategy they had followed while doing the test (simply looking at the words, reading them in their heads or reading them out loud).

Origin		**Age**		**Strategy**	
Madrid	13	< 20	1	Look	2
Galicia	12	21-30	15	Loud	30
Andalusia	12	31-40	23	Silent	23
Murcia	4	41-50	10		
Asturias	4	51-60	5		
Castille	4	> 60	1		
Aragon	2				
Basque C.	2				
Catalonia	1				
Valencia	1				

Table 3.2. Participant age, origin and strategy.

3.2.2.2 Materials

The participants were presented with orthographic stimuli on a computer screen. In a study concentrating on phonological aspects of the word-form, stimuli could have been acoustic, but this posed difficulties in an Internet-based study. The unreliability of the quality of sound data over the Internet and of the sound playing equipment in remote terminals shifted the balance in favour of orthographic stimuli. I assume that Baddeley and Hitch’s (1974) orthography-to-phonology recoding system mentioned above is at work; besides, the instructions to the participants emphasized that they should focus on the sound of the stimulus nonwords. This means that participants are accessing their idealised phonological representations, which should be conventionally equivalent for all speakers of the same language.

Participants were presented with nonword triads like the one shown in Table 3.3, containing one nonword on the left and two on the right such that the two on the right were similar to each other, and different to the one on the left, except that each of them shared one parameter with it. Since every word must be stressed on one syllable, when stress was not an issue all three words in a triad would be stressed on the first syllable (most common, unmarked stress in Spanish). All the possible parameter combinations for cvcv and cvccv words were presented. Parameter combinations that cannot occur simultaneously such as ‘sharing the stress on the first syllable vs. sharing the stress on the second syllable’ were excluded (see Appendix B for complete stimulus list).

súnta	o mélto
	o múlko

Table 3.3. An example nonword triad. In this case the top word on the right shares the third consonant (t) with the word on the left and the bottom word shares the stressed vowel in the first syllable (ú). These are the two parameters that we are comparing here.

I prepared two stimulus lists, each consisting of 83 cvccv and 39 cvcv triads (a total of 122). Each triad represented one parameter combination. The parameters are features that two words can have in common. Table 3.4 shows the parameters used in this study for the two word groups. To avoid any order effects, the two words on the right of the triad would appear in each possible order about half of the time. In order to keep the test time low and encourage participation and completion, each informant only saw a set of 45 triads that were randomly selected from one set of 122, presented in a random order. A random ordering was automatically generated each time the experiment was run, so it was different for each participant.

	cvcv		**cvccv**	
Single segment	c1	Same initial consonant	c1	Same initial consonant
			c2	Same 2nd consonant
	c2	Same 2nd consonant	c3	Same 3rd consonant
	v1	Same 1st vowel	v1	Same 1st vowel
	v2	Same 2nd vowel	v2	Same 2nd vowel
Multiple segment			tc13	Same consonants 1 and 3
			tc23	Same consonants 2 and 3
	tc	Same two consonants	3c	Same three consonants
	tv	Same two vowels	tv	Same two vowels
Syllable structure			str	Same syllabic structure (cvc-cv or cv-ccv)
Stress	s1	Same stress (on 1st syllable)	s1	Same stress (on 1st syllable)
	s2	Same stress (on 2nd syllable)	s2	Same stress (on 2nd syllable)
	sv2	Same stressed vowel in the 1st syll	sv1	Same stressed vowel in the 1st syllable
	sv2	Same stressed vowel in the 2nd syll	sv2	Same stressed vowel in the 2nd syllable

Table 3.4. Parameters used in the study for cvcv and cvccv nonwords.

Note that not all these parameters are independent of each other. (The only truly independent parameters are the single segment parameters c1, c2, c3, v1 and v2). When two words share the parameter ‘two vowels’, they necessarily share ‘vowel 1’ and ‘vowel 2’ as well. As we see in the example in Table 3.5 below, all three words share the first vowel, so when people decide which of the words on the right is more similar to *mópi*, the result is measuring the influence of parameter ‘sharing the second vowel (when they already share the first one)’.

mópi	o sóte
	o sóti

Table 3.5. Example triad comparing non-independent parameters ‘vowel 1’ and ‘two vowels’.

Stimulus nonwords were written using only letters with a transparent orthography-phonetics relationship, also avoiding the use of Spanish graphemes *ñ*, *ch* and *ll*. In order to make the nonword stimuli natural to the

Spanish ear, their phonotactic probabilities were matched to words of the same structure extracted from a corpus (Marcos Marín, 1992). The frequencies of the consonants in the stimulus nonwords mirrored those of words of the same structures from a speech corpus. For cvcv words, the similarity between the distribution of the consonants was significantly correlated with their corpus counterparts (first consonant: Pearson's r = 0.82; df = 12; p<0.001; second consonant: Pearson's r = 0.69; df = 11; p<0.01). For cvccv words, consonant cluster frequencies were significantly correlated with those in the corpus (Pearson's r = 0.70; df = 49; p<0.001), but the similarity of first consonants was not (Pearson's r = 0.46; df = 11; p<0.09). Note that given the small set of frequent consonant clusters in Spanish, it is difficult to find combinations of cluster and first consonant that are not real words for the cvccv set.

Another measure of wordlikeness (the extent to which a sound string is typical of a language) is the lexical neighbourhood density. Neighbourhood density is calculated by counting the number of words in a corpus (Marcos Marín, 1992) that sound similar to a target: (a) words of the same length that differed from the stimuli by a 1-phoneme substitution (measure used by e.g. Ziegler, Muneaux & Grainger, 2003); (b) words up to 6 phonemes (for 4-phoneme stimuli) and up to 8 phonemes (for 5-phoneme stimuli) that contained the stimulus - similar to Stoianov's (2001) approach, who consider syllables that share at least two segments as contributing towards similarity, and neighbourhood; and (c) longer words that rhymed with the stimulus – the neighbourhood density metric used by De Cara and Goswami's (2003) includes rhyme, and indeed they suggest that rhyme has a special role in the development of phonological awareness. As expected, the neighbourhood densities of the stimulus nonwords are lower than those of similar real words.

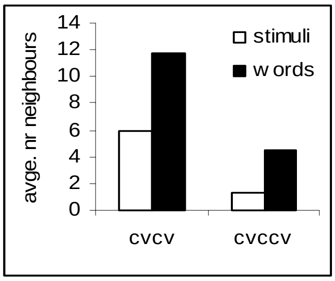


Figure 3.2. Neighbourhood density (average number of phonological neighbours) of stimulus nonwords and similar words from the corpus.

Figure 3.2 compares the neighbourhoods of cvcv and cvccv nonwords with those of all the words of the same structure found in the corpus. The phonological space of a language is to a large extent used up by the real words. Phonologically adaptive forms are used again and again, that is why nonwords cannot be expected to have as many neighbours as real words, but there are similarly shaped distributions of neighbours for cvcv and cvccv.

Even though place and manner of articulation features in the stimuli were not controlled for, in most cases the consonants used in one stimulus set are different from the consonants used in the other stimulus set for the same comparison.

3.2.2.3 Method

Participants were directed to an Internet link to the experiment web page, where they first saw the instructions. Below these were 45 stimulus triads randomly selected from the 122 from one set. Participants were asked to read the nonword triads and determine which of the two words on the right *sounded* more similar to the word on the left. Participants were directed to pay attention to stress, which was marked in all stimuli as an acute on the corresponding vowel (the usual orthographic stress mark in Spanish). Additionally, participants were instructed not to think too much, and to select their first spontaneous choice. The results, together with the demographic data, were automatically emailed back to the experimenter for analysis.

This design was intended to reveal which parts or aspects of the word people focus on more when they read the stimuli, particularly what is more salient when they look for phonological similarity. Sound similarity is mentioned in the instructions, so we were recording meditated choices, rather than automatic responses as in priming paradigms. Also, participants could take their time and read the stimuli several times, which allows for other factors to play a role in the choice. When asked which of the two stimuli on the right sounds more like the target, people activate and match the representations of the three stimuli. I used nonwords so that direct semantic representations were not available, although the semantics of the stimuli's phonological neighbours could influence the choice. In order to minimise that problem, I used two different nonword triads for each parameter comparison.

3.2.2.4 Results

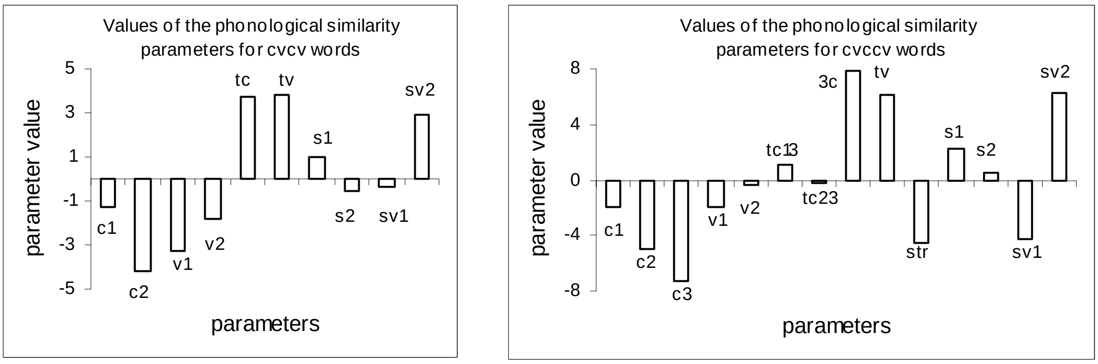
I obtained an average of 20 responses (minimum: 10, maximum: 31) for each pairwise parameter comparison. The results for cvcv and cvccv stimuli were analyzed separately. For each pairwise comparison of parameters, I counted the proportions of respondents that favoured each option to obtain the 'winner' of that comparison. For example, in the triad comparing parameter 'c1' and 'c2' for cvccv words, 14 out of 21 respondents selected 'c1' and 7 selected 'c2', so the winner is 'c1'. I then calculated a weight between zero and one that expressed the confidence of the result that the winner is 'b', such that if everybody prefers the same parameter the weight for the winner is 1; if the responses were fifty-fifty, the weight is 0, and there is no winner. I calculate that by dividing the difference between the number of people who chose 'c1' (14) minus the number of people who chose 'c2' (7) divided by the total of responses (21). In our example, (14-7)/21=0.33, meaning that 0.33 more people preferred 'c1' than 'c2', so for this comparison we would have a weight of 0.33. Tables 3.6 and 3.7 below show matrices containing the weights obtained for all the pairwise parameter comparisons for cvcv and cvccv stimuli.

cvcv	**c1**	**c2**	**v1**	**v2**	**tc**	**tv**	**s1**	**s2**	**sv1**	**sv2**
c1		-0.57	-0.48	-0.38	1	0.64	0.42	0.21	-0.09	0.5
c2	0.57		-0.07	0.58	0.56	0.8	0.65	0.48	-0.07	0.73
v1	0.48	0.07		-0.17	1	0.86	0.43	-0.22	n.a.	0.81
v2	0.38	-0.58	0.17		0.58	0.17	0.58	0.26	0.26	n.a.
tc	-1	-0.56	-1	-0.58		0.22	0.04	-0.16	-0.71	0.07
tv	-0.64	-0.8	-0.86	-0.17	-0.22		-0.4	-0.24	-0.45	-0.08
s1	-0.42	-0.65	-0.43	-0.58	-0.04	0.4		n.a.	0.74	n.a.
s2	-0.21	-0.48	0.22	-0.26	0.16	0.24	n.a.		n.a.	0.86
sv1	0.09	0.07	n.a.	-0.26	0.71	0.45	-0.74	n.a.		n.a.
sv2	-0.5	-0.73	-0.81	n.a.	-0.07	0.08	n.a.	-0.86	n.a.	
All param.	**-1.25**	**-4.21**	**-3.26**	**-1.82**	**3.68**	**3.85**	**0.97**	**-0.52**	**-0.32**	**2.88**
All segm.	**-0.20**	**-2.43**	**-2.24**	**-0.72**	**2.92**	**2.68**				
Single seg.	**1.43**	**-1.07**	**-0.38**	**0.02**						

cvccv	**c1**	**c2**	**c3**	**tc13**	**tc23**	**3c**	**v1**	**v2**	**tv**	**s1**	**s2**	**sv1**	**sv2**	**str**
c1		-0.33	-0.5	0.4	0.05	0.88	-0.5	-0.28	0.83	0.58	0.57	-0.1	0.74	-0.33
c2	0.33		-0.57	0.41	1	1	0.36	0.6	0.55	0.79	0.3	-0.11	1	-0.62
c3	0.5	0.57		1	0.69	1	0.83	0.26	1	0.65	0.6	0.08	0.83	-0.68
tc13	-0.4	-0.41	-1		0.06	0.83	-0.17	-0.13	0.43	0.33	-0.14	-0.55	0.67	-0.58
tc23	-0.05	-1	-0.69	-0.06		0.69	0.42	-0.08	0.13	0.39	0.69	-0.88	0.68	n.a.
3c	-0.88	-1	-1	-0.83	-0.69		-0.86	-0.83	-0.33	-0.09	-0.33	-0.8	-0.25	n.a.
v1	0.5	-0.36	-0.83	0.17	-0.42	0.86		-0.04	1	0.29	0.57	n.a.	0.67	-0.44
v2	0.28	-0.6	-0.26	0.13	0.08	0.83	0.04		1	-0.08	-0.27	-0.05	n.a.	-0.76
tv	-0.83	-0.55	-1	-0.43	-0.13	0.33	-1	-1		0.17	-0.55	-0.45	0	-0.64
s1	-0.58	-0.79	-0.65	-0.33	-0.39	0.09	-0.29	0.08	-0.17		n.a.	-0.74	n.a.	0
s2	-0.57	-0.3	-0.6	0.14	-0.69	0.33	-0.57	0.27	0.55	n.a.		n.a.	1	-0.14
sv1	0.1	0.11	-0.08	0.55	0.88	0.8	n.a.	0.05	0.45	-0.74	n.a.	n.a.	n.a.	0.6
sv2	-0.74	-1	-0.83	-0.67	-0.68	0.25	-0.67	n.a.	0	n.a.	-1			-1
str	0.33	0.62	0.68	0.58	n.a.	n.a.	0.44	0.76	0.64	0	0.14	-0.6	1	
All par.	**-1.99**	**-5.04**	**-7.33**	**1.05**	**-0.25**	**7.90**	**-1.96**	**-0.38**	**6.07**	**2.30**	**0.57**	**-4.20**	**6.34**	**-4.60**
All segm.	**-0.21**	**-3.06**	**-5.17**	**1.37**	**0.64**	**6.42**	**-0.43**	**-0.74**	**5.23**					
Sing. sg.	**1.61**	**-0.73**	**-2.16**				**0.73**	**0.54**						

Tables 3.6 and 3.7. Results of all parameter comparisons for cvcv and cvccv words. A positive value means that the parameter on the top row is the winner and a negative value means the parameter on the left column is the winner. The top right halves of the matrices are completed with the corresponding values (multiplied by (-1)). Missing values correspond to parameter combinations that were impossible to combine in a stimulus triad. The three bottom rows of each table show the parameter values considering all parameters (obtained by adding the values on each column), considering all but the stress and structure related parameters (obtained by adding the values on the cells corresponding to the segment-related parameters), and considering the single segments only (obtained by adding the values on the cells corresponding to the individual segments only). C1, c2, c3 = consonants 1, 2 and 3; v1, v2 = vowels 1 and 2; tc = two consonants; tv = two vowels; 3c = three consonants; s1, s2 = same stress on the 1st and 2nd syllable; sv1, sv2 = same stressed vowel on the 1st and 2nd syllable.

In order to obtain a single value for each parameter, I sum the weights for that parameter with respect to all the other parameters (add each column of Tables 3.6 and 3.7). A positive value indicates that the parameter is a net winner, that is to say it wins over more parameters than it loses against, and/or it scores higher relative to other parameters. A negative value indicates it is a net loser – it loses against more parameters than it wins over and/or it scores lower relative to other parameters. Figures 3.3 and 3.4 illustrate the parameter values obtained when we take all parameters into account for cvcv and cvccv stimuli (the general parameter values). These values are unitless, and, because they have been calculated on a square matrix, they add up to zero (for each word group).



Figures 3.3 and 3.4: Parameter values for cvcv and cvccv words.

The results show a high consistency between the two independent word groups cvcv and cvccv: the values of counterpart parameters in cvcv and cvccv are significantly correlated when we take all parameters into account (R^2=0.83, df =11, p<0.01) and when we take all the single and multiple segment parameters into account (R^2=0.88, df =7, p<0.02). They are not significantly correlated for single segment parameters only (R^2=0.71, df =2, n.s.) but this is only due to the low number of data points compared (four, which gives 2 degrees of freedom for the calculation of the significance).

This between-group consistency indicates that participants made similar choices for the two independent word groups. This adds internal consistency and robustness to the study, and validity to the methodology employed.

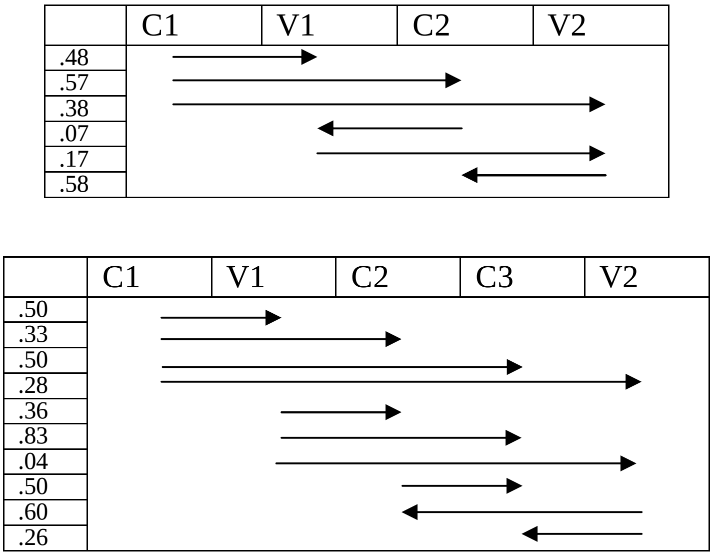
The results for each parameter in the three different measurements (taking into account all parameters, multiple segments and single segments) are well correlated, as can be seen in Figures 3.3 and 3.4. The only visible discrepancy is the relative values of the first and second vowels (*v1* and *v2*) in cvccv words. This is due to the fact that it was impossible to construct a stimulus triad for the comparison *v2* against *sv2* (same stressed vowel in the same syllable). The high value of *sv2* in this word group lowered the values of the parameters that were compared against it, but not of *v2*.

The high number of regions of origin involved for 55 participants does not allow for accurate measurements of region of origin effects. The strategy followed by participants (reading the stimuli out loud or silent - imagining the sound of it in their heads) did not have an effect on the parameter values (the correlation between the results generated with loud and silent strategies is R^2= 0.76 for both cvcv and cvccv).

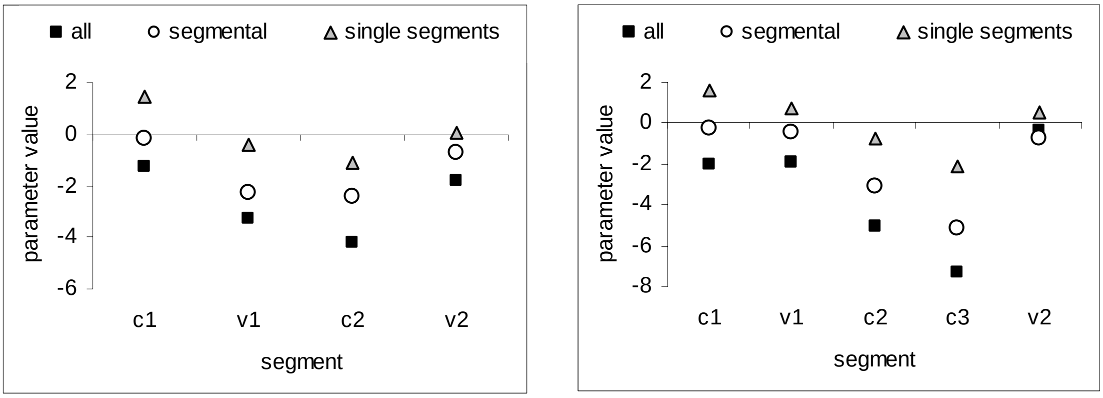
3.2.2.5 Discussion

Segment positions

The values of the individual comparisons and the general parameter values reveal different aspects of word form similarity processing. For instance, we can compare the relative importance of the different segment positions in the assessment of word similarity either by examining how each segment fares against each of the other word segments (see Figures 3.5 and 3.6) or by comparing the general values of the segmental units; the first method focuses on relationships at the segment level, and the second takes into account the more complex and subtle relationships of each segment with all segmental and nonsegmental parameters that characterize its general value.



Figures 3.5 and 3.6. Relationships between the pairs of segmental parameters for cvcv and cvccv words. The arrows depart from the winner and arrive at the loser. The left-hand column shows the weights.

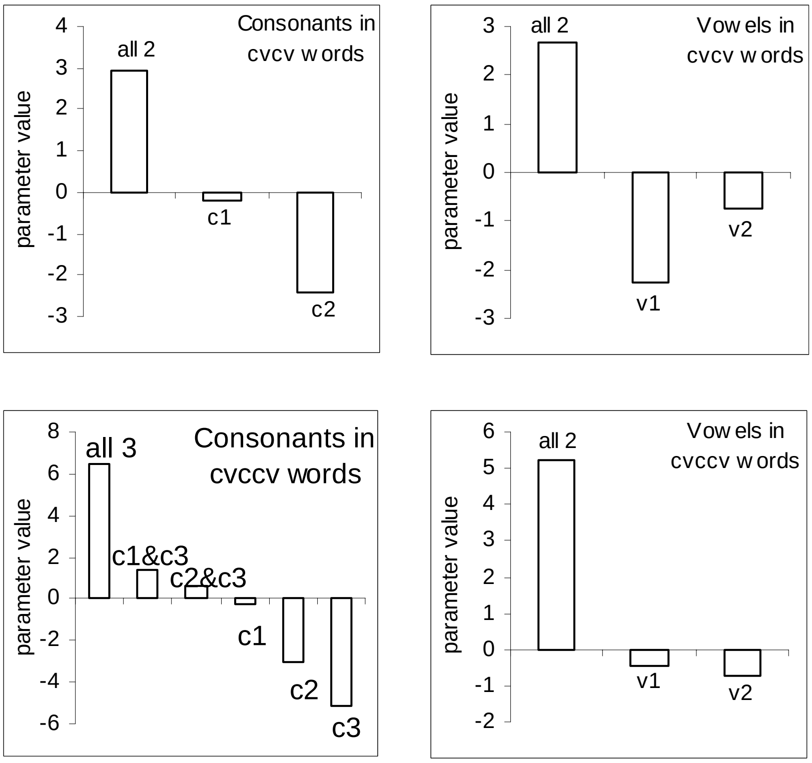


Figures 3.7 and 3.8. Parameter values of the segment positions of cvcv and cvccv words, measured considering all parameters, all segmental parameters and single segment parameters, and the single segment parameters only.

Figures 3.7 and 3.8 show the results obtained using single segment information (grey triangles), using all segment-related information (white circles), and using all parameters, including information about stress and syllabic structure (black squares) (see calculation of the values in Tables 3.7 and 3.7 above). The single segmental measurements only take into account the information shown in Figures 3.5 and 3.6 above, and miss the fact that all segments are net losers (negative parameter values) with respect to the whole parameter set. We see that similarity at word initial and final segments is perceived more readily than in the middle of the word, another expression of the ‘bathtub effect’ found by Brown and McNeil (1966) in their tip-of-the-tongue studies. (The name of this effect comes from the image of the word represented as in figures 3.7 and 3.8, as someone lying in a bathtub with only their head and feet above the surface of the water.) Brown and McNeil (1966) read to participants the definitions of relatively obscure target words and then recorded the (wrong) words they produced when they claimed to have the target ‘in the tip of their tongue’. Some of the words recorded sounded like the targets and other had similar meanings. They counted the matches between the few initial and final segments of the targets and the words recorded and found a bathtub effect, with matches at the beginning and end of the word up to 50% of the times and much less matching in the middle. Studies of malapropisms (wrongly selected similar-sounding words recorded from natural speech) show even higher matches or near-matches between errors and targets at the word initial (80%) and word final (70%) sounds, with much lower agreement levels in the middles (Aitchison & Straf, 1982; see also Fay & Cutler, 1977 and Hurford, 1981). This effect is most salient for the initial and final segment, rapidly decaying already for the second and last-but-one segments (Browman, 1978). The present results support the claim that word initial and final segments are prominent in lexical representation.

Vowels and consonants

Figures 3.9-3.12 show the relative importance of vowels and consonants using all the segment-related parameters in cvcv and cvccv words (see calculation of the values in Tables 3.7 and 3.8 above).



Figures 3.9, 3.10, 3.11 and 3.12. Parameter values for single and multiple consonants and vowels cvcv and cvccv words.

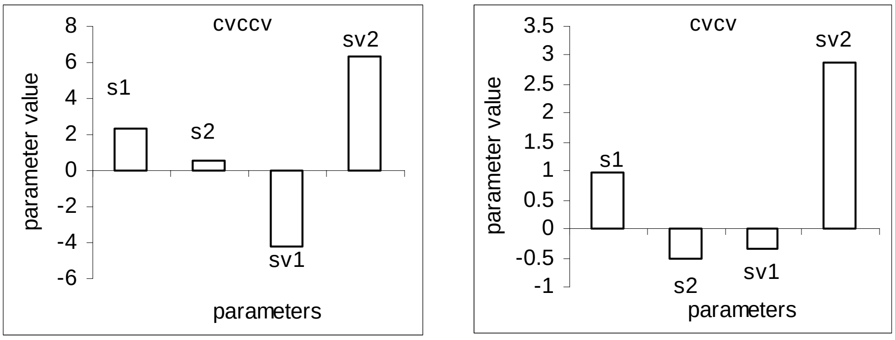
The most salient feature in Figures 3.9-3.12 is, unsurprisingly, that the more consonants or vowels two words share, the more similar they are perceived to be. Additionally, the consonant structure (sharing all consonants) and the vowel structure (sharing all vowels) have an equivalent weight in determining similarity in both word groups (after allowing for the fact that cvccv word consonant structure has three elements whereas cvcv has only two). Some studies have suggested that vowels and consonants are processed by distinct neural mechanisms at the cortical level. Caramazza, Chialant, Capazzo and Miceli (2000) described the case of two aphasics with impaired processing of vowels and consonants, respectively. 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. These studies suggest not only that consonant and vowel processing are distinct but also that the vowel structure, being processed together with tone, could be a kind of supra-segmental level exploited in speech perception. (Differences between consonant and vowel processing are further explored in chapter six of this thesis.)

In cvccv words, sharing the two syllable-initial consonants c1 and c3 (equivalent to the two consonants in cvcv words) has a higher value than sharing consonants c2 and c3. In the great majority of stimulus nonwords, consonants c2 and c3 are in different syllables (e.g. *bún-ta, kás-te*). Again, the limited stimulus cv structures precludes a full analysis, but this result indicates that syllable-initial consonants form the skeleton of the word consonant structure. Clements (1991) proposed the Sonority Dispersion Principle that the sharper the rise in sonority between the beginning of the syllable and the nucleus, the better the syllable. The paradigm of 'good' syllable would, therefore, be 'cv'. If we assume a bias to process syllables as 'cv' as a default, it makes sense for syllable-initial consonants to form the word consonant skeleton and to be more salient in processing than for instance cluster consonants or syllable-final consonants.

As far as the individual segments are concerned, the word-initial consonant in both word groups is most salient for similarity perception, followed by the vowels; other consonants obtain the lowest values in any measurement (see Figures 3.3 and 3.4 above). The final vowels (and, in cvcv words, also the first vowel) have values close to the initial consonant. The study only included stimuli starting with a consonant and ending with a vowel, so these results cannot rule out that vowels are more salient than consonants, but given that the initial segment position (in our two word-groups) is very salient, consonants appear to be more salient than they really would be in a more heterogeneous stimulus set. If we had had stimuli starting with a vowel or ending with a consonant, we might have found that the vowel structure is more salient than the consonant structure. As for the end of words, whereas single consonant values are lower towards the end of the word, the final vowel, usually a site for a gender or a verbal morpheme, shows a relatively high value. This supports the idea that there is a bias for attention to be focused on the parts of the word in Spanish where morphological information concentrates. This hypothesis could be further tested with a cross-language study involving languages with different sites for morphology.

Stress

The design of this study included four parameters related to stress, namely sharing the stress on the first syllable, on the second syllable, sharing the same stressed vowel on the first syllable and on the second syllable. Figures 3.13 and 3.14 show the general values for these parameters (calculated taking into account comparisons with all parameters).



Figures 3.13 and 3.14. General values of the stress parameters in the two word groups. s1 = stress on 1st syllable; s2 = stress on 2nd syllable; sv1 = same stressed vowel on 1st syllable; sv2 = same stressed vowel on 2nd syllable.

Sharing the stressed vowel in final position (*sv2*) obtains the highest values for both word groups, as well as third and second positions in the general parameter rankings for cvcv and cvccv words, respectively, as seen in Figures 3.3 and 3.4. This could reflect the fact that the most common Spanish verb tenses (present, simple past and future) and persons (first and third singular) are encoded by contrasts in the identity and stress of the final vowel (Table 3.8). See Appendix D for a full list of the 31 words stressed on the last syllable out of the 324 cvcv words of frequency greater or equal to 100 in a Spanish speech corpus (Marcos Marin, 1992).

	-ar			***-er, -ir***		
	pres	*past*	*fut*	*pres*	*past*	*fut*
1st	-o	-é	-ré	-o	-í	-ré
3rd	-a	-ó	-rá	-e	-ió	-rá

Table 3.8: Regular verb morphemes for first and third persons (singular) in the three most common tenses in Spanish (present, simple past and future) for verbs ending in *–ar, -er* and –*ir*.

The stimuli were nonwords, but we cannot claim that the use of nonwords precludes the perception of word-final phonemes as morphemes. E.g. the nonword *bunkí* could be perceived as the first person singular of the past tense of non-verb '*bunker*' or '*bunkir*'. If morphology perception interferes with phonology perception, in the triad [*bunkí* (*teská* or *tesmí*)], *tesmí* could be found more similar to *bunkí* because it could be perceived to be sharing the same tense and person.

In order to explore this issue, I included stimuli ending in *ú*, which is not a verbal morpheme. However, in such triads participants still found words sharing the stressed *ú* more similar than those sharing any other parameter, including sharing the three consonants. All participants responding to the triad [*kandú* (*kindá* or *pirgú*)] found *pirgú* was more similar to the base word than *kindá*. Morphology, then, cannot be directly responsible for the high score of the parameter 'same stressed vowel on the second syllable'. However, important information such as verb morphology occurs at the word final position when it is occupied by a stressed vowel. It could be adaptive to focus attention on any phonological variation in that segment position when it is stressed. Stressed final vowels, then, seem to be very salient in terms of perceived form similarity in Spanish.

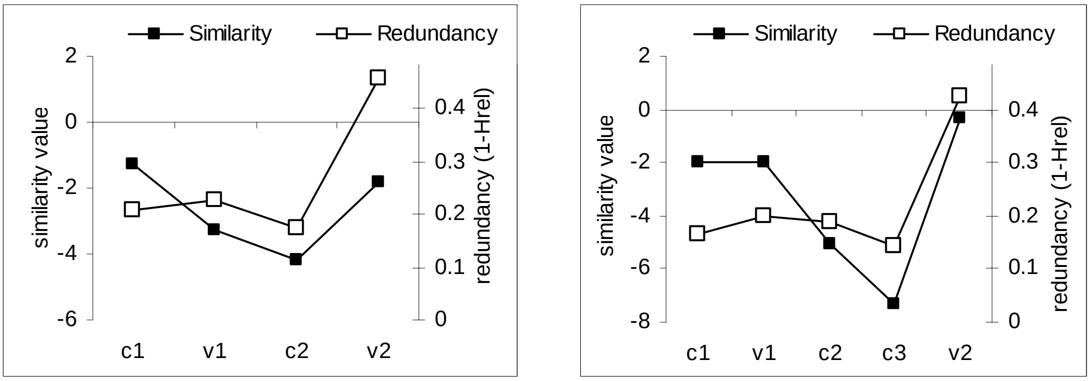
Sharing the stress on the first syllable generated high general values and ranking position (fourth for both cvcv and cvccv, see Figures 3.3 and 3.4). In Spanish, most words are stressed on the penultimate syllable: see e.g. the stress distribution of the results in Barkanyi (2002) for common word structures, or the small proportion (10%) of words stressed on the last syllable in the 324 cvcv words of frequency greater or equal to 100 in a Spanish speech corpus (Marcos Marin, 1992). This means that, even though many words share the stress on the first syllable, this parameter is salient in the perception of phonological similarity.

Syllabic structure

This parameter only applies to cvccv words, and compares two possible syllable structures: cv-ccv and cvc-cv (e.g. *da-blo* vs. *dan-go*) within the same consonant-vowel structure (cvccv). This parameter loses to every other parameter, which suggests it is of little importance for the perception of word-form similarity in cvccv words. Rouibah and Taft (2001) examined the processing units involved in the reading of French polysyllabic words and concluded that 'the syllabic structure that is so clearly manifested in the spoken form of French is not involved in visual word recognition'. Perhaps, then, visual presentation of the stimuli is obscuring the effect of this parameter on perceived word-form similarity, and auditory presentation would have resulted in a higher parameter value.

3.2.2.6 Comparison with the information profiles

Figures 3.7 and 3.8 above (single segments) can be compared with the information profiles in chapter two. The profiles resulting from the results above represent prominence in aspects of lexical access related to word-form similarity. The segmental entropy used in chapter two reflects the joint effect of all the pressures on intra-word phonological form. Van Son and Pols (2003) proposed that greater speaking effort is concentrated on more information-laden parts of the word, whereas predictable items are phonologically reduced. Mirroring this reasoning for perception, I assume that more attention is paid to the parts of the word where information tends to be concentrated. Figures 3.15 and 3.16 show the information prominence of cvcv and cvccv words (the redundancy values of the segments, redundancy being 1 – entropy, from chapter two) and the attention prominence of cvcv and cvccv word segments (the segmental parameter values: the black squares in Figures 3.7 and 3.8 above), which I interpret to be reflecting the amount of attention the corresponding segments attract (or how much people focus on the segments) when judging word-form similarity.



Figures 3.15 and 3.16. Similarity and information (redundancy) profiles of cvcv and cvccv words (lines drawn to show the profile shape). X-axis indicates parameters.

The correlations between these two measures are $R^2 = 0.21$ for cvcv words and $R^2 = 0.46$ for cvccv words. The relationship between these two profiles is that similarity should be easily detected in perceptually salient word segments, and redundancy, a measure of complexity and organization, should be higher in positions that encode aspects of the lexicon structure. Similarity and redundancy should be correlated in places where it is important to detect variability among a small number of possible segments that encode e.g. a morpheme. The closest parallels between the two profiles are the high values at the last segment preceded by the lowest values in the last-but-one segment. The main difference is the relatively high similarity and low redundancy in the first segment. The first segment in Spanish is not a usual site for morphological information, therefore it shows low redundancy, but it is important for word recognition, so it shows high similarity prominence (it is the focus of attention). The last segment, on the other hand, is the site of morphology (gender, number and verb inflections) in such short words, and shows, as expected, high redundancy values. It also shows high similarity, meaning that a lot of attention is focused on the identity of this segment.

3.2.2.7 Conclusions and future work

Different approaches to the study of phonological lexical structure and the relationships between them help to understand the functions of word-form parameters. This section has reviewed metrics of phonological similarity and has presented an empirical metric to analyse phonological similarity at the word level in Spanish that fills a gap in the literature. The empirical metric was based on a two-way forced choice between nonwords sharing different word-form parameters with a third nonword. These parameters were the same segment in the same position for individual and combinations of all word segment positions, the same stress pattern or stressed vowel and the same syllabic structure (in cvccv only).

In agreement with the findings of tip-of-the-tongue, speech error and malapropism studies, I observe that word initial and final segments are more salient than the middle ones in this similarity-judgement task. Word-initial salience and vowel and consonant structure salience could be related to phonological word representation (with implications for word production and recognition). Salience of the identity of a stressed vowel in word-final position could be explained as an enhanced attention to the usual site of morphology. Correlations between the information profile and the salience of single segments further stress the intertwined roles of morphology and lexical phonology in the perception of form similarity.

The paradigm presented may be used to establish a hierarchy of parameters of phonological similarity in other languages. A similar study applied to several languages could help establish if the relative salience of the parameters is universal or language-specific and help classify languages by the parameters configuring their phonological mental lexicon. Also, the application of the metrics reviewed at the beginning of the chapter to

Spanish could support (or otherwise) and potentially qualify the results found here by analysing other levels of phonological description of Spanish.

The quantitative parameter values obtained with the psycholinguistic study presented in this chapter will be used in chapter five to calculate word-pair similarity values in order to configure a phonological similarity lexicon structure. First, chapter four reviews and applies methods to calculate cooccurrence-based similarity between words.

Chapter 4. The structure of the mental lexicon as defined by patterns of word cooccurrence

The aim of this chapter is to obtain a representation of the syntactic-semantic level of the lexicon based on similarity between words. The chapter focuses on word-cooccurrence methods to define the position of words in the syntactic-semantic space and reviews ways of measuring similarity between them. It presents an application of one such metric to a subset of the Spanish lexicon and explores how this statistical approach to the representation of the mental lexicon performs in semantic and syntactic tasks.

4.1 A similarity-based semantic space

In the conceptual approach of semantics (Jackendoff, 1983) meaning is equated to conceptualizations, which are determined largely by the environment. I take the view that language itself is part of the environment that determines conceptualizations. Words and the way they are used in speech play a part in building the mental representations of concepts. The semantic space is configured by the structure of the world, but also by the structure of language. This interplay of mental representations and language is expressed as the grammatical constraint that Jackendoff (1983) puts on a theory of semantics: that a semantic theory must support systematicity in the relationship between syntax and semantics.

There are many approaches in the literature to configuring the semantic space (including semantics or meaning, syntax and possibly other kinds of information). Smith, Shoben and Rips (1974) proposed a feature-based semantic space where the defining dimensions are features such ‘red’, ‘living’ etc. Concepts are defined by sets of defining features, e.g. ‘pigeon’ could be defined as ‘living’, ‘flying’, ‘grey’, ‘with feathers’ etc. This model explained priming, naming delays, typicality effects and semantic deficits found in brain damaged patients, but it was criticised as being a descriptive framework that did not reflect the real underlying structure of the semantic lexicon. (See review of feature-based approaches to semantics in McNamara & Miller, 1989).

Connectionist approaches to modelling feature-based semantic (dis)similarity include Rodd, Gaskell and Marslen-Wilson's (2004) model of the effects of semantics in word recognition. They used a feedforward network architecture where words with multiple meanings had distributed representations in a high-dimensional semantic feature space; during the learning phase, each meaning formed one stable attractor basin.

Others have configured the semantic space using the structure of thesauri such as Roget’s thesaurus or WordNet. A thesaurus-based semantic space is defined by the distance between pairs of words in the thesaurus, and the basic metric is the number of links between nodes. For instance, already in the fifties Osgood, Suci, and Tannenbaum (1957) used Roget's Thesaurus to help construct bipolar scales based on semantic opposites, such as "good-bad" or "fast-slow" to measure the results of psychological experiments. Jarmasz (2003) reviews recent uses of Roget's thesaurus in natural language processing.

Budanitsky and Hirst (2001) assess the performance of methods based on WordNet in a spell-checking task. Thesaurus-based semantic spaces have been criticised because of the limited and inconsistent coverage provided by the available thesauri (Curran, 2004), and Budanitsky and Hirst (2001) point out that lexical semantic relatedness is often constructed in context and cannot be determined exclusively by resources such as WordNet.

One way to overcome these shortcomings is to substitute feature-sets and thesauri with word distributional statistics extracted from real language samples such as a large corpus. One such approach is Landauer and Dumais’ (1997) Latent Semantic Analysis (LSA). They counted occurrences of target words in whole articles of a children’s 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. They reduced the dimensionality of the word-by-article matrix using a technique called singular value decomposition. The resulting 500-dimension matrix represents a semantic space where the similarity between word types or between articles can be calculated. The LSA semantic space contains no information about word-order and hence syntax. The LSA approach has been used to explain 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 (Wiemer-Hastings, Wiemer-Hastings & Graesser, 1999; Kintsch, Steinhart, Stahl, Matthews & Lamb, 2000).

The rest of this chapter will focus on context window methods that locate words by considering what words they occur close to in text or speech. This is based on the idea that the meaning of a word is determined by the linguistic contexts in which it occurs. In context space models (also called hyperspace models because the resulting representations are multidimensional spaces) each target word is located by a vector whose components are counts of occurrences of context words in the vicinity of the target. The ‘vicinity’ is defined by the size and shape of a window, for instance, five words before or after the target word, or the preceding word only.

Figure 4.1 illustrates the process of calculating the vectors that represent the position of words in a cooccurrence-based hyperspace (see caption for explanation).

Although we are born with pretty much all the brain cells we will ever have, I believe it is the growth of the connections between these **cells** that accounts for the growth of the brain after birth. And what is amazing about the brain is that it is constantly evolving every moment we are alive, so that although born into a booming, buzzing confusion,(…)

	the	*we*	*of*	*is*	*that*	*are*
cells	3	1	1	0	1	0
brain	4	1	1	4	1	0
growth	4	0	0	1	1	0
born	1	1	0	0	1	2

Cooccurrence matrix:

Rows = target words.

Columns = context words.

Figure 4.1. Calculation of the cooccurrence vectors in a text. The counts are based on the piece of text at the top of the Figure. In this example the target words are four high-frequency content-words, and the context-words are six high-frequency function words. I consider a context window of five words before and after the target word in the text (e.g., for the first occurrence of the word ‘cell’, the window comprises the words in grey around it). The value in each cell in the cooccurrence matrix is the total number of times that the target and the context word appear within five words of each other in the text.

McDonald (2000) points out two properties of this kind of distributional statistics that make them appropriate for psycholinguistic modelling - objectivity and language independence. Distributional statistics are objective because they make minimal assumptions when exploiting the statistical patterns present in speech. As for language independence, results obtained using French, German and Mandarin corpora (Redington, Chater, Huang, Chang, Finch & Chen, 1995) mirror those obtained for English. The results of Curran (2004) also indicate that context window approaches to measuring semantic similarity yield reasonable results while being computationally cheap and orders of magnitude computationally faster than shallow parsers such as CASS, Sextant of Minipar (6-7 minutes as opposed to hours or even days to extract information from the same corpus) - see Curran (2004) for review.

Context space models have been used to categorise words syntactically (Finch & Chater, 1992; Redington, Chater & Finch, 1998; Daelemans, 1999;

Christiansen & Monaghan, in press), categorise words semantically (Levy & Bullinaria, 1998; McDonald, 2000; Curran, 2004) and model semantic and associative priming (Lund, Burgess & Atchley, 1995; Lund, Burgess & Audet, 1996; McDonald & Lowe, 1998; McDonald, 2000).

The calculation of a semantic vector space (which can be represented visually as Figure 3.1 in chapter three), requires the following elements:

- A corpus of text where the cooccurrences between targets and context-words will be counted,
- a set of target words,
- a set of context words, which provide the dimensions of the space,
- a context window around the target words, where the occurrences of context words are counted,
- a method to calculate the vectors,
- a method to calculate the distance between vectors.

The next section looks at these parameters in detail.

4.1.1 Elements and parameters of the semantic hyperspace

4.1.1.1 The corpus

The size of the corpus affects the robustness of the cooccurrence-based representations. Large corpora produce vector representations that are more immune to noise due to restricted corpus-size. Patel, Bullinaria and Levy (1998) and Curran (2004) found that an increase in the size of the corpus improved their results, even for very large corpora (Curran used a two billion word corpus). Several hyperspace studies in English use (subsets of) large corpora such as the British National Corpus (BNC, around 90 million written and 10 million transcribed spoken words) or USENET (a corpus of around 170 million word corpus of newsgroup text).

On the subject of spoken versus text corpora, McDonald (2000) gives three reasons why speech is better than text. First, speech is the primary environment for language acquisition. (I would add that speech is the primary source of human communication.) Second, the smaller type:token ratio of speech provides a more reliable source of contextual information and thus the construction of denser vectors. Third, the results he obtained with the spoken subset of the BNC (around 10 million words) fitted isolated word recognition data better than similar size text BNC subsets.

The chosen corpus may be lemmatised or otherwise prepared before counting the cooccurrences (e.g. McDonald, 2000). Lemmatisation removes all morphology and leaves only word stems, affecting the information carried by the vectors. This eliminates possible morphology-based clusters in the hyperspace. Annotated corpora can be used to disambiguate between homophones in the counts, refining the quality of the vectors, for example McDonald (2000) and Monaghan and Christiansen (2004) both took information about the syntactic category of words from the CELEX database; Curran (2004) marked up the corpus including sentence splitting, tokenization and part of speech tagging.

4.1.1.2 The context

Context window methods count occurrences of a number of context words within a window of a number of words before and/or after the target word. The target words are the nodes in the semantic space. More target words mean a more complete space.

The main variables in the context are the window size (how many words around the target word are considered) and shape (are the context-words to be counted to the left, to the right of the target, or both) and the number and choice of context words that are included in the calculation of the vector components.

