Exploring the Relationship Between Word Form and Word Meaning

Sarenne Wallbridge

4th Year Project Report
Cognitive Science
School of Informatics
University of Edinburgh

2018

Abstract

Though a word may be the most intrinsic unit of language, it can be separated into 2 distinct components - form and meaning. The goal of this project is to explore how these 2 components are related, firstly by quantifying the form-meaning relationship itself though the use of statistical analysis, and secondly by examining potential explanations for such a relationship from a cultural evolution perspective by simulating language acquisition. The first strand of this project reveals that the form-meaning relationship is extremely arbitrary but more complex that reported in previous research, highlights potential flaws in previous methodologies, and provides direction for future work. The second strand of this project presents experimental and theoretical evidence which suggests that such arbitrariness may facilitate both language acquisition and use, thus providing partial explanation for the high degree of arbitrariness in the form-meaning relationship.

Acknowledgements

First and foremost, I would like to thank my supervisor, Sharon Goldwater. I'm very grateful for all of your encouragements and guidance throughout the year which made this project an extremely valuable personal experience.

A special thanks also goes to Hanne Carlsson, with whom I completed the first half of this project, for making this project a highlight of my final year – your ideas and assistance were invaluable.

Finally, I would like to thank my parents for their constant support.

Table of Contents

Chapter 1

Introduction

1.1 Motivation

A word – arguably the most intrinsic unit of language – is made up of 2 fundamental components: its form and its meaning. Form is essentially the *shell* of a word, the way it looks and sounds, while meaning is its *core*, its denotations, connotations, and associations. Interest in the form-meaning relationship stretches back for centuries to scholars in the late 1600's such as Wilkins who attempted to create a 'perfect' systematic language [102], and even earlier to Shakespeare's infamous line:

> *'[...] That which we call a rose*
> *By any other word would smell as sweet'* — Romeo & Juliet, Shakespeare W.

The motivation for this study stems from one of the most prominent ideas within the study of the form-meaning relationship: Arbitrariness of the Sign. The notion was put forward over a century ago by the Swiss linguist De Saussure and states that a word's form is completely unrelated to its meaning [31].

There are widely-accepted exceptions to the notion, notably sound symbolisms like onomatopoeic words who's forms imitate their referent like 'bang' and 'splash', and phonaesthemes, a term coined by Firth to describe a sound systematically related to a specific meaning such as the *'gl'* set relates to an aspect of light ('glow', 'glitter', 'gleam', etc) [38]. However, the vast majority of linguistic models of human language have assumed an arbitrary relationship between form-meaning mappings and it is even considered by some linguists to be a fundamental design principle of language [79] [53].

Computational and cognitive linguistics offers the means to study the form-meaning relationship quantitatively by applying powerful statistical methods to language. Recent studies have used such methods to provide evidence against Arbitrariness of the Sign, reporting diffuse but statistically significant correlation between form and meaning across the mental lexicon, referred to throughout this study as global *phonosemantic systematicity* [47] [90] [74] [43] [93] [72]. Such findings have implications for how cognitive processes like language acquisition, use and multi-sensory integration are modelled, as well as for our understanding of the neural mechanisms which drive the organisation and storage of language.

Research with a cognitive basis suggests that a richer relationship between form and meaning is driven by a deep-seated relationship between human neural mechanisms for language and language itself [21]. Such theories compare language to a dynamic organism that adapts and evolves under the pressures human cognition, and the ever-changing environment [58]. They suggest that features of language are poised to simultaneously optimise cognitive processes like language acquisition, expression, and comprehension which are often rely on opposing properties of vocabulary structure. Gaining deeper insight into the form-meaning relationship will provide better understanding of the processes that shape it.

1.2 Outline

This project involved 2 quite distinct, but highly related components - the first involves a quantitative analysis of the relationship between form and meaning in a representative subset of English; the second explores how this relationship may impact and interact with the use and acquisition of language.

Analysing the form-meaning relationship requires a thorough understanding of form and meaning, and how they can be studied to extract information about their underlying relationship. Chapter 2 provides a summary of relevant aspects of semantics and form, as well as a critical review of similar studies which highlights some of the challenges to studying the form-meaning relation and is used to motivate the design of the current experiment. The results of the current experiment were only partially aligned with previous findings and shed light on some potential flaws in previous research, namely the applicability of the statistical analysis performed in previous studies, and the validity of conclusions drawn about the global nature of phonosemantic systematicity. Analysis of current and previous results culminate to conclusions that the global form-meaning relation is more difficult to analyse that previously reported, most likely due to oversimplification of the effects of local pockets of phonosemantic systematicity on the global form-meaning relationship. All findings provide evidence that global vocabulary structure is highly arbitrary. The experimental design, results, and conclusions are presented in Chapter 3.

The second component of this study focuses on *why* such an arbitrary vocabulary structure might exist. Chapter 4 provides an overview of theories for how cognitive processes and environmental factors have shaped, and are shaped by the form-meaning relationship from a cultural evolutionary perspective [58]. Such theories provide a basis for the experimentation presented in 5; simulations of language acquisition were designed to test 2 hypotheses regarding how acquisition is affected by the size of the vocabulary and the integration of contextual cues which encode semantic information. Learning is generally supported by structure and thus systematic vocabularies would be expected to enhance language acquisition. Both hypotheses present theories for how arbitrary form-meaning mappings may provide an advantage to language acquisition and thus provide some explanation for the high degree of arbitrariness in form-meaning mappings. Chapter 5 presents the different experimental designs, results, and conclusions. Investigations into each hypothesis demonstrated that integrating useful contextual information during acquisition produces an advantage to acquisition when learning arbitrary mappings, and that vocabulary size affects how quickly this advantage occurs; learning trends were found to be robust to different phonological feature representations but sensitive to language features like phonological set variation.