The window extends over a number of words or characters to the left and / or to the right of the target word. Some studies employ large windows of around 500 words (Yarowsky, 1992; Beeferman, 1998), but this makes the calculations computationally expensive. Others use small windows both for syntactic and semantic categorisation tasks: Finch and Chater (1996), two words to either side; Lowe and McDonald (2000), 5 words to either side; McDonald (2000), up to 10-20 words to either side; Curran (2004), combinations of 1-3 words to either side (finding the best results for one word to each side and with two words to the left). Patel, Bullinaria and Levy (1998) searched the parameter space in an attempt to optimize the window size and shape against two evaluation criteria: the ratio of mean Euclidean distances between semantically related and unrelated words, and a measure of syntactic categorisation. They found that the best results were obtained by counting the left and right contexts separately (as two components of the vector), using window sizes between two and 16 words. However, Levy, Bullinaria and Patel (1998), using different criteria for the optimisation of the parameter space - semantic and syntactic categorisation and synonym choice - found that the best results were obtained by averaging the contents of the left and right windows with window sizes between one and seven words. Monaghan, Chater and Christiansen (in press) used a window of one word to the left (the preceding word only) for a noun-verb discrimination task (carried out using both distributional and phonological clues). Mintz (2003) developed a different form of window called 'frame' consisting of a pair of words that occur separated by one intervening word, e.g. 'a _ of'. He showed that frequently occurring frames accurately predicted the syntactic category of the intervening word. Monaghan and Christiansen (2004) compared Mintz's method with Monaghan, Chater and Christiansen's (in press) preceding word window and found that while the frames had a higher accuracy for noun-verb classification, the preceding word window classified a much higher proportion of words.

The number of context words determines the dimensionality of the space. It is usually a few hundred: Finch and Chater (1992) used 150 context words; Lund and Burgess (1996) used 200, and claimed that adding more context words did not alter the results; Lowe and McDonald (2000) used 536 context words; McDonald (2000) used 446 context words..

The choice of context words defines the type of information that the space captures. Some studies simply select the most common words in the corpus (Finch and Chater, 1992; Redington, Chater and Finch, 1998), while others remove from that set a series of very frequent uninformative words such as prepositions, conjunctions, determiners, pronouns etc, which they claim are so ubiquitous that they do not help judging semantic similarity (Lowe and McDonald, 2000; McDonald, 2000; Jarmasz, 2003). Yet other studies add extra constraints to the context word set, for example McDonald (2000) and Lowe and McDonald (2000) chose the most reliable context words – those that produced the most consistent cooccurrence patterns across a number of sub-corpora. However, Levy and Bullinaria (2001) found that adding the most frequent words in the corpus (mostly functors) to Lowe and McDonald's reliable context words significantly boosted the results in a semantic categorisation task. A word context set consisting mainly of function words also seems to help categorise words syntactically (Finch & Chater, 1992 and Redington, Chater & Finch, 1998).

To sum up, syntactic categorisation tends to be best achieved with very small windows and functors in the context word set, and semantic categorisation, with larger windows and content words in the context word set.

4.1.1.3 Metrics of similarity

Vector space models of the semantic lexicon assume that semantically similar words tend to occur in similar contexts. This section reviews the most commonly used methods to measure similarity between word context vectors.

Among the geometric similarity metrics (illustrated in Figure 4.2) are the Euclidean distance, which is the distance between the two points located by vectors in a space and the City Block (also called Manhattan and Levenshtein) distance, so called because of the way you have to go from A to B in a grid-like geometry such as the Manhattan streets and avenues, in straight perpendicular lines, and turning at the corners. The City Block and Euclidean distance metrics are sensitive to vector length, but this problem can be overcome by measuring similarity as the cosine of the angle between the two position-vectors. The cosine focuses on the difference between the directions of the vectors (see Figure 4.2), and is not sensitive to vector length, which makes it appropriate to compare words of similar frequency, but it is sensitive to vector sparseness, so it should be used to compare vectors of similar sparseness.

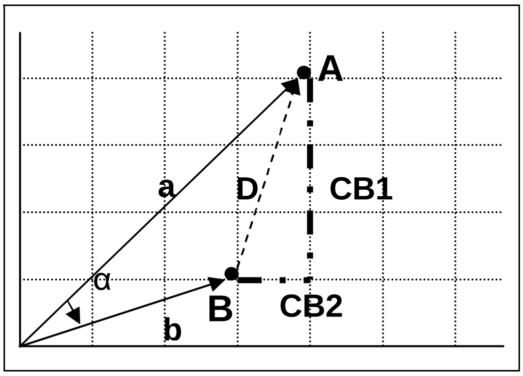


Figure 4.2. Three geometrical similarity measures between points A and B: the City Block distance is CB1 + CB2; the Euclidean distance is D; the cosine distance is the cosine of angle α.

Other metrics commonly used in information retrieval are the Dice metric (also used to measure phonological similarity, see § 3.2.1), which is twice the ratio between shared attributes and the total number of attributes for each target word, and the Jaccard metric, which compares the number of common attributes with the number of unique attributes for each pair of targets. Similarity coefficients have also been used in Internet search engines (e.g. Tudhope & Taylor, 1996). Information-theory metrics include the Kullback-Leibler divergence (or relative entropy) and Hellinger distance, both of which quantify the differences between two probability distributions.

Curran (2004) compares the behaviour of most of the metrics explained above, plus several variants including weight functions designed to assign a higher value to context words that are more indicative of the meaning. He found that Dice and Jaccard performed best in a semantic task. Levy, Bullinaria and Patel (1998) compare the Euclidean, City Block, Cosine, Hellinger and Kullback-Leibler metrics and found that the last two (the information theoretic metrics) perform best in semantic tasks.

In the studies presented in the rest of this thesis I use the cosine to measure the similarity between the cooccurrence vectors of two words (following McDonald, 2000) as the cosine of the angle they form. The cosine of the angle between the vectors locating words x and y is calculated as follows (for vectors defined by *n* components):

$$\cos(x,y)=\frac{\sum_{j=1}^{n} x_j y_j}{\sqrt{\sum_{j=1}^{n} x_j^2}\sqrt{\sum_{j=1}^{n} y_j^2}}$$

Following the same logic as the analysis of phonological similarity, the aspects of the lexicon where semantic similarity is more easily detected must correspond to the more salient structural parameters of the representational space of the semantic lexicon. In cooccurrence statistics methods, the parameters are distributional cooccurrence patterns of words. Different types of words play different parts in defining the semantic space. Section 4.2 explores a semantic hyperspace representation of the Spanish lexicon generated with cooccurrence statistics. In particular it examines the role of syntactic category (focusing on nouns and verbs), of semantics proper and of gender in the organization of the semantic hyperspace.

4.2 Exploring the Spanish semantic lexicon

Having adopted the convention to call the space generated by cooccurrence statistics "semantic", and accepted that this space not only contains properly semantic, but also syntactic and possibly other types of information, this section goes on to explore the structure of a semantic space calculated on a Spanish speech corpus. The focus of this section is to discover what word categorisations emerge from the distributional patterns in speech.

The calculation of the semantic vector space is analogous in some ways to the acquisition of the mental lexicon. Both the lexicon acquisition process and the vector space calculation count transform and categorise occurrences of items in speech and both end up with a structured collection of words. The main assumption behind semantic vector spaces is that the resulting hyperspace structure organisation is similar to the organisation of the mental lexicon.

This section will first describe the calculation of the semantic spaces on a Spanish speech corpus (with two variables: lemmatisation of the corpus and presence of functors in the context-word set); then it explores the role of the two variables in syntactic categorisation, in semantic tasks and in gender categorisation. The sections below compare directly the performance of vectors computed on a lemmatised corpus with vectors computed on a surface-form corpus in Spanish, testing the impact of inflections on syntactic categorisation and semantic tasks. The full-listing hypothesis proposed by Butterworth (1983), saying that all surface forms are individually listed in the mental lexicon, would be supported if surface forms are found to perform better than lemmas.

Christiansen and Monaghan (in press) observe that functors occur at phrase boundaries, which may reveal syntactic category, so the presence of functors in the context-word set should help syntactic categorisation, but not semantic tasks. Spanish inflected functors (determiners) in the surface-form corpus should help gender classification. Therefore I expect that gender classification should do better in the surface-form corpus, since gender morphemes are removed during lemmatisation.

4.2.1 Configuration of a Spanish semantic space

This section explains how I constructed the hyperspace representation of the Spanish semantic lexicon that is the basis of the analyses of syntactic category, gender and semantic clustering described in § 4.2.2 to § 4.2.4. In those analyses I manipulate two variables: the presence of morphology in the corpus and the presence of function words as context words in the calculation of the word vectors. Some of the other parameters are set in order to maximize vector quality given the limited size of the corpus available.

I created vectors for each of a number of target words – all the types above a certain frequency, which in practice coincides with the set of content and function context words (see Table 4.1 below). Here the corpus size constraints the number of frequent words able to generate dense vectors.

These vectors are created by counting the number of times that each context word appears within 5 words of the target word in the corpus. The frequency counts are then transformed into probability distributions to normalise for word frequency. I measured the similarity between two vectors as the cosine of the angle they form, because this metric is not sensitive to vector length, and it performs well in semantic tests (Lowe & McDonald, 2000; McDonald, 2000). The following sections describe the other elements involved in the configuration of the semantic hyperspace.

4.2.1.1 The corpus

The distributional statistics in this section are based on the same corpus used in chapter three, namely 'Corpus oral de referencia del español' an orthographical Spanish speech corpus (Marcos Marín, 1992). The words are transcribed phonetically using the same citation rules as in chapter two of this thesis. After removing all tags the corpus has 897,395 word tokens (38,847 types). This is much smaller corpus than those used in the studies mentioned in § 4.1.1.1 above. The spoken part of the BNC used in other studies mentioned above is about ten times larger[1]. Even with this important limitation, the distributional statistics provide information at the levels explored, namely syntactic category, gender and semantics. I assume that more refined vectors based on a larger corpus would provide even more detailed information including subtler nuances.

4.2.1.2 Lemmatisation

One of the variables in this study is whether the corpus contains surface forms (all word forms as found in speech, including gender, plural and verb inflections) or lemmas only (uninflected words). The corpus is not annotated, so instead of lemmatising the whole corpus by hand, I only lemmatised types of frequency greater of equal to 100, plus a few other types that added together would generate a lemma of frequency greater or equal to 100. The lemmatisation process comprised:

- Replacing feminine and plural inflections with the masculine singular form.
- Replacing all verb forms, including all persons and tenses, participles and infinitives, with the verb root: the infinitive without the final -r. Exceptions include forms of verb *ser* (be), which were replaced with the most common form, 3rd person singular of the present tense, ‘es’; forms of verb *ir* (go) were left as ‘ir’, because the forms resulting from the regular substitutions, ‘se’ and ‘i’ are homophonous with the very common impersonal pronoun ‘*se*’ and the conjunction ‘*y*’ (and), respectively.
- Removing the ending ‘-mente’ (equivalent to English ‘-ly’) from adverbs.

[1] I could not find a larger corpus of spoken European Spanish available for research, which limits the quality of the resulting vectors and therefore of the hyperspace. There are enough differences between the varieties of Spanish spoken across Latin America to make it desirable to use a single variety.

- Merging very frequent compound forms, e.g. ‘por favor’ (please) becomes ‘porfavor’ and ‘sin embargo’ (however) becomes ‘sinembargo’.

4.2.1.3 Context words and dimensionality

This is the second variable manipulated in this study. Although several studies assume that semantic information is best captured by contexts consisting of content words and syntactic information by function word contexts (Lowe & McDonald, 2000; McDonald, 2000; Jarmasz, 2003), Levy and Bullinaria (2001) found that adding functors to their context-word set significantly boosted the performance of their metric in a semantic test. This study examines the performance of two context word sets:

1) Content and function words: all word types above a certain frequency threshold.

2) Content words only: the words remaining after removing function words from set (1).

In the 'content word' condition I removed determiners, prepositions and conjunctions, plus the auxiliary verbs *ser, estar* (be) and *haber* (have) from the context-word list. Table 4.1 shows the dimensionality of the spaces generated by the different context word sets.

	surface	**lemma**
content+funct.	394 (≥200)	523 (≥100)
content only	320 (≥200)	481 (≥100)

Table 4.1. Number of context words (in brackets, threshold frequency) in the surface-form and the lemmatised corpus, when considering all words or content words only.

In a small corpus, a low number of dimensions will yield denser vectors. In order to obtain vectors of similar density with both versions of the corpus, the frequency threshold for the surface form version of the corpus is 200, and that for the lemmatised version is 100.

4.2.1.4 Window size

The cooccurrence vectors were calculated by transforming the raw cooccurrence counts within a window of five words to the left and five to the right, all conflated in a single value, into probability distributions. Window size is not a variable in this study – its effect has been extensively analyzed for English (§ 4.1.1.2). I chose a window size that generated reasonable results in most English tests, but that was not too small – again, to prevent sparse vectors given the small corpus available. Also, the eleven words contained in this window size take approximately 2.5 seconds to pronounce in a naturalistic Spanish spontaneous speech rate of 250 words per minute. This is close to the 2 seconds proposed by Baddeley, Thomson and Buchanan (1975) as the time-span of working memory. This 2.5 second window includes the five words that will be relevant for the processing of the target word, plus the five words in whose processing the target word is involved.

4.2.1.5 The vector spaces

I calculate four vector spaces using the methods and parameters above to be used in the studies presented in § 4.2.2 and § 4.2.3 below. I count the occurrences of one of two context word sets within a window of five words to the left and five to the right of the target words in two different versions of the corpus and two context-word sets. This results in four conditions:

1. Surface-form corpus, content and functors: the targets and the context words are the same: the 394 word types of frequency greater or equal to 200 in the surface-form corpus.

2. Surface-form corpus, content words only: the target words are the same as in condition 1; the context words are the 320 content words left after removing functors from the context-word set in condition 1.

3. Lemmatised corpus, content and functors: the targets and the context words are the same: the 523 word types of frequency greater or equal to 100 in the surface-form corpus.

4. Lemmatised corpus, content words only: the target words are the same as in condition 3; the context-words are the 481 content words left after removing functors from the context-word set in condition 3.

The rest of this chapter explores the performance of these four vector spaces in various syntactic and semantic categorisation tests.

4.2.2 Exploring syntactic category

In this section I use the four vector spaces calculated above and explore how different parameters contribute to syntactic word categorisation. I review approaches to syntactic categorisation using distributional cues and present an application to Spanish, focusing on the effect of corpus lemmatisation and of the presence of functors in the context word set on the categorization of all words, and then more specifically on verb-noun classification.

4.2.2.1 General syntactic categorisation

In this section I examine how distributional information can help categorize words syntactically. Frequency helps predict some parts of speech, notably function words. Figure 4.3 shows the frequency rank of syntactic categories.

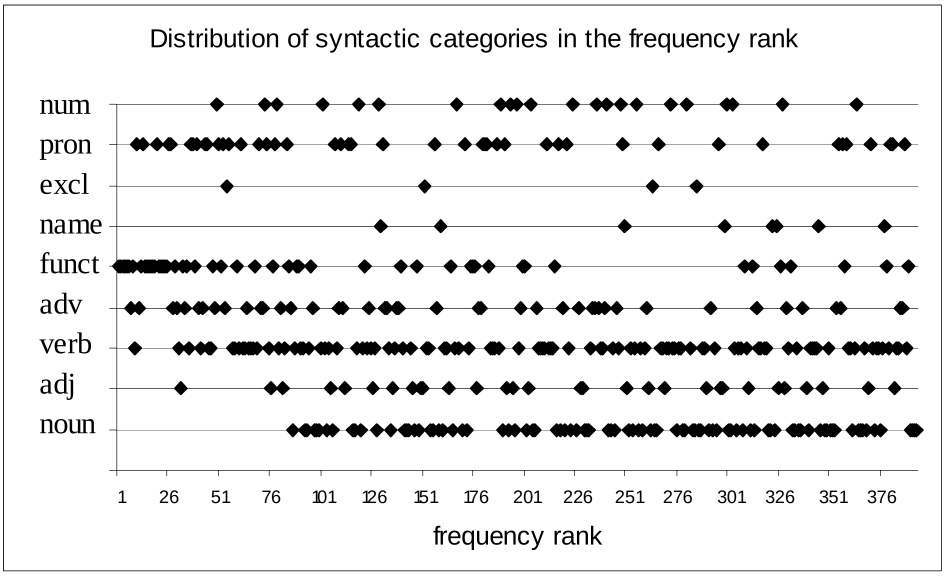


Figure 4.3. Syntactic category of the 394 surface form words of frequency greater or equal to 200 in the corpus, ranked by frequency. Each dot represents one word, and there is only one word per frequency rank position.

The most frequent words are to the left, in the higher rank positions, the more infrequent to the right of the graph, in the lower rank positions. The only obvious categorisation that could be derived from frequency information alone is that between functors and content words, since functors tend to be significantly more frequent than content words.

Simple cooccurrence statistics also reflect syntactic category. Figure 4.4 shows the distribution of words by part of speech ranked according to their average cooccurrence-based similarity with other words.

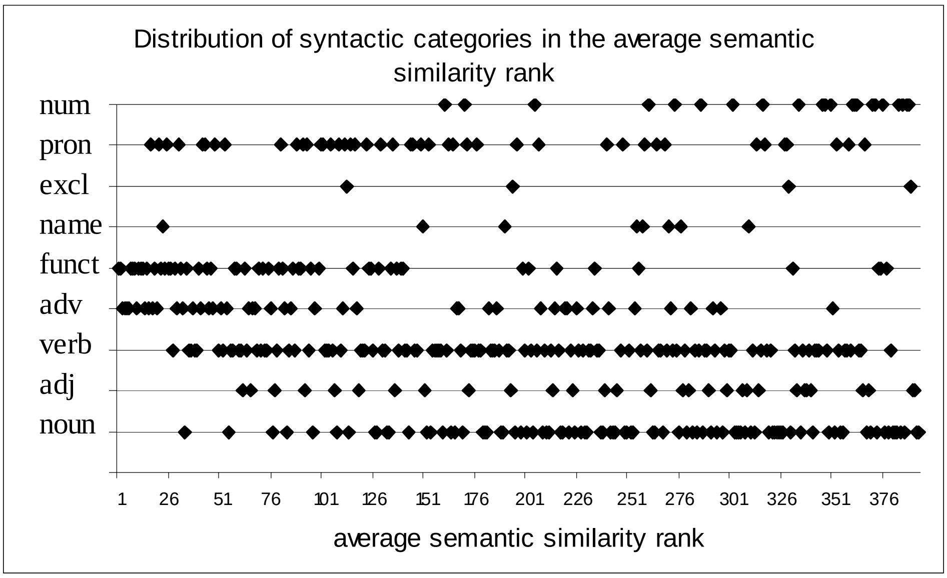


Figure 4.4. Syntactic category of the 394 surface form words of frequency greater or equal to 200 in the corpus (context including content and function words) ranked by average similarity value.

Similarity was calculated for all word pairs as the cosine of the angle formed by the two vectors representing the two words in the pair. The distances from each word to every other word were averaged, and then all words were ranked by average cooccurrence-similarity value. As in the frequency rank, function words, being so ubiquitous, cooccur with many words and cluster at the top of the similarity rank. But the cooccurrence-statistics based ranking offers more information: we also see that numerals are on average far from other words (a closer examination reveals that they are very close to each other, forming a cluster), and that verbs tend to be more similar on average to other words, while nouns tend to be less similar on average from the rest of the words.

More complex computations should achieve a more accurate syntactic categorisation of words. Section 4.2.1.5 outlined the characteristics of the four vector spaces that I will use for the tests in this section. I now investigate the effect of functors in the context-word set and of inflectional morphemes in the target-word sets on the ability of the vector space to predict the part of speech of words.

The ability to predict the part of speech or syntactic category has been tested in different ways: Levy, Bullinaria and Patel (1988) used the part-of-speech tags in the BNC to construct a syntactic categorization test. They calculated the centroid of a large number of vectors of words of each part of speech category, and then took the 100 most frequent words of each category and checked which centroid they were closest to. This method correctly categorised over 90% of the words using a window of one word to the left only or to the right and left. Redington, Chater and Finch (1998) calculated 600-dimension vectors for the 1,000 most frequent words in the corpus. They considered a window of two words to the left and right, and the information for positions two words to the left, one to the left, one to the right and two to the right were stored in separate vector components. The context words were the most common 150 words, which included a large proportion of functors. Redington, Chater and Finch's (1998) syntactic categorisation test involved hierarchical clustering of the vectors using Spearman's rank to measure vector similarity. Their method offered the possibility to introduce a cut-off point of similarity level, which they set at 0.8 to obtain the best categorisation. This unsupervised method (the syntactic category information was not provided prior to the cluster construction) correctly categorised 90% of nouns and 72% of verbs (chance baselines of 25% and 14%, respectively).

I present a supervised syntactic categorisation test also based on hierarchical clustering that categorized each word according to the category of the majority of its nearest neighbours in the space.

Method

The vectors in the four sets in § 4.2.1.5 above were manually tagged for syntactic category. Ten categories were used: noun, adjective, verb, adverb, functor, proper name, exclamation, personal pronoun, indefinite pronoun and numeral. Functors included determiners, prepositions and conjunctions; personal pronouns included possessives; indefinite pronouns included the Spanish equivalent of wh- pronouns such as *qué, quién, cómo* (what, who, how). I performed a hierarchical cluster analysis in SPSS (vector similarity metric: cosine) on each vector space and obtained a dendrogram with clusters of part-of-speech labels (See figure 4.5).

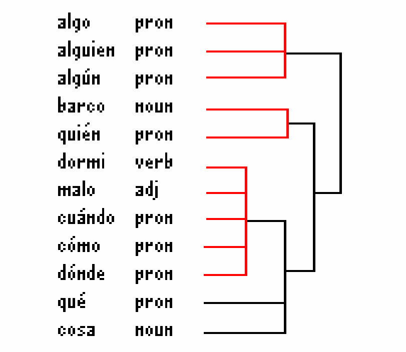
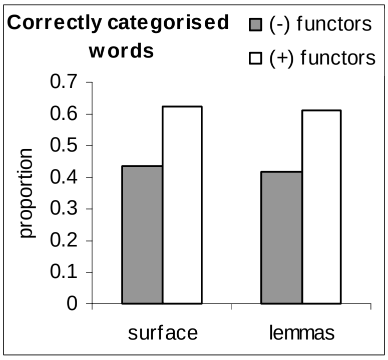


Figure 4.5 Part of a dendrogram showing hierarchical clustering (method: cosine) or words in a vector space (condition: lemmatised, functors and content. Words and part-of-speech labels are shown.

I performed a categorisation task on this dendrogram in the following way: given a new word whose position in the space (and therefore in the dendrogram) is known, it is categorised as belonging to the predominant category in its *local* cluster. I first consider each terminal-level cluster (marked in red in Figure 4.5); if there is one majority category[2] (as in the fist and third terminal-level clusters in Figure 4.5), then I count words of the majority category in that cluster as correctly categorised. If there is no majority in a cluster (as in the second terminal-level cluster in Figure 4.5), I consider that words in that cluster cannot be correctly categorised. Words clustered at the next level up (the two bottom words in Figure 4.5) count as correctly categorised if they belong to the majority category in the higher-level cluster. In the example in Figure 4.5, pronoun ‘que’ is correctly categorised because the majority of the words in the second-level cluster including the seven bottom words in the dendrogram are pronouns too. I did this only for the first two levels (considering more levels could only improve the results).

Results

This method categorised high proportions of words correctly. As seen in the summarised results in Figure 4.6, the presence of functors in the context-word set clearly improved the performance both in the surface-form (two-tailed paired t-test, t=2.23; df=9, p=0.05) and the lemmatised (two-tailed paired t-test, t=2.21, df=9, p=0.05) versions of the corpus. Surface-forms were marginally better categorised than lemmas (t-tests not significant).



[2] For 2-element clusters, there is only a majority if both items are the same category. Then, the classification algorithm will classify each of them correctly by assigning it the same category as the other item in the cluster. For larger clusters, I consider a majority of at least two items more than the next most frequent category in the cluster.

Figure 4.6. Results of syntactic categorisation task using the four vector spaces.

Figure 4.7 shows the proportion of correctly categorised words in each syntactic category, compared with chance levels. Baseline chance levels are proportional to the number of nouns, proper names, numerals etc in the target-word set. Some syntactic categories were categorised better than others, but all were categorised correctly well above chance levels, as seen in Figure 4.7.

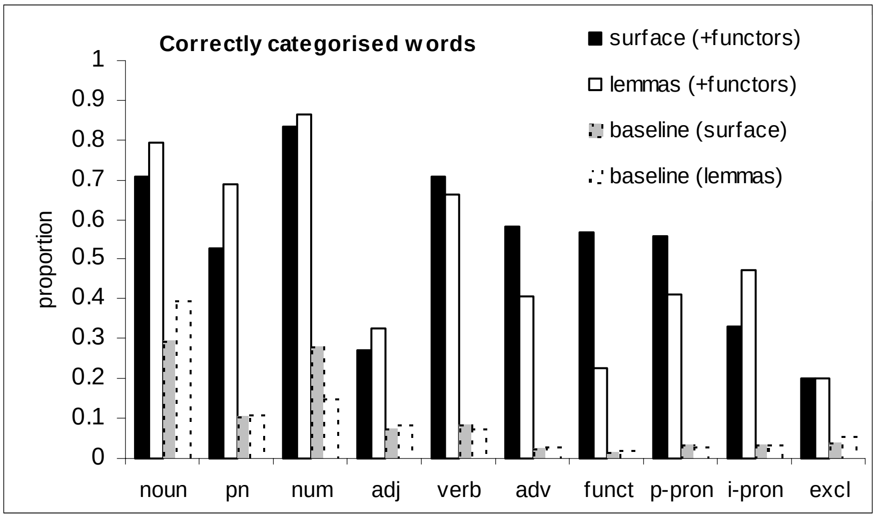


Figure 4.7. Proportion of words of the ten different syntactic categories that were correctly categorised in the two vector spaces that included functors in their context-word set. Chance baseline levels also shown. (All result-baseline two-tailed paired t-tests yield significances p<0.01).

The graph shows the proportions of words correctly categorised in the two best-performing vector spaces (those including functors in their context-word sets). This comparison shows the effect of corpus lemmatisation on a syntactic categorisation task.

Discussion

As we see in Figure 4.7, nouns, numerals, proper names, adjectives and indefinite pronouns are better categorised in the lemmatised corpus. Verbs, adverbs, personal pronouns and, particularly, functors, however, are better categorised in the surface-form corpus. This suggests that conflating all noun, adjective and indefinite-pronoun surface forms into their lemmas helps categorise them syntactically. On the other hand, conflating all surface forms of a verb into a single lemma hinders verb categorisation. Nouns, adjectives and indefinite pronouns can take gender and plural inflections. In the second group, only verbs change between versions of the corpus, with inflections removed from the lemmatised version.

Of the word categories which do not change between the surface and the lemmatised corpus versions, the largest difference is found in functors, which are better categorised in the surface-form corpus. As indicated by the results in Figure 4.6, the role of functors is to relate words to one another in the sentence, so it could be said that they *categorise*, but do not need to *be categorised*. Since relationships between words in Spanish are also signalled by agreement (in number and gender between nouns and adjectives, in number between subjects and verbs) inflected words provide a more fine-grained, and therefore more accurate, categorisation of functors than lemmas do.

These results support the idea that gender and number inflections on one hand and verb inflections on the other have different roles in syntactic categorisation. The difference between English noun (number) and verb (person and tense) morphology was pointed out by Tyler, Bright, Fletcher, and Stamatakis (2004), whose fMRI studies of noun and verb processing suggest that while noun and verb stems representations do not differ, verb and noun morpho-phonology engage different neural systems. The present results suggest that while nouns and adjectives are better categorised in a vector space based on the word root (lemma), verb categorisation is helped by the variety introduced by verb inflections.

The next section looks more closely at the classification of nouns and verbs in a vector space.

4.2.2.2. Nouns and verbs

Chiarello, Shears and Lund (2000) proposed a measure of noun-verb distributional typicality: the degree to which a word appears in similar contexts as other words of the same grammatical category. They used cooccurrence vectors for the target words (calculated using Lund & Burgess' 1996 method) and calculated the distance between all word pairs. For each target word, they subtracted the average distance between the target and all the nouns from the average distance between the target and all verbs. The resulting score, which they called noun-verb distance difference (NVDD), was high for nouns occurring in contexts similar to other nouns, and low for nouns occurring in atypical noun contexts; this score was low for verbs occurring in contexts similar to nouns and high for verbs occurring in contexts similar to other verbs. Monaghan, Chater and Christiansen (2003) found that a similar calculation of distributional typicality predicts response time in a verb/noun decision task. Christiansen and Monaghan (in press) argue that phonological and distributional information together can accurately discriminate syntactic category. They point out that where distributional cues are not reliable, for instance in function words, phonological cues are very informative. Their two experiments with a 2-word frame and a preceding-word window indicate that distributional cues classify nouns better than verbs. In the experiment in the last section I found the same with a window of five words to the left and five to the right (see Figure 4.7). Christiansen and Monaghan suggest that verb classification relies more on word-internal cues. In this section I test how good a vector space generated with a five-word window to the left and right is for noun-verb classification. By using the four vector spaces described in § 4.2.1.5, I explore the effect of inflection and of function words in the context in that categorization. I have adapted Chiarello, Shears and Lund's (2000) method in order to explore the distributional typicality of Spanish nouns and verbs. The main difference between the present study and those reviewed above is the size of the window where context-words are counted. As explained above, given the reduced size of the Spanish speech corpus available, I need a larger window in order to obtain vectors dense enough for the similarity calculations. I explain the major result divergences in terms of this difference.

Method

Using only the nouns and verbs (see Table 4.2) from the vector spaces described in § 4.2.1.5, the average similarity was calculated between each word and every other noun and each word and other every verb. The method for calculating similarity was, again, the cosine of the angle between the two vectors. Then I calculated the average similarity to verbs minus the average similarity to nouns to obtain that word's distributional typicality.

	Nouns	**Verbs**
Surface-forms	*101*	*86*
Lemmas	*207*	*80*

Table 4.2. Number of nouns and verbs in the surface-form and the lemmatised corpus versions.

Results

Figure 4.8 shows the general results for the classification of nouns and verbs. While almost 100% of verbs are classified correctly in all conditions (white bars in Figure 4.8), classification of nouns relies heavily on the presence of functors in the context-word set.

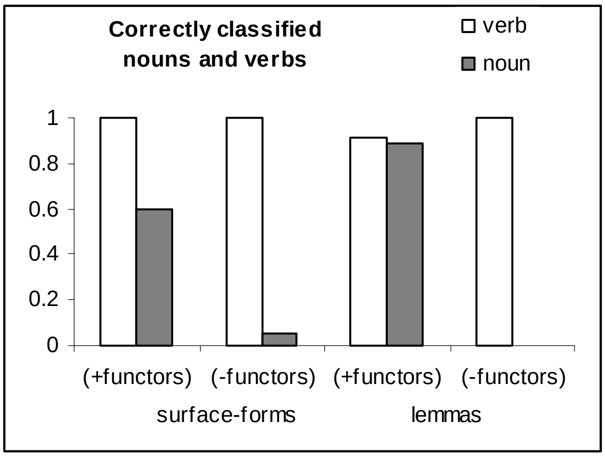


Figure 4.8. Proportions of correctly classified nouns and verbs in the four vector spaces.

Figures 4.9, 4.10, 4.11 and 4.12 show distributions of nouns and verbs in the polarized noun-verb distributional typicality space. Words with negative values are more similar on average to nouns, and words with positive values are more similar to verbs. I present the distributions of nouns and verbs in separate graphs for each of the four spaces described in § 4.2.1.5 (combinations of surface-form and lemmatised corpus and content and function word or content word only context-word sets). White bars represent correctly classified words; black bars, incorrectly classified words. (The green bars represent 'person nouns' such as '*man*', '*woman*', '*child*', '*mother*', '*father*').

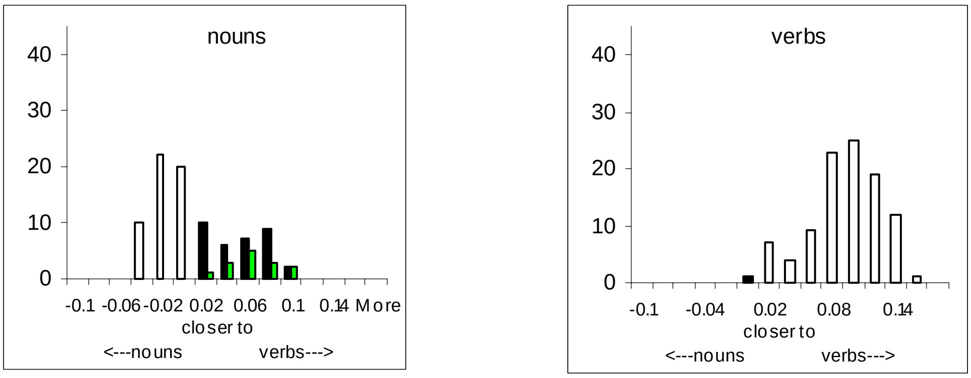


Figure 4.9. Corpus: surface-form. Context: content and function words.

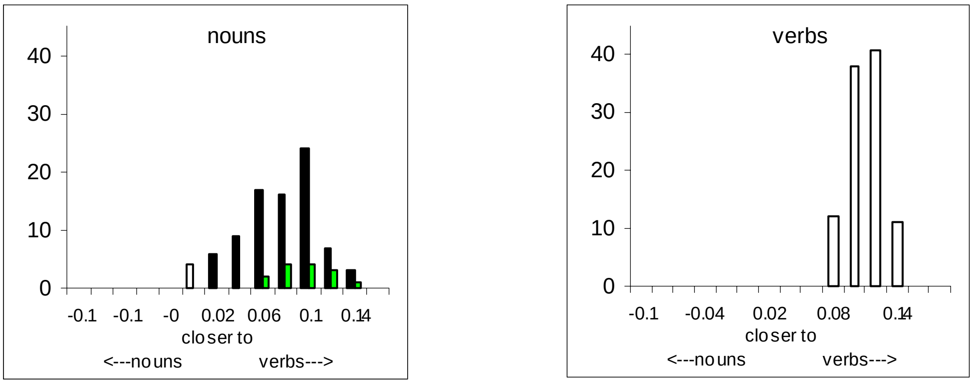


Figure 4.10. Corpus: surface-form. Context: content words only.

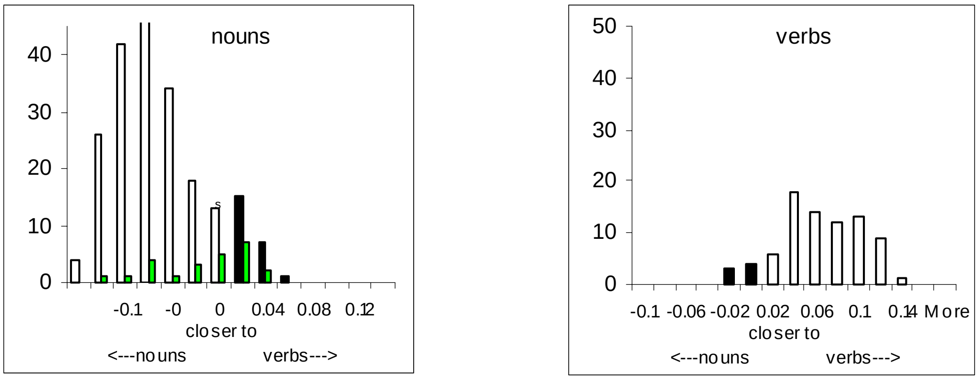


Figure 4.11. Corpus: lemmatised. Context: content and function words.

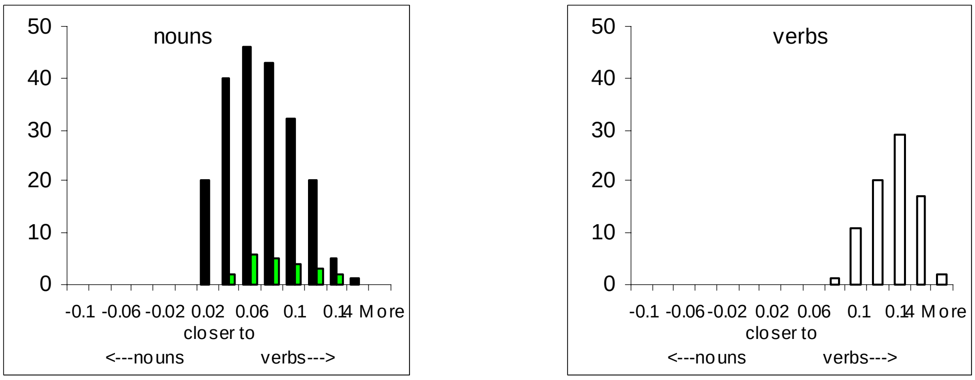


Figure 4.12. Corpus: lemmatised. Context: content words only.

In Figure 4.10, the four correctly classified nouns are part of formulaic greetings (*tardes, noches, gracias* – afternoon, night, thanks) or go with numbers (*minutes* - minutes).

The green bars in the graphs represent the proportion of nouns in the adjacent black bars that are 'person nouns'. These nouns form a clear sub-population within the nouns showing a distinct behaviour at this level of analysis. Person nouns are markedly closer to verbs than the rest of nouns in all but the last condition (lemmas, no functors). Person nouns are discussed in § 4.2.3.2 below.

Discussion

The majority of nouns and verbs are correctly classified when nearby functors are taken into account. Interestingly, the modes and shapes of the noun and verb distributions are very distinct in all cases. This indicates that, even though nouns are closer to the verb side of the distribution typicality spectrum in the spaces calculated without functors, they can still be separated from verbs. As we saw above, Christiansen and Monaghan (in press) found that a 1 or 2-word window classified nouns better than verbs. Their method is closest to our (+functor) conditions, where nouns were most accurately classified. These two results put together suggest that noun classification relies on the preceding word. In a lexicon representation based on cooccurrence, nouns may be very accurately classified by being consistently preceded by one of a reduced set of words, namely determiners. Christiansen and Monaghan's results suggest that verbs are the marked category. Our results seem to indicate the opposite, with nouns being closer to verbs in the absence of cues. This may be due to the small number of vector components that reflect the cues for nouns (determiners). The small number of components cannot reflect the very high frequency of determiners in speech which determines noun acquisition. Christiansen and Monaghan show that while phonological cues are more useful for the acquisition of the verb category, distributional cues, especially determiners (and language external cues, such as cooccurrence with objects in the environment) are more useful in the case of nouns. This means that the larger window in the present results may have introduced extra information that obscures the markedness of the noun, which was evidenced in smaller window studies.

The fact that nouns are better classified in the lemmatised corpus can be explained in terms of token frequencies: the lemmatised corpus contains only one lemma for each four surface-form determiners (combinations of masculine and feminine, singular and plural), and nouns are classified by the fewer but denser vector components corresponding to the conflated forms of the determiners.

The present results, in the context of the literature reviewed above, suggest that different parameters of the vector space reveal different speech patterns.

It is difficult to characterize them as syntactic or semantic, but the next section reviews the literature on distributional cues in semantic tasks.

4.2.3 Exploring semantics

We have seen that different levels of cooccurrence analysis, particularly different sizes of the context window reveal different aspects of the structure of the lexicon. Very small windows accurately categorise words syntactically. Very large windows, such as those used in LSA do very well in semantic tasks (Landauer and Dumais, 1997): this approach scored 64% correct in the synonym part of the TOEFL English test, where the task was to choose the closest meaning words to a target words out of four options. Smaller windows also capture semantics, as shown by Levy, Bullinaria and Patel's (1998) results for the same English test: in a larger corpus (90 million words against 4.6 million used by Landauer and Dumais) with information theoretic similarity metrics their method obtained up to 76% correct with windows of two or three words to left and right.

In 1995, Stubbs examined the semantic content of cooccurrence-based word representations. Corrigan (2004) also used cooccurrence to examine the semantic connotations of words, revealed by their statistical usage patterns. Corrigan’s case study shows that cooccurernce with reported negative events give the verb ‘happen’ negative connotations.

Other explorations of the semantic structure information contained in intermediate-size window cooccurrence vectors include the Hyperspace Approach to Language (HAL) (Lund, Burgess & Atchley, 1995; Lund & Burgess, 1996; Lund, Burgess & Audet, 1996). HAL predicted that the more similar two words were in the space - similarity measured as the Euclidean distance between vectors - the more they would facilitate each other in a lexical decision task. In a categorisation task of words belonging to semantic groups such as body parts, animals, countries etc, they distinguished words that cooccurred with each other from words that appeared in similar contexts, and claimed that the former were associated by temporal contiguity while the latter were semantically associated. Associative priming would be due to semantic association. This prediction was confirmed by Bullinaria and Huckle (1997), who found that lexical decision priming correlated with distances in a semantic vector space.

Finch and Chater (1992) found that in the cluster analysis dendrograms that represented syntactic categories, some clusters also represented semantic groupings. The limitations of dendrograms resulting from cluster analysis as a tool for rigorous comparison of semantic content of different cooccurrence spaces was pointed out by Huckle (1995).

Levy, Bullinaria and Patel (1998) used a semantic test based on Battig and Montague (1969) semantic category norms, which were collected by asking people to name, for example, 'units of time' of 'four-footed animals'. Levy, Bullinaria and Patel calculated the centroid of ten members from each category. Their classifier computed the distance between the target word and each of the centroids, choosing the closest category. They obtained the best scores (around 65% of words correctly categorised) with windows of around 10 words and information theoretic similarity measures.

McDonald (2000) used a psychologically-grounded criterion - Miller and Charles (1991) work on semantic similarity judgements - to assess the validity of cooccurrence-based semantic similarity measures. The similarity measures obtained in the vector space (with a window size of three words to each side and similarity measured as the cosine of the angle formed by two vectors) were strongly correlated with the psychologically-based ones.

4.2.3.1 Cluster analysis

These studies show that cooccurrence statistics do capture lexical semantic structure. There are no standard tests in Spanish such as the English test used by Landauer and Dumais (1997) or of semantic similarity norms for Spanish words like those of Miller and Charles (1991) that I can use to compare the effect of lemmatisation and of functors in the context word set on semantic categorisation in a cooccurrence vector space. Below are a few examples of dendrogram clusters that clearly reflect semantics (like those found by Finch and Chater, 1992). Appendix D contains some more.

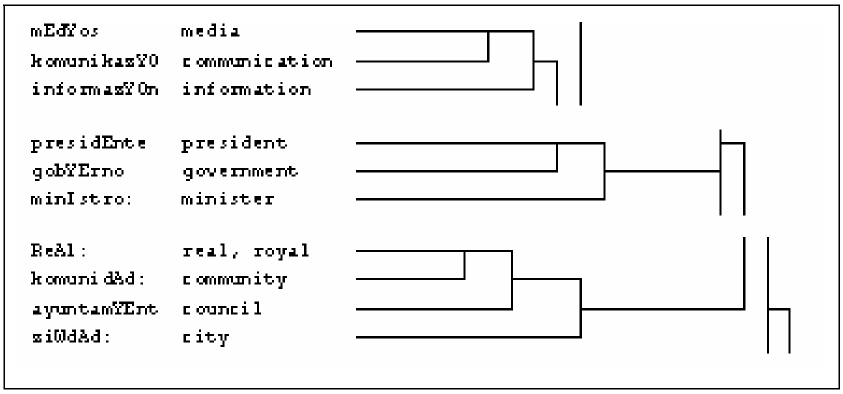


Figure 4.13. Some example semantic clusters obtained in the surface-form corpus using content words only as context words.

4.2.3.2 Person nouns

Figures 4.9 to 4.12 above show the distinct distributions of person nouns (see Appendix E for full lists of person nouns). The distributions of person nouns are in most cases significantly different from those both of verbs and of the rest of nouns. The average distribution typicalities of the person nouns are, in all cases, between those of nouns and verbs. Figure 4.14 shows the mean of the noun, verb and person noun distributions in Figures 4.9 to 4.12 above.

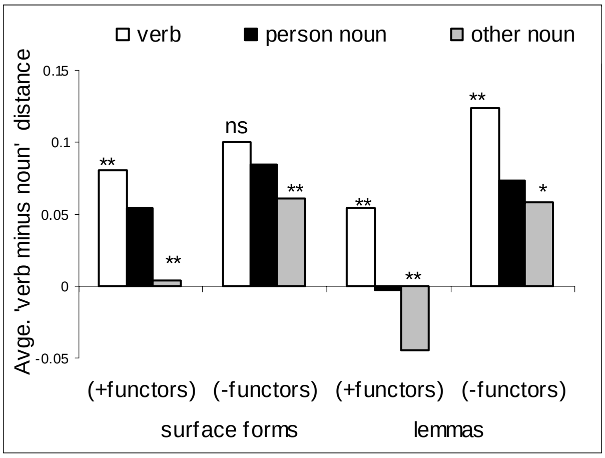


Figure 4.14. Means of the distributions of verbs, nouns and person nouns in the noun-verb polarised spaces, in the four conditions. Asterisks indicate level of significance of difference with person nouns distribution (two-tailed t-tests) (* p<0.05; ** p<0.001).

This level of cooccurrence analysis reveals the existence of a sub-group, person nouns, which behaves in a consistent way, significantly distinct from nouns and verbs. They are nouns, but they show an atypical behaviour. Person nouns can be considered a syntactic sub-class or a semantic cluster. They do behave as a separate syntactic category and additionally they are linked to a semantic class of referents in the world (people).

4.2.4 Exploring gender

Most of the published work on lexical hyperspaces generated by cooccurrence distributions is based on English, which has no grammatical gender. This study of a Spanish semantic space provides an opportunity to explore whether distributional statistics capture gender, and what they can tell us about it. This section first reviews the function of grammatical gender and then examines the effect of corpus lemmatisation and of the context word set on the categorization of masculine and feminine nouns.

4.2.4.1 The function of gender

In Spanish, all nouns are either masculine or feminine, and their determiners (with a few phonology-driven exceptions) and adjectives agree with them in gender. The function of grammatical gender is not clear. Apart from its role in designating male and female for some nouns referring to people and to animals such as *niño/niña* (boy / girl), *gato/gata* (cat masc. / fem.), gender is not related to sex. The Real Academia Española (1973) states that the gender assigned to Spanish nouns is influenced by formal, semantic, etymological and analogical factors. It seems, though, that linguistic information such as syntactic and morpho-phonological factors is more important than semantic information in the recognition of the gender of nouns (Perez Pereira, 1991). The masculine is the unmarked, generic form - the masculine form *niño* (boy) may denote either a boy or a girl; the masculine plural *hijos* (children) may include sons and daughters. The masculine form has more roles, and wider semantics than the feminine, and hence is more indeterminate than the feminine form (Real Academia Española, 1973). Alarcos (1994) states that the variety of masculine and feminine word forms and the arbitrariness in the assignation of gender to word meanings make it difficult to define the meaning of gender. He considers gender as a grammatical trait or morpheme that classifies nouns into two different combinatorial categories (masculine and feminine) not ascribed to semantics. Gender may sometimes indicate sex, size (feminine nouns usually indicate larger size) and other semantic relationships, such as general concept (feminine) vs. particular instance (masculine), but for Alarcos, the main function of gender is to signal relationships between words, and thus to make possible a flexible word ordering. In example (1) below, the gender of the adjective *viejo* (old) disambiguates whether it refers to *candelabro* (masculine), as in 1a, or to *plata* (feminine), as in 1b. In example (2), gender agreement allows contorted word orderings.

(1a) *el candelabro (m.) de plata (f.) viejo (m.)* (the old [silver candelabra])

(1b) *el candelabro (m.) de plata (f.) vieja (f.)* (the [old silver] candelabra)

(2) *del monte en la ladera por mi mano plantado tengo un huerto*
(‘of the mountain on the side by my hand planted I have an orchard’)
(I have an orchard planted by my own hand on the side of the mountain)

To sum up, the main function of grammatical gender, and of singular/plural, is a syntactic one: to classify and clarify the functions and relationships of words within a sentence. This function could have been taken by other classifications such as animate/inanimate, as is the case in other languages (Hernandez, 2001).

The next section presents a gender categorisation task based on distributional cues, and the effect on it of lemmatisation and of functors in the context-word set.

4.2.4.2 Categorisation of nouns by gender

This study uses the same categorisation method employed in § 4.2.2.2 above, but now to categorise masculine and feminine nouns in the same four vector spaces.

I consider the same two versions of the corpus, the surface-form version including masculine and feminine inflections, and the lemmatised version, where gender and number inflected forms such as *niño, niña, niños* and *niñas* are conflated into the unmarked masculine form, *niño*. The second variable is the presence of functors in the context-word set.

Results

Figure 4.15 shows the proportions of masculine and feminine nouns that were correctly classified in the four conditions. While almost 100% of masculine nouns are classified correctly in all conditions (white bars in Figure 4.15), classification of feminine nouns is greater in the surface-form corpus, and is also helped by the presence of functors in the context-word set. Gender classification without functors in the lemmatised corpus was not expected to be good, but was tested nevertheless to see to what extent cooccurring content words were able to provide clues to the gender of the word, as suggested by Boroditsky's (2001) claim that gender influences the way people think of objects, and hence the semantics of gendered nouns. No evidence of this effect is apparent in the present results.

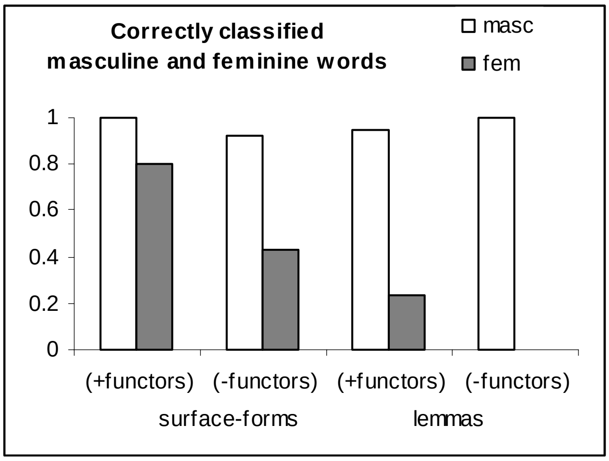


Figure 4.15. Proportions of correctly classified masculine and feminine nouns in the four vector spaces.

Figures 4.16 to 4.19 show the distribution of masculine and feminine nouns by difference between similarity to masculine minus similarity to feminine.

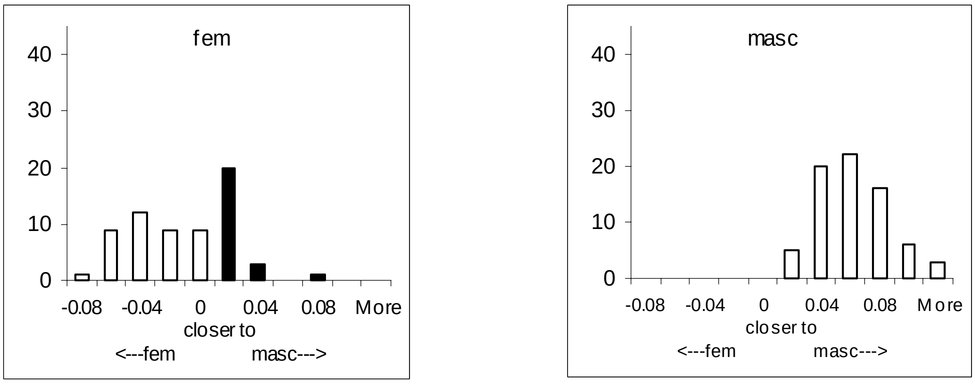


Figure 4.16. Surface-form corpus; context: content and function words.

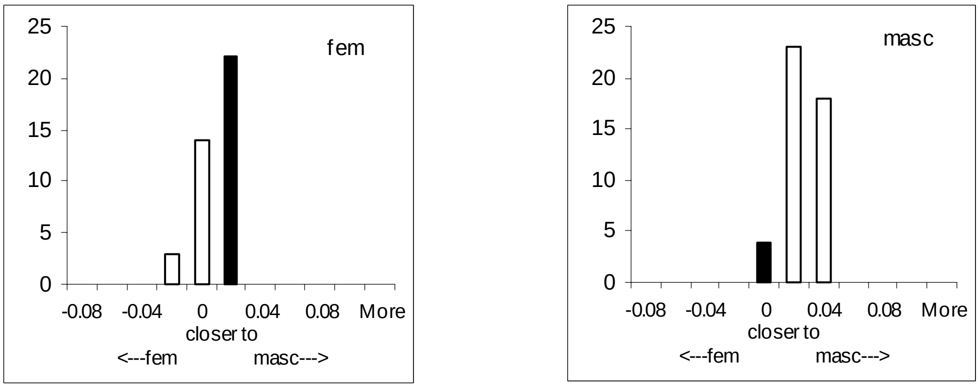


Figure 4.17. Surface-form corpus; context: content words only.

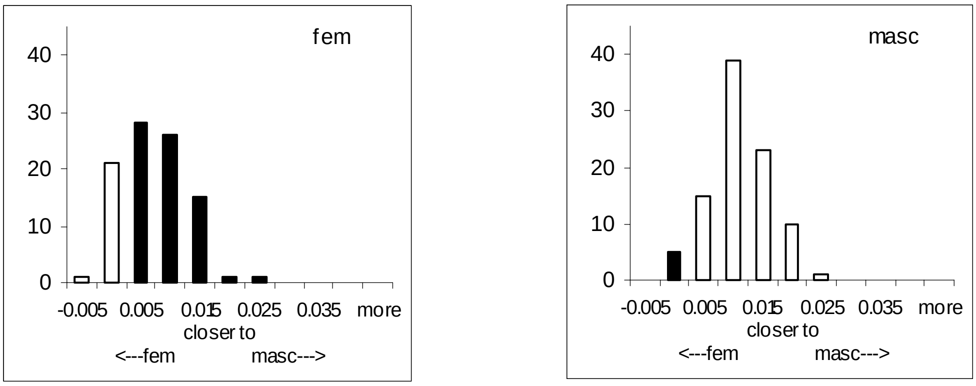


Figure 4.18. Lemmatised corpus; context: content and function words.

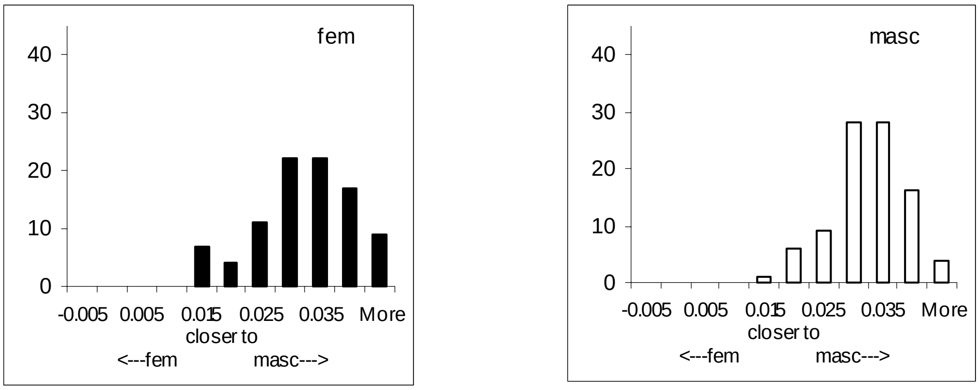


Figure 4.19. Lemmatised corpus; context: content words only.

These graphs are similar to those in § 4.2.2.2 with white bars representing correctly classified words and black bars representing words that are closer to the opposite gender.

Gender classification is better than chance in the surface-form version of the corpus, particularly when functors are included in the context-word set. There are eight wrongly classified feminine nouns in Figure 4.16: the feminine word *agua* (water), which takes the masculine article, plus seven plural nouns (out of a total of nine plural nouns in the target word set). The wrongly classified plurals are: *gracias, mujeres, veces, cosas, horas, personas* and *pesetas* (thanks, women, times, things, hours, persons and pesetas); of these, only 'women', 'things' and 'persons' are preceded in the majority of cases by determiner 'las'. The other words are mainly preceded by numerals or other words.

Discussion

The results above suggest, as predicted, that noun inflections and cooccurrence with functors provide the best cues for gender categorisation. The fact that the feminine word *agua*, which takes masculine determiners, is such an outlier indicates that the main cues for gender are determiners, agreeing in gender (and number) with the noun they precede in Spanish.

Figures 4.16 to 4.19 show that in the absence of the appropriate cues, feminine words are more similar to masculine words than to other feminine words, reflecting the fact that feminine is the marked gender.

4.3 Conclusions and future work

This section has explored the information contained in cooccurrence-based vector spaces, and has explained why these seem to be psychologically plausible mental representations of speech. I have shown that even a space generated with a fixed window using a simple similarity metric contains information leading to syntactic and semantic categorisations.

Using a corpus of Spanish transcribed speech I have tested the effects of including functors as dimensions in the vector space and of removing the inflections from the corpus. The most reliable cue for syntactic categorisation and for the binary classification of nouns and verbs and of gendered words appears to be cooccurrence with functors *and* content words. The effect of lemmatisation is mixed: gendered nouns are better categorized in a fully inflected corpus; verbs and nouns are better classified in a lemmatised corpus; in the task of categorising all words by syntactic category, the results for nouns, proper names, numerals and adjectives were better in the lemmatised corpus, while the results for verbs, adverbs and functors were better in the surface form corpus.

All these results together support the view that functors have a crucial role in the scaffolding of syntactic categorisation, and that, while nouns and adjectives are better characterized when cooccurrence with functors is taken into account, verbs and adverbs are better characterized by the distributions of verb inflections. A possible extension to the tests presented in this chapter would be to include a condition where the context-word set is composed of function words only. In that condition I would expect to see similar or improved results in syntactic tests, but worse results for semantic tests.

Designing a quantitative semantic task for Spanish such as synonym-choice would allow the comparison of cooccurrence spaces calculated with different parameter values. With such a test in place, and ideally a larger corpus, it would be possible to explore systematically the parameter space - window size, context-word set and similarity metric – in the steps of Levy, Bullinaria and Patel (1998). Knowledge about the type of information captured by the different parameter configurations could help design tools for the automatic extraction of syntactic or semantic information from speech cooccurrence patterns. This exploration would also provide theoretical insights into the way syntactic and semantic information are encoded in speech, and their interactions.

The previous chapter presented an empirical exploration of Spanish phonological similarity parameters that can be used to build a phonological similarity space. In this chapter I have studied the information captured in a cooccurrence-based syntactic-semantic similarity space. The following chapter brings these two similarity spaces together and tests the existence of systematic relationships between them.

Chapter 5. Cross-level systematicity in the lexicon

This chapter deals with the hypothesis that there is systematicity in the lexicon. Introductory chapter one proposed that the lexicon is an adaptive system where each word's phonology, semantics, and syntax is defined in terms of its relationships with those of the rest of the words. Chapter three examined parameters of phonological similarity; chapter four, a metric of semantic similarity. This chapter tests the hypothesis that the two spaces configured using phonology and semantics, two independent measures of word similarity, are systematically related to each other. Section 5.1 introduces and motivates the hypothesis and reviews the literature of the issues that it touches on. Section 5.2 presents and discusses two experiments that test form-meaning systematicity. Section 5.3 examines which types of words drive the systematicity, testing Shillcock, Kirby, McDonald and Brew's (2001, *submitted*) claim that systematicity is driven by certain 'communicatively salient words'. Finally, section 5.4 discusses the results of the chapter in the light of the literature.

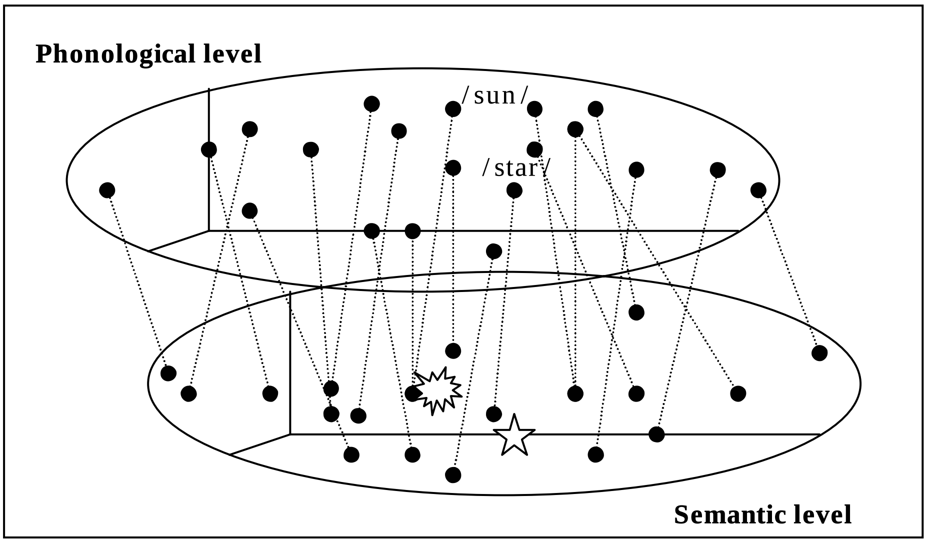


Figure. 5.1. Systematicity between the phonological and the semantic levels of the lexicon: words that are close together in the phonological level tend to be close together in the semantic space, and vice versa.

5.1 Systematicity in the lexicon

The previous two chapters looked at methods to measure phonological and semantic relationships between words. The resulting lexicon levels, defined in terms of similarity between all word pairs, follows structuralism in that words are defined not by their inherent qualities, but as elements in a system. For Saussure ([1916] 1983), language is organized as “an internal system of signs which exist in a system of relationships and differences”. Throughout this thesis I have emphasized the lexicon’s different levels of representation. Chapter one defined the lexicon as an adaptive system finding the optimal solution to the several pressures that act on it. The present chapter deals with the assumption that one of those pressures is a general bias in the brain for structure-preserving mapping between related representations, and proposes that there exists systematicity between the levels of the lexicon measured in previous chapters (phonological and cooccurrence-based).

Systematicity is a basic, pervasive property of language. The relationship between language and meanings is fundamentally systematic. Anderson’s (1991) study of categorisation concludes that the structure of the environment determines the structure of concepts. This is also evidenced in the relationships between syntactic compositionality and grammatical meaning: similar syntactic structures express similar relationships between concepts. Systematicity between morpho-phonology and meaning is less obvious, but nonetheless present, with morphemes with similar phonology denoting similar word syntactic properties – for example, as we saw in chapter three, many Spanish tenses and person morphemes are encoded in final stressed vowels. It should not come as a surprise, then, that the relationships between word forms and word meanings are also systematic. In particular, in this chapter I focus on the systematicity between the two levels examined in past chapters: phonology and semantics. Among the implications of such systematicity is the hypothesis that words with similar phonological representations tend to have similar semantic representations, and conversely, words with different phonological representations tend to have different semantic representations, already tested for English by Shillcock, Kirby, McDonald and Brew (2001, *submitted*).

Naturally, this effect is expected to be extremely small, as a multitude of other conflicting constraints act on words' phonology and semantics. Word form-meaning systematicity is a logical extension of the pervasive trend for language-referent systematicity, and only seems surprising because it is masked by the effects of other demands on the structure of the lexicon, not least the need to make words within the same semantic group sound different from each other so that they can be easily distinguished. I propose that a degree of systematicity is useful in language acquisition and comprehension, and that though not readily apparent, it is there and its effects are measurable if we use the appropriate methods.

This section first examines the background research on phonology-semantics systematicity and then addresses some issues such as what could be the function of the systematicity, why it might exist and how it relates to Saussure's arbitrariness of the sign principle.

5.1.1 Background

The work presented in this chapter and the following one is based on a study by Shillcock, Kirby, McDonald and Brew (2001), further developed in a submitted manuscript (Shillcock, Kirby, McDonald & Brew, *submitted*). Shillcock et al. looked at the structure of the English lexicon and found a small but significant correlation between the phonological and semantic distances[1] between words; specifically, they propose that certain

[1] Note that Shillcock et al. (2001) use distances where I use similarities. They are two ways of measuring the same phenomenon. A high similarity is a small distance, and vice versa. Their metric of phonological distance increases with the number of mismatches; my measure of phonological similarity increases with the number of matches. They measure semantic distance as (1 – cosine), while I measure semantic similarity as the cosine.

‘communicatively important words’ show a high correlation between their phonological and semantic distances to the rest of the words. In this chapter I apply a methodology similar to Shillcock et al.’s to two subsets of the Spanish lexicon.

Shillcock et al. considered the 1733 most frequent monosyllabic, monomorphemic English words in the British National Corpus and calculated the distance 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 the phonological distance between each word-pair, 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 semantic distance they constructed a cooccurrence-based 500-dimension vector space based on the 100 million-word British National Corpus. (The cooccurrence-based vector space method is explained and reviewed at length in § 4.1 in chapter four.) They lemmatised the corpus to reduce vector sparseness and measured the semantic distance as 1- cosine of the angle between the two word cooccurrence vectors. Finally, they obtained a correlation between the phonological and the semantic distances of Pearson’s r = 0.061, which a Monte-Carlo analysis showed to be significant (p<0.001, one-tailed).

Having shown a significant systematicity between phonology and semantics in the English lexicon, Shillcock et al. (2001) ranked the individual words according to their correlation value – they calculated, for each word, the correlation between its phonological and semantic distances to every other word. They found that 'filler' words such as *oh*, *er* and *ah* were positioned at the top of the rank. Shillcock, Kirby, McDonald and Brew (*submitted*) extend that study and find that swear-words, personal pronouns and proper names all are high in the correlation rank. They propose that these are communicatively important words that tend to be preserved in individuals with a range of language impairments. Shillcock et al. (2001, *submitted*) point out that their phonological and semantic distance metrics were separately developed for different projects, and that they are theory-independent, and therefore objective methods.

5.1.2 Preliminary issues

The notion of a systematic lexicon with high-order relationships across levels of representation raises some questions, such as: What benefit could a systematic lexicon bring to language processing? Why would it occur in the brain? Does it not contradict the Saussurean arbitrariness of the sign principle? This section addresses these three fundamental issues.

5.1.2.1 The function of the phonology-semantics systematicity

As we saw in chapter one, the structure of the lexicon is subjected to many pressures, which are sometimes opposed to each other. Here I concentrate on one of those pressures, namely the bias towards structure-preserving representations of words over the phonological and the semantic levels. This pressure may be an inevitable consequence of the nervous system representational principles. I assume that if the systematicity is there and it is measurable, it is so for a reason; if it had neutral or adverse effects, other pressures on the lexicon would have swamped its effects effectively removing it. This section presents some functions that the phonology-semantics systematicity could be serving.

In their paper on Latent Semantic Analysis, Landauer and Dumais (1997) say that young teenagers learn on average 10-15 new words a day; the authors claim that exploiting the weak distributional interrelations between words at the right level ‘can greatly amplify learning by a process of inference’. Systematicity could further help not only young learners, but also adults confronted with a novel word: the form of a word provides additional clues to its possible meaning.

The Iterated Language Model (Kirby & Hurford, 2002), based on computational simulations of language evolution, proposes that the cultural transmission of language leads to the evolution of languages that exploit structure in both the meaning and the signal spaces.

Additionally, Shillcock et al. (*submitted*) argue that the more communicatively salient words such as speech editing terms (such as *oh*, *ah*, *er*), swear-words, personal pronouns and proper nouns, assert themselves within the lexicon by preserving high phonology-semantic correlation values. This suggests the existence of a core lexicon including the strongly systematic, communicatively important words that form the scaffolding of the lexicon and provide some of the clues necessary for the inference of meaning from form. The rest of the words, not so constrained for systematicity, fill up the lexicon body.

There is even a commercial application based on the idea of form-meaning systematicity. The principle is implicit in the activity of some recently created companies specialised in creating company and product names. They invent new words aimed at conveying the desired meanings.

5.1.2.2 Structure-preserving representations

Gallistel (1990) defined a representation as a precise correspondence (an isomorphism) between objects and relations in the environment and structure-preserving systems in the brain. As Halliday (1992) explains, within each sensory area of the nervous system, objects and their relationships are represented several times at different processing stages. For example, in the visual system, the retinal image, a highly structure-preserving two-dimensional representation of the visual field, is transmitted to the Lateral Geniculate Nucleus, consisting of six retinotopically mapped layers, and from there to the visual cortex, also retinotopically mapped.

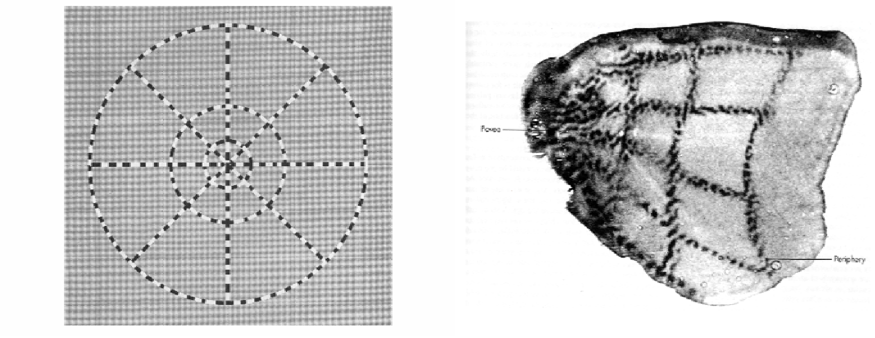


Figure 5.2. Retinotopic representation of a stimulus (left) on the striate cortex (right) of a monkey. From Tootell, Silverman, Switkes and De Valois (1982).

This means that two points that are close together in the proximal stimulus (the first representation in the retina) will be represented by cells that are close together at every stage of processing. In Figure 5.2 we see how the representation on the striate (visual) cortex maintains the fundamental structure of the stimulus. The same happens in hearing, with structure-preserving tonotopic representations of sounds. Different frequencies are perceived by different areas of the cochlea in the inner ear, so that similar frequencies in the acoustic stimulus are represented together in the proximal stimulus. The structure of the proximal stimulus is maintained in successive representations in the primary auditory cortex and in the associative cortex.

In the somatosensory cortex and in the primary motor cortex, adjacent parts of the body end up being represented close together in the cortex (see Figure 5.3). Additionally, these two maps, which lie along each other in the cortex, also map each other. The somatosensory and the motor homunculi (cortex representations) are somehow distorted representations of the human anatomy. Apart from the tendency towards isomorphism, other constraints affect the representation, such as the fact that the homunculus proportions are driven by the number of sensory receptors in the skin, or the limitations of the projection of the 3D human body surface onto the 2D cortex. However, behind these other constraints it is easy to see that general anatomical structure is preserved.

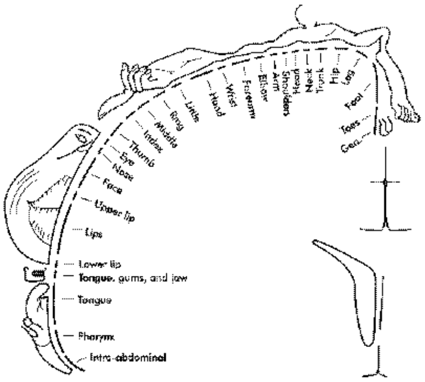


Figure 5.3. Structure-preserving map of the body surface represented on the somatosensory cortex; this map is called the homunculus ('little man' in Latin). From Penfield and Rasmussen (1950).

Structure preserving representations are pervasive in the mammalian brain and present strong processing advantages (Halliday, 1992). First, they make possible analytic processes such as breaking down the stimulus into a number of different types of information, such as colour, orientation and motion in the visual system, which are also represented in a structure-preserving way (see Figure 5.4). The visual proximal stimulus is initially processed by different retinal neuron systems that transform it into a series of jigsaw-like representations, each of them concerned with one aspect of the stimulus. These are the three planes in Figure 5.4.

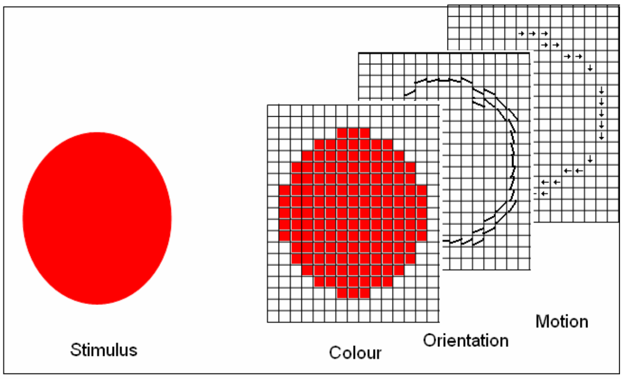


Figure 5.4. Three modality topographical representations of a rotating red oval stimulus. For each visual field (each square in the grids), different modalities of information are processed.

Second, they allow synthetic processes to occur, such as grouping, concerned with building large-scale descriptions, and integration of the different modalities of information. Grouping consists of considering all the pieces of information about one aspect of the stimulus. Elements that share some physical similarity such as colour or orientation are grouped together. For example, taking into account the orientation of all the pieces of the visual field (centre plane in Figure 5.4) allows inference of the shape of the stimulus object. Visual grouping makes descriptions more concise: for example the motion plane in Figure 5.4 can be described as a long list of X points each moving in a different direction, or, if grouped together, as rotation movement. Integration is concerned with putting together information from the different aspects of the stimulus (colour, motion etc) allowing, for instance, the perception of objects as sets of features: in Figure 5.4, ‘one red rotating oval’. These processes make descriptions more manageable by focusing on general properties of the objects and allow generalisation and inference.

Factors such as preserving structure and making generalisations define the systematicity of representations. In the case of language, a representation is a precise correspondence between words and the relationships between them and structure-preserving systems in the brain. The lexicon is represented over different modalities: phonology, syntax, semantics etc. I assume that the faculty of language, like other processing modalities, presents systematicity across representations. This chapter deals fundamentally with the systematicity between two types of information contained in speech: phonological and semantic information. Systematicity implies that language processing involves generalizing from and integrating the different types of information present in the linguistic stimulus (mainly speech, but also text).

The requirements for systematicity, then, are structure-preserving representations and mechanisms to extract, integrate and generalize over different modalities of information contained in the stimulus (the latter have been reviewed under statistical learning in chapter one). Summing up, systematicity of mental representations is ubiquitous in the nervous system and it provides a tool for generalisation and inference.

5.1.2.3 Arbitrariness of the sign

One consequence of the systematicity between word forms and meanings is that it presupposes an intralinguistic determinism of word forms and meanings - given the meaning of word X (its distributional patterns in use), there is a bias for word X to contribute to the overall lexicon systematicity. In other words, there is a bias for word X to have a form that contributes to a phonological level of the lexicon that systematically relates to the semantic level. Therefore its form is not arbitrary. This relates to Saussure's arbitrariness of the sign principle. For Saussure ([1916] 1983) a linguistic sign is a sound pattern linked to a concept. He proposes that signs are involved in two types of relationships: signification, or the link between the form and the concept, and value, determined by the relationships between the signs that form part of a system (Figure 5.5). The following words by Saussure point to a complex lexicon where relationships between words are crucial: "To think of a sign as nothing more than a combination of a sound pattern with a concept would be to isolate it from the system to which it belongs, it would be to suppose that a start could be made with individual signs, and a system constructed by putting them together. On the contrary, the system as a united whole is the starting point, from which it becomes possible, by a process of analysis, to identify its constituent elements".

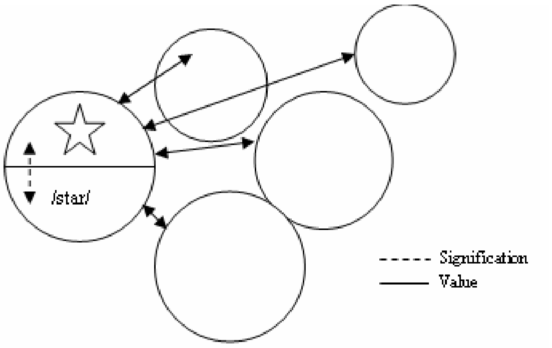


Figure 5.5. Schematic representation of Saussurean relationships of signification (within the sign) and of value (between signs).

Saussure proposed as the first principle of language the arbitrariness of the sign, or the fact that there is no necessary, intrinsic, direct or inevitable relationship between the form and the meaning of a sign. The arbitrariness of the sign was already noted by Aristotle and by Plato (in the Cratylus dialogue). In the present study we are also looking at parameters relating to form and concepts: phonological and semantic word representations. Arbitrariness, then, refers to the signification.

The arbitrariness of signification is not without critics. The sound symbolism literature assumes that there are universal associations between certain sounds and certain meanings. Sound symbolism proposes the opposite to the arbitrariness of the sign principle, namely the idea of a correlation between the form and the meaning of words; and in particular, the claim that phonemes bear information about or are associated with certain meaning (e.g. Magnus, 2001). Sapir (1929) observed correlations between back and front vowels and the notions of big and small, respectively, and Ultan (1978) found that these associations occur cross-linguistically. Kelly, Leben and Cohen (2003) found that certain obstruent consonants are perceived as hard and masculine while sonorants are perceived as soft and feminine. This kind of studies, among others, are carried out and applied today in firms specialising in naming new products to characterize the product and to appeal to different consumer groups.

Shillcock et al. (*submitted*) argue that clusters of similar-meaning words containing similar consonant clusters such as *street, strip, stream, stripe, strap, etc*, which could be the most visible examples of phonology-semantic systematicity, in fact do not contribute to that systematicity, and tend to appear towards the bottom of their systematicity ranking (perhaps because they form a cluster of self-sustained systematicity that can afford to do without systematicity with respect to the rest of the lexicon).

Jespersen, a proponent of sound symbolism or phonosemanticism, wrote: ‘Is there really much more logic in the opposite extreme which denies any kind of sound symbolism (apart from the small class of evident echoisms and 'onomatopoeia') and sees in our words only a collection of accidental and irrational associations of sound and meaning? ...There is no denying that there are words which we feel instinctively to be adequate to express the ideas they stand for.' (Jespersen, 1922).

Jespersen's last observation can be related to the non arbitrariness of the *value* of the sign. Already Sapir (1929) and Firth (1935) felt that speech sounds carried meaning, but suggested their meaning was not inherent to them. Rather, this was a result of "phonetic habit", a tendency to give similar meanings to words with similar sounds. Chandler (2001) also points out that "the principle of arbitrariness does not mean that the form of a word is accidental or random (...). Whilst the sign is not determined extralinguistically, it is subject to intralinguistic determination".

This is consistent with systematicity, which implies that while any one word's phonology is independent from its semantics, the relationships between words' phonological representations are not independent from the relationships between their semantic representations. In the systematic lexicon a dog could suddenly be called 'caterpillar', or someone could use the word 'tree' as a verb, but not without consequences: the rest of the lexicon would need to modify itself to accommodate the change. In an adaptive lexicon always juggling the pressures it is subjected to, such a change could bring instability. This would trigger a chain-reaction of events in the general direction of increasing the stability of the whole system.

In section 5.1 I have described the methods and results of the studies of Shillcock et al. (2001, *submitted*), which show that there is a small but significant correlation between the structure of the form and of the meaning representational spaces of English words. This correlation arguably reflects communicatively important words, and could facilitate word perception, help the acquisition of new words in childhood and the understanding of novel words in adulthood. I have shown that such systematicity across representations is pervasive in the nervous system, and it presents important advantages for processing. This model of the lexicon is not in conflict with the arbitrariness of the sign principle, since the relationship between each form and its meaning remains arbitrary; the systematicity applies between the space of word forms and the space of word meanings in a given language. The next section applies a methodology similar to Shillcock et al.'s to test systematicity in the Spanish lexicon.

5.2 Testing phonology-semantic systematicity in the Spanish lexicon

In this section I test the hypothesis that there is systematicity between phonology and semantics in Spanish. All the principles discussed above are universal: the learning requirements, the neural structure and the philosophical characteristics of symbols apply to all languages. This section tests, among other things, the universality of the phonology-semantic systematicity. If it exists in Spanish as well as in English, there are more grounds to suppose that it is a universal phenomenon and to expect to find it in other languages.

Systematicity implies that words that are phonologically similar will tend to be semantically similar. In order to test this hypothesis, I configure a phonological space and a semantic space by calculating the phonological and the semantic similarity distances between all the word pairs in two different subsets of the lexicon. The hypothesis to be tested is that for a set of word pairs, their phonological and semantic similarity values will be significantly correlated. Section 5.2.1 describes the methodology employed in this test. Section 5.2.2 describes the implementation of the systematicity test using those methods, and presents and discusses the results. Section 5.2.3 is an attempt to remove particularly syntactic information from the calculation of the correlation in order to test the correlation between word form and word meaning.

5.2.1 Methodology

In this section I first describe the metrics of phonological and semantic similarity employed. Then I describe at length and motivate the use of an information-based correlation measure, Fisher divergence, which I use later to calculate the correlation between the phonological and the semantic spaces. Finally, I describe the test used to determine the significance of the correlation: a Monte-Carlo analysis.

5.2.1.1 Phonological similarity

For the phonological similarity I use a parameter-based method that applies the results of the study in chapter three. In that study participants had to select which of two test non-words sounded more similar to a target nonword. The test nonwords shared different parameters with the target, for example they shared the same initial consonant, same final vowel, same two vowels, stress on the same syllable etc. The result was a scale of values that reflected the relative impact of each parameter on the participants' perception of phonological similarity.

cvcv	
c1	-1.25
c2	-4.21
v1	-3.73
v2	-1.83
tc	3.68
tv	3.85
s1	0.97
s2	-0.52
sv1	0.14
sv2	2.88

cvccv	
c1	-1.99
c2	-5.04
c3	-7.33
v1	-1.96
v2	-0.34
tc13	1.05
tc23	-0.25
3c	7.90
tv	6.07
str	-4.60
s1	2.30
s2	0.57
sv1	-4.20
sv2	6.34

Table 5.1 Parameter values for cvcv and cvccv words, from chapter three. (See § 3.2.2.4 for calculation of the values.) C1, c2, c3 = consonants 1, 2 and 3; v1, v2 = vowels 1 and 2; tc = two consonants; tv = two vowels; 3c = three consonants; s1, s2 = same stress on the 1st and 2nd syllable; sv1, sv2 = same stressed vowel on the 1st and 2nd syllable.

In chapter three's analysis of the *relative* importance of these values, the fact that some of them were positive and some negative was informative. For the metric of phonological similarity we need to make all values positive so that sharing a parameter always makes two words *more* similar, to a degree proportional to the parameter value. In order to do that, I recalculate the parameter values from the empirical result matrices obtained in chapter three (shown in Tables 3.6 and 3.7). For the calculation of a given parameter value, I add together the positive values in its column. This way I only take into account the parameters it wins over. Because Fisher divergence is sensitive to the magnitude of the values, I transform the obtained parameter values into a probability distribution, shown in Table 5.2.

cvcv	
c1	0.085
c2	0.008
v1	0.022
v2	0.032
tc	0.223
tv	0.214
s1	0.117
s2	0.053
sv1	0.082
sv2	0.165

cvccv	
c1	0.049
c2	0.031
c3	0.016
v1	0.080
v2	0.065
tc13	0.188
tc23	0.050
3c	0.048
tv	0.156
str	0.076
s1	0.068
s2	0.002
sv1	0.157
sv2	0.014

Table 5.2 New parameter values for cvcv and cvccv words (transformed into probability distributions).

The similarity metric algorithm used in the present test is illustrated in Table 5.3 below and works as follows: first the values from the study in chapter three are transformed into a probability distribution (so that their sum equals 1); then, for each word pair, the algorithm checks whether the two words share each of the parameters in the study in chapter three. If they do, it adds the value for that parameter to the similarity value of the pair (e.g. pair '*pAra*

– *pEro'* shares consonants one and two, the two consonants at the same time, and the accent on the first syllable, so it has marks in the c1, c2, 2c and a1 cells. The resulting phonological similarity value for each word pair, shown on the right-hand column in Table 5.3, is the sum of the values of the parameters the two words share.

	Empirical param.	0.085	0.008	0.022	0.032	0.223	0.214	0.117	0.053	0.082	0.165	
	Param:	*c1*	*c2*	*v1*	*v2*	*tc*	*tv*	*s1*	*s2*	*sv1*	*sv2*	Phon.Sim.
pAra	pEro	x	x			x		x				0.232
kOmo	pEro				x			x				0.150
kOmo	pOka			x				x		x		0.221

Table 5.3. Calculation of phonological similarity of three example word pairs

5.2.1.2 Semantic similarity

Semantic similarity is calculated using the same vector space approach used in § 4.2.1, in chapter four, applied to the surface-form version of the 'Corpus oral de referencia del español' (Marcos Marin, 1992). Each word's position vector is calculated by counting the cooccurrences with a set of context-words. The metric of similarity between two word vectors is the cosine of the angle formed by the two word's position vectors (the cosine as a measure of similarity is explained at length in chapter four, § 4.1.1.3).

5.2.1.3 Correlation between similarity spaces: Fisher divergence

I need a tool to calculate a correlation between the phonological and the semantic spaces, which are defined by the similarity between every word pair in a subset of the lexicon (see Figure 5.6). In this section I will work through an imaginary example whose starting point are the fictitious semantic and phonological spaces in Table 5.4 represented as matrices of distances between pairs of the words 0,1,2,3 and 4. (Note that using distances and similarities should produce the same results, as long as the same measure is used in both the phonological and the semantic spaces.)

SEM	0	1	2	3	4
0	0				
1	0.34	0			
2	0.78	0.98	0		
3	0.86	0.34	0.17	0	
4	0.13	0.79	0.56	0.26	0

PHON	0	1	2	3	4
0	0				
1	0.13	0			
2	0.48	0.56	0		
3	0.44	0.62	0.21	0	
4	0.35	0.59	0.66	0.5	0

Table 5.4. Fictitious semantic and phonological spaces for a lexicon consisting of words 0,1,2,3 and 4.

Shillcock et al. (2001, *submitted*) measure the correlation between the phonological and the semantic distances with Pearson's r. Pearson's correlation assumes data normality and independence. The sets of pairwise distance (or similarity) measures are not necessarily normally distributed, indeed, on many occasions they are multimodal distributions. The data are not independent either. Independence means that one value cannot be predicted from observation of other values, but in our phonological and semantic architectures, each pairwise similarity value depends on all the other similarity values, and if we change one value, the other values will be affected.

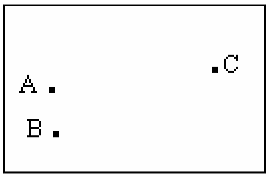


Figure 5.6. Three points in a 2D space.

For instance, Figure 5.6 shows three points in a two-dimensional space. If we change the distance between A and C, the distance between B and C may also change. This seems to imply that Shillcock et al.'s calculations are fundamentally flawed. However, some preliminary tests with Pearson's r on the Spanish data showed significant results similar to those obtained for English. The statistical significance levels attained imply that it is very unlikely that this would have happened by chance, so the possibility remains open that correlating the phonological and the semantic spaces with

Pearson’s r is measuring some potentially different aspect of systematicity between them.

Measuring the correlation with Spearman rank coefficient misses multimodal distributions (which do occur in the phonological space). Given that we will be dealing with large data sets, a two-sample z-test would also be appropriate, and indeed z-scores are highly correlated to Fisher divergence values (R^2>0.96).

I measure the correlation between phonological and semantic similarities using Fisher divergence, a symmetric variant of Fisher information used by Ellison and Kirby (*in preparation*) in a similar task, namely measuring the divergence of distance matrices between the phonologies of different languages.

The calculation of Fisher divergence involves converting the distance values in each space into probability distributions, calculating the geometric mean for each word-pair and then computing for each word the difference in information in the two spaces (the confusion probability) multiplied by the geometric mean. Fisher divergence presents several advantages over other correlation metrics: it correlates matrices; it takes unitless probability distributions as input and it relates to Information theory. Also, the confusion probability for each word-pair can be interpreted as a psychometric measurement, namely the probability that one word is mistaken for the other.

The first step in the calculation of Fisher divergence is transforming the sets of distances into probability distributions: each pairwise distance between two words becomes the probability of confusion of word y with word x in each space C(y|x; space). Intuitively, the more distant two items, the lower their confusion probability. Note that these are theoretical confusion probabilities, different from the actual probability that one word is misheard as another one in a conversation, and also different from the values obtained in chapter three from the comparison of similar non-word stimuli. For the calculation of the confusion probabilities we need *n*, a normalising constant to make the sum of confusion probabilities for each word x equal to 1 (see Table 5.5):

$$n(x; space) = \sum_{\forall y} e^{\text{-K.dist(x, y; space)}}$$

	y=0		y=1		y=2		y=3		y=4		sum
SEM											
n(0)	1	+	0.79	+	0.58	+	0.55	+	0.91	=	3.837
n(1)	0.79	+	1	+	0.51	+	0.79	+	0.58	=	3.665
n(2)	0.58	+	0.51	+	1	+	0.89	+	0.68	=	3.656
n(3)	0.55	+	0.79	+	0.89	+	1	+	0.84	=	4.065
n(4)	0.91	+	0.58	+	0.68	+	0.84	+	1	=	4.006
PHON											
n(0)	1	+	0.91	+	0.72	+	0.74	+	0.78	=	4.153
n(1)	0.91	+	1	+	0.68	+	0.65	+	0.66	=	3.907
n(2)	0.72	+	0.68	+	1	+	0.86	+	0.63	=	3.893
n(3)	0.74	+	0.65	+	0.86	+	1	+	0.71	=	3.959
n(4)	0.78	+	0.66	+	0.63	+	0.71	+	1	=	3.789

Table 5.5. Calculation of *n* for each word.

Now we can calculate the semantic and phonological confusion probability distributions (Table 5.6):

$$C(y \mid x; space) = \frac{1}{\text{n(x)}} e^{\text{-K.[dist(x, y; space)]}}$$

C(y|x; sem)	0	1	2	3	4	sum
0	0.2606	0.2059	0.1518	0.1436	0.2382	1
1	0.2155	0.2728	0.1383	0.2155	0.1578	1
2	0.1593	0.1387	0.2735	0.2431	0.1855	1
3	0.1355	0.1944	0.2187	0.246	0.2054	1
4	0.2281	0.1444	0.1693	0.2085	0.2497	1
sum -->	0.9991	0.9561	0.9516	1.0567	1.0365	

C(y|x; phon)	0	1	2	3	4	sum
0	0.2408	0.2201	0.1727	0.1775	0.1889	1
1	0.2339	0.2559	0.1736	0.1665	0.17	1
2	0.1842	0.1743	0.2569	0.2221	0.1626	1
3	0.1862	0.1643	0.2183	0.2526	0.1786	1
4	0.2071	0.1753	0.167	0.1866	0.2639	1
sum -->	1.0521	0.9899	0.9885	1.0053	0.9641	

Table 5.6. Confusion probability distributions.

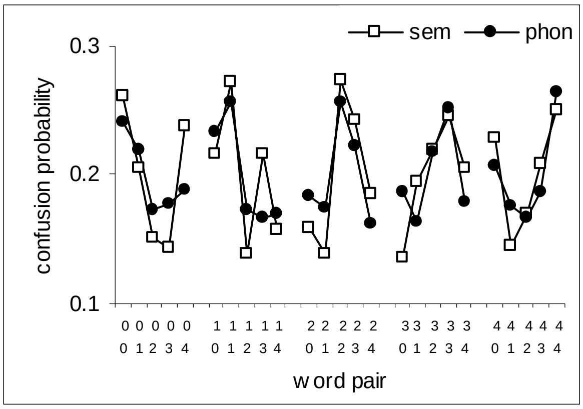


Figure 5.7: Confusion probabilities between each word x (top row of x-axis) and word y (bottom row of x-axis) in the context space and the phonological space. We have separate C's for each word.

Let us focus on the confusion probabilities of word 0 in the phonological space (the leftmost five black circles in Figure 5.7). Given word 0 (e.g. *cat*), these values are the *theoretical* probabilities that when a speaker says *cat*, the listener will hear each of the words 0, 1, 2, 3, and 4, given their phonology. The sum of all these probabilities (all the possible outcomes) is 1. The first value (pair 0, 0: *cat*, *cat*) is the probability that the listener will hear *cat* when the speaker says *cat*, and, as expected, it is the highest of all. The second word must sound rather similar to cat, because it has a high confusion probability. The third one is the most different, and so on for the rest of word 0, and for the rest of the words. Note that in this example, for each word, its confusion probability with itself (0,0) (1,1) (2,2) (3,3) (4,4) is the highest of all, which is to be expected, and reflects the fact that the distances of every word from itself is 0 (Table 5.4). The probability distributions C are in asymmetric square matrices since *n*(x) will not always equal *n*(y).