1.3 Contributions

Contributions can be split into the 2 components of this study. The first component involved collaboration with Hanne Carlsson, as detailed below.

Phonosemantic Systematicity This section was based on Shillcock et al. [90] which introduced a number of novel methods for quantifying the relationship between form and meaning.

- Implemented Shillcock et al.'s methods for measuring lexicon-wide phonosemantic systematicity, and testing it's significance with several modifications including enhanced representation of word semantics and alternative phonological representation
- Designed 5 phonological form representations which required the production of linguistically-sound phonological feature vectors; the best metric involved adapting the Levenshtein algorithm to produce edit distances between sets of vectors.
- Established strong linguistic and computational basis for implementational modifications and design choices.
- Obtained results that oppose several similar papers, highlighting complexities in the form-meaning relationship that weren't considered in previous research [74] [93].
- In-depth, critical analysis of the methodology and conclusions from previous studies based on the current results and related research.

Parts of this section were completed in collaboration with Hanne Carlsson, specifically the basic implementation of Shillcock et al.'s method, and the implementation of vector-based phonological form representations.

Language Design This section focused on the cognitive aspect of vocabulary structure and its relation to the language acquisition process.

- Successfully reimplemented language acquisition simulations to test the hypothesis that arbitrary vocabulary mappings maximise the effect of contextual information on language acquisition [73].
- Tested the validity and robustness of simulations by making phonologically-based improvements.
- Experimentally evaluated the hypothesis that vocabulary size indirectly affects vocabulary structure in conjunction with theoretical evidence.
- Provided evidence that arbitrary form-meaning mappings may facilitate language acquisition.
- Critically analysed simulations, especially with respect to realistic language acquisition.

Chapter 2

Background: Phonosemantic Systematicity

The overall aim of this project is to examine the links between how words are represented in 2 respective feature spaces, form and meaning, both of which can be defined in many ways.

The first strand of this study involves quantifying this relationship. This requires a deep understanding of form and meaning, and how they can be studied to extract information about their relationship. This section begins by detailing *word form* and corresponding representational methods, followed by *semantics*, particularly how they can be represented, and which aspects of semantics are important in the study of form and meaning, and finally the notion upon which this study is based (*Arbitrariness of the Sign*). The chapter finishes with a review of related research to inspire the experiment design detailed in the following chapter.

2.1 Form

Form acts as a word's superficial *shell* - the way it looks and sounds. The procedures for quantifying form used through out this study involve the use of phonological and orthographic features, both of which will be detailed below.

2.1.1 Phonology

Though language has multiple mediums, spoken sounds must be the most elementary mode of communication. Phonology involves the study of systematic use of sounds as linguistic items across and between languages [61]. Though closely related, it is important to draw a distinction between phonetics and phonology; phonetics is concerned with the physical aspect of sound production and perception, while phonology deals in a more abstract component of sound, specifically how different sounds in a language can establish different meanings [61].

Phonemes are language-specific sound elements that distinguish words from one another. Though they are often often used as the basic building blocks of phonology, phonology can be studied at representational levels both above and below phonemes – syllables and distinctive features, respectively [61]. Both units are used throughout this study to build phonological word representations and are described below.

2.1.1.1 Syllables

Though phonologists from a range of domains have long accepted the importance of syllable units, there is little agreement regarding the specific definition of a syllable. This highlights the complexity imbued throughout natural language - there is evidence that neural mechanisms for processing syllabic information are already established in children of only a few months old and that syllabic segmentation is easier than phoneme segmentation [41] [63], however definite characterisation of the concept remains blurry.

A general interpretation of syllables used throughout this project is as the structural units that determine the melodic organisation of phonological strings, influencing rhythm, tone, and stress patterns of a language [45]. Though this study is interested in the form-meaning relationship across all language, examining all words of a language is impossible. To obtain a representative set of words, an important aspect of the strategy used throughout this study is to consider frequently-used words (further details in section 3.1.1) [90] [74]. Syllabic filtering is a useful tool here; empirical evidence demonstrates that 50 of the most frequent words in the 100-million-word British National Corpus are monosyllabic [36]. Further theoretical support comes from Zipf's Law of Abbreviation, which states that frequent words are shorter [81] [104]. He argued that such a property is characteristic of human lexical systems to optimise communicative efficiency by maximising concision while minimising communicative effort expended.

Syllables are comprised of combinations of phonemes. Though syllable structure varies between languages, a typical syllable model includes 2 components: the onset consonant which can be obligatory, optional, or restricted, and the rime depicted in figure 2.1. The rime consists of the nucleus vowel, optionally followed by a coda consonant [96].

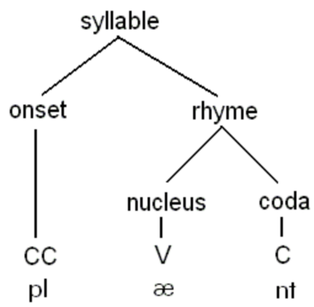


Figure 2.1: Syllable Structure of the word *plant* [23]

2.1.1.2 International Phonetic Alphabet

The International Phonetic Alphabet (IPA) is a system of phonetic notation devised to allow a standardised representation of sound in languages [6]. The framework currently includes 107 letters and 52 diacritic for describing pronunciation accents, combinations of which can be used to represent phonemes and syllables.

2.1.1.3 Distinctive Features

Each IPA symbol can be uniquely identified by a particular setting of phonological features specifying the associated articulation and acoustics [24]. First formalised by Jakobson in 1941 and further developed by Chomsky and Halle, distinctive features are one of the most prominent set of phonological features [19]. Though variations exist, traditional features are binary to indicate feature presence or absence. This project based phonological feature representations on Riggle's Phonological Feature Chart which includes 23 distinctive features for 72 IPA symbols, as well as additional information about articulatory classes, symbol variants, vowel structures, and diacritics [84]. Features of the chart take values of $\{+,-,0\}$, which were encoded for this project as $\{1,-1,0\}$ (full chart is included as figure A.2 in Appendix A)

Distinctive features are most commonly used to characterise phonemes but the features themselves can be grouped into subcategories: major class, laryngeal, manner, and place [60].