Note that the matrices in Table 5.6 are completely full, showing the confusion probabilities between each word and every other word, while the original semantic and phonological distance matrices only contained one value for every pair. The confusion probability matrices are not symmetrical, as they consider each word independently. This fact is unique to the calculation of correlations between distance matrices.

The normalising constant K tells us how ‘clear’ each word is. It affects the weight given to the inequalities between the confusion probabilities of each pair. If K=0, all C(y|x) =1/(nr of words); this means that when someone says *cat*, the listener is as likely to understand *cat* as any of the other words (very inefficient communication). As K increases, the pair differences that equal 0 go to 1 and the rest will decrease in value and tend to 0 (see Figure 5.8). This means when someone says *cat*, the listener will have better and better chances of hearing *cat*. In this thesis I have used K = 1 in all calculations.

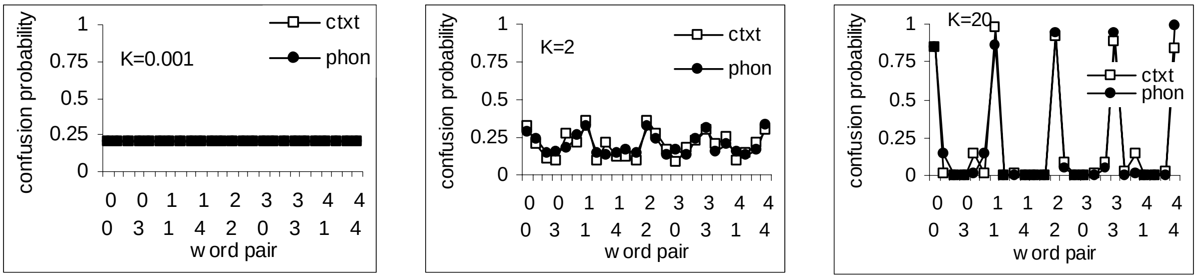


Figure 5.8. Confusion probabilities C(y|x) obtained with different K’s. (Note that not all word pair labels appear on the x axes.)

Fisher divergence tells us how different the two distributions shown in Figure 5.7 are. In order to calculate Fisher divergence F for each word, we first need to define the geometric mean distribution Q(y|x; sem, phon) of C(y|x; sem) and C(y|x; phon). First we calculate the normalising constant *k* for each word (Table 5.7).

$$k(x;sem,phon)=\sum_{\forall y}\sqrt{(\,\mathrm{C(y\mid x;sem)\;C(y\mid x;phon)}\,)}$$

k 0 =	0.997029
k 1 =	0.996984
k 2 =	0.997787
k 3 =	0.996874
k 4 =	0.998588

Table 5.7. Normalising constants *k* for each word

Now I use *k* in the calculation of the geometric mean distribution Q (Table 5.8):

$$Q(y\mid x;sem,phon)=\sqrt{(\,\mathrm{C(y\mid x;sem)\;C(y\mid x;phon)}\,)/\,\mathrm{k(x;sem,phon)}}$$

Q(y|x; SEM, PHON)

	0	1	2	3	4	sum
0	0.2513	0.2135	0.1624	0.1601	0.2128	1
1	0.2252	0.265	0.1554	0.19	0.1643	1
2	0.1717	0.1558	0.2656	0.2329	0.1741	1
3	0.1593	0.1793	0.2192	0.25	0.1921	1
4	0.2177	0.1593	0.1684	0.1975	0.2571	1
sum -- >	1.0251	0.9729	0.971	1.0306	1.0003	

Table 5.8. Geometric mean distribution Q.

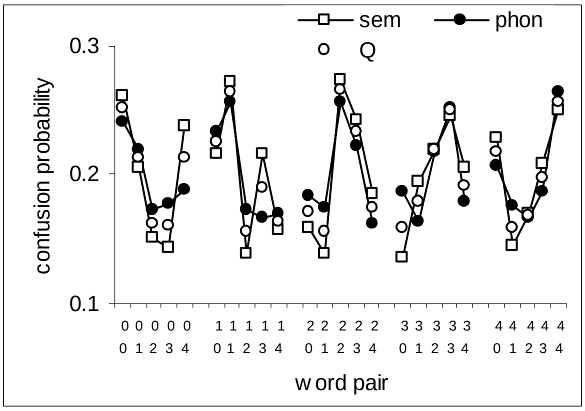


Figure 5.9: As Figure 5.7, but showing also the Geometric mean Q for each word pair.

We calculate the Fisher Divergence F for each word (Table 5.9):

$$F(x;sem,phon)=\sum_{\forall y}(\mathrm{log(C(y\mid x;sem)\text{ - }log(C(y\mid x;sem))}^{2}\;\mathrm{Q\,(y\mid x;sem,phon)}$$

For each word (x), for every other word (y) we take the square of the difference of logs of the confusion probability for that pair. This is the difference in information content (bits) of the pair in the semantic space and the phonological space, squared. The Geometric mean Q for that pair tells us how much weight we should give to that squared difference.

F(0) =	0.0496
F(1) =	0.0503
F(2) =	0.0369
F(3) =	0.0521
F(4) =	0.0235

Table 5.9. Fisher Divergence for each word.

The Fisher Divergence is the sum of all word divergences:

$$F(sem, phon) = \sum_{\forall y} (\mathrm{x; sem, phon})$$

In the present example, Fisher divergence = 0.212

A high Fisher divergence value indicates a low correlation, and a low value, a high correlation.

5.2.1.4 Significance of the correlation

Fisher divergence is a unitless measure of how different (or divergent) the phonological and semantic spaces are. One way to determine the significance of this unitless measure is a Monte-Carlo analysis, which quantifies the probability that the Fisher divergence obtained could have occurred by chance.

The Monte-Carlo analysis to determine the significance of a correlation between two variables is carried out like this: first I calculate the veridical correlation, and then I calculate the correlation between one of the variables and the scrambled values of the other variable a number of times. The idea behind a Monte-Carlo analysis is that if the two variables are correlated, scrambling one variable tends to worsen the correlation value obtained. On the other hand, if the two variables are not really correlated, scrambling is as likely to increase as to decrease the correlation. The significance is determined by the position of the veridical correlation value in the distribution of random correlations, for example, if the veridical correlation falls in the 4th place of a list of 100 random results, its p value is estimated to be 4 / 100 = 0.04.

There is a type of Monte-Carlo analysis called the Mantel test (Legendre & Legendre, 1998) which calculates the significance of the correlation between two distance matrices, but it usually employs Pearson's *r* or Spearman's rank as correlation measures. I adapt Mantel's test by using Fisher divergence instead, and follow Mantel's randomisation method. In order to randomize the values of one of the matrices (e.g. the phonological similarity matrix), I permutate its rows and columns and I calculate the correlation between the original semantic similarities and the scrambled phonological similarities. Permutating the rows and columns has the same effect as scrambling the word pairs before calculating the pairwise phonological similarities. In this study I have compared the veridical Fisher divergence values with 1000 randomisations to obtain robust significances.

5.2.2 Measuring the phonology-semantic correlation

I now apply the methods explained above to a two independent subsets of a corpus of Spanish transcribed speech, and present and discuss the results.

5.2.2.1 Materials

The words to configure the semantic and phonological spaces were extracted from the same Spanish transcribed speech corpus used in preceding chapters. The surface-form version of the corpus was used, meaning that the word-forms contained some morpho-phonological information. The semantic word vectors were calculated counting cooccurrences with function and content words, as this factor combination had the best performance in the experiments presented in chapter four. I consider two separate subsets of the phonetically transcribed Spanish lexicon: the 252 words of structure cvcv and the 146 words of structure cvccv of frequency greater or equal to 20 in the surface-form corpus.

5.2.2.2 Procedure

The correlation between the phonological and the semantic space is tested separately for the cvcv and the cvccv word groups. For each group I calculate the phonological similarity and the semantic similarity between all the possible word-pairs. I calculate the phonological similarity using the method described in § 5.2.1.1 above. For the semantic similarity, words are extracted from the surface-form corpus, and vectors are calculated on the same corpus, counting the cooccurrences with function and high frequency content words (see § 5.2.1.2). Table 5.10 shows the empirically obtained values for both word groups, transformed into probabilities in such a way that they add up to one.

cvcv	
c1	0.070
c2	0
v1	0.023
v2	0.057
tc	0.187
tv	0.191
s1	0.123
s2	0.088
sv1	0.092
sv2	0.169

cvccv	
c1	0.053
c2	0.023
c3	0
tc13	0.083
tc23	0.070
3c	0.151
v1	0.053
v2	0.069
tv	0.132
s1	0.095
s2	0.078
sv1	0.031
sv2	0.135
str	0.027

Table 5.10. Phonological similarity parameter values used in the calculation of the phonological space configuration.

Having obtained a semantic and a phonological similarity value for all word pairs, I calculate the correlation between them using Fisher divergence, and I perform a Monte-Carlo analysis to test its significance.

5.2.2.3 Results

Table 5.11 shows the correlation values (Fisher divergence) for the veridical (unscrambled) phonological and semantic pairwise similarities for the cvcv and the cvccv word groups. It also shows the number of words configuring the spaces, and the significance of the correlation, calculated with a Monte-Carlo analysis with 1000 randomisations. Figure 5.10 shows histograms of the Fisher divergence values obtained with random word pairings, indicating the position of the Fisher divergence obtained with the veridical pairs.

	Fisher divergence	Nr. words	Significance
cvcv	5.03	252	p<0.05
cvccv	2.18	146	p<0.001

Table 5.11. Correlation values (Fisher divergence) between phonological and semantic similarity for the cvcv and cvccv word groups, and its significance. (The lower the Fisher divergence values, the more correlated phonology and semantics are.)

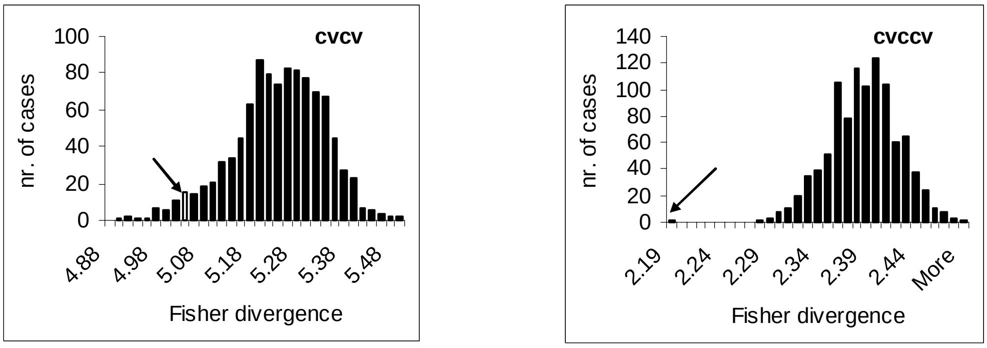


Figure 5.10. Histogram plots showing the results of the Monte-Carlo analysis for cvcv and cvccv words. The veridical results are in the white bins, also indicated by the arrows.

5.2.2.4 Discussion

The results in Table 5.11 and Figure 5.10 show significant correlations between phonological and semantic similarity in cvcv and cvccv words Spanish. A close analysis of how each similarity space was calculated can help understand what drives this correlation. Phonological similarity is calculated with the parameter values obtained in the study presented in chapter three. The choice of parameters is debatable, and that study could have included more parameters, such as sharing not only the same segment in the same position, but also the same phoneme in a different position; or include feature-based instead of segment-based parameters, for instance sharing the same voicing, manner and place of articulation in the same or in a different position. I examine the phonological similarity side of the correlation more closely in the next chapter. The main point is that the phonological similarity metric is psychologically informed.

The metric of semantic similarity is based on the condition from chapter four where semantic similarity was measured on a surface-form (non-lemmatised) corpus, and the cooccurrences with both function and content words were computed (surface, functors + content words, see § 4.2.1.5). As we saw in chapter four, these two conditions together best capture syntactic aspects such as part of speech or gender. Let us examine the relationship of each of them to syntax. First, surface forms contain morphemes such as verb endings and gender markings. The correlation could be driven by syntactic factors, such as the fact that feminine words, plurals and past tenses occur in similar contexts. Second, the main role of functors is to organize syntactic relationships, to signal which word relates to which other: in the phrase 'a bag of chips', 'a' indicates that 'bag' is a noun, and 'of' indicates that 'chips' is connected to 'bag'. In the calculation of the position vectors, I counted each time a word cooccurred with 'a', 'of' etc, giving us clues to the word's syntactic category.

All of this indicates that the correlation between phonology and our measure of semantics may be driven by syntax, at least to some extent. In the next section I attempt to eliminate syntax from the similarity metrics and so discern the influence of other factors such as meaning on the correlation.

5.2.3 Distilling the correlation between word form and meaning

The correlation found between the phonological and the semantic distances in § 5.2.2 could be driven solely by syntax, reflecting the match between the morphosyntactic information contained in word phonology and syntactic information captured by words' cooccurrence with functors. Another contributing factor to the correlation could be phonological typicality, the fact that different syntactic classes have different phonological characteristics: Kelly (1992, 1996) shows phonological differences between English nouns and verbs. For example, disyllabic nouns tend to have initial stress whereas disyllabic verbs tend to have final stress; on average, nouns have more segments, more syllables and longer duration than verbs; and nouns tend to have more low vowels and more nasal consonants than verbs. (See also Durieux & Gillis, 2000, and Monaghan, Chater & Christiansen, 2003, for reviews.)

Another factor could be phonological priming, the putative tendency to produce words containing sounds that are similar to recently uttered or heard words. The effect of (short-range) phonological priming could be eliminated from the correlation metric by using very large context windows such as those of Landauer and Dumais (1997). Phonological priming can be considered as a reflection of the similarity-based structure of the phonological lexicon on speech. An uttered word activates similar-sounding words more than different-sounding words, so the former are more likely than the latter to be uttered soon after.

Among the more tentative contributing factors to the correlation found in § 5.2.2 is the bias towards systematicity between the phonology and the *meaning* levels of the lexicon discussed above. We saw in chapter four that cooccurrence-based semantic similarity spaces do capture meaning, as shown by the facts that they model semantic priming and that they perform above average in semantic tests (§ 4.2.3).

This section aims to test the correlation between word form and word meaning by removing the influence of syntax from the semantic similarity metric. One way to eliminate the influence of syntax in the correlation would be to use the lemmatised corpus and remove the functors from the context word sets in the calculation of the vectors. That condition performed worst of all in syntactic classification tasks: part of speech (§ 4.2.2.1), nouns and verbs (§ 4.2.2.2) and masculine and feminine nouns (§ 4.2.2.4); but it performed well in a semantic task such as noun classification of 'person nouns' (§ 4.2.2.3). However, during lemmatisation, as well as losing their morphemes, certain words have their root changed, and this affects their position in the phonological similarity space. For instance, feminine inflections are an integral part of words and cannot be removed without losing phonological information about the word ending, syllabic structure and length. Lemmatisation replaces irregular forms of verbs by their (regular) stem. Verbs present an additional problem. The canonical verb form, the infinitive, has one of three very characteristic endings: stressed *-ar*, *-er* or *-ir*. My lemmatisation removes the final -r, but still leaves a syntactically conspicuous final stressed *-a*, *-e* or *-i*.

An alternative way of eliminating the effect of syntax on the correlation is to use the surface forms, but to exclude parameters that may pick up on the morphology from the phonological similarity metric. I do not remove the parameters directly related to the last segment, site of the gender morpheme, for several reasons. The last segment is a site of important phonological information, as we saw in chapters two and three, and dispensing with it altogether leaves an incomplete picture of the word’s phonology. Feminine endings are not always inflections of a masculine stem: most feminine words are uninflected (in the aggregate cvcv and cvccv words, only 22% are inflections of a masculine stem), and the ending is arguably part of their phonological identity. Besides, it is not always the case that feminine words end in *-a*, and masculine in *-o*, with about 15% of masculine and feminine words ending in *-e* (see Figure 5.11).

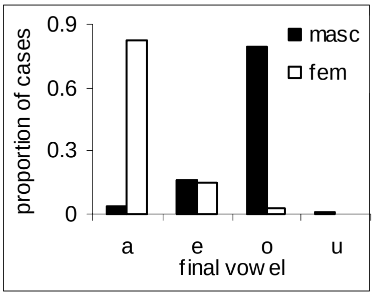


Figure 5.11. Final segment of the aggregate cvcv and cvccv gendered words.

(Note that plural inflections are not an issue, since the two word-groups at hand both end in a vowel, and are all singular.) As explained in chapter three (§ 3.2.2.5.3), the stress-related parameters – sharing the stress on the same syllable and sharing the same stressed vowel on the same syllable – reflect morphological similarity related to verb tense and person. Therefore, removing the stress-related parameters should eliminate most of the morphosyntactic information from the phonological similarity metric.

Summing up, I attempt to remove the effects of syntax by eliminating cooccurrences with functors in the semantic space and by eliminating stress-related parameters from the phonological similarity metric. The next section presents a measurement of a correlation with the new, relatively syntax-free data.

5.2.3.1 Materials

As in § 5.2.2, I use the 252 cvcv and the 146 cvccv phonetically transcribed words of frequency greater or equal to 20 in the surface-form corpus. The position vectors for the semantic similarity calculations take into account cooccurrences with content words, but not with functors.

5.2.3.2 Procedure

The procedure is essentially the same as that of the last section, with a few crucial differences. For the semantic similarity, the calculation of each word's position vector considers the cooccurrences of the target word with the content words - but not with the functors - of frequency greater or equal to 200 in the corpus. The phonological similarity metric calculates the parameter values in the same way as in § 5.2.2.2, but now excluding the parameters related to stress (stress in the same syllable and same stressed vowel in the same syllable) and to syllabic structure. See the new parameter values in Table 5.12. Note that these values are different and not completely correlated with the values in Table 5.10 above, because the removed parameters did not intervene in their calculation.

cvcv	
c1	0.178
c2	0.009
v1	0.021
v2	0.072
tc	0.388
tv	0.332

cvccv	
c1	0.081
c2	0.028
c3	0
tc13	0.105
tc23	0.094
3c	0.321
v1	0.082
v2	0.043
tv	0.246

Table 5.12. Phonological similarity parameter values used in the calculation of the correlation.

5.2.3.3 Results

Table 5.13 shows the correlation values (Fisher divergence) for the cvcv and the cvccv word groups, the number of word pairs configuring the spaces, and the significance, calculated with a Monte-Carlo analysis of 1000 randomisations. Table 5.13 and Figure 5.12 show the results of the Monte-Carlo analysis, indicating the position of the Fisher divergence obtained with the veridical pairs.

	Fisher divergence	**Nr. words**	**Significance**
cvcv	7.79	252	p<0.05
cvccv	3.69	146	p=0.09

Table 5.13. Correlation value (Fisher divergence) and significance for the cvcv and cvccv word groups after removing syntactic cues from phonological and semantic similarity metrics.

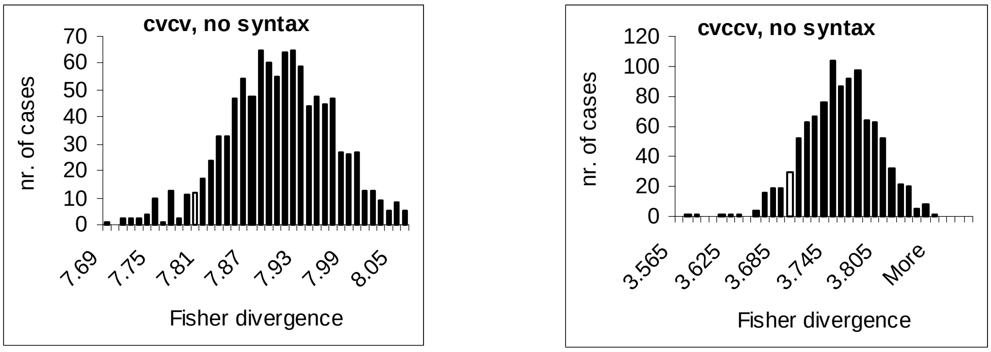


Figure 5.12. Histogram plots showing the results of the Monte-Carlo analysis for cvcv and cvccv words. The veridical results are in the white bins.

5.2.3.4 Discussion

The results obtained after eliminating syntactic information from the data are significant for cvcv words, but only marginally significant for cvccv words. However, the fact that near significance values are obtained in two independent word-groups adds robustness to the results. This indicates that word form may be correlated with word *meaning*, but the results are not totally conclusive. Nevertheless, they are encouraging, given the rough phonological and semantic similarity metrics employed and the relatively small samples of the lexicon tested. It would be interesting to test the correlation with a phonological similarity metric including more parameters and a more robust semantic similarity based on a larger corpus and perhaps using a larger context window. Chapter six will offer some insight in some of these directions.

These results, together with those of section 5.2.2, show that there is a measurable significant correlation between the cooccurrence-based and the phonological levels of representation of the Spanish lexicon. I have shown that part of this correlation can be attributed to syntax, but a small part may rely on the meaning of the concepts denoted by words. The next section looks at the word classes that drive the phon-sem systematicity.

5.3 The systematicity of different word classes

Shillcock et al. (2001, *submitted*) obtained a measure of the phon-sem systematicity for each word, reflecting how well each word fits in with the rest of the lexicon. Shillcock et al. (2001, *submitted*) found that certain communicatively important word classes tended to obtain very good correlation values, and they proposed that these words reflect the pressure towards systematicity to a greater extent than the rest of the lexicon.

In this section I calculate the correlation (Fisher divergence) between the phonological and semantic similarity of each word with every other word, and so can rank them by how well words fit in a phonology-semantics systematic lexicon. I examine the effect of syntactic category and of gender on word fitness in a systematic lexicon, and also look at some of the communicatively important word groups proposed by Shillcock et al.

5.3.1 Method

I replicate Shillcock et al.'s methodology to calculate each cvcv and cvccv word's phon-sem correlation, with the difference that I use Fisher Divergence (see § 5.2.1.3) as a measure of the phonology-semantics (phon-sem) correlation, instead of Pearson's r. As in Also, as in § 5.2., I measure the correlation between phonological and semantic *similarity* (instead of *distance*). In the example explaining the calculation of Fisher divergence in § 5.2.1.3, the correlation values for individual target words appear in Table 5.9. I calculate the rankings of cvcv and cvccv words by Fisher divergence in the 'syntax' and the 'no syntax' conditions, with the similarity calculations explained in § 5.2.2 and § 5.2.3, respectively. (The complete rankings of cvcv and cvccv words by Fisher divergence in the 'no syntax' condition are shown in Appendix F.)

In order to determine how each class of words behaves in terms of the phon-sem correlation, I examine the distribution of the word classes in the list of words ranked by their correlation value.

5.3.2 Results

5.3.2.1 Syntactic factors: noun-verb and gender

This section examines the behaviour of syntactic classes with respect to phon-sem systematicity, looking at how nouns and verbs on the one hand, and masculine and feminine words on the other hand behave in the phon-sem correlation ranks. I examine both the ‘syntax’ and the ‘no syntax’ rankings.

Monaghan, Chater and Christiansen (2003) proposed that the phonological and the collocational typicality of a word with respect to its syntactic category enhance word processing. The words at the top of the correlation-ranked list have, obviously, a stronger match between their phonology and their semantics (which includes meaning *and syntax*), which is another expression of typicality. I now test whether nouns and verbs have different degrees of phonological typicality by looking for any differences in their distributions in the ranked list.

In line with the tests presented in chapter four, I will also look for an effect of gender on word systematicity. Table 5.14 shows the results of two-tailed t-tests applied to the comparisons between nouns and verbs on one hand, and masculine and feminine words on the other hand.

	'syntax'		**‘no syntax’**	
	N-V	gender	N-V	Gender
cvcv	**V>N**	**M>F**	**N>V**	M=F
	t=3.36	t=3.00	t=2.49	t=0.52
	df=90	df=58	df=90	df=58
	__p=0.001**__	__p=0.004**__	__p=0.01*__	p=0.60 (n.s.)
cvccv	V=N	M=F	**N>V**	M>F
	t=1.18	t=1.04	t=3.88	t=1.76
	df=31	df=39	df=31	df=39
	p=0.5 (n.s.)	p=0.6 (n.s.)	__p<0.001**__	p=0.08 (n.s.)

Table 5.14. Results of two-tailed t-tests for the distributions of syntactic category (verbs and nouns) and gender (masculine and feminine) in the phon-sem correlation word rankings. Statistically significant results in bold. Fist line states which word type distribution is higher in the rank; t=t-value; df=degrees of freedom; p=significance.

Although not all results are statistically significant, some trends are apparent in Table 5.14. For the verb-noun distinction, the more heavily inflected verbs tend to present higher phon-sem correlation values in the 'syntax' condition. In this condition, where the measurements of both phonology and semantics in this condition are laden with syntax, the phon-sem correlation is a manifestation of phonological typicality of syntactic categories. Therefore, these results support the idea that verbs have greater phonological typicality than nouns (although at least part of the phonological typicality must be based on the similarly-sounding word inflections).

In the ‘no syntax’ condition where the phonological and semantic similarity metrics remove a great deal of the syntax (§ 5.2.3), the phon-sem correlation cannot be equated with typicality of syntactic categories. Differences in the distributions of nouns and verbs are more likely to be related to word *meaning* as captured by cooccurrence statistics. In this condition, nouns present better systematicity than verbs.

Christiansen and Monaghan’s (in press) studies for English suggest that while cooccurrence statistics alone classify nouns better than verbs, classification of verbs relies more on word-internal cues. In agreement with those suggestions for English, the results presented in Table 5.14 above for Spanish also indicate that cooccurrence statistics, combined with phonological information, classify nouns better than verbs. Verb classification seems to rely on morphology (encoded in word-final phonological regularities) and also on patterns of cooccurrence with functors, as in the 'syntax' condition.

As far as gender is concerned, results are less clear, although there is a general trend for masculine words to be more systematic than feminine words. This suggests that the interaction between gender and systematicity is weaker than that of syntactic category and systematicity, and perhaps a larger set of data would reveal finer aspects of it.

Still on the subject of gender, in the cvcv word group there are six gender-incongruous words - words that have the typical ending of one gender, while grammatically taking the opposite gender, such as *mano* (hand), which looks masculine because it ends in the typical masculine phoneme -*o*, but it is actually feminine: *una mano blanca* (a white hand). The gender-incongruous words are *mano* (hand), *moto* (motorbike), *foto* (photograph), *tema* (subject), *cura* (priest) and *sida* (AIDS). This incongruity can be kept only in relatively frequent words, and some gender-incongruous words which fell into disuse actually changed their grammatical gender to match their form. Differences in systematicity between the genders could help explain gender-incongruous words. In the 'syntax' condition, because cooccurrence with (gendered) determiners and other functors is part of the metric, gender incongruous words are expected to group with the words of the same grammatical gender (e.g. *la mano* would group with feminine words). In the 'no syntax' condition, if incongruous words are grouped with words of the same grammatical gender (*la mano* grouping with *la casa, la mesa* etc), then we can infer that their 'semantic' gender is grammatically encoded by the determiners; if, on the other hand, they group with words with the same ending and opposite syntactic gender (*la mano* grouped with *el perro, el pato*), then we can infer that their 'semantic' gender is determined by their form.

In the 'syntax' condition, the six incongruous words are in the bottom half of the ranked word list, with the three masculine-form, grammatically feminine words at the very bottom (positions 203, 244 and 248 out of 252). In the 'no syntax' condition, while the three feminine-form, grammatically masculine words stay in similar rank positions, the masculine-form, grammatically feminine words go up in the correlation ranking to group with the more systematic, normal masculine words. This is true particularly of the less frequent *moto* (goes up to position 154 from 203) and *foto* (goes up to position 91 from 248), while mano is still quite close to the bottom in position 206 out of 252.

This small effect supports the claim that semantic gender is driven by word form. Studies on a larger section of the lexicon in a larger corpus, using the gender differences in phon-sem systematicity could help determine what drives the gender of words – form or syntax.

This section has shown that there are differences in the degree of systematicity between syntactic classes, and that those differences depend on the way systematicity is measured. If we include syntax in the similarity metrics, inflected forms are more systematic, as expected. If we remove syntax from the similarity metrics, the unmarked syntactic classes show higher phon-sem systematicity.

The next section examines the behaviour with respect to systematicity of Shillcock et al's (2001, *submitted*) proposed communicatively salient words, in Spanish.

5.3.2.2 Communicatively salient words

Shillcock et al.'s (2001, *submitted*) studies in English found that what they termed communicatively important word classes, namely speech editing terms (such as *oh, ah, er*), swear-words, personal pronouns and proper nouns, tended to appear high in correlation ranking. The word groups considered in this chapter contain very few words belonging to those classes, but I nevertheless attempt to find out whether the communicatively salient classes available rank high in Spanish, mirroring the results for English.

There are no editing terms or personal pronouns in the sets, and only two swear-words in the cvcv group. The swear-words are *puta* and *coño*, and perhaps the less offensive *culo* could be added to this group, because it is part of a few very rude expressions. They rank 31st (puta) 35th (*culo*) and 247th (*coño*) out of 252, respectively in the 'syntax' condition, and 11th, 70th and 163rd, respectively, in the no syntax. There is no indication, then, that swear words are particularly systematic in Spanish.

There are 15 cvcv proper nouns (nine person's first name, three place names, one family name and one month name), and 10 cvccv proper nouns (6 person's first names and three place names). Figure 5.13 shows the distribution of cvcv proper nouns in the 'no syntax' condition as an illustration. Table 5.15 shows the results of two-tailed t-tests between the distributions of all words and the distributions of proper nouns in the two word-groups and the two conditions.

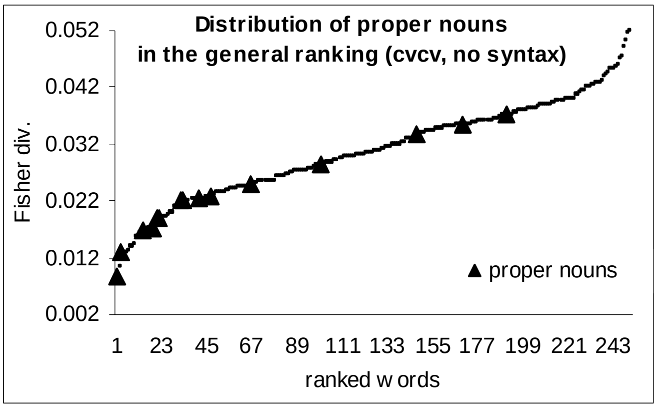


Figure 5.13. Illustrative example of the distribution of proper nouns in the ranking of words by phon-sem correlation (Fisher divergence).

	'syntax'	‘no syntax’

cvcv	t=17.55
	df=14
	p<0.001**
		t=2.74
		df=14
		p=0.01**

cvccv	t=0.81
	df=8
	p=0.44 (n.s.)
		t=4.29
		df=8
		p<0.01**

Table 5.15. Comparisons of the proper noun distributions with the distribution of the rest of the words. t=t-value; df= degrees of freedom; p=statistical significance.

Figure 5.13 shows how proper nouns cluster at the top of the ranking, with low Fisher divergence values. Table 5.15 shows that they do so significantly in three out or four cases. In the case where they do not cluster at the top it is worth noting that the four proper nouns placed at the end of the ranking that stop the clustering being significant are the three place names in the set. In the 'syntax' condition, *place* proper nouns go down in the rank, becoming more differentiated from person proper nouns in both cvcv and cvccv words.

Spanish proper nouns, then, particularly person first names (as opposed to place names), are significantly clustered at the top of the ranking, supporting the results found by Shillcock et al. for English.

An examination of the distribution of other word classes in the Spanish word phon-sem systematicity ranking revealed that numerals have distinct distributions. There are two numerals in each word group: *doce* (twelve), *cero* (zero) in the cvcv group and *quince* (fifteen) and *cinco* (five) in the cvccv group. Figure 5.14 shows the position of the words divided by the total number of words.

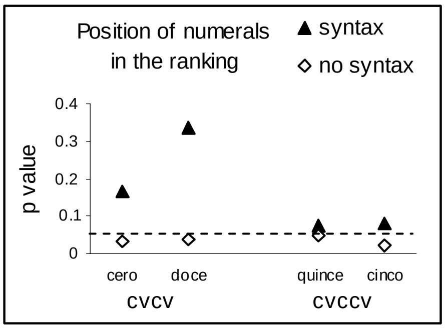


Figure 5.14. Significance of the position at the top of the rankings of the two cvcv and the two cvccv numerals. Dashed line marks the p=0.05 significance threshold.

Numerals are found towards the top of the rank in all cases, statistically significantly so in the 'no syntax' condition in both word groups.

It might be argued that these results with these particular numerals is due, on the phonological side, to the fact that they sound very similar and, on the semantic side, to the fact that numerals are tightly clustered in the cooccurrence space, as we saw in chapter four (Figure 4.4). This is not a complete explanation, however, since the phonological similarity metric employed, based on sharing the same segments *in the same position*, does not pick up much of the phonological similarity between *doce* and *cero*, and *quince* and *cinco*. Therefore, we cannot rule out that high systematicity is a general property of numerals.

This section has confirmed that it is possible to measure the contribution to systematicity of individual words it also has presented results that support Shillcock et al.'s claim that certain word classes contribute more than others towards systematicity in the lexicon. It has also shown that the similarity metrics greatly affect the behaviour of word classes in terms of systematicity.

The 'syntax' condition can be considered to represent an aggregate syntax-semantic level of the lexicon, and the 'no syntax' condition, the semantic level of the lexicon. The results for syntactic categories and for gender (Table 5.14) and for proper nouns (Table 5.15) suggest that the phonology-syntax systematicity can lead to different classifications of words than phon-sem systematicity, showing once again a complex multilevel multidimensional lexicon.

5.4 General discussion

The tests reported in § 5.2 show there is a significant correlation between the phonological and the cooccurrence-based levels of the Spanish lexicon. When syntax is removed from the paradigm (§ 5.3), the correlation is significant for one word group, but only marginally significant for the other. In this discussion I analyse how my results for Spanish fit in with those of Shillcock et al. for English and offer possible explanations as to why their result’s statistical significance was better. Table 5.16 summarises the results of my two tests and Shillcock et al.’s study, and the similarities and differences between them.

	Shillcock et al.	**Exp 1 (§ 5.2.2)**	**Exp 2 (§ 5.2.3)**
Result (significance of the correlation)	p<0.001	p<0.05 ; p=0.001	p<0.05 ; p<0.1
Nr. of words	1733	252 ; 146	252 ; 146
Nr. of pairs	1,500,778	31,626 ; 10,585	31,626 ; 10,585
Phon. similarity metric	Wagner-Fisher edit distance with psychologically motivated penalty values	Psychologically motivated parameters (incl. some reflecting morphology)	Psychologically motivated parameters, (excl. those reflecting morphology)
Lemmatisation of the corpus	Yes	No	No
Context words (in sem. similarity metric)	Content	Content+functors	Content
Word structure	Different length and syllabic structure (monosyllabic only)	Same length and syllabic structure (bisyllabic)	Same length and syllabic structure (bisyllabic)

Table 5.16. Comparison of the results and experimental variables of Shillcock et al. (2001, *submitted*) and the two tests presented in this chapter.

Both Shillcock et al's English study and Experiment 2 for Spanish made an effort to remove syntax, if the approaches were slightly different. Shillcock et al.'s results claim to capture meaning but not syntax is based on the lemmatisation of the corpus. As I explained in § 5.2.3 above, while lemmatisation of English words is relatively straightforward and leaves the stem unchanged in the majority of cases, lemmatisation of Spanish words has unwanted consequences for word phonology. My approach to removing syntax was to exclude the phonological similarity parameters that captured verb tense and person, the main morphosyntactic elements in my data. Both Shillcock et al. and my 'no syntax' condition removed syntax from the semantic representations by excluding functors from the context word set in the calculation of the vectors.

Shillcock et al.'s phonological similarity metric, an edit distance algorithm, used penalties for mismatches between words based on psychologically motivated perceptual differences between phonemes. Crucially, they also introduced an unmotivated high penalty for word length mismatch. My data consisted of two word sets of equal length and syllabic structure, so length could not play a role in my phonological similarity metric. All my phonological similarity parameter values were psychologically plausible: they were the direct result of a study of people's phonological similarity judgements (reported in chapter three). However, they were very limited. They only compared segments in the same position (first with first, second with second etc), ignoring cases where the same segments appeared in a different position. This metric does not pick up, for example, the phonological similarity between *mato* and *toma*. My parameters were based on phoneme identity, so unlike Shillcock et al.'s, my metric did not take into account feature sharing.

The size of the corpus and of the lexicon sample studied also makes Shillcock et al.'s study more reliable. Their word set was one order of magnitude larger than my cvcv and cvccv sets, and their pair set, i.e. the number of pairwise phonological and semantic distances, *two* orders of magnitude larger than mine. Shillcock et al.'s semantic vectors were calculated on a 100 million word corpus, whereas my corpus was only 1 million words. This means that both their phonological and semantic spaces are more fine-grained than the ones used in this chapter.

All this suggests that even though my phonological parameters were more psychologically plausible, Shillcock et al.'s were more fine-grained and, together with a larger, higher-definition data-set, they may have produced more refined spaces. Additionally, it is possible that word length, absent from my metric, plays an important role in the phonological space (this last possibility is further explored in chapter six).

On the other hand, Shillcock et al. calculated the correlation between the phonological and semantic spaces using Pearson's r, which, as we have seen, is not appropriate in cases like this where measures are not normally distributed and, crucially, independent of each other. I used an information-theory measure appropriate to measure the correlation of two similarity matrices.

Section 5.3 has shown that, in Spanish, certain word classes support phonology-semantics systematicity. These results are consistent with those of Shillcock et al.'s (*submitted*) for English.

This chapter has combined the methodologies presented in chapters three and four to build a method to measure the correlation between the phonological and the semantic levels of the lexicon, and has indicated that there is a significant correlation between them, at least partly brought about by word meaning. We have also seen that certain word classes seem to drive this correlation. An in-depth analysis has shown that there is scope for refinement in the methods. Taken together, the results for Spanish presented here and Shillcock et al.'s (2001, *submitted*) results for English support the universality of the correlation.

The next chapter introduces a new methodology to explore the phonological similarity parameter space, based on the correlation between phonology and semantics measured in the present chapter. The results further support the existence of a pressure for systematicity in the lexicon, and also reveal traces in the phonological lexicon of the opposed pressure for word intelligibility.

Chapter 6. The phonological lexicon structure and systematicity

Chapter five showed evidence supporting the existence of a systematic relationship between the phonological and the semantic levels of the lexicon - the latter including word morphosyntax as well as meaning. This systematicity is the basis of the new approach to the study of the phonological level of the lexicon. I examine the impact of different parameters of phonological similarity on systematicity in an attempt to reveal the pressures that configure the phonological structure of the mental lexicon. I show how while certain parameters seem to contribute to systematicity, others seem to respond to opposed pressures that go against systematicity, but help word recognition.

6.1 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), syntactic structure (Kelly 1996, Christiansen & Monaghan, in press) or intra-word organisation (chapter two of this thesis). This chapter presents theory-independent corpus-based methods that aim at discovering aspects of the phonological mental lexicon. These methods assume and are based on a systematic relationship between the phonological and the syntax-semantic levels of the mental lexicon. This means that the *semantic* lexicon level, through its systematic relationship with the phonological level, plays a part in the evaluation of the parameters that configure the *phonological* lexicon.

In chapter five I presented evidence supporting the existence of a pressure for systematicity across levels of representation of the lexicon, particularly of the tendency for word phonological similarity to correlate with word semantic similarity. This ‘phon-sem’ correlation may be driven by the phonological space, by the semantic space, or by both. It may be the case that the phonology of words adapts to match their semantic and syntactic relationships, or that the meanings adapt to match the words' phonological relationships. In a complex adaptive lexicon it is more likely that both spaces have coevolved under the pressure for systematicity that links them. In this chapter I concentrate on the phonological side of the correlation and attempt to answer questions such as: How well do phonological spaces configured with different parameter sets correlate with the semantic space? Is the empirical, psychologically plausible set of parameter values particularly good for the correlation? Can we use the phon-sem correlation to predict the values of parameters of phonological similarity?

The methodology employed involves evaluating a phonological parameter space in terms of its correlation with the semantic space (the phon-sem correlation), in two ways: first, a random search of the parameter space returns a general measure of how each phonological parameter affects the phon-sem correlation; second, a hill-climbing search returns the parameter configuration that obtains the best phon-sem correlation. This information is contrasted and compared with the empirical parameter values from chapter three, and the results are discussed.

I also describe the application of the above methods to new word groups, for which I do not have empirical parameter values. In one of those cases I use the methodology to make a testable prediction of what the empirical parameter values might be in the new word group. Finally, I discuss the combined results of the different methodologies employed in the chapter and draw some conclusions. The hypotheses to be tested are:

- that the psychologically plausible parameter configuration produces a better phon-sem correlation than most randomly generated configurations, because I assume a pressure towards phon-sem systematicity in the lexicon;

- that, nevertheless, the parameter values yielding the optimal phon-sem correlation will be different from the empirical values, and this can be explained in terms of pressures on the lexicon structure other than phon-sem systematicity;
- that we can make testable predictions based on the phon-sem correlation, for instance, predict what the empirical phonological parameter values will be for a word-group.

6.2 A random search of the phonological lexicon

A random search provides a general idea of the behaviour of dependent variables with respect to an independent variable. In this case, we are interested in the behaviour of parameters of phonological similarity with respect to phon-sem systematicity.

I search a phonological parameter space using a random search algorithm: I generate a random configuration of values of parameters of phonological similarity. The random configuration is used to calculate the phonological similarity in all the word pairs in a sample of the lexicon. I correlate these pairwise similarity values with the semantic similarity values for the same word group. (As in chapters four and five, semantic similarity is based on cooccurrence, and the phon-sem correlation is measured with Fisher divergence.) I use the *phon-sem correlation* value obtained to evaluate the initial *phonological* parameters - a high correlation indicates that the random parameter values tend to contribute to systematicity, and a low correlation indicates that the random parameter values tend to go against systematicity.

6.2.1 Data

I perform random searches in three independent phonological spaces: those formed by cvcv, cvccv and cvcvcv words. I already used the cvcv and cvccv word groups in chapter five: the 252 words of structure cvcv and the 146 words of structure cvccv of frequency greater than or equal to 20 in the surface-form version of the ‘Corpus oral de referencia del español’ (Marcos Marín, 1992). For the third group I extract the 148 cvcvcv words of frequency greater than or equal to 20 from the same corpus version. The semantic similarity values are calculated exactly in the same way as those of the cvcv and cvccv word groups, with the same context words for the ‘syntax’ and ‘no syntax’ conditions as in cvcv and cvccv words. As in chapter five, the 'syntax' condition includes stress parameters in the calculation of phonological similarity and functors and content words in the semantic similarity algorithm; the 'no syntax' condition excludes stress parameters from the phonological similarity algorithm and functors from the semantic similarity algorithm. For the phonological similarity metric of cvcvcv words, I extend the parameter set to accommodate the different word structure, and include, for instance, 'sharing three vowels', 'sharing the stress on the antepenult syllable' (see all parameters in Figure 6.4 below).

	cvcv	**cvccv**	**cvcvcv**
words (freq >20)	252	146	148
nr. param. (syntax)	6	9	14
nr. param. (no syntax)	10	14	20
empirical param. values	yes	yes	no

Table 6.1. Some information on the three independent lexicon subsets tested in this chapter.

Any similarities between the parameter impact values obtained in these three independent spaces would further support the existence of systematicity between phonological and semantic lexical relationships.

6.2.2 Method

6.2.2.1 The hyperspace

The general mechanism of this random search consists of measuring the correlation between many randomly generated phonological similarity spaces and the semantic similarity space. An analysis of the covariance of the random phonological similarity parameters with the correlation values will reveal which parameters are driving the correlation.

c1	c2	v1	v2	tc	tv	Fisher d.
0.087517	0.212833	0.055019	0.084772	0.29637	0.263489	7.774749
0.17769	**0.008853**	**0.021369**	**0.071753**	**0.388484**	**0.331852**	**7.790784**
0.210902	0.203684	0.071964	0.105185	0.280549	0.127717	7.803025
0.220101	0.214613	0.037353	0.035377	0.233557	0.258999	7.808966
0.161624	0.178805	0.078445	0.067063	0.268136	0.245927	7.812247
0.058123	0.314918	0.019214	0.01777	0.281783	0.308192	7.813697
0.260471	0.246922	0.101917	0.038547	0.309919	0.042223	7.815148
0.109106	0.318201	0.087451	0.067963	0.265865	0.151413	7.841004
0.253304	0.274158	0.073371	0.04138	0.26691	0.090877	7.842624
0.075941	0.00366	0.130133	0.109758	0.482546	0.197963	7.844313
0.125452	0.138836	0.099334	0.039778	0.309391	0.287209	7.84574

Table 6.2. Eleven points, extracted from the top of the systematicity-ranked 2000 random points in the 6-dimensional phonological parameter space for cvcv words (no syntax). The first six columns show the random parameter values and the last column, the Fisher divergence between the phonological space calculated with those parameters and the semantic space. In bold, the empirical values and their corresponding Fisher divergence.

The random search of the parameter space follows the following steps:

1. I generate a set of random parameter values (like the non-bold lines in Table 6.2) and normalise them in such a way that they add up to one[1].
2. I use these parameter values to compute the phonological similarity values for all the word pairs in a set.
3. I calculate the Fisher divergence between those pairwise phonological similarity values and the veridical semantic similarities (those calculated in chapter four and also used in chapter five) for the same word pairs (the phon-sem correlation).
4. I keep a record of the random parameter values (first six columns in Table 6.2) and the phon-sem correlation obtained with them (right-hand column in Table 6.2).
5. I repeat steps 1 to 4 2000 times.

[1] In order to counter Fisher divergence’s sensitivity to the magnitude of the data, all the random parameter sets are normalised. This way, the sum of all values is always one, and the fluctuation in phonological similarity value magnitude is a function of the relative parameter values only.

The result of the random search approach 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 systematicity value, determined by the phon-sem correlation – the correlation between the phonological level of the lexicon calculated with the random parameter value set, and the semantic level of the lexicon. To help visualize this hyperspace, Figure 6.1 shows just two of the many dimensions involved in its configuration (*tc* and *v1*).

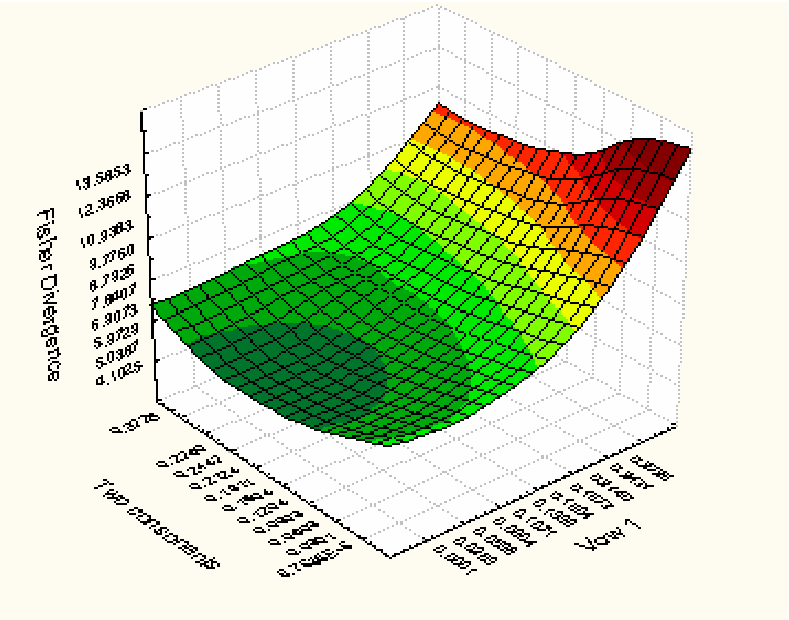


Figure 6.1: Surface plot of the phon-sem correlation showing two phonological parameters: first vowel (*v1*) and two consonants (*tc*) (cvcv words, 'syntax' condition). The surface is created by 2000 3D points. The horizontal position of each point is given by the values of phonological parameters *tc* and *v1*. The height is given by the phon-sem correlation (Fisher divergence) obtained with the parameter value combination.

The valleys in the surface correspond to the best phon-sem correlation (low Fisher divergence). In the example in Figure 6.1, if we hold all other parameters constant, the best correlation (dark green valley) is obtained by the lowest values of *v1* combined with intermediate *tc* values. The worst correlation is obtained with very low *tc* and very high *v1* values (red corner).

Note that the phonological similarity parameters are not always independent of each other. In the random search, however, parameter values are random and independent in each run of the algorithm described above; for the calculation of the linear parameter impact values each parameter was compared independently with Fisher divergence.

6.2.2.2 Extracting parameter impact values from the hyperspace

The effect of each parameter on systematicity may be assessed with a regression analysis, which answers the question: to what extent the values of one parameter predict the Fisher divergence values? (In Table 6.2, a regression tells us to what extent the values in each of the first six columns predict the values in the right-hand column). Regression analysis can be used to determine the nature of the effect of each parameter on systematicity – is it linear or non-linear? What model fits the effect best? It also returns a quantitative measure of the impact of each parameter on systematicity.

Linear relationships between the parameters and systematicity are revealed by the linear covariance of each parameter with the phon-sem Fisher divergence. Linear covariance can be measured with a number of tools: the regression linear r^2, the covariance, and the correlation coefficient Pearson's *r*. These three measures correlate perfectly with each other, and in this study I choose to use Pearson's *r* because it is the only one that indicates whether a parameter value is directly or inversely proportional to the phon-sem Fisher divergence. For example, in order to measure the effect of the first consonant (*c1*) on the phon-sem correlation I calculate Pearson's *r* for the first and the last columns in Table 6.2.

In a linear relationship, a *low* Fisher divergence value indicates a *high* phon-sem systematicity, so I use the negative of Pearson’s *r* as the measure of the linear covariance of the parameter with systematicity. A positive covariance indicates that high values of the parameter in question improve the systematicity measured. Conversely, a negative covariance indicates that low value of the parameter improve the systematicity. A covariance near zero indicates that the parameter does not greatly affect the systematicity.

The linear covariance may be only an approximation, since I do not know whether the parameters affect the correlation linearly. I test all the **nonlinear** functions available in SPSS (2003) on each parameter and obtain an r^2 value for each nonlinear model:

$$r^2 = \frac{SSR}{SST}$$

SSR = regression sum of squares

SST = total sum of squares

This r^2 is a measure of how well each model fits the data. I run a curve estimation of the following regression models: linear, logarithmic, inverse, quadratic, cubic, compound, power, sigmoid, growth, exponential and logistic, and examine the r^2 obtained by each. Myers (1990) warns of one problem of such exploratory use of regression: 'Several models can be fit that would be of nearly equal effectiveness. Thus the problem that one deals with is the selection of one model from a pool of candidate models'. To deal with this problem, Stevens (1992) suggests cross validating the models on different data sets. Similar performance of the models across the independent word groups cvcv, cvccv and cvcvcv will help identify the most reliable regression function.

6.2.2.3 The empirical parameter values

The random search and analysis of covariance will return values representing the phonological parameter's impact on the phon-sem correlation (this will be expanded in § 6.2.3.1); these parameter values may be consistent across the independent word-groups, supporting phon-sem systematicity. I now describe a method to substantiate the claim that these parameters are exploiting *phonological similarity*, and not some other information pattern in the lexicon. The test is based on comparing the random-search, corpus-based parameter impact values with the empirical parameter values obtained in chapter three. The empirical values measured the impact of each parameter on perceived phonological similarity, so a correspondence between them and the impact values obtained with the random search will support the claim that we are indeed exploring a phonological similarity space.

The psycholinguistic study in chapter three returned empirical values for the parameters of phonological word similarity for cvcv and cvccv words. In chapter five I calculated the phon-sem correlation based on those parameter values and found it to be significant. I will use the data from chapter three to ground the random search: I expect to find that the parameter impact values obtained with the random search are similar in some ways to the empirical ones. This would mean that phonological structure of the lexicon predicted by the phon-sem correlation is similar in some ways to the phonological structure of the lexicon derived from the empirical data in chapter three, and would provide extra evidence in favour of the systematic lexicon.

Note that in chapter three I only tested cvcv and cvccv words, so I do not have empirical parameter values for cvcvcv words. Later in this section I will use information from the other word-groups as well as the results of the random and the hill-climbing searches to predict empirical parameter values for cvcvcv words.

The empirical values employed here are calculated in a slightly different way than the ones shown in chapter three. Here we need a set of positive values that can be normalised so that they add up to one (see Footnote 1 in pag. 174), so in the matrices in chapter three (§ 3.2.2.4) I only add, for each column, the *positive* values. That means that, for each parameter, I only count the values related to the parameters it wins over.

The empirical parameter values calculated in this way are shown in Figure 6.2.

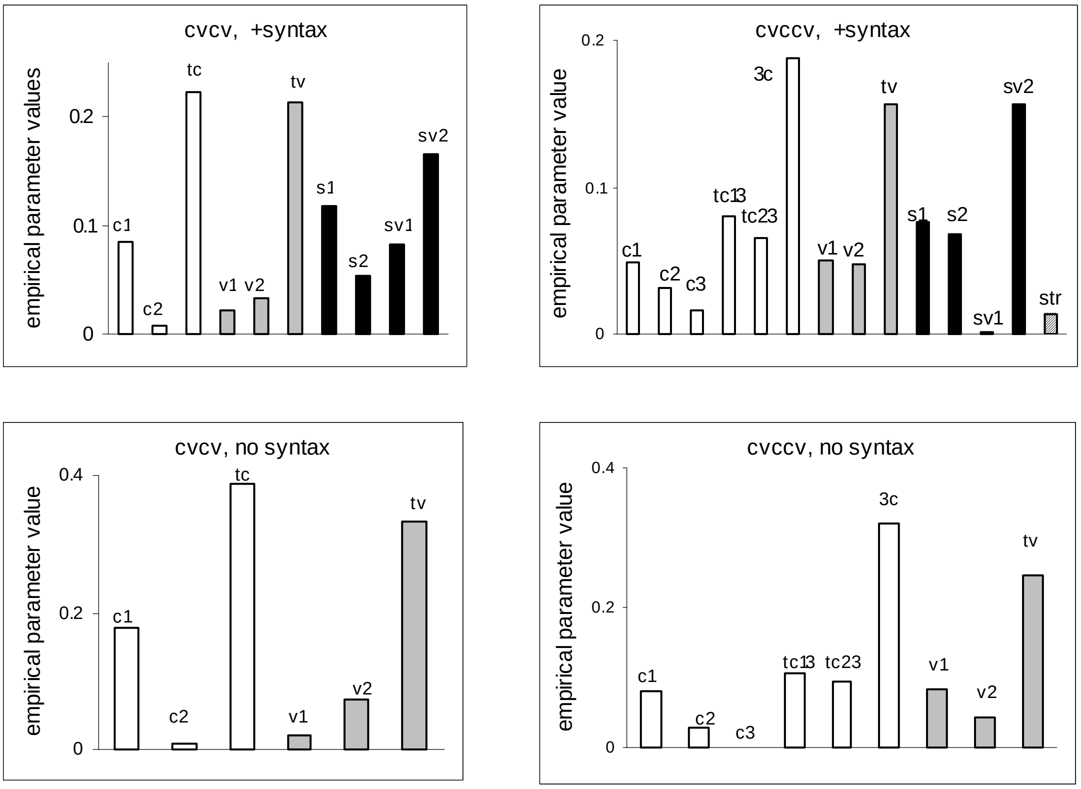


Figure 6.2. Empirical parameter values obtained from psycholinguistic testing (chapter three) for cvcv and cvccv words, calculated taking stress into account ('syntax' condition) and not taking it into account ('no syntax' condition).

These empirical parameter values are very similar to those obtained in chapter three with a slightly different calculation, and their main features are the same: the more consonants or vowels shared, the more similar two words are perceived to be; the initial consonant is the most salient single segment; and sharing the stressed final vowel greatly increases perceived similarity. They are also the same used in chapter five, but this time they have not been transformed into a probability distribution.

6.2.3 Results for cvcv, cvccv and cvcvcv words

I calculate 2000 random points for each of the three phonological spaces: cvcv, cvccv and cvcvcv, each in two conditions, 'syntax' and 'no syntax'.

I calculate first the linear impact parameter values and then examine how linear and nonlinear models fit the relationship of each parameter with systematicity.

6.2.3.1 Systematicity-driven linear parameter impact values

The following Figures show the linear impact values of the phonological parameters in the cvcv and the cvccv (Figure 6.3) and the cvcvcv (Figure 6.4) word groups, in the 'syntax' and 'no syntax' conditions. The bars represent the covariance of each parameter with the phon-sem systematicity. (Note that these bars represent the effect of the *joint* application of all parameters to the phonological space calculation.)

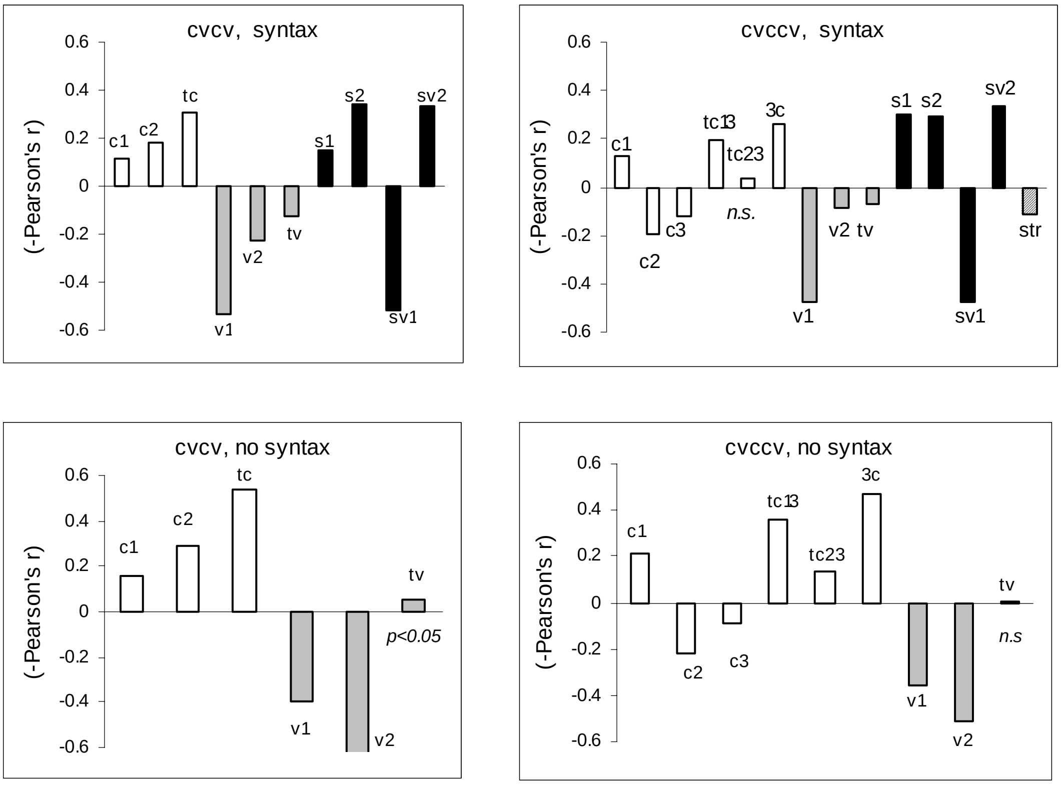
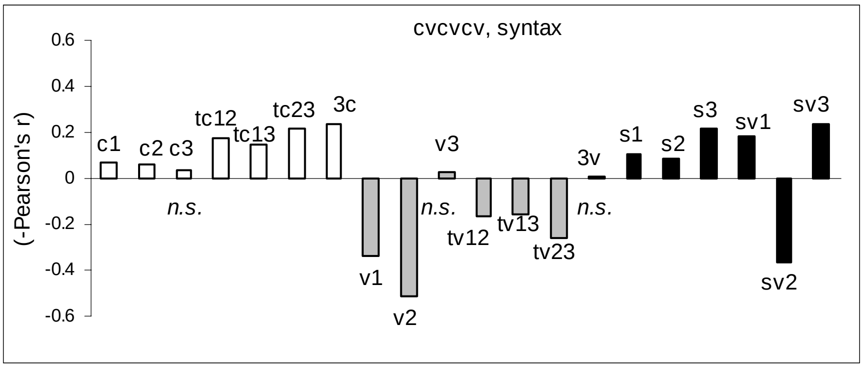


Figure 6.3: Linear 'parameter impact' values, representing the impact of the phonological similarity parameter on the phon-sem correlation. Two conditions, syntax and no syntax are shown for the cvcv and cvccv word groups. White bars= consonant-related parameters; grey bars=vowel-related parameters; black bars=stress-related parameters; striped bar=structure-related parameter. Unless otherwise stated, p<0.01.



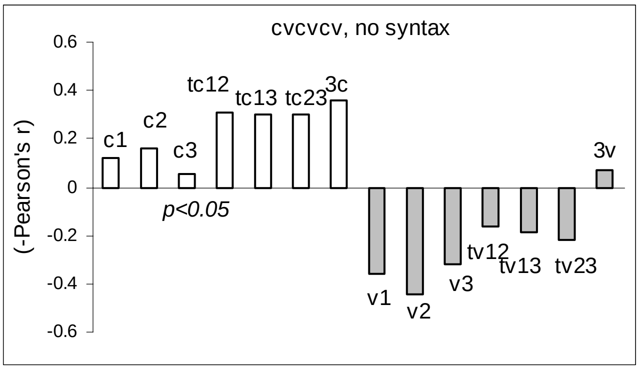


Figure 6.4: Same as Figure 6.3 above, but for cvcvcv words. Unless otherwise stated, p<0.01.

The parameter impact values show a high level of coherence across the three word groups. Comparable parameters across groups are highly correlated, as shown in Table 6.3.

	'syntax'	
	cvcv (10)	cvccv (14)
cvccv (14)	0.86	
cvcvcv (20)	0.90	0.94

	'no syntax'	
	cvcv (6)	cvccv (10)
cvccv (10)	0.84	
cvcvcv (14)	0.95	0.90

Table 6.3. Consistency across word-groups in the 'syntax' and 'no syntax' conditions: R^2 of counterpart parameter impact values indicates covariance of the parameters with respect to lexicon systematicity in three independent word groups - cvcv, cvccv and cvcvcv. The number of parameters in each condition is shown in brackets. All p<0.01.

These high across-group correlations provide support for the methodology employed, indicating that the phonological parameters have the same impact on phon-sem systematicity in three independent subsets of the lexicon. In other words, this consistency supports the existence of systematicity between the phonology and the semantics of the lexicon - if there was no phon-sem systematicity in the lexicon, the phon-sem correlation would not have generated the same phonological parameter values in three independent word groups.

One similarity between the three word groups is that sharing all consonants or all vowels (*tc, 3c, tv, 3v*) tends to have greater impact on systematicity (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. The morphosyntactic information encoded by the final vowel may explain its positive impact in the ‘syntax’ condition.

Figures 6.3 and 6.4 show that most consonant parameters (in white) have positive impact on systematicity, while vowel parameters (in grey) have a negative impact. The only exceptions are negative *c2* (the syllable-final consonant) and *c3* (the second-syllable initial consonant) in cvccv words. Other exceptions are sharing all vowels in cvccv and cvcvcv words, and, as mentioned earlier, the last vowel in cvcvcv words. Note that impact value of the final vowel is much lower in the 'no syntax' condition than in the 'syntax' condition in all three word groups. This may be explained by the fact that the final vowel carries in many instances morphosyntactic information: when correlated with syntax-laden semantic representations, the phonological representations are more influenced by the weight of the last 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 *semantic* similarity space too.

Sharing the same stressed *vowel* has very different effects depending on the syllable. The same stressed vowel on the final syllable makes words very phonologically similar. The stressed final vowel, as explained in § 3.2.2.5, encodes important verb morphosyntactic information. 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 semantic space both endorses the methodology and confirms the syntax-phonology space correlation proposed by the phonological typicality literature (Durieux & Gillis, 2000; Kelly, 1992, 1996; Monaghan, Chater & Christiansen, 2003). Enhancing final stressed vowel distinction at the phonological level greatly improves the phon-sem correlation, so this parameter must be driven by verb endings, with their highly systematic relationships between phonological and cooccurrence-based representations.

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

This suggests that while the identity of the final stressed vowel organises the systematic lexicon on a morphosyntactic dimension, the identity of the vowel in the stressed penultimate syllable may be crucial for word differentiation and recognition.

The linear impact values have given us an idea of the role each phonological similarity parameter plays on the phon-sem correlation. The next section explores which regression models best predict the behaviour of the parameters.

6.2.3.2 Regression curve estimation parameter models

The linear covariance values have provided us with a rough idea of how each parameter affects the systematicity between lexical phonology and semantics, particularly of whether their impact is positive or negative. This section looks at how different regression equations model the relationship between the phonological parameters and the phon-sem correlation.

I run the linear and ten nonlinear standard regression functions (logarithmic, inverse, quadratic, cubic, compound, power, sigmoid, growth, exponential and logistic) available in SPSS (2003) on the data. Appendix G shows the r^2 for all the functions in the cvcv word-group in the 'syntax' and the 'no syntax' conditions. The functions' fit for the different parameters is highly consistent across groups, satisfying Stevens' (1992) test to find the most reliable regression function. Figure 6.5 shows an illustration of the r^2 for one word-group (cvcv) in the 'no syntax' condition. An examination of the r^2 (measuring how well models fit the data) reveals some connections with the linear impact parameter values shown in Figures 6.3 and 6.4 above.

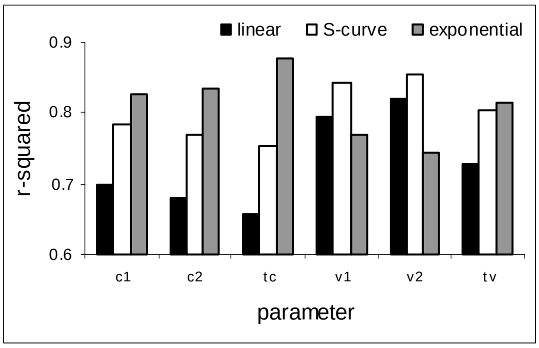


Figure 6.5: Measure of the fit of three regression models to the cvcv parameters in the 'no syntax' condition to the phon-sem systematicity.

- The linear function is never the best predictor of the phon-sem correlation given the parameter data, but it obtains its best r^2 values for negative-impact parameters.

- The growth, exponential, logistic and compound functions return very similar r^2 values. I take one representative from this group: the exponential function.
- 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.
- The worst single predictor of negative-impact parameters is the inverse function.
- 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.

The combined results from § 6.2.3.1 and § 6.2.3.2 supports the claim that there are two classes of parameters with respect to phon-sem systematicity. The second hypothesis stated in § 6.1 predicted that the study of the parameters of phonological similarity with respect to the phon-sem systematicity would reveal pressures in the lexicon different than those contributing to systematicity. Here we have two classes of parameters, one contributing to and the other working against systematicity. Figure 6.6 shows one illustrative example of each class.

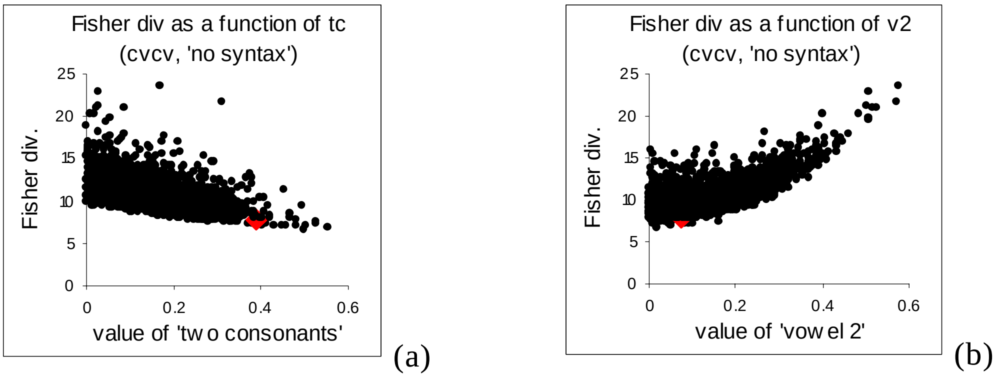


Figure 6.6. Scatter plot of Fisher divergence against individual parameter values. Two parameters are shown: (a) two consonants and (b) vowel 2 , in the cvcv word group, 'no

syntax' condition. The empirical parameter values are also shown as larger red points.

The two classes of phonological similarity parameters with respect to the organisation of the mental lexicon are:

1. Class one parameters: Individual and groups of consonants, the stressed syllable and the identity of the final stressed vowel have positive impact values on phon-sem systematicity and are best modelled by an exponential function $Y = b0 * (e^{(b1*t)})$. High systematicity (low Fisher divergence values) is brought about by high parameter values - see Figure 6.6 (a). This suggests that words sharing these phonological traits tend to be closer together in the cooccurrence-based semantic space.
2. Class two parameters. Individual vowels and the identity of the penultimate-syllable stressed vowel have negative impact values on phon-sem systematicity, possibly playing a role in word differentiation and identification. They are best modelled by a sigmoid curve function $Y = e^{(b0 + (\frac{b1}{t}))}$ and also reasonably well modelled by a linear function Y = b0 + b1t. High systematicity (low Fisher divergence values) is brought about by low parameter values (see e.g. Figure 6.6 (b)). This suggests that words sharing these phonological traits tend to be far apart in the cooccurrence-based semantic space.

6.2.3.3 The function of class one and class two parameters in Spanish

Class one parameters either are closely linked to narrow niches of syntactic function (such as the final stressed vowel encoding verb tense and person) or offer many combinatorial possibilities (such as the consonants in a word). These two characteristics are desirable in parameters that drive systematicity 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 (there are only five vowels in Spanish compared with 18 consonants) and may be related to word differentiation, the pressure opposed to systematicity 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* (see Figure 6.7). (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.)

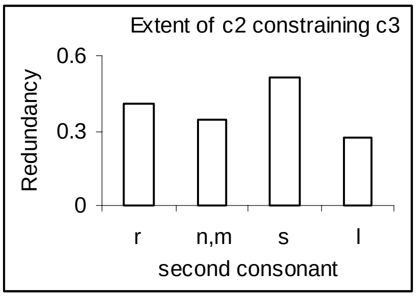


Figure 6.7. Redundancy measures the extent to which c3 can be correctly guessed once c2 is known, in cvccv words. Redundancy is 1 – relative entropy (see chapter two for definitions of entropy and redundancy).

In the discussion of the results of chapter three's study of parameters of phonological similarity I mentioned several studies suggesting the differential processing of vowels and consonants. I expand that review here linking it to systematicity. If consonants and vowels work for and against systematicity, respectively, this may indicate that certain neural mechanism(s) contribute to systematicity while other(s) work against it, perhaps contributing to word recognition.

Cole, Yan, Mak, Fanty and Bailey (1996) presented participants with English speech where either consonants of vowels had been rendered incomprehensible. They found that vowels are clearly more important for recognition than obstruent consonants in test sentences where both were equally represented. They studied extreme cases where either consonants or vowels were not available to the listener; in natural speech, however, they claim that there is a mutual interaction of consonants and vowels, and that we recognize a word thanks to its vowel structure given its consonant structure or vice versa.

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 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). In that study it was clear that the differences were brought about by the vowel-consonant distinction, and not by a distinction in the degree of voicing. Monaghan and Shillcock's (2003) connectionist model of Caramazza et al.'s effect showed that separable processing of vowels and consonants is an emergent effect of a divided processor operating on feature-based representations.

In another 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*). 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 preserve more of the sound of the original than the transposition of two vowels, and mention as supporting facts 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). These results suggest that, at least in languages like Spanish and Italian, vowels and consonants are processed separately and might contribute to the lexicon structure in different ways.

Chapter three mentioned the ‘phonological similarity effect’ (PSE) found by Conrad and Hull (1964) - 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 consonant-vowel-consonant nonwords with Norwegian participants, and analysed the impact on PSE of three parameters of phonological similarity - sharing middle vowels (*mal, sar, tas*), sharing the consonant frames (*kal, kol, kul*) and sharing 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 an absence or 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 recognised *better* than distinct word lists. What is most relevant to the present discussion is the fact that consonant frame lists (*kal, kol, kul*) were recalled and recognised 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 the systematic consonant-based dimension (in the *k_l* position). It is then easy to memorise which of the few possible vowel (Norwegian has 11 vowels) were present; the order can be expected to be easily recalled considering that most transposition speech errors involve consonant transpositions, and very few vowel transposition (e.g. the first 40 phonological substitution errors in Italian returned by the online Max Plank speech error database[2] comprise 27 consonant

[2] Max Plank speech error database online at http://www.mpi.nl/world/corpus/sedb/.

substitutions, 7 vowel substitutions and 6 other errors; in English, 24 consonant substitutions, 10 vowel substitutions and 6 other errors), indicating that the vowel order is more easily memorised.

Some papers support the hypothesis that my proposed 'class two' parameters may be 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 *shwas* than native English speakers when speaking English, probably because there is no reduction to schwa in Spanish.

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' 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. French speakers therefore do not pay attention to stress information when judging similarity.

In Spanish, among other languages, stress information cannot be predicted from segmental information. In these 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, Cutler and Sebastian-Gallés (1997) found that segmental judgements are more affected by stress in Spanish than in Dutch.

In this section I have reviewed evidence that class one parameters have links with syntactic function; that class one parameters have higher combinatorial power than class two parameters; that different neural mechanisms may underlie processing of consonants (class one) and vowels (class two); and finally, that class two parameters vowel identity and stress may be important for word recognition in Spanish.

These studies, together with the results presented in past sections, support the division of function between class one parameters (help maintain systematicity, which in turns helps generalisation and inference) and class two parameters (help word recognition in a systematic lexicon).

6.2.4 The empirical parameter values against the systematicity-driven impact parameter values

In this section I tackle the question of how well adapted the empirical parameters are to the pressure for systematicity. As stated above, I expect that the empirical parameters will prove to be well, but not perfectly, adapted to the systematicity pressure. The linear impact values for each parameter obtained in the last section are a measure of the parameter's influence on the systematicity between phonology and semantics. A match between this influence and the empirical parameter values grounds the existence of a link between perceived word phonological similarity and word semantic similarity.

I approach this problem by comparing the phon-sem correlation obtained with the empirical parameters, and the phon-sem correlations obtained with random parameter values. I do this in same the two conditions as in the previous section ('syntax' and ‘no syntax’) for the two word groups for which I obtained empirical values, cvcv and cvccv.

I insert the empirical parameter values along with the Fisher divergence obtained with them (e.g. the bold line in Table 6.2 above) among the 2,000 random parameter configurations and corresponding Fisher divergences. I calculate the rank of the empirical Fisher divergence in the 2,000 list. In this Monte-Carlo analysis the statistical significance of the empirical Fisher divergence can be calculated from its position in the ranking. Figure 6.8 shows those positions in the two conditions for the two word groups, together with the empirical Fisher divergence values and their significance.

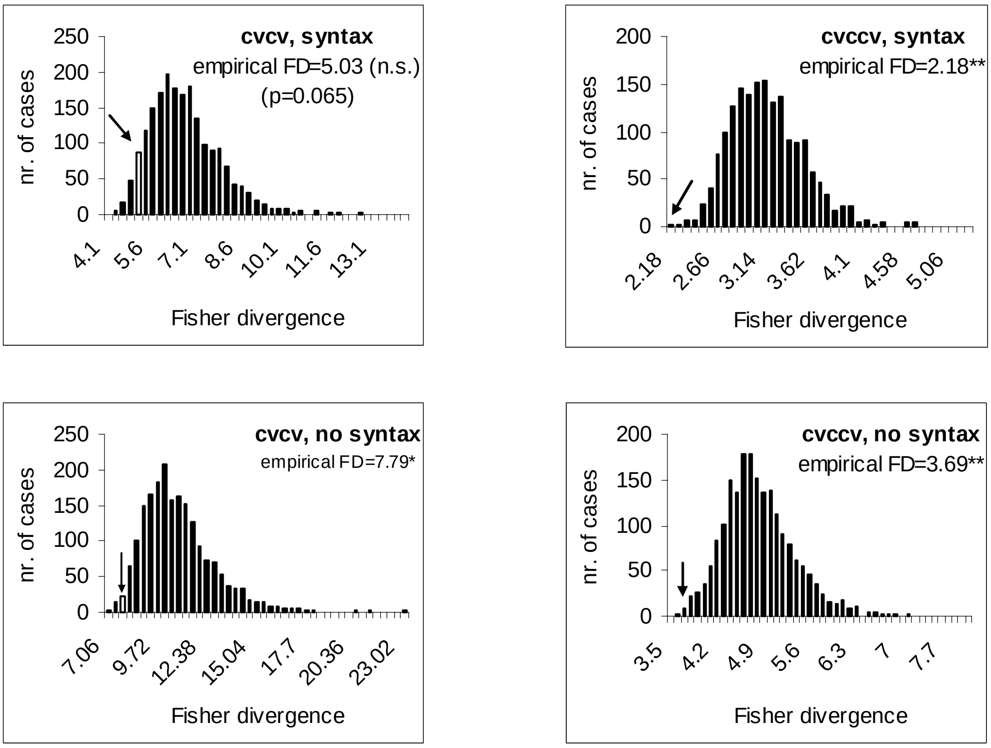


Figure 6.8. Position of the Fisher divergence obtained with the empirical phonological parameter values in word groups cvcv, cvccv in the 'syntax' and 'no syntax' conditions.

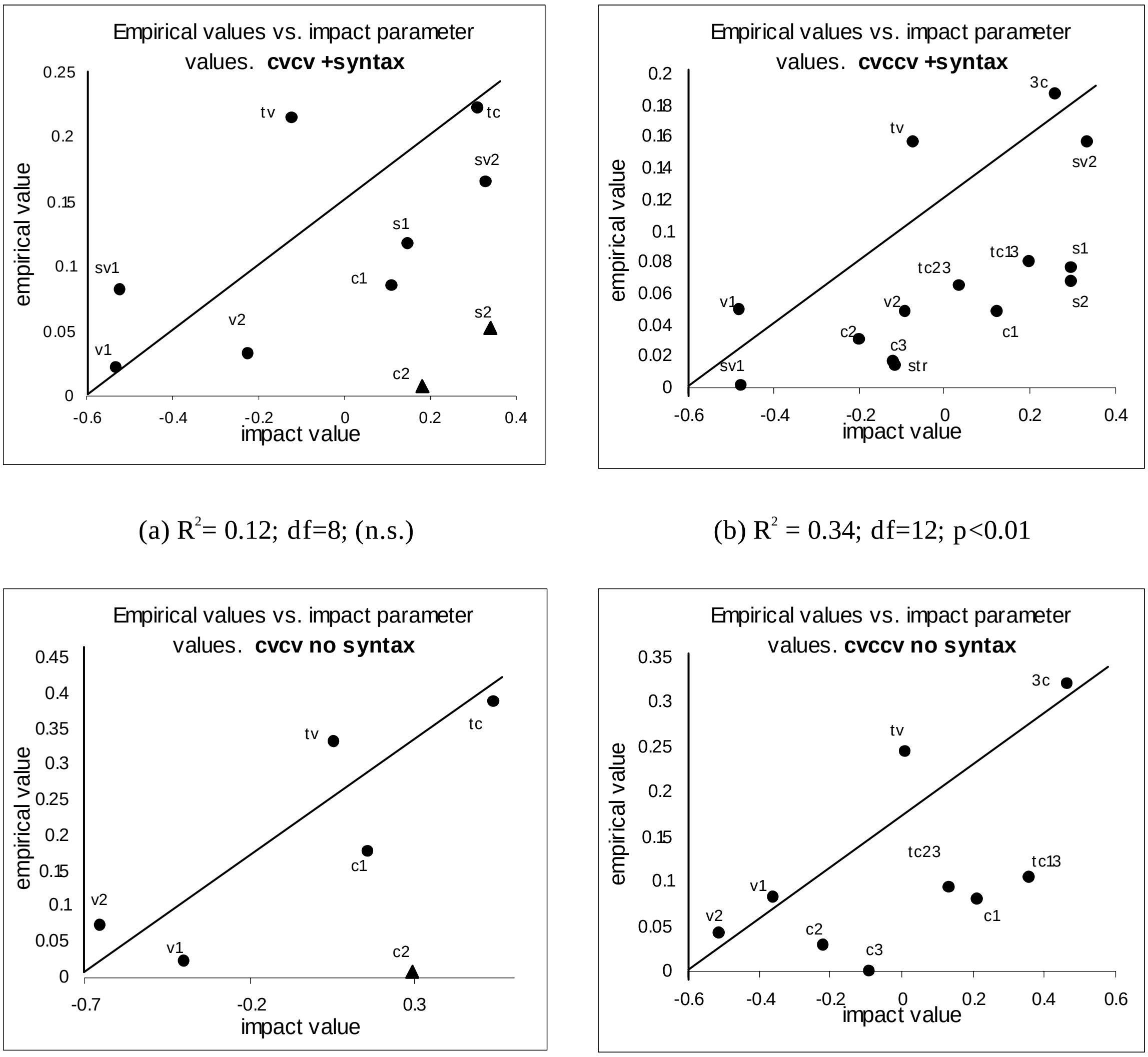
Empirical Fisher divergence value shown; significance: * p<0.05, **p<0.01

In three out of the four conditions the empirical parameter configuration yields significantly high Fisher divergences (measuring the phonology-semantic correlations). In the syntax condition for cvcv words, the correlation is marginally significant (p=0.065). This means that the empirical parameter configurations tend to yield a significantly high phon-sem systematicity; I explain this in terms of the pressure for systematicity across levels of representation in the lexicon. Additionally, the skewed-to-the-left graphs in Figure 6.8 illustrate the fact that randomly generated parameter configurations are more likely to yield a low systematicity lexicon. The empirical parameters yield a highly systematic lexicon, suggesting that the phonological lexicon is under a strong pressure for systematicity.

An illustration of this significance can be seen in Figure 6.6 (a) and (b) above. The large red dots in the Figures represent the position of the empirical parameter value and the Fisher divergence it helps obtain. In both cases, Fisher divergence is low, indicating high systematicity.

In order to determine the extent of the pressure for systematicity, I measure how unlikely it is that the empirical parameters would generate systematicity. Class one parameters - such as *two consonants*, illustrated in Figure 6.6 (a) – show the effect of a strong pressure for systematicity: the position of the parameter in the random space is highly significant (one-tailed Monte-Carlo, p<0.01). In the case of class two parameters - such as vowel two, illustrated in Figure 6.6 (b) – the empirical parameter value is not significant (one-tailed Monte-Carlo, p=0.21 n.s.). This means that, while both parameters contribute to systematicity, class one parameters are under strong pressure to do so, but class two parameters only support it because it is easy for them to do so - because of their low combinatorial power, many words will be similar to each other along those parameters.

Discrepancies between the empirical values and the systematicity-driven impact values can be explained at least in two ways. One, when asked to judge phonological similarity, people show evidence of pressures on the lexicon other than systematicity, for example, they may be focussing on aspects of the word that help differentiate between similar-sounding words. Alternatively, the parameter impact values obtained above may be overfitting to the data set. The second explanation can be ruled out if there is consistency across word groups. Figure 6.9 shows how the empirical values covary with the correlation-driven parameters.



(a) R^2= 0.12; df=8; (n.s.)

(b) R^2 = 0.34; df=12; $p<0.01$

(c) R^2= 0.31; df=4; (n.s.)

(d) R^2 = 0.34; df=8; $p<0.05$

Figure 6.9. Scatter plot of the parameter impact values obtained in § 6.2.3.1, against the empirical parameter values shown in Figure 6.2, based on the study in chapter three.

Parameters on the top-left half are more important for phonological similarity in the empirical task than in the systematicity-driven metric. Parameters on the bottom-right half are more important for phonological similarity in the systematicity driven metric than in the empirical task. Correlation between the two axes (R^2) is also shown; the triangles correspond to the parameters that stop $R^2$from being significant.

Most of the correlation-driven 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 the correlation is more driven by consonants, people focus more on vowels when asked to tell how similar words sound.

Additionally, in the cvcv group, parameters *c2* and *s2* do not correlate well, with corpus-based values much higher than empirical values. The two counterpart parameters in cvccv - *c3* (last syllable-onset consonant) and *s2* - are among the worst correlating parameters, supporting the fact that the misalignment is not due to overfitting. The low correlation of these two parameters may be due to the onset consonant of the final syllable being important for syntax-semantic word categorization, but somehow people not consciously noticing it when directed to compare the way words *sound.*

These two facts seem to indicate a divergence in the two methodologies employed in the test of systematicity between the phonological and the semantic levels of the lexicon. Across-group consistency in the correlation-driven parameter values indicates that the parameters behave in the same way in all word groups, supporting the robustness of the methodology. Across-group consistency in the empirical parameters (§ 3.2.2.4) supports the robustness of the empirical study methodology.

The fact that the empirical values differ from the correlation-driven ones may be due to the nature of the psycholinguistic task presented in chapter three. Participants were asked which of two pseudo-words *sounded* more similar to a third pseudo-word. This may have biased the choices. Perhaps if they had been asked which of the two pseudo-words they thought *meant something more similar* to the third pseudo-word, they would have drifted away from the concrete word form, and allowed access to a more holistic word representation. Such a task would elicit judgements of systematicity more effectively, and it might have produced parameter values more similar to the systematicity-driven impact values. This poses an empirical question that can only be resolved with further tests.

The analysis of the results of the random search of the phonological parameter space with respect to systematicity, together with the empirically obtained parameter values returns the conclusions summarised in Table 6.4.

	Class one parameters	**Class two parameters**
parameters	consonants; stress; stressed final vowel	vowels, stressed penultimate-syllable vowel
impact on systematicity	positive	negative
best regression model	exponential	sigmoidal
systematicity-driven vs. empirical values	systematicity > empirical	empirical > systematicity
function	maintain systematicity	word identification

Table 6.4. Two classes of phonological similarity parameters with respect to phon-sem systematicity.

The random search of the phonological parameter space suggests that there are two classes of parameters with respect to phon-sem systematicity - consonant parameters, stress placement and the stressed final vowel all contribute to systematicity, while the identity of the stressed vowel in the penultimate syllable works against it. These two classes are also best modelled by different regression functions, exponential for class one and sigmoidal curve for class two. I suggest that the class two parameters are crucial in the differentiation of words with similar phonology that tend to occur in similar contexts. Contrasting the empirical parameter values derived from chapter three with the random-search values obtained above supports this different-function hypothesis. While the parameter values obtained with the empirical study and the random-search are generally well correlated, the systematicity-driven random search accorded higher values to class one parameters, and lower values to class two parameters than people did.

6.2.5 Random search of an extended phonological parameter space

In this section I apply the methodology in § 6.2.1 to search a more fine-grained phonological similarity parameter space adapted to words of any structure and length. This space takes into account word syllable structure and includes parameters related to the identity and the features of the segments in different syllable positions; to stress; to the rhyme and the word-onset (the first few segments); to one word being contained in the other; and finally to word length, which studies in English (Shillcock et al., 2001, *submitted*) suggest have a strong effect on phonological similarity.

I suggested at the end of chapter five that word length might play a crucial role in the phon-sem correlation in Spanish, but word length effects could not be tested with the equal-length sub-sets of the corpus employed (cvcv and cvccv words) . In order to test this, and at the same time possibly reveal other important parameters of phonological similarity, I explore systematicity in an extended sample of the lexicon: the 516 highest-frequency words from the same corpus of Spanish transcribed speech used throughout this thesis. Unlike the length- and consonant-vowel structure-homogeneous cvcv, cvccv and cvcvcv groups, this word set includes short and long words of various CV structures.

The methodology in § 6.2.2.2 is applied to the phonological and semantic spaces generated with the 516-word group. The semantic similarity between the 132,870 word-pairs is calculated in the same way as in chapters four and five, with the same context words for the 'syntax' and ‘no syntax’ conditions. The phonological similarity metric required a new parameter set adapted to the varying word length and structure, and is explained in detail in the next section.

6.2.5.1 The phonological similarity metric

This exploration of a phonologically heterogeneous lexicon includes parameters related to the whole word, such as difference in word length, one word being contained in the other, or sharing initial segment features; other parameters are related to the syllables, such as sharing the syllable onset and coda or having the same word CV structure; yet others relate to stress, rhyme and vowel features. Such a general approach applied to a varied lexicon sample yields a general guide as to what parameters affect the phon-sem systematicity in the whole lexicon.

	Segment
0	onset consonant
1	cluster consonant
2	vowel
3	glide
4	coda consonant

Table 6.5. Organisation of the syllable segments.

For the calculation of the phonological similarity of each word-pair, the two words were divided up into syllables and then the segments in each syllable were placed into a fixed template (see Table 6.5) so that each element could be compared with its counterparts in other syllables.

The two words to be compared were aligned along the stressed syllable[3] (see Figure 6.10) and parameters related to syllable-position were compared between aligned syllables and their adjacent syllables (e.g. in Figure 6.10, syllable *mo* in the top-centre word is compared with all three syllables in the bottom-centre word: the one directly below it, and those to its right and left; syllable *ni* in the top centre word is compared with *ri* and *ta* in the bottom-centre word).

[3] This is motivated by the fact that in Spanish poetry metrics the stress of the last syllable in a line alters the count of the number of syllables. Spanish rhyme needs the end of the lines to be stress-aligned.

di	rek	**ti**	ba	ar	**mo**	ni	ka	ba	**ul**	
		Ro	ka	ga	**ri**	ta			**me**	sa

Figure 6.10. The two words to be compared are aligned along their stressed syllable (stressed syllables in bold).

The phonological similarity metric compares pairs of words and computes a similarity value based on the following parameters:

- same manner of articulation, place of articulation, sonority, voicing (6 categories, following Burquest & Payne, 1993) in the syllable-onset, syllable-coda, and word initial consonants;
- same phoneme identity in the syllable-onset, syllable-coda and word initial consonants;
- same vowel openness (open, medium, closed) and position (anterior, central, posterior);
- presence of a glide (semivowel or semiconsonant);
- presence of the same cluster consonant, i.e. the second consonant in a 2-consonant cluster;
- one word being contained in the other in terms of syllables - e.g. *ce-ga-to* (blind) and *ga-to* (cat), but not *glo-bo* (balloon) and *lo-bo* (wolf), because in the second pair there is a discrepant syllable boundary.
- one word being contained in the other, but not in terms of syllables - e.g. *glo-bo* (balloon) and *lo-bo* (wolf);
- same CV word structure;
- same vowel structure (sharing all the vowels in the same order);
- sharing the stress on the same syllable (last, penultimate, antepenultimate);
- similar word onset (number of common segments in the first syllable);

- similar rhyme (number of common segments in the last syllable);
- same final-syllable vowel;
- different length (this parameter penalises word length discrepancies between the two words in the pair, measured in segments).

6.2.5.2 The linear parameter impact values

Figure 6.11 shows the linear parameter impact values calculated with 135 random parameter configurations, following the same method described in § 6.2.2 and also employed in § 6.2.3.1.