IPA Consonants The manner class of distinctive features describes the configuration of speech organs (the tongue, lips, palate) to create a specific sound and is an important feature set for classifying consonants [50]. Common categories of IPA consonants based on the manner features include:

- Plosive: consonants which cause an interruption to airflow and sound
- Fricative: consonants which involve a severe narrowing of the airflow
- Affricate: combinations of plosive and fricative sounds
- Nasal: consonants involving some airflow through the nose to maintain sound throughout oral airflow interruptions.

IPA Vowels Similarly, the set of Place features are often used to distinguish vowel classes using high, low, front, back, and tensity features so describe the position of the tongue during articulation [51].

Chart A.2 displays such categorisations for both consonants and vowels.

2.1.2 Orthography

Rather than sound, orthography is concerned with visual, written language. A large component of orthography revolves around the ordering of orthographic units - graphemes. Different writing systems use graphemes to symbolise varying degrees of representation: logographic systems such as Chinese involve graphemes that represent morphemes, syllabic system graphemes map to syllables of languages such as Japanese or Cherokee, while most European languages make use of alphabetic graphemes that roughly correspond to phonemes [67].

An interesting characteristic of alphabetic orthographies is orthographic depth. This notion characterises the correspondence between the graphemes of an orthography and the phonemes of the corresponding phonological language; shallow orthographies like Spanish consist of letters isomorphic to phonemes, while the letter-phoneme relationship is much less direct in deeper orthographies like English, where letters can have multiple pronunciations, and phonemes can be written in varying ways [42]. DeFrancis argued that the function of writing systems is to represent spoken language, and thus the psychological mechanisms required to process a writing system are constrained by those required for processing the spoken language. The set of phonemes and morphemes of a spoken language are therefore major constraints to the development of corresponding writing systems [32]. Though phonological form is the main focus of this study, orthographic form was also examined as a comparison.

2.2 Semantics

Following the analogy of form as the *shell* of a word, the semantic *core* enables the transfer of information and ideas between people and is a crucial component of this study, specifically how semantics can be represented, and which aspects of semantics should be representation. An official definition from Encyclopaedia Britannica reads ‘semantics is the scientific study of meaning in natural and formal languages’ [34]. Though processed by humans subconsciously, defining semantics in such a way that computers can interpret them is a difficult task [27]. Some projects like WordNet and FrameNet involve manually-constructed databases of semantic-based word groupings by hierarchical relations, grammatical properties, and other features [71], [86], however the most prevalent method, and the one used throughout this study, relies on context.

The idea was initially credited to English linguist, John Firth, as early as 1957. Firth’s quote can be found in nearly every piece of Natural Language Processing research:

> *‘You shall know a word by the company it keeps’* — John Firth [37]

Firth’s quote, sometimes referred to as the distributional hypothesis, encapsulates an area of linguistics essential to this project: distributional semantics [49]. Though seemingly simply, the concept has proven to be extremely powerful as a foundation for word embeddings like Google’s Word2vec and Stanford’s Glove vectors which represent words as vectors with features corresponding to contextual information such that words with similar contexts appear close in vector space [78], [70].

Word meaning is supported by a number of language aspects such as syntax and morphology. Interestingly, Firth's idea captures this feature – it can be applied to both syntactic and semantic information extraction by changing the size of the contextual window used to construct semantic vectors [56]. This study attempts to examine word meaning at a very fundamental level and required careful consideration of which aspects of semantics to retain and which to ignore. The most important decisions concerned how word meanings would be represented, and which words to represent; concepts involved in these decisions are detailed below.

2.2.1 Word Embeddings

Distributional semantics aims to quantify the semantics of linguistic matter using distributional properties in language data [49]. Though distributional vector spaces have been used to represent language data since the 1960s [88], word embeddings as we know them were originally coined by Bengio et al. in their series of papers 'A Neural Probabilistic Language Model' in the early 2000s [12]. Nearly a decade later, work embeddings have gained extensive popularity across a vast array of language processing domains including sentiment and bias detection, machine translation, and search optimisation, as well as areas of cognitive science such as the study of communication and language acquisition [55]. The rapid inflation of interest and use is largely attributed to similar increases to the availability of computational power and massive corpora like the Google Book Corpus, Google N-Gram Corpus, and the Wikipedia Corpus [68] [29] [97] [9] which made training reliable, domain-tailored vectors feasible [8].

Embedding algorithms can be personalised to specific tasks, and trained on different sets of text to capture domain-specific information, making them an ideal tool to explore the relationship between word form and meaning. There is a huge range of representations and implementations - this section provides an brief overview of popular methods.

2.2.1.1 Definition

Machine/Deep Learning methods are constrained to dealing with numerical data [56]. Converting language to numbers offers the opportunity to apply powerful analytical methods to information previously thought to be strictly qualitative, but also to concepts and ideas which are fundamentally human. Word embeddings provide a mechanism for such translation [70]. An embedding function W maps words to high-dimensional vectors such that:

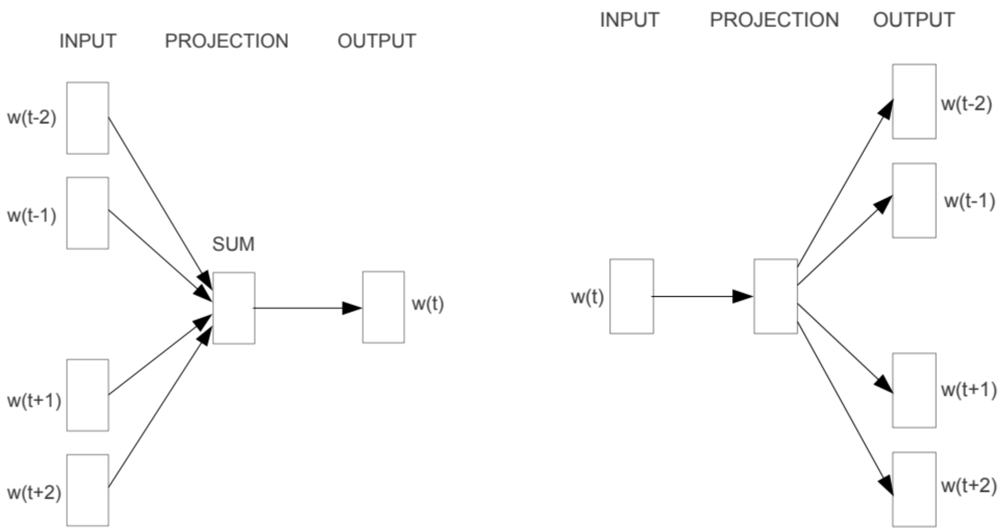
$$W : word \to \mathbb{R}^n$$

Mapping procedures vary in application and complexity; in the simplest of forms, a word embedding can be a 1-hot-encoding specifying the presence of a particular word in a vocabulary. Popular mapping processes can be separated into 2 broad classes - frequency-based and prediction-based embeddings. Modern implementations of both classes perform similarly [44]; this study employs prediction-based embeddings, specifically Word2Vec [70], the details of which are described below.

Prediction-Based Embeddings While frequency-based methods use deterministic operations to construct vectors, prediction-based embeddings are obtained by training a neural network to maximise its predictive capabilities. The network's inputs and outputs are 1-hot encoded vectors of size $(1 * V)$ where V is the vocabulary size; like other unsupervised feature-learning tasks, the value of the task is in the weights learned by the neural network which become the target word's context vector [55]. Two important training strategies are Continuous Bag of Words (CBOW) and Skip-gram, both of which are shallow neural networks.

CBOW predicts the probability of a target word occurring in a given context. Given a contextual window of size n, the network is fed n 1-hot representations of the words before and after time step t, and predicts the probability of each word in the vocabulary occurring at t. With appropriate optimisation, CBOW training is generally faster than Skip-gram [8].

Skip-gram uses the opposite strategy - words are used to predict their neighbours. Given the 1-hot vector representation of a word, the network learns to predict the probability that each word in the vocabulary is a neighbour of the input word. Skip-gram often out-performs CBOW when representing rare, ambiguous words [11].



(a) CBOW Model (b) Skipgram Model

Figure 2.2: Predition-based Methods [69]

2.2.1.2 Word2Vec

Put forward by Mikolov and a team at Google, Word2Vec is an extensive toolkit for creating word embeddings using either of the 2 predictive-based mapping methods described above [70]. A popular extension of Word2Vec methods is Facebook's fastText library for generating language representations based on the skip-gram mapping procedure which treats words as sets of n-grams rather than atomic entities [14]. Along with tools for constructing representations, fastText also offers a range of pre-trained word vectors in a number of languages.

This project makes use of fastText pre-trained vectors to construct semantic representations. FastText was chosen as it offers a set of vectors trained on articles from Simple English Wikipedia, an edition of Wikipedia written in a simplified subset of regular English aimed at English learners [26]. The lexicon selected in this study is constrained to relatively simple, common words (as detailed in section 3.1.1) and thus well-suited to the Wikipedia Simple English vectors.

2.2.2 Morphology

Morphology involves the study of word structure and relationships between words of the same language [4]. Morphological patterns play important roles in the organisation of language and provide communicative cues regarding word meaning [65]. For example, plurals are often marked with an inflectional morpheme like *'s'* in English, or through morpheme reduplication in languages like Somali, while grammatical category can also be inferred from morphological construction, such as inflectional endings in English like *'ly'* for adverbs and *'ing'* for verbs [73].

To explore the form-meaning relationship, the initial components of this study involved teasing apart different language aspects that contribute to meaning, namely the effects of morphology on semantics. Significant morphological concepts used through this study will be discussed below.

Words An important distinction in morphology is between word forms and lexemes. Word forms refer to specific inflection of a words while lexemes encompass the set of inflections of a particular word, represented by a lemma [92]. Though word forms offer interesting information about syntactic relations, lemmas provide more concentrated semantic information and thus are the focus of this study.

Morphemes On a lower level, Linguistics consider words to be constructed from indivisible units of meaning - morphemes. Morphemes can function independently as words (free morphemes), or can act on each other to alter the semantics or grammatical function of a word (bound morphemes) [92]. The morphological typology of a language describes the general patterns of word construction from morphemes; typological categories include *Synthetic* languages which rely heavily on the affixation of bound morphemes to free morphemes to convey meaning, and *Analytic* languages where sentences can be composed entirely of free morphemes, making morphology less important for meaning depiction than features like word-order [99]. English contains properties of both typological categories and thus morphological word construction was important aspect of this study [2].

Word Formation Every language contains specific morphological rules for constructing meaningful semantics; such rules can often be divided into 2 categories - formational and inflectional [54]. For the purpose of this project, formational rules are more interesting as they involve the generation of semantics by relating and creating new lexemes through compounding or blending [92] [46]. On the other hand, inflectional rules are generally used to express grammatical categories, such as tense, person, and gender using prefixes, suffixes, and infixes which can produce inflective forms of a lexeme [54]. These provide syntactic regularities which are important components of language, but are not of interest for this project. One way to remove the effects of bound morphemes is to focus on words with a single, free morpheme.