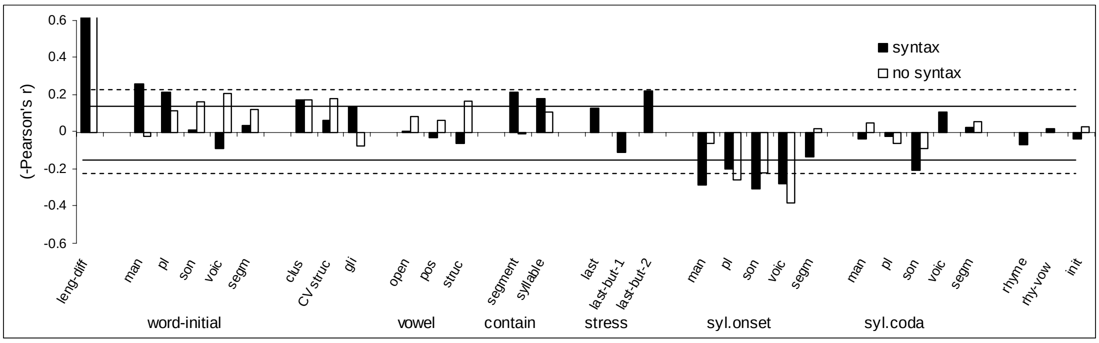


Figure 6.11. ‘Parameter impact’ values. Two conditions, ‘syntax’ and ‘no syntax’, are shown for the ‘all word’ group. (Note that length difference is a penalisation parameter.) Continuous line indicates p=0.05; discontinuous line indicates p=0.01.

Table 6.6 shows the significant parameter impact values that contribute to (class one) and work against (class two) the phon-sem systematicity.

	Class one parameters	Class two parameters
'syntax'	word length; word initial manner and place of articulation; stress in the antepenultimate syllable; one word containing the other; cluster consonant	syllable onset sonority, manner of articulation, voicing, and place of articulation; syllable-coda sonority
'no syntax'	word length; word initial voicing and sonority; cluster consonant, word CV structure; word vowel structure	syllable-onset voicing and place of articulation

Table 6.6. Parameter impact values that reach significance in the 'all word' group.

The highest positive impact value corresponds to word length difference, with an impact value of 0.62 in both conditions. This parameter works as a penalisation in the calculation of word similarity. In the phonological similarity metric, the 'length-difference' parameter value (multiplied by the length in segments in the longest word minus the length in segments in the shortest word in the pair) is subtracted from the pair's phonological similarity. In other words, the higher the parameter value, the more length difference is *penalised* in terms of word similarity. This means that, as suggested in chapter five, length difference plays a crucial role on the phonological similarity side of the phon-sem systematicity, and may be one of the reasons why the phon-sem correlation obtained in chapter five (§ 5.2.3) did not reach statistical significance.

Word initial consonant features contribute to systematicity, with the 'syntax' condition placing more emphasis on manner and place of articulation, and the 'no syntax' in voicing and sonority. The initial consonant is, as reviewed in chapter three, crucial in lexical representation; its positive impact on the phon-sem correlation suggests this parameter contributes to systematicity. The cluster consonant also seems to contribute to systematicity in this group of words of any CV structure (but not so in cvccv words, where we saw that the cluster consonant *c3* worked against systematicity). The measurement in the two word groups is different in that in cvccv words the cluster consonant was always in the same position in the word, whereas here, comparisons across different syllables are also taken into account. Further studies would be necessary to determine the role of the cluster consonant, for instance, a more detailed phonological similarity metric testing the effects on systematicity of parameters 'same consonant cluster in the aligned syllable' against 'same cluster consonant in a different syllable'. Sharing the stress in the last and antepenultimate syllables also contributes to systematicity, consistent with the cvcv, cvccv and cvcvcv studies. Containment of one word by the other also works in favour of systematicity. In their paper about the possible word constraint in word segmentation, Norris, McQueen, Cutler and Butterfield's (1997) showed that, in English, it is easier to detect e.g. the word *apple* when embedded in *vuffapple* (where *vuff* could be an English word) than when embedded in *fapple* (where *f* could not be a word in English). In Spanish, the first condition is similar to our parameter 'one word contained by the other in terms of syllables', and the second condition can be assimilated to our parameter 'one word contained by the other in terms of segments, but not syllables' e.g. *glo-bo* and *lo-bo*, where *g* (or any sequence containing part of a syllable) cannot be a word. Norris et al. used a word spotting task, and in the present case, the two words are being compared in terms of their similarity, but the present results suggest that, in Spanish, containment is recognized and contributes to the phonological lexicon organization in both conditions.

Class two parameters, working against systematicity, and possibly helping word recognition, relate to syllable onset and syllable coda features. These parameters affect all syllables in the word, not only in the initial syllable. Sharing the stress on the penultimate syllable also negatively impacts systematicity. Because the phonological similarity metric compared not only aligned syllables in the two words, but also each syllable with the one before and after (Figure 6.10), I cannot make assumptions as to the relevance of these parameters in different word positions. The results suggest that syllable onset and coda consonant features, independent of position in the word, help to distinguish otherwise similar sounding words which tend to appear in similar contexts.

These results generally agree with those from the cvcv, cvccv and cvcvcv word study. Single vowel impact values do not reach statistical significance, but the word vowel structure (sharing all the vowels in the same order) does. This parameter is comparable to 'two vowels' in cvcv and cvccv words, and 'three vowels' in cvcvcv words, and, like them, has a positive impact on systematicity. Stress-related parameter values reflect those obtained in the cvcv, cvccv and cvcvcv word groups, namely positive impact value for the last and last-but-two syllables, negative for the last-but-one, although only the impact of shared stress in the last-but-two syllable reaches significance.

One discrepancy is the mentioned difference between the behaviour of the cluster consonant in cvccv words against the present group of words of any CV structure.

Among the parameter impact values not reaching significance is sharing the full segments in any position - a common theme in the results is the fact that consonant features' impact on systematicity is stronger than that of the corresponding full segment. The presence of glides and vowel features, most syllable-coda consonant features (except sonority in the 'syntax' condition), rhyme and sharing initial syllable segments are not significant either.

This random search included many interdependent phonological variables, and this may have affected the results. Additionally, only 135 points of the space were calculated (against 2,000 for the cvcv, cvccv and cvcvcv spaces). This means that these results are only preliminary, but, together with those from the larger random searches, they indicate that this methodology has potential to reveal interesting aspects of the structure of the phonological lexicon.

6.2.6 Summary of section 6.2

The random-search methodology has quantified the impact of individual phonological parameters on the systematicity between phonological and context-based similarity in three subsets of the lexicon (cvcv, cvccv and cvcvcv words). Two classes of parameters were apparent:

Class one parameters are best modelled by an exponential function; these parameters either allow many phonological combinations or are linked to morphosyntax. These parameters contribute to systematicity: words that tend to occur in similar contexts in speech also tend to share consonants and the final stressed vowel.

Class two parameters are best modelled by a sigmoid function and also by a linear function. These parameters have a low combinatorial power. They work against systematicity: words that tend to occur in different contexts in speech tend to share vowels and the stressed syllable.

This suggests that while class one phonological parameters contribute to systematicity in the lexicon, class two parameters might be helping the identification and recognition of otherwise similar sounding words which tend to occur in similar contexts.

The empirical parameter values obtained from the psycholinguistic study reported in chapter three correlated well with the random search parameter values, suggesting that some parameters (class one) are under pressure to promote systematicity in the lexicon while others (class two) oppose systematicity to help word identification.

An exploration of an extended parameter space showed the strongly positive impact of word length similarity on systematicity.

The high consistency of the parameter values across independent lexicon subsets supports that the phonological organisation of the lexicon is the consequence of the interaction of the pressure for systematicity and the opposed pressure for word intelligibility.

The next section looks for the parameter configurations that yield the best possible phon-sem correlations, and again compares them with the empirical parameter values to extract conclusions relevant to the lexicon's phonological structure.

6.3 A hill-climbing search of the phonological lexicon

In § 6.2 I generated random phonological parameter configurations, calculated a phonological similarity space with them and measured how well the phonological space correlated with an independently-measured semantic space. All these randomisations could be visualised as a surface (Figure 6.1) with peaks of low correlation and valleys of high correlation. In this case I use a parameter optimisation technique called hill-climbing[4] that goes directly to the areas of high correlation. It is an algorithm designed to find the phonological parameter configuration that obtain the optimal phon-sem correlation.

Comparing the 'optimal' parameter values that obtain the most systematic lexicon with the empirical parameter values can help determine what other constraints are acting on the mental lexicon and their effect on systematicity. For example, if the best phon-sem correlation is obtained by placing a lot of emphasis on consonants in the processing of phonological similarity, why did people actually place more emphasis on vowels in the study reported in chapter three?

6.3.1 Method

In order to attempt to find the parameter configuration that returns the best phon-sem correlation (the lowest point in the surface in Figure 6.1) I draw a method from the field of Artificial Intelligence called hill-climbing search. The general principle behind it is that a random parameter configuration is evaluated according to some metric; then one random change is made to one of the parameters, and the changed parameter configuration is evaluated again. If the evaluation is better, the random change is kept, otherwise it is discarded. This process is repeated until a stable state is reached, signalled by the fact that no change in any of the parameters improves the evaluation criterion.

Figure 6.12 shows a graphic representation of a hill-climbing search in the two-parameter space already shown in Figure 6.1. The following explanation

[4] The standard name ‘hill-climbing search’ seems to indicate that we are looking for the highest point in a search space. In the present case, however, the best result is the lowest point (low Fisher divergence means high phon-sem correlation). I keep the method name, hill-climbing, but ask the reader to bear in mind that in this study we are actually talking about ‘valley descending’.

of my implementation of the hill-climbing search will use this Figure as an illustrative example.

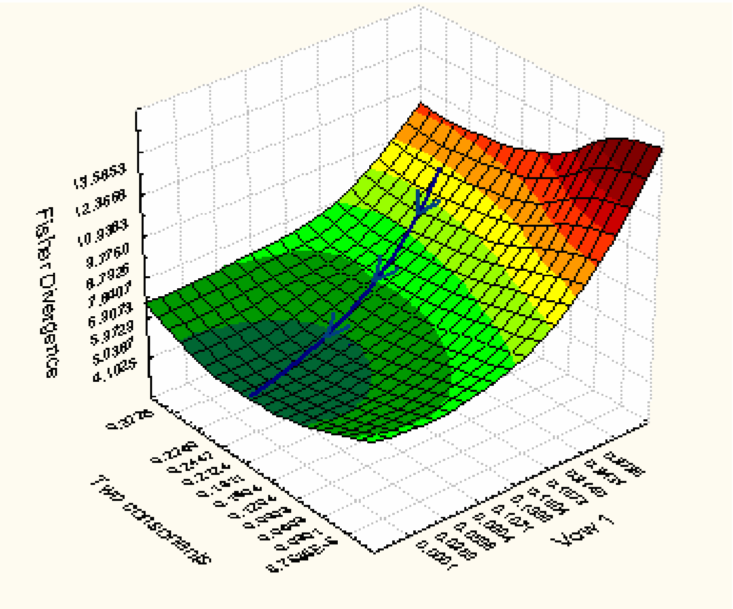


Figure 6.12. The blue line represents the path of the hill-climbing search, from the yellow area towards the green area. Note again that the term ‘valley-descending’ would be more appropriate in this particular case.

In the present case, the evaluation metric for the informed search is the phon-sem correlation measured with Fisher divergence, and the algorithm works like this:

1. I start off with a randomly generated phonological parameter set (the top end of the blue line, defined by the parameter values *tc*=0.28 and *v1*=0.26). Since Fisher divergence is sensitive to the magnitude of the data, this random parameter set is converted into a probability distribution (as I did in the random search).

2. Using these parameter values, I calculate the distances between all the word-pairs in the group to produce a phonological similarity space.

3. I calculate the correlation (using Fisher divergence, or FD) between this phonological similarity space and the cooccurrence-based semantic similarity space (the same used in section 6.1.2, and in chapter five).

4. I change one of the parameters: I randomly add or subtract a small amount to one of the parameters (0.05 for 200 iterations, 0.02 for a further 200 iterations and 0.01 for the rest of the iterations).
5. I repeat steps 2 and 3 with the new parameter set. If the new FD is higher than the old one, I discard the change and make another random change. If the new FD is lower than the old one, I keep the change and go to step 4.

These steps are repeated until no random changes return a better FD, that is, until the blue line in Figure 6.12 reaches the lowest point of the valley. In practice, I repeated the algorithm until no change is detected in FD for 50 iterations.

An informed search like this does not tell us about the overall shape of the surface; it only shows downhill paths. One problem of this kind of algorithm is that the path could end up in a local minimum, a point which is lower than its surrounding area, but it is not the overall lowest point in the surface. The ever-descending path cannot escape from local minima, and going into one prevents us from finding the lowest valley representing the best parameter configuration. This potential problem can be ameliorated by doing several runs of the algorithm with different initial random configurations, the equivalent of starting in different points in the surface in Figure 6.12. I ran the search twice for each space and arrived at practically the same parameter configurations. The following results show one of them.

I applied this algorithm to the cvcv, cvccv and cvcvcv words in both the 'syntax' and the 'no syntax' conditions. The results are shown and discussed in the next section.

6.3.2 Results

Table 6.7 shows the FD's obtained with the phonological parameter configurations for cvcv, cvccv and cvcvcv words with the methodology explained above, in the 'syntax' and the 'no syntax' conditions (the FD's obtained with the empirical parameter values are shown for comparison).

	syntax		no syntax	
	FD (hill-climb.)	FD (empiric.)	FD (hill-climb.)	FD (empiric.)
cvcv	3.34	5.03	5.76	7.79
cvccv	1.64	2.18	2.80	3.69
cvcvcv	1.94	n.a.	3.21	n.a.

Table 6.7. Fisher divergence correlation values obtained with the empirical and the hill-climbing parameter values in the 'syntax' and the 'no syntax' conditions.

These FD's are well below those found in the random space searches of past sections, because the hill-climbing algorithm actively looks for the best possible parameter configuration, and refines it to obtain such optimal values that would be very unlikely to occur by chance.

To illustrate this improbability, we can place the results from the informed search in context, in the systematicity spaces from § 6.2.4.

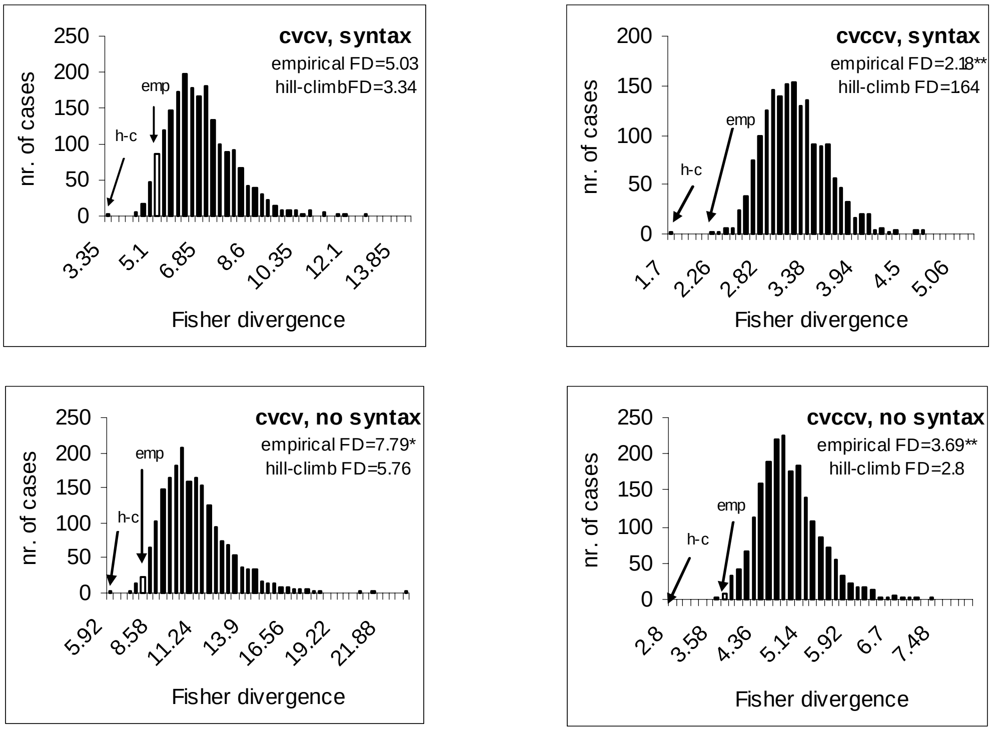


Figure 6.13. Distribution of FD values obtained with 2000 random parameter sets, with the empirical parameters (emp, in the white bin; *p<0.05, **p<0.01) and with the parameter values from the hill-climbing algorithm (h-c).

Figure 6.13 shows the distribution of Fisher divergence values obtained with 2000 random parameter sets, and the position of the empirical values (already shown in Figure 6.8) and of the parameter values resulting from the hill-climbing search. It is clear that the hill-climbing search obtains a very good phon-sem correlation, far better than any of the 2000 obtained with random parameter configurations.

The parameter value configurations that obtained the lowest Fisher divergence for cvcv and cvccv words are shown in Figure 6.14 and for cvcvcv in Figure 6.15.

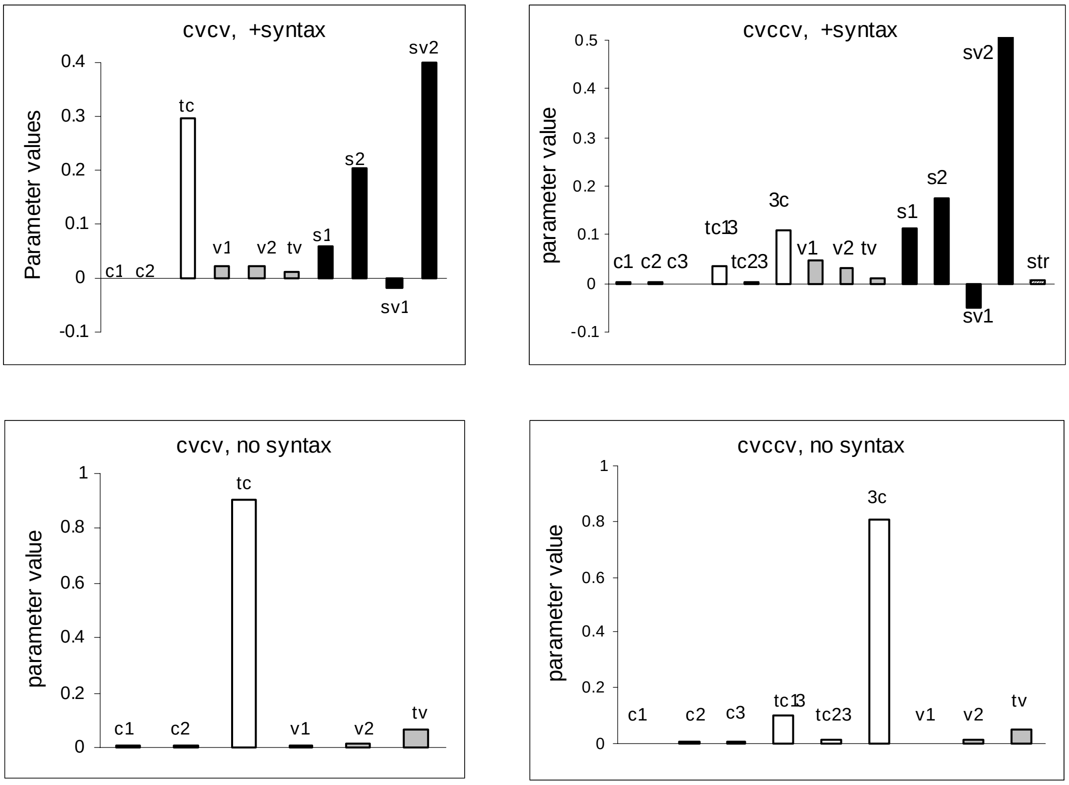


Figure 6.14: Parameter values obtained with a directed search of the phonological parameter space. These are the parameter configurations that obtained the best phon-sem correlation values after 650 iterations of the search algorithm. Two conditions, syntax and no syntax are shown for the cvcv and cvccv word groups. White bars= consonant-related parameters; grey bars=vowel-related parameters; black bars=stress-related parameters; striped bar=structure-related parameter.

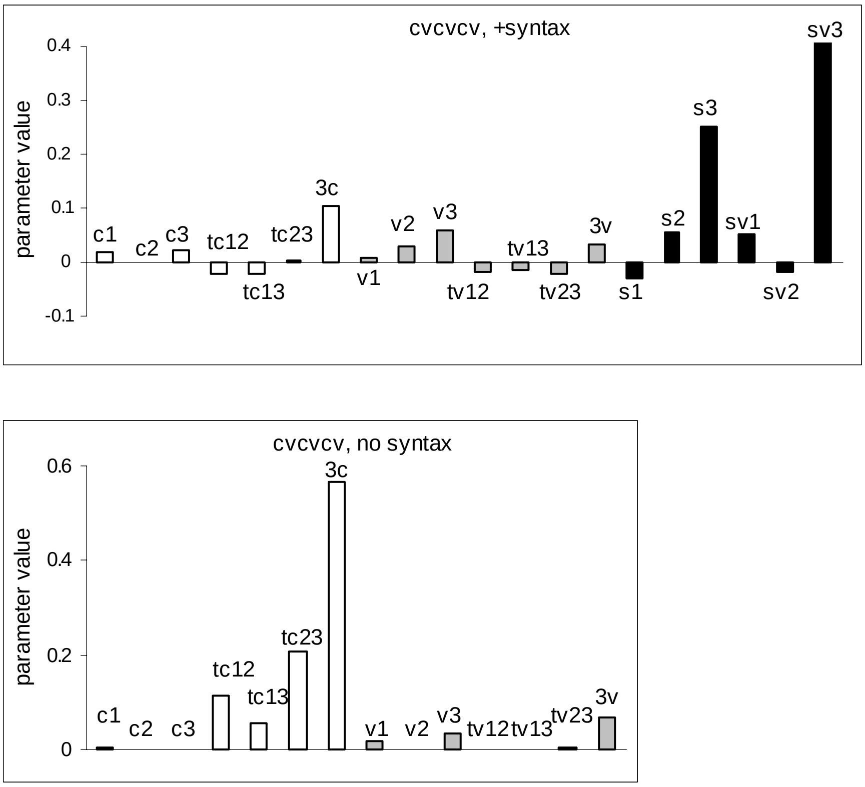


Figure 6.15 The same as Figure 6.14, but for cvcvcv words.

The configurations returned by the hill-climbing search and shown in Figures 6.14 and 6.15 consistently rely heavily on the same small number of parameters - a combination of stressed vowel on the last syllable, stress on the last syllable and all consonants in the 'syntax' condition, and simply all consonants in the 'no syntax' condition.

The hill-climbing methodology strongly relies on a parameter (sharing all consonants) affecting a minute proportion of the word pairs constituting our sample lexicons (0.007% of cvcv words, 0.006% of cvccv words and 0.001% of cvcvcv words). The phonological parameters analyzed form a surface like that depicted in Figure 6.1, but multidimensional. This hypersurface may be such that one parameter has a much steeper gradient than the rest, so that the search (blue line in Figure 6.12) goes down in the direction of that parameter so fast that the effects of the gradients of other parameters are obscured. This may be the case with the parameter 'sharing all consonants'.

I ran the search twice for each word group and also examined the across-group consistency to double-check that the search did not end up in a local minimum. Additionally, as seen in Table 6.8, the consistency across word-groups is remarkably high.

	'syntax'	
	cvcv (10)	cvccv (14)
cvccv (14)	0.78	
cvcvcv (20)	0.87	0.94

	‘no syntax’	
	cvcv (6)	cvccv (10)
cvccv (10)	0.99	
cvcvcv (14)	0.99	0.95

Table 6.8. Consistency across word-groups in the 'syntax' and ‘no syntax’ conditions: R^2 of counterpart parameter values. The number of parameters is shown in brackets. All p<0.01.

These R^2 values show a high degree of convergence between the three parameter configurations, indicating that the hill-climbing algorithm finds similar phonological organisation with respect to systematicity in three independent subsets of the lexicon. This convergence supports the reliability of the methodology. Let us now examine how the parameter values obtained with the hill-climbing method correlate with the parameter impact values and with the empirical values (Table 6.9).

R^2		Impact	Empirical
'syntax'	cvcv	0.43 (df=8)*	0.25 (df=8)
	cvccv	0.35 (df=12)*	0.32 (df=12)*
	cvcvcv	0.15 (df=18)*	n.a.
'no syntax'	cvcv	0.36 (df=4)	0.50 (df=4)*
	cvccv	0.35 (df=8)*	0.61 (df=8)**
	cvcvcv	0.30 (df=12)*	n.a.

Table 6.9. Correlations (R^2) of the parameter values obtained with the hill-climbing search with the parameter linear *Impact* values, and with the *Empirical* values. (* p<0.05; **p<0.01.)

The parameter impact values resulting from the random search and the hill-climbing values were based on the systematicity between the phonological and the semantic levels of the lexicon. The empirical parameters were obtained from psycholinguistic tests on word-form data alone. Table 6.9 shows that eight out of ten correlations are statistically significant, and the two that are not have very few degrees of freedom (df=4). In particular, the empirical values correlate best with the values returning the best phon-sem correlation in the 'no syntax' condition, where word representations rely most on word meaning.

A scatter plot of the hill-climbing parameter values against the empirical parameter values will reveal more about the correlations, as well as the discrepancies between the two metrics in a similar way as the scatter plots between the linear parameter impact values (Figure 6.3) and the empirical values (Figure 6.2). An analysis of the behaviour of the empirical values against the correlation-driven values might help reveal pressures other than systematicity affecting the structure of the lexicon.

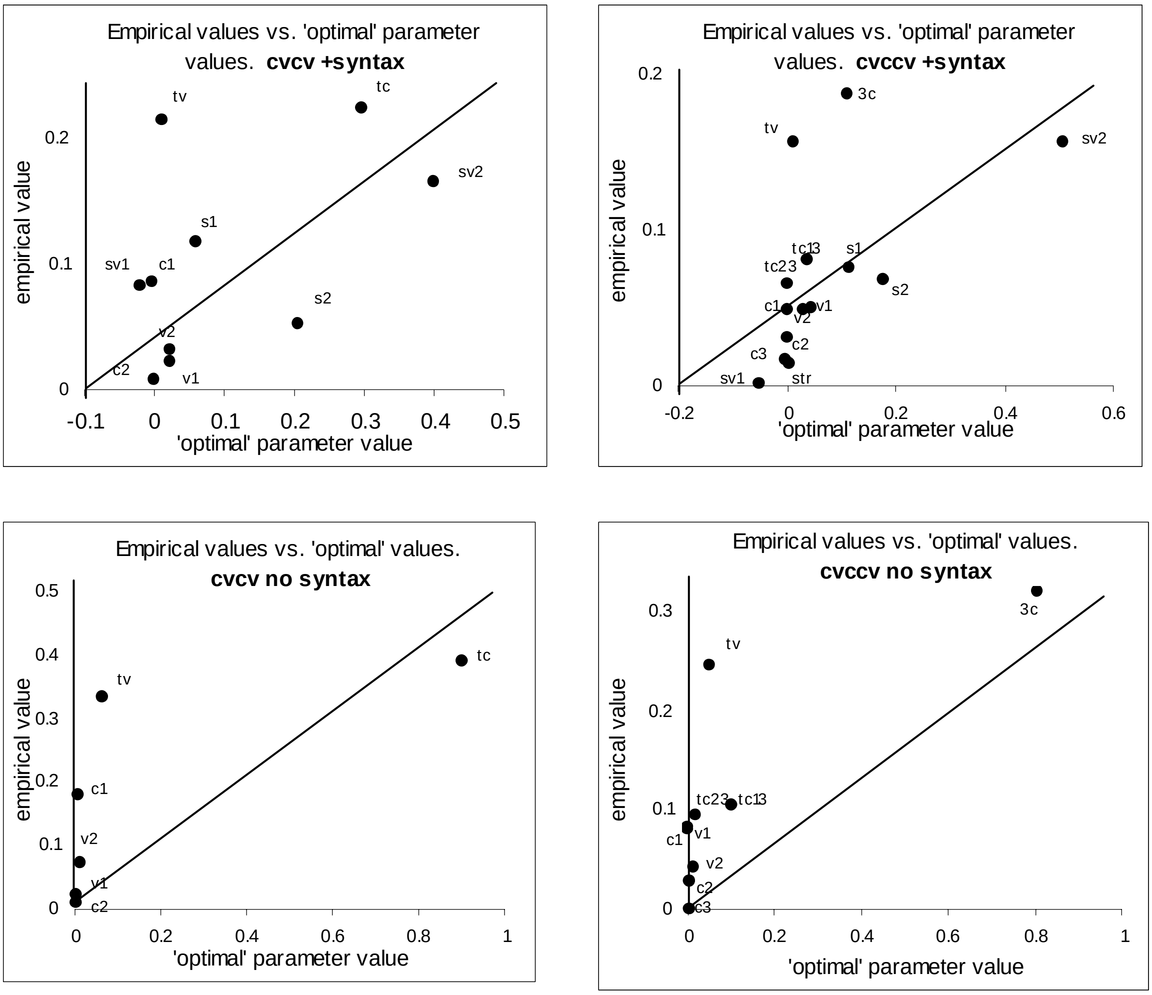


Figure 6.16. Scatter plot of the 'optimal' values against the parameter values obtained empirically in chapter three.

Figure 6.16 and Figure 6.14 above show that the phon-sem systematicity is driven to a very high extent by similar last-syllable stressed vowels and similar word consonant structures. The empirical measurements presented in chapter three suggest that people also rely mainly on the stressed final vowel for phonological similarity. However, it also suggests that people attach almost as much importance to vowels as to consonants, which is not reflected in the hill-climbing results.

This further supports the distinction between class one and class two parameters proposed above, the former (mainly consonants and the stressed final vowel) contributing to phon-sem systematicity, and the latter (vowels and the stressed syllable), focused on by people, working against systematicity. As suggested above, class two parameters could be helping to distinguish individual words from others that may be used in similar contexts.

Figures 6.14 and 6.15 show that the hill-climbing phonological space relies strongly on a few parameters, almost dismissing the others. The selected parameters, not surprisingly, are related to syntax (stress) in the 'syntax' condition. In the 'no syntax' condition, systematicity is driven in all word groups mainly by 'sharing all consonants'. In the surface-form words in our data-sets, only half or less of the word pairs that share all consonants also share the same root (45% of cvcv pairs, 37% of cvccv pairs and 50% of cvcvcv pairs) (e.g. forms of the same verb, masculine and feminine forms of the same noun or adjective). In a lemmatised corpus, words sharing the same root would be conflated into the same lemma, and there would be no pairs of words sharing the same root. One problem with lemmatisation is that the vowel structure would also be altered, affecting the phonological representations of words.

The fact that the systematicity-driven metric relies heavily on sharing all consonants might simply be telling us that words sharing the same root have similar cooccurrence patterns in speech. But over half of the pairs sharing the three consonants have different roots, indicating that sharing the consonant structure is the main contributor to systematicity between phonology and semantics in the lexicon.

As suggested above, the reliance on few parameters can be an artefact of the hill-climbing algorithm, favouring the parameter with the steepest gradient. Removing the ‘winning’ parameter should reveal the next best parameter. Indeed, this seems to have happened in the two conditions. In the ‘syntax’ condition, sv2 is the clear winner, obscuring the contribution of other parameters towards the correlation. In the ‘no syntax’ condition, when stress is removed, the important role of the consonant structure is revealed.

6.3.3 Summary of section 6.3.

The hill-climbing search optimised the parameter configuration to obtain the best phon-sem systematicity in the lexicon. In the 'syntax' condition, systematicity is driven by stress parameters. In the 'no syntax' condition, by sharing all consonants.

The random and the hill-climbing searches of the phonological parameter space have returned different quantitative information about the parameters. The next section integrates all this information to make a testable prediction about the empirical values for cvcvcv words.

6.4 Predicting empirical parameter values for CVCVCV words

We now know how the phonological parameters behave with respect to the phon-sem correlation (impact values, § 6.2.3.1) and which parameter configurations obtain the optimal phon-sem correlation (hill-climbing values, § 6.3.2), for cvcv, cvccv and cvcvcv words. I also have empirical parameter values for cvcv and cvccv words. I can now integrate all this information and use it to predict the empirical values of some of the phonological parameters for cvcvcv words.

Table 6.10 shows the parameters that can be considered as counterparts in the three word-groups. (Note that syllables are counted from the end of the word.)

cvcv	**cvccv**	**cvcvcv**	
c1	c1	c1	word-initial consonant
c2	c2	c3	2nd syllable-initial consonant
tc	3c	3c	all consonants in the word
v1	v1	v2	last-but-one syllable vowel
v2	v2	v3	last syllable vowel
tv	tv	3v	all vowels in the word
s1	s1	s2	stress on last-but-one syllable
s2	s2	s3	stress on last syllable
sv1	sv1	sv2	stressed vowel on last-but-one syllable
sv2	sv2	sv3	stressed vowel on last syllable

Table 6.10. Counterpart parameters for the cvcv, cvccv and cvcvcv word-groups.

For each parameter, I have eight values. For instance, for the word-initial consonant *c1*, one value for each measure in each word group (Table 6.11).

c1	**cvcv**	**cvccv**	**cvcvcv**
hill-climbing	0.0080	-0.002	0.0027
parameter impact	0.1593	0.212	0.1228
empirical value	0.1776	0.098	x

Table 6.11. Known values of phonological value *c1* obtained with different methodologies.

I can combine these known values to predict the empirical values for cvcvcv parameters. This can be done 'manually' by looking at the values for the same parameter in different word groups and across methods, and extrapolate a value for *x*. The advantage of a manual method is that I can take into account factors such as the fact that cvcvcv words' syllabic structure is more similar to that of cvcv words. This method returns the predicted 'empirical' parameter values for cvcvcv words shown in Figure 6.17.

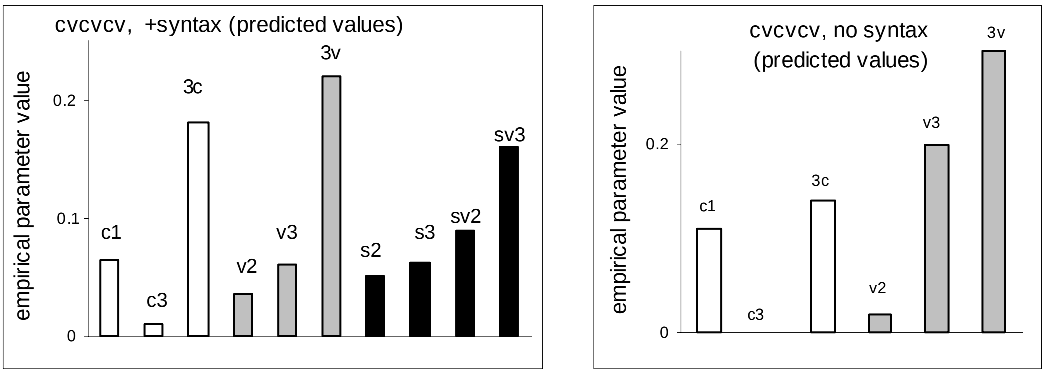


Figure 6.17. Predicted empirical parameter values for the cvcvcv word group.

These are theory-driven testable values which, if confirmed by an empirical study of cvcvcv words similar to that described in chapter three, would further support the pressure for systematicity between the phonological and the semantic levels of the lexicon.

6.5 Conclusions

The present chapter is based on the hypothesis, tested in chapter five, that there are systematic relationships between the phonological and the syntax-semantic levels of representation of the lexicon. It has explored the measure of such systematicity as a tool to study and predict the organization of the phonological level of the lexicon. This exploratory analysis had some limitations, for instance, the corpus size limits the accuracy of the cooccurrence-based word representations. It must also be noted that the parameters included in the metrics of the phonological (segment-based against, for example, feature-based) and the cooccurrence-based (window-size, similarity metric etc) spaces were not ad-hoc, but were used in independent previous studies; using parameters specifically selected for finding, say, systematicity might have yielded clearer results. However, in an exploratory study such as this, the impact of the different parameters is only discovered *a posteriori*. In this exploration of a new paradigm the high levels of convergence between the results obtained with three independent subsets of the lexicon support their reliability and the robustness of the methods.

I set out to test the hypotheses that the pressure for systematicity between the phonological and the semantic levels of the lexicon would make empirical phonological parameter values return a significant phon-sem correlation; that the phon-sem correlation obtained with the empirical parameters would not be the best possible, because pressures other than systematicity affect the phonological structure of the lexicon; and finally, that the phon-sem correlation could be used to make testable predictions about the phonological space.

The random search method returned parameter impact values, a measure of how each parameter influences the phon-sem correlation, for four different lexicon subsets: cvcv, cvccv and cvcvcv words, and all words. The results are significantly consistent across the first three word groups, indicating that the phon-sem correlation (a measure of the systematicity in the lexicon) is based on the same types of phonological characteristics, such as stress and consonant identity and position, for different word groups.

An analysis of the parameter impact values for a word group containing words of all lengths and structures revealed the great importance of word length difference for phonological similarity, confirming the conjecture at the end of chapter five and consistent with the results of Shillcock et al.'s (2001, *submitted*).

The empirically obtained parameter values (for cvcv and cvccv words) turned out to correlate significantly in most cases with the random-search results, supporting my first hypothesis.

The hill-climbing search obtained optimal parameter configurations that obtained better phon-sem correlations than the empirical values, which supports my second hypothesis.

This optimal parameter configuration, which also correlates well with the empirical parameters, is based on few parameters with very high values, namely sharing the word-final stressed vowel, and sharing the word consonant structure.

I proposed that while these parameters reflect the pressure for systematicity on the lexicon, other important parameters in the empirical configuration, such as the stressed vowel, reflect the opposed pressure for easy identification and intelligibility of words that (because of systematicity) sound similar and have similar cooccurrence patterns in speech.

Finally, the parameter information obtained with the random and the hill-climbing searches of the cvcv, cvccv and cvcvcv groups, together with knowledge of the empirical parameter values of the cvcv and cvccv groups was used to predict empirical parameter values for cvcvcv words, the testable prediction I anticipated in my third hypothesis.

In all, this chapter has offered a new approach to the study of the phonological organisation of the lexicon that transcends phonology and includes other lexical dimensions such as syntax and semantics.

Chapter 7. Conclusion

This thesis set out to explore the complex, adaptive nature of the mental lexicon. Using corpus-based methodologies, it has examined the internal phonological structure of words (chapter two), relationships of phonological and cooccurrence-based similarity between words (chapters three and four) and a higher level of organisation based on systematicity between the phonological and the cooccurrence-based representations of the lexicon (chapters five and six). In each case, the lexicon organisation was explained in terms of adaptations to pressures that can ultimately be related to language as a tool for human communication, and to the fact that language has to be easily acquired by successive generations of people.

7.1 The adaptive lexicon

In the past chapters I have quantified relationships both within and between words, always finding evidence that the lexicon is organised along many different dimensions. I have focused on patterns of lexical organisation that only emerge when large subsets of the lexicon are taken into account. The analysis of the results of chapters two to six of this thesis suggests that the mental lexicon is an adaptation that responds to the multiple, often conflicting pressures acting on it. These pressures ultimately relate to human communication and to the learnability of language by human infants.

In chapter two I examined the degree of phonological information (measured as entropy) found in the different word segment positions. The resulting information profile is an emergent property of a system of words. The profile of the words uttered in speech showed a left-to-right decreasing information level that may be an adaptation to the need to segment speech - words tend to begin at points of high phonological information content and finish at redundant, more predictable points. The information profile of the word types, with information evenly spread across the word length, suggested that it is adapted for optimal storage, to make the most of its representational space. The information profile of the child-directed lexicon did not show this adaptation, suggesting that the first words to be learned are not so tightly packed in terms of information, perhaps configuring a scaffolding upon which subsequent words are stored.

In chapter three I presented an empirical study measuring the relative impact of different phonological parameters on perceived word similarity in Spanish. In agreement with other findings in the literature of lexical structure, I found, for instance, that two words sharing the initial consonant are perceived to be more similar than two words sharing a word-internal consonant. I also found evidence of an interference of morphology in the judgement of phonological similarity – the stressed final vowels encode several verb tense and person morphemes, and two words that share the same stressed final vowel are judged to be more similar than if they share any other parameter.

In chapter four I measured similarity between words based on the words they cooccur with in speech. I constructed a lexicon representation using this measure of similarity and showed the emergence of categories such as parts of speech, noun-verb, feminine-masculine and semantic categories. Patterns of cooccurrence with closed-class words defined the word's syntactic identity, while patterns of cooccurrence with determiners influenced the word's gender, and patterns of cooccurrence with open-class words affected the word's semantic classification.

Chapter five tested the existence of a systematic mapping between the representations of the lexicon obtained in the previous two chapters - phonological and cooccurrence-based. Systematicity is a manifestation of the general nervous system tendency for structure-preserving mappings, which naturally leads to generalisation and inference - it provides useful links between concepts and words while exploiting the natural tendencies of the human nervous system. I measured a small but statistically significant degree of systematicity between the phonological and the cooccurrence-based levels of the lexicon. I removed most syntactic information from the similarity metrics on which the two representations were based in order to measure the systematicity between word form and word meaning. This had been found previously for English (Shillcock, Kirby, McDonald & Brew, 2001, *submitted*), and the present study extended the effect to Spanish, in support of the hypothesis that systematicity between levels of representation is a universal trait of language.

In chapter six I explored the relationships between different parameters of phonological similarity and the 'phon-sem' systematicity. The results revealed another pressure on the lexicon, opposed to that of systematicity. In a purely systematic lexicon, words with similar meanings, used in similar contexts, would tend to sound similar. This poses a problem for communication: two words that sound the same are usually distinguished by the context, but if their contexts are also similar, they will be easily confused. The pressure that works to solve this problem tries to make words with similar meanings have different forms so they can be easily distinguished from each other. The methods applied in chapter six revealed that different parameters of phonological similarity behaved in different ways: words sharing the same consonant structure tended to be close together in the semantic space, supporting systematicity; however, words sharing the same stressed vowel (in the penultimate syllable) tended to be far apart in the semantic space, opposing systematicity. This suggests that, at least in Spanish, systematicity is based on the words' consonant space, while the stressed vowel might be serving the function of distinguishing potentially ambiguous words.

The explanations in this thesis have emphasized the systematic nature of the lexicon. It makes no sense to speak of the information contained in one word, and it is irrelevant to define how the form of one word relates to its meaning or how similar two words in an isolated pair are to each other. Information, systematicity and similarity are properties of a large set of words. I have focused on the relationships between words, the identity of the words themselves becoming irrelevant. In the complex adaptive mental lexicon, a change in the phonology of a single word has effects on the information structure of all words; its phonological similarity to the rest of the words is changed, and, because of the pressure for systematicity between phonological and syntax-semantic relationships, its syntax and semantics will be under pressure to change too.

In sum, this thesis supports the view that the lexicon has evolved a robust, complex structure that accommodates an ever-changing balance of pressures. The next section briefly presents a theoretical framework of the evolution of the adaptive lexicon.

7.2 An evolutionary theoretical framework for the adaptive lexicon

Throughout this thesis I have stressed the idea that the lexicon is a complex adaptive system. One of the characteristics of a CAS is that it evolves over time by a mechanism of selection, through continuous adaptations to pressures. The two main pressures I have proposed are that the lexicon has to allow human communication of concepts, and that it has to be learnable by human infants.

In this final section I sketch a theoretical framework to study the lexicon as embodying human language capacity, evolving in an environment that includes the human brain, human communication interactions, and the concepts to be communicated.

According to Hull (1988), the essential mechanism of evolution by selection includes a phase of stable information that evolves over the generations (there has to be variation in that information); a phase of contingent instantiations of that information that interact with the environment; and a cycle of replication of the information and development of instantiations whose interaction with the environment determines the differential replication of information. An example of selection is the natural selection of living organisms: the information is encoded in the genes, and it codes for the organism. Organisms interact with their particular environment, with the result that adaptive gene variants prevail over time while maladaptive ones die out.

Language selection has been studied from this point of view in the past. Some authors proposed that the stable information is I-language or internal language, and pieces of E-language (external language, such as speech or text) are the contingent instantiations (e.g. Kirby & Hurford's 2002 Iterated Learning Model). Others argued the opposite: information is found in speech, and it develops contingent instantiations in people's brains, which, in turn, produce more speech (e.g. Croft, 2000; Mufwene, 2001). I follow the latter trend and propose that *linguistic* information (syntax, phonology) resides in speech. Linguistic information evolves over the generations through change in the proportions of the information variants (such as sound variants, syntactic structure variants) in the speech of a linguistic community. The individual mental lexicons (I-language) are the instantiations of linguistic information, and they interact with an environment that includes human brains, the concepts to be communicated and speech coming from other humans.

In this framework, concepts, the contents of lexical *semantics*, are not part of the linguistic information, but rather of the environment where linguistic information evolves. Concepts are part of a different system with a different dynamics: semantic information is found in people's brains, and speech acts are contingent instantiations of that information. The expressions of concepts can be seen as memes[1] (Dawkins, 1986) that interact with other memes, and that interaction determines their success to stay in the meme pool.

Since one of the main pressures acting on the lexicon is that it has to help people communicate, the (linguistic) lexicon needs a way to capture the (semantic) concepts in its linguistic structure, that is to say, to maintain symbolic associations. The pressure for systematicity across representations may have had a role in organising the lexicon's syntactic and phonological structure, and the structure of concepts around each other.

I propose the evolutionary relationship between linguistic and semantic aspects of the lexicon illustrated in Figure 7.1:

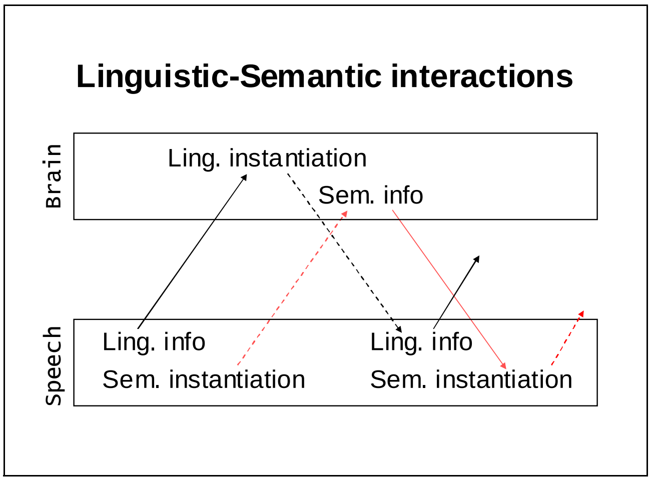


Figure 7.1: Interactions between the evolution of linguistic (black lines) and semantic (red lines) aspects of the lexicon.

Information encodes the structure of the instantiations, and each instantiation can only produce more of the same information that encoded it, or, as illustrated in Figure 7.1, 'reflect' the information back. In other words, for a given person, the linguistic input (the phonology and syntax they hear) equals the linguistic output (the phonology and input they use when they speak); for a given piece of speech, the semantic input (the meaning intended by the speaker) equals the semantic output (the meaning understood by the

[1] Meme: term coined by Dawkins. Memes are the units of cultural evolution, in the same sense as genes are the units of biological evolution. Memes include tunes, ideas, values and skills, and they replicate when they are learned by a new person.

hearer). At the linguistic level, I have been assuming this theoretical framework throughout the thesis, in that I have analyzed patterns of information in speech and assumed that they are the trigger for the development of language in new generations of infants.

Variation in the linguistic pool comes from mutation (errors that ‘catch’) and contact between languages; variation in the population comes from the fact that information from different instantiations, containing different variant combinations, is mixed together to produce a new combination of variants in each new instantiation. This means that syntactic and phonological variant combinations from the speech of many speakers contribute to form the unique mental lexicon of each new speaker. Social factors such as the prestige of the variants and the patterns of contact between speakers of a language affect the differential spread of variants, and hence, language evolution.

This thesis has shown systematic relationships between phonology on one hand and cooccurrence-based representations on the other; we have also seen that cooccurrence encodes for both syntax and semantics. However, the results presented and reviewed in chapter four suggest that syntax and semantics are encoded by very different cooccurrence patterns – syntax is best captured by small windows that take into account the exact position of words, semantics by much larger windows. Another difference between the syntactic and the semantic levels is that syntax is encoded mainly by language-internal relationships, whereas semantics needs to have links with the realm of concepts. The theoretical framework presented in Figure 7.1 could be tested by a paradigm that considered the evolving interrelations between linguistic information (i.e. phonology and syntax) on one hand, and semantic information on the other hand.

This framework offers an explanation to the symbiotic relationship between humans and language. It also attempts to explain the relationships between semantics and the other aspects of language, taking into account the pressure for language to capture meaning so that it can help people communicate. In exchange, humans help language survive and replicate: when people communicate their ideas, the speech they produce also carries the information necessary to create new linguistic instantiations in human infants.

7.3 Contributions and implications of this thesis

7.3.1 Original research

This thesis has offered support for the hypothesis that the lexicon is a complex structure that responds adaptively to pressures derived from its relationships with humans. It has presented a collection of approaches to the study of the mental lexicon that provide new evidence for previously unexplored aspects of the organisation of the mental lexicon, such as:

- the adaptive nature of the phonological information profile,
- the impact of aspects of phonology on perceived word similarity in Spanish,
- gender classification in a cooccurrence similarity space.

This thesis also has supported other findings, mainly by presenting evidence from Spanish, such as

- syntactic categorisation in a cooccurrence similarity space,
- systematic relationships between phonological and cooccurrence-based similarity between words.

Finally, it has introduced a new paradigm for the study of the phonological structure of a language that takes into account the systematic relationships between phonological and cooccurrence-based similarity.

7.3.2 Theoretical implications

I have applied concepts like adaptation and complexity to the study of language to focus on systematic properties that only emerge when we consider large sets of linguistic data.

The corpus-based approach adopted supports the statistical learning hypothesis by indicating that the lexicon structure is developed, among others, due to sensitivity to within-word (e.g. entropy) and between-word (e.g. phonologically-encoded morphology) phonological statistical patterns, and also to word cooccurrence patterns (e.g. cooccurrence-encoded syntactic classes). In most of the metrics employed, particularly in the cooccurrence statistics and in the calculation of the information profile of the word tokens, every utterance of each word contributed to the lexicon representation. The similarity-based lexicon model I have adopted means this thesis is best understood within an analogy-based framework (e.g. Skousen, 1995) where new word exemplars are processed, stored and retrieved in the form of phonological, contextual and other information, and this information is then related to analogous exemplars stored at the same levels of information, and also across levels, in the rest of the lexicon.

7.3.3 Open-ended research

This thesis was not intended to provide the last word on the mental lexicon, but rather has presented an overview of how a diverse collection of new quantitative approaches can contribute novel insights to the study of this vast subject.

The explorations presented in this thesis can be extended and refined by using a larger corpus or several corpora of different languages or E-language modalities (speech, text, emails). Additionally, the results could be improved by tailoring the metrics of similarity and the phonological and cooccurrence parameters to the different tasks. For instance, patterns of cooccurrence with adjacent words seems to return the most accurate syntactic categorisation

(Mintz, 2003; Monaghan & Christiansen, 2004), whereas patterns of cooccurrence with the words in the same paragraph or document are best suited to encode words' semantic identity (e.g. the LSA approach, Landauer & Dumais, 1997). Focusing on the words' stressed vowel seems to bear on lexical individuation, reflecting lexical contrast; focusing on the consonant structure seems to reflect how words fit in with other words in the structure of the lexicon, reflecting lexical integration and systematicity.

The methods presented in the thesis may also be applied to various fields and open new lines of research:

Phonology

Chapter six explores the relationship between certain aspects of word phonology (phonological parameters) and the systematicity between a phonological-similarity and a cooccurrence-similarity representation of the lexicon. The results presented suggest that while some of these phonological parameters support phon-sem systematicity, others have been recruited by the opposite pressure to make words that occur in similar contexts sound different from each other for more unambiguous recognition. In chapter six I presented preliminary results for an extended phonological parameter set applicable to words of all lengths. Different combinations of phonological parameters can reveal the relationships between the phon-sem systematicity and various aspects of phonology (from features to word-length or prosody; using acoustic speech representations, orthography or other representations of E-language). A cross-linguistic comparison of the results of such studies might reveal universal properties of the phonological systems with respect to systematicity (as well as language-specific ones).

Syntax and semantics

Chapter six introduced a method to quantify the impact of parameters of phonological similarity on the systematicity between the phonological and the cooccurrence similarity-based spaces, the latter including syntactic and semantic information. This added an extra dimension to the description of the phonology of a language. Similarly, we could better describe the syntax and the semantics of a language by taking into account the impact of parameters of cooccurrence-based similarity on the phon-sem correlation. This could be another test for the hypothesis that certain word classes (such as proper names, swear-words, or certain syntactic categories) support the systematicity. Cross-language comparisons of such studies could, again, reveal universal (as well as language-specific) properties of syntax and semantics with respect to systematicity.

Language acquisition

Studies similar to the ones presented in the thesis, but based on corpora of child-directed and child-produced speech may provide clues to the sequential involvement of different pressures in the development of the mental lexicon. In chapter two I did compare the information profiles of words from an adult and a child-directed corpus, and I explained the results in terms of the differential impact of the pressures on the lexicon during language development. A psycholinguistic test to quantify the impact of parameters of phonological similarity carried out with children of different ages could help study the development of the phonological mental lexicon structure and of its relationships with morphology, among others. Looking at the levels of syntactic and semantic categorisation achieved by a cooccurrence space based on corpora of child-directed speech could reveal the sequentially incremental syntactic and semantic structure of the developing mental lexicon. Patterns of phon-sem systematicity found in such corpora again could show the time-course of pressures for systematicity and the opposed pressure for phonological differentiation of words occurring in similar contexts.

Language change

Change in a complex system has far-reaching consequences. The existence of phon-sem systematicity introduces a new level of complexity in the study of the lexicon that implies that change in one domain will have an impact on the other. This opens the way to explorations of the effects of semantics and syntax on phonological change, and of phonology on semantic and syntactic change.

Glossary

For the purposes of this thesis, these terms are defined as follows, except when otherwise stated.

Complex adaptive system (CAS): An organized system of agents that evolves over time in order to maximize some mesure of fitness. CAS have emergent properties that could not be derived from the sum of its parts but which arise from their complexity. Organic life is a CAS evolving to maximize the reproducibility of organisms in a particular environment; other systems that have been described as CAS include the global economy, the stock exchange, the immune system, society, culture, and language.

Content word: Also called open-class words. Defined by opposition to functors, content words have lexical meaning. Content words include nouns, verbs, adjectives and adverbs.

Cooccurrence statistics: Corpus-based definition of a word in terms of other words it occurs close to in speech or text. One word is defined by how often the defining words (usually, high-frequency words, or function words) appear inside a window of a given number of words around (in front, after or both) the target word.

Entropy: Entropy is a measure of the information or the uncertainty that each segment position in a set of words carries. The probability of each phoneme and allophone occurring in each segment position of a set of words are calculated. For probabilities (p_1, p_2, p_3...p_n), the entropy (H) is: $H = - \Sigma (p_i \cdot \log p_i)$.

Functor: Also called function words or closed class words, functors have very little lexical meaning, but serve to express grammatical relationships between other (content) words. Functors include prepositions pronouns, conjunctions, articles and auxiliary verbs.

Information profile (*also* **information contour**): Shape obtained over the whole word when we plot the entropy level of each word segment taken in isolation calculated over a set of words.

Lexicon: the set of words in a language, together with the relationships between them at all levels of description: phonological, syntactic, semantic etc. It embodies language competence.

Monte-Carlo analysis: Statistical analysis based on a comparison of a veridical result with many results obtained with random parameter configurations. The position of the veridical result in the distribution of all results is a measure of its statistical significance.

Parameters of phonological similarity: Phonological aspects that two words may share. We may consider segmental parameters (e.g. both words having the same initial segment, both words containing segment /m/, both words having the same number of segments), feature-based parameters (e.g. both words starting by or containing a coronal consonant) or suprasegmental parameters (e.g. being stressed on the final syllable; being unstressed; having the same number of syllables).

Redundancy: Redundancy is a measure of the predictability carried by each segment position in a set of words. Redundancy (R) is: R = 1 – H.

Slope of the information profile (***m***): Measure of the steepness of the information profile linear trendline. In the trendline equation $y = \boldsymbol{m}x + n$, the slope is ***m***.

Structure-preserving: *See* Systematicity.

Systematicity: A structure-preserving relationship between structured representations. The structure of one representation can be inferred from the structure of another.

Token: Each of the occurrences of a word type. For instance, in a given corpus, we can have 245 tokens of the word type ‘of’.

Type: Each of the different words occurring in a corpus. For instance, ‘of’ is a type (for which there are 245 tokens in a given corpus).

REFERENCES

Aitchison, J. 2001. Language change. Cambridge: Cambridge University Press.

Aitchison, J. & Straf, M. 1982. Lexical storage and retrieval: a developing skill. In A. Cutler Slips of the tongue and language production. Berlin: Mouton.

Alarcos Llorach, E. 1994. Gramática de la lengua española. Madrid: Espasa Calpe.

Amsler R.A. & 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.

Anderson, J.R. 1991. Is human cognition adaptive. Behavioral and Brain Sciences, 14(3): 471-484.

Andrews, S. 1997. The role of orthographic similarity in lexical retrieval: resolving neighbourhood conflicts. Psychonomic Bulletin and Review, 4: 439-461.

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

Baddeley, A.D. & Hitch, G.J. 1974. Working memory. In Bower, G.H. (ed.), Recent Advances in Learning and Motivation. Vol. 8. New York: Academic Press.

Baddeley, A.D, Thomson, N. & Buchanan, M. 1975. Word length and the structure of short-term memory. Journal of Verbal Learning and Verbal Behavior, 14: 575-589.

Baldwin, D.A. 2000. Interpersonal understanding fuels knowledge acquisition. Current directions in psychological science, 9(2): 40-45.

Bammesberger, A. 1992. The place of English in Germanic and Indo-European. In R.N. Hogg (ed.) The Cambridge history of the English language. Vol. I. The beginnings to 1066. Cambridge: Cambridge University Press.

Bárkányi , Z. 2002. A fresh look at quantity sensitivity in Spanish. Linguistics, 40(2): 375-394.

Battig. W.F. & Montague, W.E. 1969. Category norms for verbal items in 56 categories: A replication and extension of the Connecticut category norms. Journal of Experimental Psychology Monograph, 80: 1-46.

Beckwith, R., Fellbaum, C., Gross, D. & 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.

Beeferman, D. 1998. Lexical discovery with an enriched semantic network. In Proceedings of the Workshop on Usage of WordNet in Natural Language Processing Systems, 358-364.

Berger, A.L. Della Pietra, S.A. & Della Pietra, V.J. 1996. A maximum entropy approach to natural language processing. Computational Linguistics, 22 (1): 39-71.

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.

Bod, R., Jay, J.H. & Jannedy, S. 2003. Probabilistic linguistics. Cambridge, MA: MIT Press.

Boroditsky, L. 2001. Does language shape thought? English and Mandarin speakers' conceptions of time. Cognitive Psychology, 43(1): 1-22.

Bozsahin, C. 2002. The combinatory morphemic lexicon. Computational linguistics, 28(2): 145-186.

Brew, C. & McKelvie D. 1996. Word-pair extraction for lexicography. In K. Oflazer and H. Somers (eds.) Proceedings of the Second International Conference of New Methods in Language Processing, 44-45.

Broe, M. 1993. Specification Theory: The treatment of redundancy in generative phonology. Unpublished PhD dissertation, University of Edinburgh.

Browman, C.P. 1978. Tip of the tongue and slip of the ear. Implications for language processing. UCLA Working Papers in Phonetics, 42.

Brown, R. & McNeil, D. 1966. The “tip of the tongue” phenomenon. Journal of Verbal Learning and Verbal Behaviour, 5: 325-337.

Budanitsky, A. & Hirst, G. 2001. Semantic distance in WordNet: An experimental, application-oriented evaluation of five measures. In Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics, Pittsburgh, PA.

Bullinaria, J.A. & Huckle, C.C. 1997. Modelling lexical decision using corpus derived semantic vectors in a connectionist network. In J.A. Bullinaria, D.W. Glasspool & G. Houghton (eds.), Fourth Neural Computation and Psychology Workshop: Connectionist representations. London: Springer, 213-226.

Burquest, D.A. & Payne, D.L. 1993. Phonological analysis: A functional approach. Dallas, TX: Summer Institute of Linguistics.

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

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

Cachin, C. 1997. Entropy measures and unconditional security in cryptography. Vol 1: ETH Series in Information Security and Cryptography. Konstanz, Germany: Hartung-Gorre Verlag.

Cairns, P., Shillcock, R.C., Chater, N. & Levy, J. 1997. Bootstrapping word boundaries: A bottom-up corpus-based approach to speech segmentation. Cognitive Psychology, 33(2): 111-153.

Caplan, D. & Waters, G.S. 1999. Verbal working memory and sentence comprehension. Behavioral and Brain Sciences, 22(1).

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

Chandler, D. 2001. Semiotics: the basics. London: Routledge.

Chiarello, C., Shears C. & Lund, K. 2000. Distributional typicality: A new approach to estimating noun and verb usage from large scale text corpora. Brain and Cognition, 43(1-3): 94-98.

Chomsky, N. 1986. Knowledge of language. New York: Praeger.

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.

Christiansen, M.H., Allen, J. & Seidenberg, M. 1998. Learning to segment speech using multiple cues: A connectionist model, Language and Cognitive Processes, 13(2-3), 221-268.

Christophe, A., Dupoux, E., Bertoncini, J. & Mehler, J. 1994. Do infants perceive word boundaries? An empirical study of the bootstrapping of lexical acquisition. Journal of the Acoustical Society of America, 95(3): 1570-1580.

Clements, G.N. 1991. The role of the sonority cycle in core syllabification, in G. N. Clements, J. Kingston, M.E. Beckman (eds.) Papers in Laboratory Phonology. Vol.1: Between the Grammar and Physics of Speech. Cambridge, UK: Cambridge University Press.

Coady, J. A. & Aslin, R.N. 2003. Phonological neighbours in the developing lexicon. Journal of Child Language, 30(2): 441-469.

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

Cole, R., Yan, Y., Mak, B., Fanty, M. & 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.

Collins A.M. & Loftus E.F. 1975. A spreading activation theory of semantic processing. Psychological Review, 82: 407-428.

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

Corrigan, R. 2004. The acquisition of word connotations: Asking “What happened?”. Journal of Child Language, 31: 381-398.

Crestani, F. 2003. Combination of Similarity Measures for Effective Spoken Document Retrieval. Journal of Information Science, 29(2): 87-96.

Croft, W. 2000. Explaining language change: an evolutionary approach. Longman.

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. 1990. Exploiting prosodic probabilities in speech segmentation. In G.T.M. Altmann (ed.), Cognitive models of speech processing: Psycholinguistic and computational perspectives. Cambridge: MIT Press, 105-121.

Cutler, A. & Carter, D.M. 1987. The predominance of strong initial syllables in English vocabulary. Computer Speech and Language, 2: 133-142.

Cutler, A. & Butterfield, S. 1992. Rhythmic cues to speech segmentation: Evidence from juncture misperception. Journal of Memory and Language, 31: 218-236.

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.

Daelemans, W. 1999. Machine learning approaches. In Hans vanHalteren (ed.), Syntactic wordclass tagging. Dordrecht: Kluwer Academic Publishers.

Dawkins, R. 1986. The selfish gene, 2nd ed. Oxford: Oxford University Press.

De Cara, B. & Goswami, U. 2003. Phonological neighbourhood density effects in a rhyme awareness task in 5-year-old children. Journal of Child Language, 30: 695-710.

De Rosnay, J. 1997. Homeostasis: resistance to change. Principia Cybernetica Web http://pespmc1.vub.ac.be/HOMEOSTA.html.

Deacon, T. 1997. The symbolic species: the co-evolution of language and the human brain. London: Penguin Books.

Dooley, K.J. 1997. A complex adaptive systems model of organization change. Nonlinear Dynamics, Psychology, and Life Sciences, 1(1): 69-97.

Dumay, N., Benraiss, A., Barriol, B., Colin, C., Radeau, M. & Besson, M. 2001. Behavioral and electrophysiological study of phonological priming between bisyllabic spoken words. *Journal of cognitive neurscience*, 13(1): 121-143.

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

Durieux, G. & Gillis, S. 2000. Predicting grammatical classes from phonological cues: an empirical test. In H. J. Weissenborn (ed.) Approaches to bootstrapping: phonological, syntactic and neurophysiological aspects of early language acquisition. Amsterdam: Benjamins, 189-232.

Eddington, D. 2000. Spanish stress assignment within the Analogical Modeling of Language. Language, 76(1): 92–109.

Eddington, D. 2002. Spanish diminutive formation without rules or constraints. Linguistics, 40(2): 395-419.

Elliot, A.J. 1981. Child language. Cambridge, England: Cambridge Textbooks in Linguistics.

Faust, M.E. & Gernsbacher, M.A. 1996. Cerebral mechanisms for suppression of inappropriate information during sentence comprehension. Brain and Language, 53(2): 234-259.

Fay, D. & Cutler. 1977. A. Malapropisms and the structure of the mental lexicon. Linguistic Inquiry, 8: 505-520.

Fernald, A. & Kuhl, P. 1987. Acoustic determinants of infant preference for Motherese speech. *Infant Behavior and Development*, 10: 279-293.

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

Finch, S.P. & Chater, N. 1992. Bootstrapping syntactic categories. Proceedings of the Fourteenth Annual Conference of the Cognitive Science Society. Indiana: Bloomington, 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. & 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. & Landauer, T.K. 1999. The Intelligent Essay Assessor: Applications to educational technology. Interactive Multimedia Education. Jounal of computer-ehnanced learning 1(2).

Forrester, J.W. 1975. The collected papers of Jan W. Forrester. Wright-Allen Press.

Frege, G. 1892, 1960. On Sense and Reference, in *Translations from the Philosophical Writings of Gottlob Frege*, P. Geach and M. Black (eds.) Oxford: Basil Blackwell.

Friederici, A.D. & Wessels, J.M.I. 1993. Phonotactic knowledge of word boundaries and its use in infant speech perception. Perception and Psychophysics, 54(3): 287-295.

Frisch, S.A. 1996. Similarity and Frequency in Phonology, PhD dissertation, University of Illinois.

Frisch S.A., Pierrehumbert J.B., Broe M.B. 2004. Similarity avoidance and the OCP. Natural language and linguistic theory 22(1): 179-228.

Gallistel, C.R. 1990. Representations in animal cognition – An introduction. Cognition, 37(1-2): 1-22.

Gaskell, M.G. & Marslen-Wilson, W.D. 2002. Representation and competition in the perception of spoken words. Cognitive Psychology, 45: 220-266.

Gell-Mann, M. 1994. The Quark and the Jaguar. New York: Freeman & Co.

Girin, L., Schwartz, J.L., Feng, G. 2001. Audio-visual enhancement of speech in noise. Journal of the Acoustical Society of America,109(6): 3007-3020.

Goldsmith, J. 2000. On information theory, entropy and phonology in the 20th century, Folia Linguistica, 34(1-2), 85-100.

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

Gómez, R.L. & Gerken, L. 1999. Artificial language learning by 1-year-olds leads to specific and abstract knowledge. Cognition, 70: 109–135.

Grice, P. 1975. Logic and conversation (in:) P.Cole and J.L. Morgan (eds.) Syntax and semantics. Vol 3: Speech acts. New York: Academic Press

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

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

Guthrie, L., Pustejovsky, J., Wilks, Y. & Slator, B.M. 1996. The role of lexicons in natural language processing, Communications of the ACM, 39(1): 63-72.

Halliday, T. (ed.) 1992. Biology: Brain and Behaviour. Book 6: The senses and communication. Milton Keynes: The Open University.

Harrap Compact Spanish Dictionary. 1999. Edinburgh: Chambers Harrap Publishers.

Hernández Montoya, R. 2001. El género del género. Caracas: Venezuela analítica. http://www.analitica.com/bitblioteca/roberto/genero.asp.

Hinton, G.E. & Shallice, T. 1991. Lesioning an attractor network: Investigations of acquired dyslexia. Psychological Review; 98: 74-95.

Hinton, L., Nichols, J. & Ohala, J. 1994. Sound Symbolism, Cambridge, England: Cambridge University Press.

Hoff, E. & Naigles, L. 2002. How children use input to acquire a lexicon. Child development, 73(2): 418-433.

Holland, J.H. 1995. Hidden Order. Reading, MA: Addison-Wesley.

Houston-Price, C. 2004. Infants use of cues to word meaning. Paper presented in the Linguistic Circle, Department of Theoretical and Applied Linguistics, University of Edinburgh.

Huckle, C.C. 1995. Grouping Words Using Statistical Context. Proceedings of the Seventh Conference of the European Chapter of the Association for Computational Linguistics. Morgan Kaufmann, San Francisco CA, 278-280.

Hull, D.L. 1988. Science as a process: an evolutionary account of the social and conceptual development of science. Chicago: University of Chicago Press.

Hurford, J.R. 1981. Malapropisms, left-to-right listing and lexicalism. Linguistic Inquiry 12: 419-23.

Hurford, J.R. 1989. Biological evolution of the Saussurean sign as a component of the Language Acquisition Device. Lingua, 77: 187-222.

Ikeno, A., Pellom, B., Cer, D., Thornton, A., Brenier, J.M., Jurafsky, D., Ward, W., Byrne, W. 2003. Issues in recognition of Spanish-accented spontaneous

English, in ISCA & IEEE ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition, Tokyo, Japan.

Jackendoff, R. 1975. Morphological and semantic regularities in the lexicon. Language, 51: 639-671.

Jackendoff, R. 1983. Semantics and cognition. Cambridge, MA: MIT Press.

Jantsch, E. 1980. The Self-Organizing Universe, Oxford: Pergamon Press.

Jarmasz, M. 2003. Roget’s thesaurus as a lexical resource for natural language processing. Masther’s thesis, Ottawa-Carleton Institute for Computer Science, University of Ottawa.

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

Johnson, K. 1997. Speech Perception without speaker normalization. In K. Johnson and J.W. Mullennix (eds.) Talker Variability in Speech Processing. San Diego: Academic Press, 145-66.

Johnson, E.K., Jusczyk, P.W., Cutler, A., Norris, D. 2003. Lexical viability constraints on speech segmentation by infants. Cognitive psychology, 46(1): 65-97.

Jusczyk, P.W. 1999. How infants begin to extract words from speech, Trends in Cognitive Sciences, 3(9): 323-328.

Jusczyk, P.W., Luce, P. & Charles-Luce, J. 1994. Infants' sensitivity to phonotactic patterns in the native language. Journal of Memory and Language 33: 630–645.

Jusczyk, P.W., Luce, P.A. & Charles-Luce, J. 1994. Infants’s sensitivity to phonotactic patterns in the native language. Journal of Memory and Language, 33: 630-645.

Keller, R. 1994. On language change. The invisible hand in language, London: Routledge.

Kelly, B. Leben, W. & Cohen, R. 2003. The meanings of consonants. Proceedings of the 29th Berkeley Linguistics Society Annual Meeting.

Kelly, M .H. 1992. Using sound to solve syntactic problems: The role of phonology in grammatical category assignments. Psychological Review, 99: 349-364.

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.

Kemler Nelson, D.G., Hirsh-Pasek, K., Jusczyk, P.W. & Wright-Cassidy, K. 1989. How prosodic cues in motherese might assist language learning. Journal of Child Language, 16: 55-68.

Kempe, V. & Brooks P.J. 2001. The role of diminutives in the acquisition of Russian gender: Can elements of child-directed speech aid in learning morphology? Language Learning 51 (2): 221-256.

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

Kintsch, W. 2001. Predication. Cognitive Science, 25: 173-202.

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

Kirby, S. 1999. Function, selection and innateness. Oxford: Oxford University Press.

Kirby, S. & Hurford, J.R. 2002. The emergence of linguistic structure: an overview of the iterated learning model. In A. Cangelosi & D. Parisi (eds.), Simulating the evolution of language. London: Springer, 121-148.

Kirby, S., Smith, K. & Brighton, H. 2004. From UG to Universals: Linguistic adaptation through iterated learning. Studies in Language, 28(3).

Kirby, S. & Ellison, M.E. (*in preparation*). Quantifying lexical structure: An information-theoretic approach to language comparison..

Kjellmer, G. 2000. Potential words (Neologisms). Word- Journal of the International Linguistic Association, 51(2): 205-228.

Knight, C., Studdert-Kennedy, M. & Hurford, J.R. 2000. The evolutionary emergence of language. Cambridge: Cambridge University Press.

Komarova, N.L. & Kondrak, M.A. 2003. Language, learning and evolution. In M.H. Christiansen and S. Kirby (eds.), Language Evolution. Oxford: Oxford University Press.

Kondrak, G. 2000. A new algorithm for the alignment of phonetic sequences. Proceedings of the First Meeting of the North American Chapter of the Association for Computational Linguistics, 228-295.

Krott A., Schreuder R. & Baayen, R.H. 2002. Linking elements in Dutch noun-noun compounds: Constituent families as analogical predictors for response Latencies. Brain and Language, 81(1-3): 708-722.

Kuhl P.K., Andruski J.E., Chistovich I.A., Chistovich L.A., Kozhevnikova E.V., Ryskina V.L., Stolyarova E.I., Sundberg U. & Lacerda F. 1997. Cross-language analysis of phonetic units in language addressed to infants. Science, 277(5326): 684-686.

Labov, W. 1994. Principles of linguistic variation. Vol. I: Internal factors. Oxford: Basil Blackwell.

Lacerda, F. 1995. The perceptual-magnet effect: An emergent consequence of exemplar-based phonetic memory. In K.Elenius, and P. Branderud (eds.) Proceedings of the XIIIth International Congress of Phonetic Sciences, 2. Stockholm: KTH and Stockholm University, 140-147.

Landauer, T. K. & 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.

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

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

Levy, J.P. & Bullinaria, J.A. 2001. Learning lexical properties from word usage patterns: Which context words should be used?. In R.F. French and J.P Sougne (eds.), Connectionist models of learning, development and evolution: Proceedigs of the Sixth Neural Computation and Psychology Workshop, London: Springer, 273, 282.

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.

Lopez Ornat, S. 1994. La adquisición de la lengua española. Madrid: Siglo XXI.

Lowe, W. & McDonald, S. 2000 The direct route: Mediated priming in semantic space. Proceedings of the 22nd Annual Conference of the Cognitive Science Society (pp.806-811). Lawrence Erlbaum Associates.

Luce, P.D., 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. Cambridge, MA: MIT Press, 122-147.

Lund, K. & 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. & Atchley, R. 1995. Semantic and associative priming in high-dimensional semantic space. In Proceedings of the 17th Annual Conference of the Cognitive Science Society. Mahwah, NJ: Erlbaum, 660-665.

Lund, K., Burgess, C. & Audet, C. 1996. Dissociating semantic and associative word relationships using high-dimensional semantic space. In Proceedings of the 18th Annual Conference of the Cognitive Science Society. Mahwah, NJ: Erlbaum, 603-608.

Macnamara, J. 1982. Names for things: a study of human learning. Psychological Review, 79: 1-13.

Magnus, M. 2001. What's in a Word? Studies in Phonosemantics. PhD thesis. University of Trondheim, Norway. www.trismegistos.com/Dissertation/

Manning, C.D. 2003. Probabilistic syntax. In R. Bod, J. Hay & S. Jannedy (eds.), Probabilistic Linguistics. Cambridge, MA: MIT Press.

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

Markman, E.M. 1989. Categorization and naming in children: problems of induction. Cambridge, MA; MIT Press.

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

Mattys, S.L. & Jusczyk, P.W. 2001. Phonotactic cues for segmentation of fluent speech by infants, Cognition, 78(2): 91-121.

Mattys, S.L., Jusczyk, P.W., Luce, P.A. & Morgan, J.L. 1999. Phonotactic and prosodic effects on word segmentation in infants. Cognitive Psychology, 38(4): 465-494.

Maturna, H. & Varela, F. 1992. The Tree of Knowledge. Boston: Shambhala.

Maye, J. & Gerken, L. 2000. Learning phonemes without minimal pairs. In: S.C. Howell, S.A. Fish and T. Keith-Lucas, Editors, Proceedings of the 24th Boston University Conference on Language Development, Somerville, MA: Cascadilla Press, 522–533.

Maye, J., Werker, J.F. & Gerken, L. 2002. Infant sensitivity to distributional information can affect phonetic discrimination. Cognition 82(3): B101-B111.

McDonald, S. 2000. Environmental Determinants of Lexical Processing Effort. PhD thesis. University of Edinburgh.

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

McGurk, H. & MacDonald, J. 1976. Hearing lips and seeing voices, Nature, 264: 746-748.

McMahon, A. & McMahon, R. 2003 Finding families: quantitative methods in language classification, Transactions of the Philological Society, 101(1): 7-55.

McNamara, T.P. & Miller, D.L. 1989. Attributes of theories of meaning. Psychological Bulletin, 106: 377-394.

McQueen, J. 1998. Segmentation of continuous speech using phonotactics. Journal of Memory and Language, 39: 21-46.

Melamed, I.D. 1999. Bitext maps and alignment via pattern recognition. Computational Linguistics 25(1): 107-130.

Melzi, G. & King, K.A. 2003. Spanish diminutives in mother-child conversations. Journal of Child Language, 30(2): 281-304.

Merriman, W.E. & Bowman, L.L. 1989. The mutual exclusivity bias in children's word learning. Monographs of the society for research in Child Development, 53(3-4).

Meyer, D.E. & 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.

Miller, G.A. & Charles, W.G. 1991. Contextual correlates of semantic similarity. Language and cognitive processes, 6: 1-28.

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. & Christiansen, M.H. (in press). The differential contribution of phonological and distributional cues in grammatical categorisation. Cognition.

Monaghan, P., Chater, N. & Christiansen, M.H. 2003. Inequality between the classes: Phonological and distributional typicality as predictors of lexical processing. Proceedings of the 25th Annual Conference of the Cognitive Science Society. Mahwah, NJ: Lawrence Erlbaum.

Monaghan, P. & Christiansen, M.H. 2004. What distributional information is useful and usable in language acquisition? Proceedings of the 26th Annual Conference of the Cognitive Science Society. Mahwah, NJ: Lawrence Erlbaum.

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

Morgan, J. & Newport , E. 1981. The role of constituent structure in the induction of an artificial language. J. Verbal Learning and Verbal Behaviour, 20: 67-85.

Mueller, S.T., Seymour, T. Krawitz, A., Kieras, D.A. & Meyer, D.E. 2003. Theoretical implications of articulatory duration, phonological similarity and phonological complexity. Journal of Experimental Psychology: Learning, Memory and Cognition, 29(6): 1353-1380.

Mufwene. S.S. 2001. The Ecology of Language Evolution. Cambridge: Cambridge University Press.

Mullennix, J.W., Bihon, T., Bricklemyer, J., Gaston, J., Keener, J.M. 2002. Effects of variation in emotional tone of voice on speech perception. Language and Speech 45: 255-283.

Murphy, G. L. 2002. The big book of concepts. Cambridge, MA: MIT Press.

Myers, R. 1990. Classical and modern regression with applications (2nd ed.) Boston, MA: Duxbury Press.

Nakamura, J. & Nagao, M. 1988. Extraction of semantic information from an ordinary English dictionary and its evaluation. Proceedings of the 12th International Conference on Computational Linguistics (COLING-88), Budapest, 459-464.

Neely, J.H. 1991. Semantic priming effects in visual word recognition: A selective review of current findings and theories. In D. Besner & G.W. Humphreys (eds.) Basic processes in reading: Visual word recognition. Hillsdale, NJ: Lawrence Erlbaum Associates, 234-336.

Nettle, D. 1999. Linguistic diversity. Oxford: Oxford University Press.

Ng, K. 1999 Towards robust methods for spoken document retrieval. In Proceedings of the International Conference on Spoken Language Processing, 3: 939-942.

Norris D., McQueen J.M., Cutler A. & Butterfield S. 1997. The possible-word constraint in the segmentation of continuous speech, Cognitive Psychology, 34(3): 191-243.

Nowak M.A. & Komarova N.L. 2001. Towards an evolutionary theory of language, Trends in Cognitive Science 5(7): 288-295.

Osgood, C., Suci, G. & Tannenbaum, P. 1957. The Measurement of Meaning. Urbana, IL: University of Illinois Press.

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

Patel, M., Bullinaria, J.A. & Levy, J.P. 1998. Extracting semantic representations from large text corpora. In J.A. Bullinaria, D. Glasspool & G. Houghton (eds.) Proceedings of the 4th Neural Computation and Psyschology Workshop. London: Springer-Verlag.

Peirce, C.S. 1897-1903.Logic as semiotics: The theory of signs. In J. Buchler (ed.), The philosophical writings of Peirce (1955). New York: Dover Books.

Penfield, W. & Rasmussen, T. 1950. The cerebral cortex of man: a clinical study of localization of function. New York: Macmillan.

Peperkamp, S. 2003. Phonological acquisition: Recent attainments and new challenges. Language and Speech, 46: 87-113.

Peperkamp, S. & Dupoux, E. 2002. Coping with phonological variation in early lexical acquisition, in I. Lasser (ed.) The Process of Language Acquisition. Berlin: Peter Lang Verlag, 359-385.

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.

Perez Pereira, M. 1991. The acquisition of gender: What Spanish children tell us. Journal of Child Language, 18(3): 571-590.

Phillips, B.S. 1999. The mental lexicon: Evidence from lexical diffusion. Brain and language 68(1-2): 104-109.

Pierrehumbert, J.B. 2001a. Exemplar dynamics: Word frequency, lenition, and contrast. In Bybee, J. & P. Hopper (eds.) Frequency effects and the emergence of linguistic structure. John Benjamins, Amsterdam, 137-157.

Pierrehumbert, J.B. 2001b. Stochastic Phonology. Glot International, 5(6): 195-207.

Pierrehumbert, J.B. 2003a. Phonetic diversity, statistical learning, and acquisition of phonology. Language and Speech, 46: 115-154

Pierrehumbert, J. B. 2003b. Probabilistic Phonology: Discrimination and Robustness. In R. Bod, J. Hay and S. Jannedy (eds.), Probabilistic Linguistics. Cambridge, MA: MIT Press.

Pinker, S. 1984. Language learnability and language development. Cambridge, Mass: Harvard U. Press.

Pinker, S. & Bloom, P. 1990. Natural language and natural selection. Behavioral and Brain Sciences 13: 707-784.

Plaut, D.C., McClelland, J.L., Seidenberg, M.S. & Patterson, K.1996. Understanding normal and impaired word reading: Computational principles in quasi-regular domains. Psychological Review, 103: 56-115.

Prigogine, I. & Stengers. I. 1984. Order Out of Chaos. (New York: Bantam Books).

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

Quesada, J.F, Kintsch, W. & 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.

Radeau, M., Morais, J. & Segui, J. 1995. Phonological priming between monosyllabic spoken words. Journal of experimental psychology: Human perception and performance, 21 (6): 1297-1311.

Real Academia Española. 1973. Esbozo de una nueva gramática de la lengua española, Madrid: Espasa Calpe.

Redington, M. & Chater, N. 1997. Probabilistic and distributional approaches to language acquisition. Trends in Cognitive Science 1(7): 273-281

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

Redington, M., Chater, N., Huang, C., Chang, L., Finch, S. & Chen, K. 1995. The universality of simple distributional methods: Identifying syntactic categories in mandarin chinese. In Proceedings of the 4th International Conference of the Cognitive Science of Natural Language Processing. Dublin City University.

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.

Robert-Ribes, J., Schwartz, J.L., Lallouache, T. & Escudier, P. 1998. Complementarity and synergy in bimodal speech: Auditory, visual, and audio-visual identification of French oral vowels in noise. Journal of the Acoustical Society of America, 103(6): 3677-3689.

Rodd, J.M., Gaskell, M.G. & Marslen-Wilson, W.D. 2004. Modelling the effects of semantic ambiguity in word recognition. Cognitive Science, 28: 89-104.

Rosch, E. 1978. Principles of categorization. In E.Rosch & B. Loyd (eds.) Cognition and categorization. Hillsdale, NJ: Lawrence Erlbaum Associates, 27-48.

Rouibah, A. & Taft, M. 2001. The role of syllabic structure in French visual word recognition. Memory and Cognition, 29(2): 373-381.

Rumelhart, D. & McClelland, J. 1986. On learning the past tenses of English verbs. Implicit rules or parallel distributed processing. In McClelland, J., Rumelhart, D. (eds.), Parallel distributed processing. Vol. 2. Cambridge, MA: MIT Press, 216-271.

Saffran J.R. 2003. Statistical language learning: Mechanisms and constraints. Current directions in psychological science, 12(4): 110-114.

Saffran, J.R. 2001. The use of predictive dependencies in language learning. Journal of Memory and Language 44: 493–515.

Saffran, J.R., Aslin, R.N. & Newport, E.L. 1996. Statistical learning by 8-month-old infants. Science, 274: 1926–1928

Saffran, J.R., Newport, E.L. & Aslin, R.N. 1996 Word segmentation: the role of distributional cues. Journal of Memory and Language, 35: 606–621

Sanchez-Casas, R., Igoa, J.M., Garcia-Albea, J.E. 2003. On the representation of inflections and derivations: Data from Spanish. Journal of Pshycholinguistic Research, 32(6): 621-668.

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.

Schütze, H. 1993. Word Space. In S.J. Hanson, J.D. Cowan & C.L. Giles (eds.). NIPS 5. San Mateo, CA: Morgan Kaufmann.

Segui, J. & Grainger, J. 1990. Priming word recognition with orthographic neighbors – Effects of relative prime target frequency. Journal of Experimental psychology – Human perception and performance,16(1): 65-76.

Seidenberg, M. & McClelland, J. 1989. A distributed developmental model of word recognition and naming. Psychological Review, 96: 523-568.

Sereno, M.I. 1991. Four analogies between biological and cultural/linguistic evolution. Journal of Theoretical Biology, 151: 467-507.

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

Shannon, C.E. 1951. Prediction and entropy of printed English, Bell Systems Technology Journal, 30: 50-64.

Shen, J., Hung, J. & Lee, L. 1998. Robust entropy-based endpoint detection for speech recognition in noisy environments, Proceedings of the 5th International Conference on Spoken Language Processing, 3. Melbourne: The Australian Speech Science and Technology Association Incorporated, 1015-1018.

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. Hillsdale, NJ: Lawrence Erlbaum Associates, 233-238.

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.

Shillcock, R.C., Kirby, S., McDonald, S. & Brew, C. (*submitted*). Exploring the structure-preserving mental lexicon.

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

Slaney, M. & McRoberts, G. 2003. BabyEars: A recognition system for affective vocalizations. Speech Communication, 39(3-4): 367-384.

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.

Smith, A.D.M. 2005. Mutual exclusivity: communicative success despite conceptual divergence. In M. Tallerman (ed.) Language origins: Perspectives on evolution. Oxford University Press.

Smith, A.D.M. and Vogt, P. 2004. Lexicon acquisition in an uncertain world. Proceedings of 5th Evolution of language conference.

Smith, E.E., Shoben, E.J. & Rips L.J. 1974. Structure and process in semantic memory – featural model for semantic decisions. Psychological Review, 81(3): 214-241.

SPSS® 2003. (Computer software). Version 12.0 for Windows.

Stemberger, J. 1991. Apparent anti-frequency effects in language production: the addition bias and phonological underspecification. Journal of Memory and Language, 20: 161-185.

Stevens, J. 1992. Applied multivariate statistics for the social sciences (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Stoianov, I.R. 2001. Connectionist lexical processing. PhD thesis. University of Groningen.

Stubbs, M. 1995. Collocations and semantic profiles. On the cause of the trouble with quantitiative studies. Functions of Language 2(1): 23-55.

Tamariz, M. & Shillcock, R.C. 2001. Real world constraints on the mental lexicon: Assimilation, the Speech Lexicon and the information structure of Spanish words. Proceedings of the 23rd Annual Conference of the Cognitive Science Society. Hillsdale, NJ: Lawrence Erlbaum Associates, 1058-1063.

Tootell, R.B.H, Silverman, M.S., Switkes, E. & De Valois, R.L. 1982. Deoxyglucose Analysis of Retinotopic Organization in Primate Striate Cortex. Science, 218: 902-904.

Tse S.K., Kwong S.M., Chan C. & Li H. 2002. Sex differences in syntactic development: Evidence from Cantonese-speaking preschoolers in Hong Kong. International Journal of Behavioral Development, 26(6): 509-517.

Tudhope, D. & Taylor, C.A 1996. Unified Similarity Coefficient for Navigating Through Multi-Dimensional Information. Proceedings of the 1996 American Society for Information Science and Society Annual Conference.

Tyler, L.K., Bright, P., Fletcher, P. & Stamatakis, E.A. 2004. Neural processing of nouns and verbs: the role of inflectional morphology, Neuropsychologia, 42(4): 512-523.

Ultan, R. 1978. Size-sound symbolism. In J. Greenberg (Ed.) Universals of Human Language, Vol. 2: Phonology, Stanford University Press.

Van Droogenbroek, M. & Delvaux, J. 2002. An entropy based technique for information embedding in images. Proceedings of the 3rd IEEE Benelux Signal Processing Symposium. Leuven, Belgium: IEEE Benelux Signal Processing Chapter, 81-84.

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

Vogt, P. & Coumans, H. 2003. Investigating social interaction strategies for bootstrapping lexicon development. JASSS, The journal of artificial societies and social simulation, 6(1): U83-U110.

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

Wagner, K., Reggia, J.A., Uriagereka, J. & Wilkinson, G.S. 2003. Progress in the Simulation of Emergent Communication and Language. Adaptive Behavior, 11(3): 37-69.

Wiemer-Hastings, P., Wiemer-Hastings, K. & 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 (Proceedings of the AIED'99 Conference). IOS Press, 535-542.

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

Williams, G.C. 1992. Natural Selection: domains, levels, and challenges. New York/Oxford: Oxford University Press.

Worden, R.P. 2000. Words, memes and language evolution, in C. Knight, M. Studdert-Kennedy & J.R. Hurford (eds.). The evolutionary emergence of language, Cambridge University Press.

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

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

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.

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

Yarowsky, D. 1992. Word-sense disambiguation using statistical models of Roget’s categories trained on large corpora. In Proceedings of the 14th International Conference on Computational Linguistics, 454-460.

Zamuner, T. S. 2001. Input-based phonological acquisition. Unpublished doctoral dissertation, University of Arizona, Tucson.

Ziegler J.C., Muneaux, M. & Grainger, J. 2003. Neighborhood effects in auditory word recognition: Phonological competition and orthographic facilitation. Journal of Memory and Language 48 (4): 779-793.

Zwitserlood, P. 1996. Form Priming. Language and Cognitive Processess, 11: 589-596.

APPENDICES

Appendix A

Finite schemes used to calculate the information profiles by feature. Vowels are taken to be individual elements of the finite scheme.

Manner of articulation		Place of articulation	
1	plosive – voiced	1	bilabial – voiced
2	plosive – voiceless	2	bilabial – voiceless
3	nasal – voiced	3	labiodental – voiced
4	vibrant (tap) – voiced	4	labiodental – voiceless
5	vibrant (trill) – voiced	5	interdental – voiced
6	fricative – voiced	6	interdental – voiceless
7	fricative – voiceless	7	dental – voiced
8	lateral – voiced	8	dental – voiceless
9	affricate – voiceless	9	alveolar – voiceless
10	approximant – voiced	10	alveolar – voiceless
11	approximant – voiceless	11	palatal – voiced
12	glide	12	palatal – voiceless
13	vowel a	13	velar – voiced
14	vowel e	14	velar – voiceless
15	vowel i	15	vowel a
16	vowel o	16	vowel e
17	vowel u	17	vowel i
		18	vowel o
		19	vowel u

Appendix B

The two sets of stimulus nonwords used in the empirical study described in chapter three.