2.3 Arbitrariness of the Sign

De Saussure was one of the first linguistics to make a distinction between word form (the *signifier*) and concept (the *signified*). This contribution forms the basis for this study. Put forward by De Saussure over a century ago, Arbitrariness of the Sign specifies that the orthographic and phonological forms of a word are unrelated to its meaning [31]. Though the existence of localised sets of exceptions to De Saussure's notion such as onomatopoeic words and phonaesthemes (reoccurring phonetic clusters in words of related meaning [38]) have long been accepted, the property became a psycholinguistic convention as well as a foundational assumption in many linguistic theories. Some even consider the property as a defining characteristic of well-designed communication systems [53].

2.4 Previous Works - Phonosemantic Systematicity

Computational linguistics offered the means to test De Saussure's property by quantifying the form-meaning relationship with powerful statistical analysis. Most directly, this involved the study of how relationships between linguistic items in a form-based space map to the corresponding relationships in a semantic space [43]. Multiple papers in this vein have reported evidence against the Arbitrariness of the Sign notion in the form of correlations between phonological and semantic distance mappings than are greater than would be expected by chance - a property coined *phonosemantic systematicity* [47] [90] [74] [43] [93] [72].

Phonosemantic systematicity has been reported on both local and global levels: strong correlations in localised subsets of words, and a weak systematicity across the entire lexicon [47]. Though the validity and effects of local phonosemantic systematicity in the form of sound symbolism have long been accepted, relatively little is known about lexicon-wide systematicity [74].

The first goal of this project was to quantitatively analyse the form-meaning relationship, specifically the existence of global phonosemantic systematicity. A thorough review of related literature and similar studies was completed to examine the advantages and potential pitfalls of previous methods. 3 significant papers studying global and local phonosemantic systematicity and the corresponding interactions were key for guiding and justifying the design of Experiment 1 (presented in section 3.1). All 3 made use of the general methodology introduced by Shillcock et al., demonstrating the robustness of the method, and consistency of corresponding results.

Even so, consensus was not complete; papers provided opposing evidence regarding the global nature of the reported systematicity, and some doubts about the general methodology arose during the current study. Details of findings, methodologies, and conclusions from all 3 papers are discussed below.

2.4.1 Shillcock et al.

Shillcock et al.'s research [90] was the first paper to claim global phonosemantic systematicity through a small but statistically significant correlation between form and meaning using a representative subset of English. For every pair of words in their set of monomorphemic, monosyllabic words, semantic and phonological similarities were calculated to compute a correlation value between the 2 similarity score sets; the correlation was deemed statistically significant using a randomisation test - comparing the true correlation to the distribution of correlations obtained by repeating the procedure after randomly reassigning form-meaning mappings. Details of their implementation and methodology, as well as some criticisms, can be found in section 4.1 *Methodology*.

Their experimentation introduced 3 novel methods; the first for studying phonosemantic systematicity through the correlation between phonological edit distances (*form*) and distributional semantic distances (*meaning*), and the second for asserting statistical significance using a domain-tailored randomisation test.(see section 3.3 for further details), both of which have been re-implemented [74] [47] [93]. The consistency of results obtained in subsequent papers using both methods initially supported their use as a basis for our experimental design [47]. Potential weaknesses in the method related to the correlation analysis were encountered during experimentation and are detailed below (section 3.2).

Their third contribution was a per-word systematicity score based on correlation values of only word pairs including the target word. Per-word scores demonstrated that systematicity is unevenly distributed across the lexicon and that a large proportion of the correlation mass is conserved within 4 word categories - speech editing terms like 'oh', 'er', and 'ah', pronouns, proper names, and swear words. Shillcock asserted that high-correlation words are 'communicatively important' as higher correlation indicates that the rest of the mental lexicon contributes heavily to cementing the meaning of a word based on its phonological form.

2.4.2 Monaghan et al.

Monaghan et al. [74] examined both global and local phonosemantic systematicity, as well as the relationship between a particular word's systematicity score and age of acquisition. The first component of their paper is particularly relevant to the design of Experiment 1 as both methods introduced by Shillcock are used to examine the form-meaning relationship.

The Shillcock correlation procedure was tested over different sets of English words, and implemented using multiple representations of phonology and semantics to assess the robustness of the reported phonosemantic systematicity and ensure that it is independent of how phonology and semantics are represented. Monaghan highlighted the fact that 70.9% of English word uses are monosyllabic, a strong argument for applying monosyllabic filtering to our dataset. Some interesting phonological distance measures were employed, however the same correlation measurement issues as the Shillcock paper remained. To test the significance of correlations between the different similarity sets, Monaghan et al. used the Mantel randomisation test to recreate Shillcock's randomisation and similarly found more systematicity than expected by chance for all word sets and representations [66]. Their consistent results over phonological and semantic representations, as well as different vocabulary sets, strongly endorsed Shillcock's mechanisms for measuring phonosemantic systematicity, and its respective significance.

Monaghan et al. derived their own per-word phonosemantic systematicity score based on the effect of removing the target word on the overall form-meaning correlation. The primary use of such scores was to study language acquisition; using scores in conjunction with age of acquisition ratings, Monaghan et al. found that systematicity is more pronounced in words learned earlier and concluded that vocabulary structure promotes early language acquisition using systematicity [74]. However, they noted that though statistically significant, global phonosemantic systematicity could be a side-effect of localised pockets of highly correlated words. To this end, the secondary use of phonosemantic systematicity scores was to produce a distribution of systematicity across the lexicon. When compared to the topology of systematicity distributions in randomised vocabularies, Monaghan et al. concluded that global phonosemantic systematicity, as measured with Shillcock et al.'s method, is not a consequence of local phonosemantic systematicity but rather a feature of the entire vocabulary.