STIMULUS SET 1			
1_5_c1_c2	búnta	bísko	línko
2_5_c1_c3	káste	kíndo	bínto
3_5_c1_tc13	pósta	púrke	púrte
4_5_c1_tc23	rásli	rónte	bósle
5_5_3c_c1	kórdu	kírda	kósla
6_5_c1_v1	sárke	sónti	pánti
7_5_c1_v2	mínde	mórka	kórke
8_5_c1_tv	fínto	fáste	kísto
9_5_c1_a1	kórpa	kengú	méngu
10_5_c1_a2	sultó	sánde	pandé
11_5_c1_av1	tárbo	túnte	kánte
12_5_c1_av2	kurtá	kombé	sondá
13_5_c1_str	bésto	búgra	túnka
14_5_c2_c3	lórdi	pérku	péndu
15_5_c2_tc13	mórfa	sérpo	mélfo
16_5_c2_tc23	kínte	gándo	gánto
17_5_3c_c2	méska	músko	pústo
18_5_c2_v1	linká	bentó	bistó
19_5_c2_v2	gúsmi	tésba	térbi
20_5_c2_tv	pósti	tésto	tórti
21_5_c2_a1	bésta	tusgó	túlgo
22_5_c2_a2	tuská	nósde	nordé
23_5_c2_av1	mólka	gálpe	góspe
24_5_c2_av2	pustó	leská	lenkó
25_5_c2_str	dákme	mókri	mónsi
26_5_c3_tc13	mónke	díska	míska
27_5_c3_tc23	mindó	saldé	sandé
28_5_3c_c3	gálti	gólte	pónte
29_5_c3_v1	pórda	mésdi	mósti
30_5_c3_v2	társe	bínso	bínde
31_5_c3_tv	ménto	sárti	sérmo
32_5_c3_a1	lúmpe	jospá	jósta
33_5_c3_a2	bundó	tálde	talpé
34_5_c3_av1	súnta	mélto	múlko
35_5_c3_av2	tonké	perká	perté
36_5_c3_str	bísle	dáblo	dángo
37_5_tc13_tc23	lésta	lónti	kósti
38_5_3c_tc13	dínke	dúnko	dúlke
39_5_tc13_v1	tíngu	sírka	tórga
40_5_tc13_v2	rósta	bónde	búste
41_5_tc13_tv	bísna	búlne	tílka
42_5_tc13_a1	férna	falnó	páldo
43_5_tc13_a2	jentó	júlta	pulká
44_5_tc13_av1	bárke	búnko	gánto
45_5_tc13_av2	rendá	risdó	tisbá
46_5_tc13_str	mínle	máklo	dárso
47_5_3c_tc23	básme	búsmo	túsmo
48_5_tc23_v1	tónse	lúrde	túrsa
49_5_tc23_v2	sáldi	pérbi	példo
50_5_tc23_tv	túrke	mórka	múnze
51_5_tc23_a1	pánte	luntí	lúsdi
52_5_tc23_a2	fustó	mésta	melgá
53_5_tc23_av1	méspa	bíspo	bérto
54_5_tc23_av2	pulká	golké	gorbá
55_5_v1_3c	pónda	górti	péndi
56_5_3c_v2	sínte	sónta	mórke
57_5_3c_tv	tárlo	tírle	másto
58_5_3c_a1	púnke	pinká	lísma
59_5_3c_a2	dintá	dénto	pergó
60_5_3c_av1	sólfi	sálfe	tóske
61_5_3c_av2	kandú	kindá	pirgú
62_5_v1_v2	párti	lánde	lóndi
63_5_v1_tv	jélbo	sénta	sénto
64_5_v1_a1	tílpa	kindá	kúnda
65_5_v1_a2	pirbó	tínka	tenká
66_5_v1_av2	sinká	mistó	mestá
67_5_v1_str	gánti	mágle	móske
68_5_v2_tv	málde	tórne	tárne
69_5_v2_a1	sórga	mendá	méndi
70_5_v2_a2	bondé	tálke	talkí
71_5_v2_av1	tónde	rúspe	róspa
72_5_v2_str	múlde	kábre	kánfo
73_5_tv_a1	góspi	toldí	tálde
74_5_tv_a2	randé	tárge	torgú
75_5_tv_av1	bírko	timpó	tímpa
76_5_tv_av2	miské	dínte	danté
77_5_tv_str	kónda	bótra	búste
78_5_a1_av1	pésta	dúrko	dérko
79_5_a1_str	mésda	portí	pótri
80_5_a2_av2	kustó	perká	perkó
81_5_a2_str	tinká	púrde	pugré
82_5_av1_str	kéndo	mírga	mégra
83_5_av2_str	fasté	turpó	tublé
84_4_c1_c2	kátu	kóbe	róte
85_4_c1_v1	sípo	sáne	kíne
86_4_c1_v2	máke	míto	Líte
87_4_c1_tc	díja	dóme	Dóje
88_4_c1_tv	pína	pébo	Tíba
89_4_c1_a1	lóga	lasé	Máse
90_4_c1_a2	pité	púro	Kuró
91_4_c1_av1	dúka	dóse	Lúse
92_4_c1_av2	letí	lomé	Bomí
93_4_c2_v1	lóri	péru	Póku
94_4_c2_v2	kábu	díbe	Dípu
95_4_c2_tc	tíso	késa	Teas
96_4_c2_tv	bóra	kíre	Kóna
97_4_c2_a1	síre	maró	Mádo
98_4_c2_a2	bagú	rígo	Risó
99_4_c2_av1	lúko	dáke	Dúre
100_4_c2_av2	daké	pokí	Pore
101_4_v1_v2	súla	múte	Mile
102_4_v1_tc	zúki	púna	Zóka
103_4_v1_tv	mópi	sóte	Sóti
104_4_v1_a1	kéla	bedó	Bído
105_4_v1_a2	tiká	piré	Pore
106_4_v1_av2	masó	palé	Puló
107_4_v2_tc	búse	táre	Báso
108_4_v2_tv	síka	bóra	Bíra
109_4_v2_a1	táro	buló	Búle
110_4_v2_a2	dolú	séru	Serí
111_4_v2_av1	mále	róse	Rási
112_4_tc_tv	kúte	káto	Dúbe
113_4_tc_a1	káli	keló	Péjo
114_4_tc_a2	puné	póna	Kodá
115_4_tc_av1	síto	sáte	Mile
116_4_tc_av2	milá	molé	Botá
117_4_tv_a1	néko	tejó	Túja
118_4_tv_a2	kasí	dári	Deró
119_4_tv_av1	ména	ketá	Kéto
120_4_tv_av2	golé	móke	Mike
121_4_a1_av1	séli	túka	Téka
122_4_a2_av2	siró	kaní	Kanó

STIMULUS SET 2			
1_5_c2_c1	lárde	pórti	Lónti
2_5_c3_c1	méldo	búsda	músta
3_5_tc13_c1	dénko	dárku	dárgu
4_5_tc23_c1	fáste	jísto	fúlgo
5_5_3c_c1	móste	másta	málka
6_5_v1_c1	bírte	mílko	bálko
7_5_v2_c1	tásli	rónti	tónte
8_5_tv_c1	kólpa	tógra	kúgre
9_5_a1_c1	léngo	mástu	lastú
10_5_a2_c1	purdá	kentí	pénti
11_5_av1_c1	sáski	tánde	súnde
12_5_av2_c1	rilkó	fengó	rengú
13_5_str_c1	móndi	pérga	mégra
14_5_c3_c2	ménto	dálti	dánsi
15_5_tc13_c2	kánde	kúldo	múnko

16_5_tc23_c2	bólda	sélde	sélte
17_5_3c_c2	tínde	tónda	mónga
18_5_v1_c2	fóste	pórgu	pásgu
19_5_v2_c2	gúlda	pónka	pólke
20_5_tv_c2	túnsa	múrka	mónke
21_5_a1_c2	sórta	mínke	mirké
22_5_a2_c2	tesní	golbá	gósba
23_5_av1_c2	bínte	dílgo	dángo
24_5_av2_c2	gandé	sulté	suntó
25_5_str_c2	lágdo	púsme	púgre
26_5_tc13_c3	pórdo	pálde	kálde
27_5_tc23_c3	túspe	góspo	górpo
28_5_3c_c3	sérgo	sárga	tásga
29_5_v1_c3	pólsi	kórma	kérsa
30_5_v2_c3	bángu	télku	télga
31_5_tv_c3	pánde	tárbe	tírdo
32_5_a1_c3	mílno	kérsa	kerná
33_5_a2_c3	nordé	mastú	másdu
34_5_av1_c3	jínfe	tílso	tálfo
35_5_av2_c3	bunkí	tesmí	teská
36_5_str_c3	tópla	gúbre	gúnle
37_5_tc23_tc13	kólta	gúlte	kúste
38_5_3c_tc13	kásla	kósle	kórle
39_5_v1_tc13	rénko	téspa	rúska
40_5_v2_tc13	másti	nóldi	mólte
41_5_tv_tc13	bírno	tísko	básne
42_5_a1_tc13	tálbe	górti	torbí
43_5_a2_tc13	kusté	milpá	kílta
44_5_av1_tc13	lírte	pínko	lánto
45_5_av2_tc13	duntá	noská	dosté
46_5_str_tc13	gábli	púkro	gúnlo
47_5_3c_tc23	nélte	nálto	pálto
48_5_v1_tc23	lórba	kónte	kírbe
49_5_v2_tc23	kólde	gánte	gáldi
50_5_tv_tc23	jándo	bálto	búndi
51_5_a1_tc23	míska	térbo	teskó
52_5_a2_tc23	kelpá	bintó	bílpo
53_5_av1_tc23	dénko	pésta	púnka
54_5_av2_tc23	perbó	fistó	firbá
55_5_v1_3c	dásli	támpe	dósle
56_5_v2_3c	tólga	sémpa	télgu
57_5_tv_3c	bálpe	tánde	bólpo
58_5_a1_3c	dólko	társe	dalké
59_5_a2_3c	lispá	fontó	lóspo
60_5_av1_3c	séngo	bésta	sangá
61_5_av2_3c	bolgó	tespó	bélga
62_5_v2_v1	gálke	mórpe	márpi
63_5_tv_v1	téspa	méspa	méspo
64_5_a1_v1	búrpo	kásde	kusdé
65_5_a2_v1	gurká	lenfí	lúnfi
66_5_av2_v1	parbó	jeldó	jaldí
67_5_str_v1	júldo	bírta	bútra
68_5_tv_v2	lónje	bósde	básde
69_5_a1_v2	dálmo	pérbi	perbó
70_5_a2_v2	tulgá	rinkó	rínka
71_5_av1_v2	bárte	sángo	singé
72_5_str_v2	tásgu	lórte	lótru
73_5_a1_tv	kánde	múspo	maspé
74_5_a2_tv	korbé	tankí	tónke
75_5_av1_tv	pálte	sánsi	sansé
76_5_av2_tv	pargá	poltá	pálta
77_5_str_tv	lósti	gárde	gódri
78_5_av1_a1	bálte	sánko	sínko
79_5_str_a1	dárko	melgá	mégla
80_5_av2_a2	kansí	poldí	poldé
81_5_str_a2	tublí	gátre	ganté
82_5_str_av1	tíbra	mókre	mírke
83_5_str_av2	muspá	kertó	ketrá
84_4_c2_c1	góbe	mábi	gáfi
85_4_v1_c1	bátu	lájo	béjo
86_4_v2_c1	túka	méla	téli
87_4_tc_c1	túka	téke	tépe
88_4_tv_c1	méso	lébo	míba
89_4_a1_c1	lémo	kúbi	lubí
90_4_a2_c1	korí	madú	kádu
91_4_av1_c1	pábe	gári	póri
92_4_av2_c1	siná	delá	seló
93_4_v1_c2	dúte	súra	síta
94_4_v2_c2	lábe	jóne	jóbi
95_4_tc_c2	bálo	béli	séli
96_4_tv_c2	dóke	mópe	múka
97_4_a1_c2	míne	bója	boná
98_4_a2_c2	kudí	tepó	tédo
99_4_av1_c2	rúba	múto	mébo
100_4_av2_c2	kabí	nerí	nebú
101_4_v2_v1	téra	múga	mégo
102_4_tc_v1	púka	póke	dúle
103_4_tv_v1	pále	káme	kámo
104_4_a1_v1	téga	níbo	nebó
105_4_a2_v1	buré	kotí	kúti
106_4_av2_v1	mogá	lipá	lopé
107_4_tc_v2	súti	sáto	gáli
108_4_tv_v2	mílo	subjo	síjo
109_4_a1_v2	fóre	náki	naké
110_4_a2_v2	pefó	dulá	dúlo
111_4_av1_v2	sáre	tálu	tolé
112_4_tv_tc	jíne	kíle	jáno
113_4_a1_tc	náse	póbi	nosí
114_4_a2_tc	kepú	fanó	kápo
115_4_av1_tc	méja	pébo	mújo
116_4_av2_tc	telí	madí	tuló
117_4_a1_tv	gópe	dúsa	dosé
118_4_a2_tv	padí	tojé	táji
119_4_av1_tv	goté	póla	polé
120_4_av2_tv	betá	tósa	tésa
121_4_av1_a1	múna	túpe	típe
122_4_av2_a2	badé	romé	romí

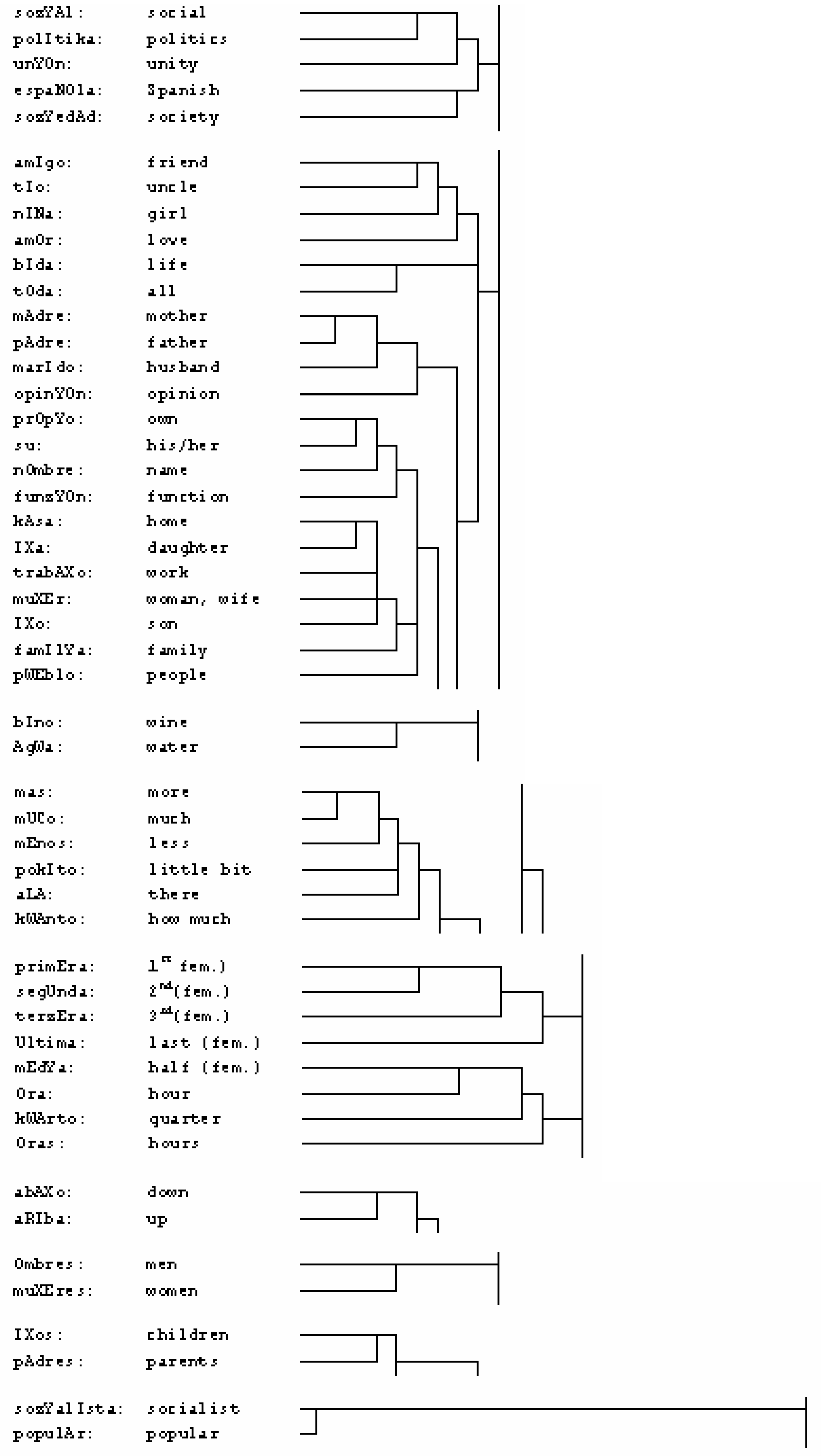
Appendix C

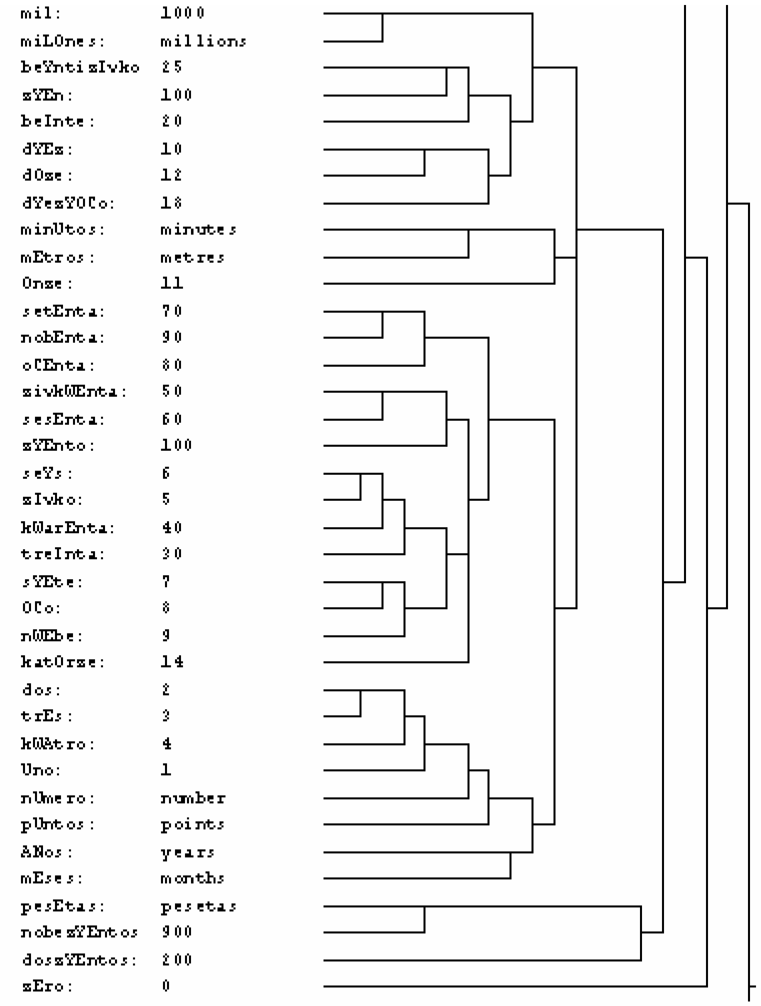
The 31 cvcv words stressed on the last syllable in the 324-word list (74% verbs, 16% nouns, 6% proper nouns, 3% adverbs) used for comparison in chapter four.

word	**ps**	**tense**	**translation**
pasó	v	past	*it happened*
llegó	v	past	*he arrived*
quedó	v	past	*he stayed, remained*
tocó	v	past	*he touched*
llevó	v	past	*he carried*
llamó	v	past	*he called*
dejó	v	past	*he let, left*
ganó	v	past	*he won*
cayó	v	past	*it/he fell*
miró	v	past	*he looked at*
sacó	v	past	*he took out*
José	pn		*Jose (man's name)*
chalé	n		*chalet*
café	n		*coffee*
diré	v	fut	*I will say*
pasé	v	past	*I passed*
llamé	v	past	*I called*
llegué	v	past	*I arrived*
quedé	v	past	*I stayed, remained*
papá	n		*daddy*
mamá	n		*mummy*
quizá	adv		*perhaps*
será	v	fut	*it will be*
verá	v	fut	*he will see*
dirá	v	fut	*he will say*
dará	v	fut	*he will give*
cogí	v	past	*I took*
metí	v	past	*I put into*
salí	v	past	*I went out*
menú	n		*menu*
Perú	pn		*Peru*

Appendix D

Examples of semantically related words captured in dendrogram clusters. Hierarchical clustering of vectors based on cooccurrence in the surface-form corpus using content words only as context words.




mil: 1000
miLOnes: millions
beYntizIvko 25
zYEn: 100
beInte: 20
dYEz: 10
dOze: 12
dYezYOCo: 18
minUtos: minutes
mEtros: metres
Onze: 11
setEnta: 70
nobEnta: 90
oCEnta: 80
zivkWEnta: 50
sesEnta: 60
zYEnto: 100
seYs: 6
zIvko: 5
kWarEnta: 40
treInta: 30
sYEte: 7
OCo: 8
nWEbe: 9
katOrze: 14
dos: 2
trEs: 3
kWAtro: 4
Uno: 1
nUmero: number
pUntos: points
ANos: years
mEses: months
pesEtas: pesetas
nobezYEntos 900
doszYEntos: 200
zEro: 0

Appendix E

Lists of ‘person nouns’ referred to in 4.2.3, in the surface-form and the lemmatised versions of the corpus, with their English translations.

Surface-form

gente	*people*
hija	*daughter*
hijos	*children*
hombre	*man*
madre	*mother*
mujer	*woman*
mujeres	*women*
niño	*child*
niños	*children*
padre	*father*
persona	*person*
personas	*people*
señor	*sir, man*
tío	*uncle*

Lemmatised

abogado	*lawyer*
alcalde	*mayor*
amigo	*friend*
chico	*boy*
ciudadano	*citizen*
don	*Mr*
doña	*Mrs*
gente	*people*
hermano	*brother*
hijo	*son*
hombre	*man*
madre	*mother*
marido	*husband*
ministro	*minister*
mujer	*woman*
niño	*boy, child*
padre	*father*
pareja	*couple*
persona	*person*
presidente	*president*
pueblo	*people*
rey	*king*
santo	*saint*
señor	*sir, man*
tío	*uncle*

Appendix F

Rankings by phon-sem correlation (measured as Fisher divergence *FD*; low values indicate high correlations) of the 252 cvcv and the 146 cvccv words (‘no syntax’ condition). Part of speech information also shown.

CVCV WORDS			
rank	**word**	**ps**	**FD**
1	XosE	pn	0.0085
2	fAse	n	0.0103
3	XOse	pn	0.0128
4	CalE	n	0.0129
5	ganO	v	0.0129
6	sEde	n	0.0133
7	fIla	n	0.014
8	dOze	num	0.014
9	zEro	num	0.0144
10	mitA	n	0.0157
11	pUta	n	0.0158
12	sUbe	v	0.0165
13	dAme	v	0.0167
14	sOto	pn	0.0167
15	mIre	v	0.0167
16	LEge	v	0.017
17	lUna	n	0.0173
18	fECa	n	0.0174
19	sERa	pn	0.0178
20	berA	v	0.0186
21	XAbi	pn	0.0188
22	pEpe	pn	0.0189
23	deXO	v	0.0191
24	tokO	v	0.0192
25	lUCa	n	0.0194
26	LamO	v	0.0197
27	mOdo	n	0.0199
28	LEna	adj	0.0201
29	kUya	p-pr	0.021
30	pIko	n	0.021
31	tIra	v	0.0214
32	RAma	n	0.0217
33	CIna	pn	0.022
34	ROma	pn	0.0221
35	tUbe	v	0.0222
36	kOpa	n	0.0222
37	lIga	n	0.0223
38	kUyo	p-pr	0.0224
39	lUXo	n	0.0225
40	kOXa	v	0.0225
41	mAyo	pn	0.0225
42	kOma	v	0.0225
43	LegO	v	0.0226
44	bEso	n	0.0226
45	bOto	n	0.0227
46	tOke	v	0.0228
47	mAri	pn	0.0228
48	lObo	n	0.023
49	lAdo	n	0.0234
50	nOCe	n	0.0234
51	dirE	v	0.0234
52	sUya	p-pr	0.0234
53	ROka	n	0.0236
54	zIne	n	0.0237
55	tIro	n	0.0239
56	bIbe	v	0.0241
57	tORe	n	0.0241
58	bANo	n	0.0242
59	pAgo	v	0.0243
60	WEko	n	0.0244
61	dAma	n	0.0245
62	bAle		0.0245
63	zIta	n	0.0246
64	pIde	v	0.0247
65	mAsa	n	0.0247
66	XEfe	n	0.0247
67	lOli	pn	0.0251
68	dANo	n	0.0253
69	dAto	n	0.0254
70	kUlo	n	0.0255
71	nINa	n	0.0256
72	dIme	v	0.0256
73	dUra	adj	0.0257
74	kOno		0.0257
75	bAse	n	0.0257
76	LEbe	v	0.0257
77	mUCa	adj	0.0257
78	kIlo	n	0.0261
79	bIbo	v	0.0263
80	LEno	adj	0.0263
81	sIge	v	0.0264
82	bALe	n	0.0264
83	kayO	v	0.0266
84	pAse	v	0.0267
85	kAsi	adv	0.027
86	pUso	v	0.0271
87	RIsa	n	0.0272
88	LebO	v	0.0272
89	CIka	n	0.0273
90	bIda	n	0.0274
91	fOto	n	0.0274
92	YElo	n	0.0275
93	menU	n	0.0275
94	tIpo	n	0.0277
95	tUya	p-pr	0.0279
96	kafE	n	0.0279
97	bEte	v	0.0282
98	gERa	n	0.0283
99	pEso	n	0.0283
100	ROLo	n	0.0283
101	bEra	pn	0.0284
102	bAXo	v	0.0285
103	kAda	adj	0.0288
104	kAbo	n	0.0288
105	gAto	n	0.0289
106	kedO	v	0.0289
107	sAka	v	0.0292
108	tEla	n	0.0293
109	mEsa	n	0.0294
110	papA	n	0.0295
111	pEga	v	0.0297
112	mOda	n	0.0297
113	tOda	i-pr	0.0298
114	kAro	adj	0.0298
115	tOno	n	0.0299
116	kALe	n	0.0299
117	kUra	n	0.03
118	LEba	v	0.0302
119	zOna	n	0.0302
120	RIko	adj	0.0303
121	tApa	n	0.0303
122	mUCo	adv	0.0304
123	RIka	adj	0.0305
124	dUro	adj	0.0306
125	bIno	v	0.0307
126	sEko	adj	0.0309
127	sILa	n	0.0309

128	koXI	v	0.031
129	tUbo	n	0.031
130	mamA	n	0.0312
131	mEte	v	0.0313
132	RAto	n	0.0315
133	nINo	n	0.0315
134	kizA	adv	0.0316
135	tEma	n	0.0318
136	serA	v	0.0318
137	tEle	n	0.0319
138	sIgo	v	0.0319
139	pOne	v	0.0319
140	dIXe	v	0.0324
141	tUyo	p-pr	0.0324
142	CIko	n	0.0329
143	LEbo	v	0.033
144	sAko	v	0.0331
145	mOto	n	0.0332
146	pOka	adj	0.0332
147	tEre	pn	0.0336
148	sAle	v	0.0339
149	sUyo	p-pr	0.034
150	lECe	n	0.0342
151	LEga	v	0.0344
152	mIsa	n	0.0344
153	kALa	v	0.0344
154	pAga	v	0.0345
155	pIso	n	0.0345
156	dEbo	v	0.0347
157	sEpa	v	0.0348
158	bAXa	v	0.0348
159	mAlo	adj	0.0349
160	ROpa	n	0.0349
161	dIga	v	0.0349
162	dIze	v	0.035
163	kONo	n	0.035
164	pAro	v	0.035
165	pIdo	v	0.0352
166	sAla	n	0.0353
167	gAna	v	0.0353
168	kApa	n	0.0353
169	zEna	n	0.0354
170	RAfa	pn	0.0354
171	mIra	v	0.0354
172	bOka	n	0.0355
173	pEna	n	0.0356
174	dIXo	v	0.0356
175	sIga	v	0.0357
176	dICo	v	0.036
177	dEbe	v	0.0361
178	dAba	v	0.0361
179	pERo	f	0.0362
180	sIda	n	0.0362
181	pasO	v	0.0363
182	RAro	adj	0.0363
183	dEXo	v	0.0363
184	dIgo	v	0.0364
185	lOka	adj	0.0364
186	mAta	v	0.0365
187	kEso	n	0.0369
188	kOla	n	0.0371
189	sAbe	v	0.0372
190	dEXe	v	0.0372
191	pAko	pn	0.0372
192	kAXa	n	0.0372
193	mAXa	adj	0.0373
194	mOno	n	0.0374
195	ROto	adj	0.0377
196	dONa	n	0.0378
197	dUda	v	0.0379
198	kOXe	v	0.038
199	kOko	n	0.038
200	tOdo	i-pr	0.038
201	fALa	v	0.0382
202	LAma	v	0.0383
203	sIdo	v	0.0384
204	lOko	adj	0.0384
205	kOCe	n	0.0384
206	mAno	n	0.0386
207	sOla	adj	0.039
208	kEda	v	0.039
209	bAya	v	0.039
210	mACo	n	0.039
211	pOko	adv	0.0391
212	kEdo	v	0.0391
213	LAmo	v	0.0392
214	ROsa	adj	0.0394
215	lAta	n	0.0395
216	bOda	n	0.0396
217	dAdo	v	0.0397
218	kOXo	v	0.0398
219	tOma	v	0.04
220	mAla	adj	0.0401
221	pAlo	n	0.0401
222	pUro	adj	0.0402
223	kOme	v	0.0402
224	ROXa	adj	0.0407
225	ROXo	adj	0.0409
226	pAso	v	0.0413
227	mEto	v	0.0415
228	kAma	v	0.0416
229	YERo	n	0.0421
230	kAza	v	0.0422
231	pAra	f	0.0423
232	dEXa	v	0.0424
233	pUra	adj	0.0426
234	kIta	v	0.0427
235	pElo	n	0.0429
236	kOsa	n	0.043
237	pUdo	v	0.0433
238	kAsa	n	0.0438
239	sIno	f	0.0442
240	kEde	v	0.0444
241	nOta	n	0.0452
242	LEgo	v	0.0453
243	tOka	v	0.0454
244	fALo	n	0.0458
245	nAda	i-pr	0.046
246	kAso	n	0.0472
247	kAbe	v	0.0474
248	pAsa	v	0.0493
249	pEro	n	0.0501
250	sOlo	f	0.0517
251	kAra	n	0.0522
252	kOmo	i-pr	0.0529

CVCCV WORDS

rank	word	ps	FD
117	moskU	pn	0.0084
105	REnfe	pn	0.0103
24	kInze	num	0.0111
93	lInze	n	0.0112
43	gOlfo	n	0.0115
114	XOrdi	pn	0.0116
7	zIvko	num	0.0118
90	tOrno	n	0.0123
72	zIfra	n	0.0126
69	sIgno	n	0.0129
120	bAsko	adj	0.0134
143	dUlze	adj	0.0135
145	RItmo	n	0.0137
36	madrI	pn	0.0148
132	karnE	n	0.0152
119	pArla	pn	0.0152
99	bOlsa	n	0.0155
33	sIglo	n	0.0158
127	REkta	n	0.0159
2	dEsde	f	0.0164
26	bIsta	n	0.0164
134	kOnCa	pn	0.0168

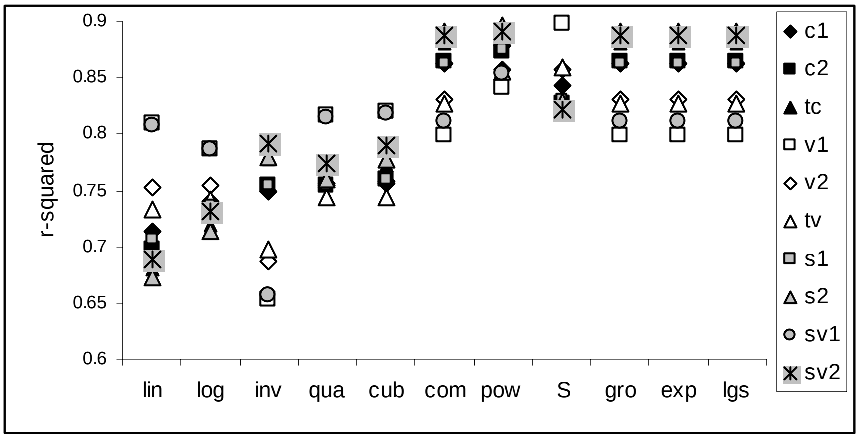
68	pAblo	pn	0.0172
128	pensE	v	0.0175
144	sUsto	n	0.0179
131	XEsto	n	0.0183
123	pOlbo	n	0.0187
95	dIsko	n	0.0192
133	sObra	n	0.0192
100	gOrdo	adj	0.0193
4	sObre	f	0.0195
110	kArgo	n	0.0195
96	bOmba	n	0.0197
31	zErka	adv	0.0197
23	pAdre	n	0.0198
42	kUrso	n	0.0203
6	pArte	n	0.0206
113	pUnta	n	0.0206
18	mAdre	n	0.0212
115	tUrno	n	0.0214
122	gustO	n	0.0214
130	pOnte	v	0.0214
92	mEnte	n	0.0216
20	tArde	n	0.0216
63	bErbo	n	0.0216
59	XUnto	adj	0.0217
102	fIvka	n	0.0218
80	kOsta	n	0.0219
111	kostO	v	0.022
81	REnta	n	0.0222
37	REsto	n	0.0223
73	bUska	v	0.0224
64	gOlpe	n	0.0225
118	nEgra	adj	0.0226
45	mArCa	n	0.0228
53	dArle	v	0.023
29	lArgo	adj	0.023
89	bAnda	n	0.023
125	bErla	v	0.0232
11	mUndo	n	0.0232
15	bEvga	v	0.0232
40	nOrte	n	0.0232
146	kontO	v	0.0235
104	nOrma	n	0.0236
60	lIbre	adj	0.0236
32	pEdro	pn	0.024
129	XOrXe	pn	0.024
75	pIsta	n	0.024
84	sAnto	n	0.0241
94	dArse	v	0.0244
71	bErde	adj	0.0244
49	mArta	pn	0.0245
54	sAnta	n	0.0246
83	sAlga	v	0.0251
142	gOrda	adj	0.0251
52	XUnta	n	0.0254
108	bEnde	v	0.0255
67	kOrte	n	0.0258
112	CIste	n	0.0262
126	pAkto	n	0.0263
76	mArko	n	0.0263
56	pObre	adj	0.0267
12	fOrma	n	0.0267
106	sEkso	n	0.0268
121	kAsko	n	0.0268
25	tEvga	v	0.027
91	pAlma	n	0.0272
19	gUsta	v	0.0273
136	lIsto	adj	0.0273
57	dOble	n	0.0273
5	tEvgo	v	0.0276
3	dOnde	i-pr	0.0276
51	mEtro	n	0.0276
141	ROmpe	v	0.0276
77	bEvgo	v	0.0277
50	gUsto	v	0.028
140	lEtra	n	0.0281
70	zInta	n	0.0284
61	podrA	v	0.0285
35	pOvgo	v	0.0286
85	bErlo	v	0.0289
88	kOrta	v	0.0289
138	kInta	adj	0.0292
116	mAndo	v	0.0295
62	nEgro	adj	0.0295
97	mAnCa	n	0.0296
30	fOndo	n	0.0296
79	kInto	adj	0.0308
22	mIsma	i-pr	0.0311
82	porkE	f	0.0312
17	fAlta	v	0.0312
86	sAlbo	adv	0.0315
9	mIsmo	i-pr	0.0315
109	tArda	v	0.0316
139	kArga	n	0.0317
78	kOrto	adj	0.0317
8	XEnte	n	0.0321
21	berdA	n	0.0321
13	pUnto	n	0.0323
46	bEnta	n	0.0324
65	lArga	adj	0.0325
107	kAlma	n	0.0325
28	kAmpo	n	0.0326
103	tOnto	adj	0.0328
34	bArko	n	0.0329
16	nUvka	adv	0.0329
101	mAnda	v	0.033
41	sIrbe	v	0.0341
66	kUlpa	n	0.0345
135	sErlo	v	0.0347
27	lIbro	n	0.0348
58	kArne	n	0.0349
48	lIsta	adj	0.0355
98	bAvko	n	0.0356
47	tAnta	adj	0.0357
87	pInta	n	0.0357
39	dAndo	v	0.0357
38	kArta	n	0.0363
74	pOvga	v	0.0377
44	XUsto	adv	0.0383
1	pOrke	f	0.0389
55	mArka	n	0.039
124	kAnta	v	0.042
137	bAsta	v	0.0439
14	bIsto	v	0.0452
10	tAnto	adv	0.0463

Appendix G

A measure of the prediction power of all the regression functions in the cvcv word-group parameters. Numerical values and plots shown for the 'syntax' condition (this page) and the 'no syntax' condition (next page).

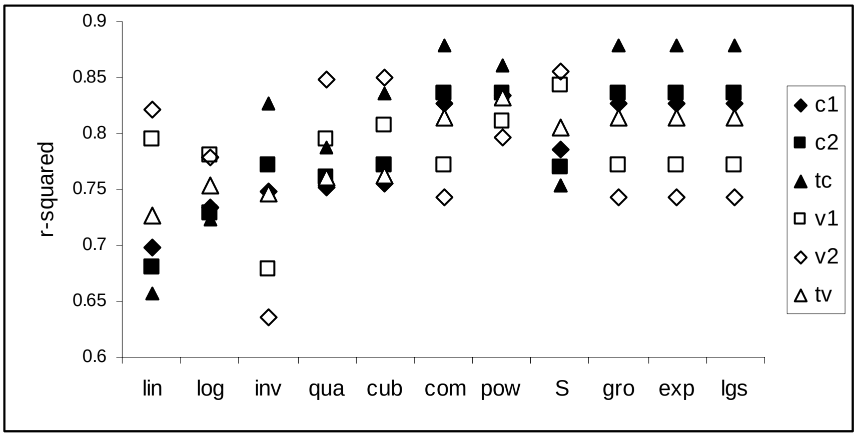
'Syntax' condition:

r^2	c1	c2	tc	v1	v2	tv	s1	s2	sv1	sv2
lin	0.713	0.698	0.679	0.809	0.753	0.733	0.706	0.672	0.808	0.689
log	0.74	0.73	0.719	0.787	0.755	0.742	0.736	0.714	0.786	0.731
inv	0.749	0.754	0.777	0.653	0.687	0.697	0.754	0.78	0.657	0.792
qua	0.755	0.754	0.762	0.817	0.756	0.743	0.755	0.759	0.815	0.774
cub	0.758	0.759	0.773	0.82	0.756	0.743	0.76	0.777	0.818	0.79
com	0.863	0.864	0.881	0.799	0.831	0.828	0.863	0.892	0.811	0.887
pow	0.878	0.874	0.887	0.842	0.857	0.855	0.875	0.897	0.854	0.891
S	0.843	0.827	0.823	0.899	0.858	0.859	0.831	0.829	0.905	0.822
gro	0.863	0.864	0.881	0.799	0.831	0.828	0.863	0.892	0.811	0.887
exp	0.863	0.864	0.881	0.799	0.831	0.828	0.863	0.892	0.811	0.887
lgs	0.863	0.864	0.881	0.799	0.831	0.828	0.863	0.892	0.811	0.887



'No syntax' condition:

r^2	c1	c2	tc	v1	v2	tv
lin	0.698	0.68	0.657	0.795	0.822	0.727
log	0.734	0.728	0.723	0.781	0.779	0.753
inv	0.748	0.772	0.826	0.679	0.636	0.747
qua	0.752	0.76	0.787	0.795	0.848	0.761
cub	0.755	0.771	0.835	0.807	0.85	0.762
com	0.827	0.835	0.878	0.771	0.743	0.814
pow	0.834	0.835	0.861	0.811	0.797	0.832
S	0.785	0.77	0.754	0.842	0.855	0.805
gro	0.827	0.835	0.878	0.771	0.743	0.814
exp	0.827	0.835	0.878	0.771	0.743	0.814
lgs	0.827	0.835	0.878	0.771	0.743	0.814



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.

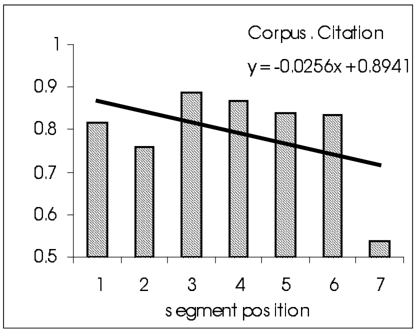


Figure 1: Information profile of 7-segment words from the citation transcription of the speech corpus.

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/, /ɟ/, /χ/, /l/, /ʎ/, /tʃ/.

Fast-speech transcription: The above plus semivowel /i/, /u/, voiced approximants /β/, /ð/, /ɣ/, voiceless approximants /β/, /ð/, /ɣ/, labiodental /m/, dental /n/ and /l/, palatalised /n/ and /l/, velarized /n/, /z/, dental voiced /s/, dental /s/, fricative /ɾ/, 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.

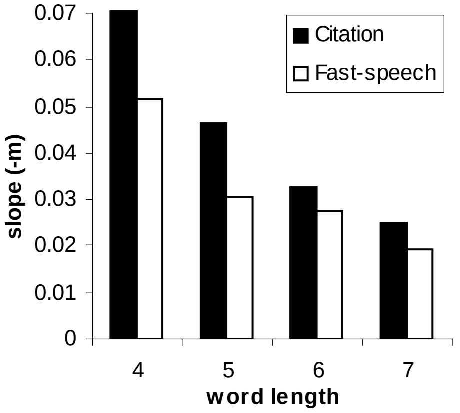


Figure 2: Slopes of the information profiles of the citation and the fast-speech transcriptions applied to the corpus, over the four word lengths.

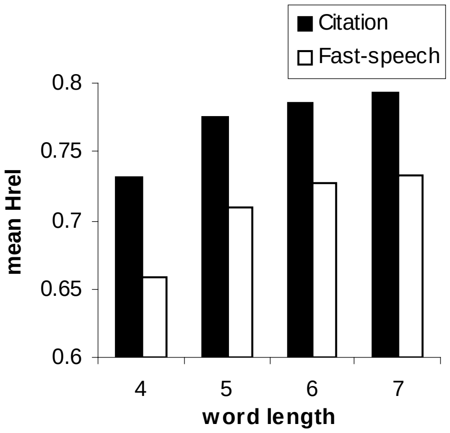


Figure 3: Mean relative entropy of the citation and fast-speech transcriptions over the four word lengths.

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.

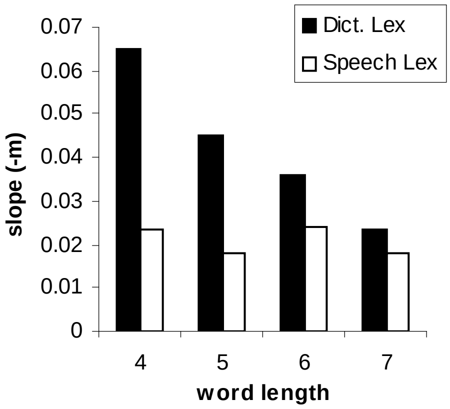


Figure 4: Slopes of the information profiles of the dictionary lexicon and the speech lexicon over the four word lengths.

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.

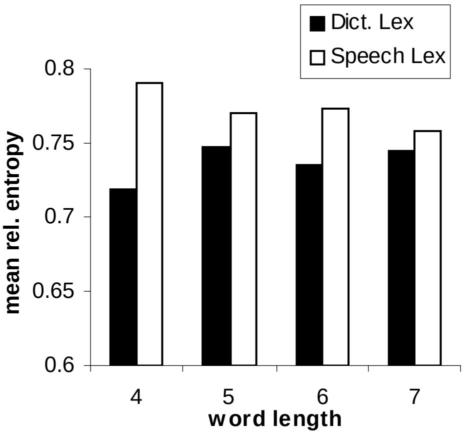


Figure 5: Mean relative entropy of the dictionary lexicon and the speech lexicon over the four word lengths.

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.

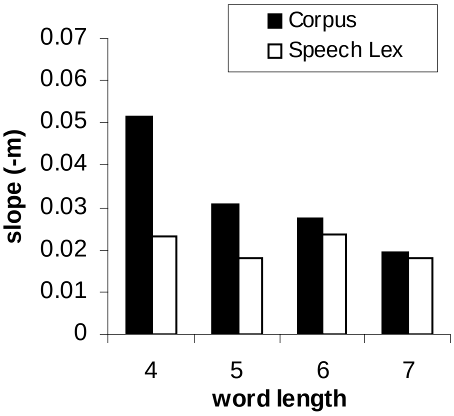


Figure 6: Slopes of the information profiles of the corpus and the speech lexicon across the four word lengths.

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.

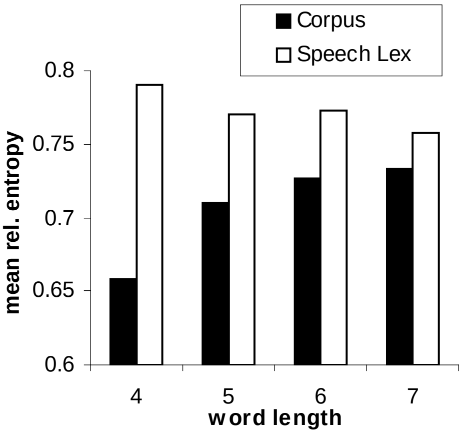


Figure 7: Mean relative entropy of the corpus and the speech lexicon across the four word lengths.

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.