2.4.3 Gutierrez et al.

Based on evidence for local and global phonosemantic systematicity from 2 separate veins of research - behavioural studies, and statistical analysis - Gutierrez et al. aimed to combine both research strands by means of 2 innovations: kernel regression to enable flexible lexicon-wide analysis that can better account for the existence of local phonosemantic systematicity, and a metric-learning algorithm for learning weighted edit distances between word-form representations and optimised to minimise kernel regression errors [47]. Though Gutierrez et al. primarily measured systematicity using kernel regression through the quality of semantic predictions given form-representations, Shillcock et al.'s systematicity measurement procedure was also employed.

Aside from these original contributions, Gutierrez et al.'s experiments included other deviations from Shillcock et al. and Monaghan et al.; form representations were based on orthography rather than phonology, and semantic context vectors used by Shillcock et al. [90] and Monaghan et al. [74] were replaced with Word2Vec embeddings. When using the Shillcock et al. methodology, Gutierrez et al. achieved comparable, though slightly higher, correlation to Monaghan et al. on the same dataset using simple orthographic edit distance. Interestingly, their correlation coefficient increased by a factor of 5.7 when optimised weighted edit distances were used [47]. Similarly to Shillcock et al. and Monaghan et al., Gutierrez et al. confirmed the significance of all of their correlation scores using a variation of the Mantel randomisation test.

The use of kernel regression allowed a more flexible modelling of global phonosemantic systematicity by accounting for local systematicity, and produced predictions of semantic representations given form representations. Such predictions were used to estimate semantic distances between words. Gutierrez et al. found that their estimates and the true distances were much more highly correlated ($r = 0.1028$) and statistically significant. This result provides further support for both the of phonosemantic systematicity, however highlighted potential weaknesses in previous research in the handling of local phonosemantic systematicity.

Gutierrez et al. developed another variation of per-word systematicity score - words with lower regression errors had higher systematicity scores. As would be expected, words with high scores included phonaesthemes like 'fluff', 'flutter', and 'flick' as well as onomatopoeic words, however Gutierrez et al. noted that contrary to Monaghan et al.'s results, systematicity did not appear to be randomly distributed through the lexicon. Further discussion of these results is presented in section 3.5

Chapter 3

Experiment 1: Phonosemantic Systematicity

For the past century, linguistic theories have leant on De Saussure's notion of Arbitrariness of the Sign, however research from related fields suggests a more nuanced relationship between form and meaning where arbitrariness is complemented by systematicity to enhance functions of language processing [31] [33] [59]. Given that language is replete with structure on grammatical, syntactic, and morphological levels, it would not be surprising to expect systematicity between the phonology and semantics. Such a property would greatly facilitate language acquisition by allowing known words to guide learning and understanding of new words [74]. However, systematicity reduces the ability to discriminate between word meanings, introducing potential for confusion and inability to convey novel meaning by crowding mappings between form and meaning [43] [73]. Given that one of the powers of language is enabling communication, pressure to maintain discriminability is also to be expected.

The onset of this project involved quantifying the relationship between form and meaning under the competing pressures for systematicity and discriminability, specifically examining the existence of the global phonological systematicity referred to by Monaghan et al. among others [74]. The current implementation was based on the research of Shillcock et al. who reported global phonosemantic systematicity as a small, but statistically significant correlation between phonological and semantic distances of words pairs in an important subset of English [90]. Our aim was to make implementational modifications to the dataset, and form and semantic representations to improve the reliability and accuracy of results.

This chapter describes the procedure used to examine the form-meaning relationship, specifically the lexicon selection, algorithms for defining form and meaning spaces, and the corresponding findings. Results are somewhat inconsistent with previous research and highlight potential flaws in both the current implementation and previous methods.

3.1 Methodology

To analyse the relationship between form and meaning, a representative subset of the English language was extracted from CELEX, including only monomorphemic, monosyllabic words. CELEX is one of the largest annotated lexical databases and contains morphological, orthographic, and phonological information for English, Dutch, and German [7]. A number of phonological and semantic similarity scores were computed for each word pair to create a sets of

$$(semanticsimilarity, phonologicalsimilarity)$$

tuples used to compute phonosemantic correlation. The statistical significance of the correlation was assessed using a randomisation test.

This experiment was based on Shillcock et al.'s research [90], with modifications made to a number of elements. The Shillcock et al. experimentation involved 5 major components:

1. *Lexicon selection*: A subset of English words (monomorphemic and monosyllabic) was extracted from CELEX to maximised representativeness and minimise the effects of morphologically related pairs on the global correlation.
2. *Phonological similarity computation*: phonological distance between the CELEX IPA representations for every word pair was computed using a dynamic programming implementation of the Levenshtein minimum edit distance algorithm (Wagner-Fischer [100]) applied to a penalty scheme between pairs of phones.
3. *Semantic similarity computation*: semantic distance between pairs was defined as the cosine distance between vector representations. These consisted of 500-dimensional co-occurrence vectors trained on the British National Corpus (BNC).
4. *Correlation calculation*: Phonosemantic systematicity was reported as the Pearson correlation score between the sets of similarity scores.
5. *Statistical significance calculation*: once the correlation for true form-meaning mappings had been calculated, the statistical significance was assessed using a domain-specific randomisation test

This section describes our own implementation, as well as justifications for design modifications.

3.1.1 Lexicon Selection

Language can be a tricky form of data to manipulate and analyse. The aim of this research is to examine the form-meaning relationship across an entire language, however accounting for all words is impossible for a number of reasons. A subset of English similar to the Shillcock et al. lexicon was extracted from CELEX to provide a small, representative sample of the average mental lexicon: the monosyllabic, monomorphemic subset of English [74].

Interestingly, Monaghan et al. also removed historically-related words by disregarding word pairs with corresponding entries in etymological dictionaries [74]. Historical dependencies play an important role in vocabulary construction and thus were included in the current experiment to maintain as representative and large a sample as possible.

The set of 52447 English words in the CELEX database was reduced to 3284 monosyllabic, morphemic words before creating all possible word pairs. Duplicate pairs were removed, resulting in 5387143 pairs. All preprocessing procedures are detailed below.

3.1.1.1 Syllabic Filtering

The `epl.cd` file of the CELEX database contains at least 8 fields with phonological information pertaining to each English CELEX word, as shown in table 3.1. The 8th field, `PhonSylBCLX`, contains a phonetically syllabified word representation with syllables separated intro list elements (`[..]`) of CELEX IPA characters [7]. Extracting the monosyllabic words involved filtering such representations, retaining only those containing 1 syllable.

Table 3.1: Example of `epl` entry for `"smile"`

ID Num	**Head**	**Cob**	**Pron Cnt**	**Pron Status**	**Phon Strs DISC**	**Phon CVBr**	**Phon Syl BCLX**
42844	`smile`	2892	1	`P`	`'sm2l`	`[CCVC]`	`[smaIl]`

Although this filtering process reduced our dataset from 52447 words to 6760, previous research argues that monosyllabic words are a representative sample of the average mental lexicon [74] [104].

3.1.1.2 Morphemic Filtering

Monomorphemic words were extracted from the 4th field of the `eml.cd` file, `MorphStatus`, where monomorphemic words can be identified by the code `M` as displayed for the monomorphemic word ‘smile’ in table 3.2. The CELEX English word set contains 7401 monomorphemic words, of which 3284 were also monosyllabic.

Table 3.2: Example of `eml` entry for `"smile"`

ID Num	**Head**	**Cob**	**Morph Status**	**Lang**	**Morph Count**	**...**
42843	`smile`	1488	`M`	–	`1`	...

Morphemic filtering was an equally important filtering step as morphological regularities and language-specific syntactic constraints discussed in section 2.2.2 would have likely dominated the form-meaning correlation. Tamariz noted that including syntactic information in her dataset substantially increased the resulting phonosemantic systematicity in an subset of Spanish, though the importance of such filtering depends on the morphological typology of the language in question [93] [99].

3.1.1.3 Duplicate Filtering

Though the aforementioned filtering processes produced a representative lexicon, a final stage of filtering was deemed necessary after exploring the similarity score tuples for all word pairs in the lexicon. CELEX stores multiple copies of a particular word form if it correspond to different meanings (also known as homonyms), however Word2Vec only assigns each word form a single vector, regardless of the number of corresponding meanings. For example, words like 'bank' (the edge of a river) and 'bank' (to depend on something) are semantically distinct and thus the form 'bank' would appear multiple times in the filtered lexicon but each would be assigned the same semantic vector. Therefore, they were removed to avoid skewing the correlation score with ambiguous semantic representations.

This final filtering process removed 260 word pairs, resulting in a final set of 5387143 words pairs.

3.1.2 Form Space

Previous research used of a range of methods to define similarity between word forms with each paper using slight variations [74] [47]. This experiment involved 5 original variants based on previous research, separated into 2 classes:

- *Direct Distance metrics*: methods that measure form similarity directly.
- *Vector representations*: methods that convert forms into vector representations before measuring vector similarity

To account for phonology, all methods must convert words into sequences of phonological feature vectors corresponding to IPA symbols before computing phonological distance. The conversion from strings to feature-sets is detailed below, before descriptions of each metric and their respective strengths and weaknesses.

Distinctive Feature Set Generation This process first involved creating a feature vector for each CELEX IPA symbol. CELEX symbols were first converted to IPA using chart A.1; feature vectors were then assigned using Riggle's Phonological Features Chart (v11.02) [84] (further chart details in section 2.1.1.3). Most CELEX IPA symbols could be assigned a feature vector directly from the Riggle chart A.2, however there were a few cases that required additional consideration:

3.1.3 Phonological Features Chart

- `':'` : IPA contains additional suprasegmentals to indicate stress and intonation. `':'` is applied to vowels to indicate the vowel is long. In the current implementation, `'A:'` indicated that the proceeding vowel should be duplicated.
- Certain CELEX vowel symbols, such as `3` and `V`, were not explicitly contained in Riggle's feature chart. Therefore, they were extracted manually from the vowel-relation chart which indicates specific vowel features (`[front, back, tense, rnd, high, low]`). The 17 other features are consistent across all vowels. Full vectors for both symbols are displayed in table 3.3

Table 3.3: Distinctive Feature Vectors manually extracted for `3` and `V`

Feature	3	V
Cons	-1	-1
Son	1	1
Syll	1	1
Lab	1	1
Rnd	-1	-1
Cor	-1	-1
Ant	0	0
Dist	0	0
Dor	1	1
High	-1	-1
Low	-1	-1
Back	-1	1
Tense	-1	-1
Phar	1	1
Voi	1	1
Sg	-1	-1
Cg	-1	-1
Cont	1	1
Stride	-1	-1
Lat	-1	-1
Drel	-1	-1
Nasal	-1	-1
Long	-1	-1

Orthographic representations were converted into sets of phonological feature vectors corresponding to each CELEX IPA symbol. For example, the word 'smile' would be converted into a set of 4 feature vectors for the CELEX IPA symbols `[s, m, aI, l]`

3.1.3.1 Direct Distance Metrics

The direct form-distance metrics implemented in this experiment were:

- Minimum Edit Distance (Orthographic)
- Minimum Edit Distance (Phonological)
- String Substitution Edit Distance (Phonological)

This section presents the basis, design, and justification for each metric.

A common method for measuring the distance between sequences of items is the *Levenshtein* algorithm [62]. Generally applied to strings of characters, Levenshtein determines the minimum number of edits (MED) required to transform one string into another. Edit operations include character insertions (`ins`), deletions (`del`), and substitutions (`sub`); operation costs can vary, but traditionally, `ins` and `del` cost 1 while `sub` costs 2.

Shillcock et al. measured phonological edit distances between word pairs using a dynamic-programming implementation of Levenshtein (Wagner-Fischer [100]) where substitution costs were determined using a manually-devised penalty scheme between phonological feature vectors from the Festival Speech System [94]. Feature differences such as vowel length and consonant voicing were assigned small penalties while vowels-consonant conversions received high penalties. Shillcock noted that the penalty scheme could be more 'psychologically realistic' [90]. Rather than attempt an improved feature-based penalty scheme, our metrics act directly on distinctive features which encode differences between features, consonants, and vowels explicitly.

The definitions of Levenshtein operations were modified such that `sub` cost is base on the number of features differing between 2 phonological feature vectors as demonstrated [75]:

$$\texttt{sub}\left(\begin{bmatrix}1\\1\\0\end{bmatrix},\begin{bmatrix}1\\-1\\1\end{bmatrix}\right)=\sum\left|\begin{bmatrix}1\\1\\0\end{bmatrix}-\begin{bmatrix}1\\-1\\-1\end{bmatrix}\right|=\sum\left|\begin{bmatrix}0\\2\\-1\end{bmatrix}\right|=3$$

Costs for `ins` and `del` are more difficult to define. 2 phonological variations of the Levenshtein algorithm were implemented; each models insertions and deletions differently. Traditional Levenshtein was also applied to orthographic word representations.

Of the 5 current metrics, the direct phonological distance (Phonological MED and MSD) metrics have the strongest theoretical basis for capturing meaningful phonology than the vector representations for 2 reasons. Our phonological Levenshtein variations require fewer assumptions than both Shillcock et al. and Monaghan et al. distance metrics, but most importantly, the phonological similarity scores from Levenshtein feature edit distance have been found to match such human similarity judgements [89].

Orthographic MED The original Levenshtein algorithm was applied directly to orthographic representations of words to produce a simple baseline for measuring word similarity based on form, the results of which could be compared to Gutierrez's control results [47].

Phonological MED The first modified Levenshtein metric involved defining a constant cost for `ins`, `del` operations based on $1/2$ the mean cost of substitution between all IPA symbols to maintain the traditional balance of costs between `ins`, `del`, and `sub`. Similar phonological distance measures have been found to match human judgements of phonological similarity [89].

$$\texttt{ins},\texttt{del} = \frac{1}{2}\left(\frac{\sum_{i=1,j=i+1}^{n} \texttt{sub}(s_i, s_j)}{\frac{n(n-1)}{2}}\right)$$

$$\texttt{ins},\texttt{del} = \frac{\sum_{i=1,j=i+1}^{n} \texttt{sub}(s_i, s_j)}{n(n-1)}$$

where $i, j \in set(IPAsymbols)$, $n = |set(IPAsymbols)|$.

Phonological Minimum Substitution Distance (MSD) The second metric was implemented to further reduce the assumptions involved in computing distance by removing the constant cost constraint on *ins* and *del* operations [75]. Instead, `ins` and `del` operations were modelled as substitutions between the target symbol and a zero vector.

Similarity A simple conversion was applied to obtain similarity scores from distances:

$$similarity(u, v) = \frac{1}{1 + distance(u, v)}$$

3.1.3.2 Vector Representation

Representing form as a vector enables different aspects of phonological form to be captured in vector magnitude and direction. To convert phonological feature vectors to a single vector, 2 composition functions were employed:

- Mean of feature vectors
- Concatenation of feature vectors

Mean Feature Vector Used as a baseline, this operation encoded the presence of phonological features and resulted in 23-dimensional vectors, allowing form comparison in a relatively low-dimensional space.

However, mean phonological feature vectors do not provide reliable form representations. The largest source of information loss is due to disregard for the ordering of phonological symbols which results in distinct words sharing phonological representations such as `d0g` and `g0d`. Though this metric includes some notion of the phonology, time is a key component of language and must be accounted for.

Concatenation of Feature Vectors Another composition function that encodes some information about phone ordering is concatenation. All feature vectors were concatenated before being padded with zeros to the length of the longest concatenation word vector. The experimental lexicon contains relatively short words, with the longest word containing 7 phones; thus all vectors were of dimension $7 * 23 = 161$.

Though order is accounted for, simple concatenation weights orderings too strictly. Consider the simple example - words 'trip' and 'rip' (`trIp`, `rIp`) are as phonologically similar as 'trip' and 'trap' (`trIp`, `tr&p`) however in the concatenation space, the latter pair will match on all dimensions representing `[t, r, p]` while the former pair will not align exactly on any.

3.1.4 Semantic Space

The semantic space in our study is based on the same distributional principles as used by Shillcock et al., however the current implementation made use of fastText's Wikipedia Simple English pre-trained vectors in the hopes of building a more robust and representative feature space [14]. The two most important modifications were the vector features themselves, and the corpus used to create them.

To construct semantic representation, Shillcock et al. used a set of 500 context words from the BNC such that vector features corresponded to the occurrence frequency of each context word within a window of ± 5 words from the target word. Shillcock et al. were required to reduce their lexicon to ensure reliable semantic representations, noting that vector reliability diminishes with frequency. FastText vectors were chosen in part to circumnavigate this problem as words are represented in 300 dimensions by combinations of n-grams, thus reducing the effects of word frequency on reliability.

The Wikipedia Simple English corpus [26] was particularly well-suited to this task as it's vocabulary consists of common, relatively short words which matches the monosyllabic phonological constraint of this experiment and could result in more reliable semantic representations of such words [104].

3.1.4.1 Semantic Similarity

Similarly to the Shillcock et al. implementation, semantic similarity between vectors u, v was computed using $1 - cosine(u, v)$ [83]. A common problem in high dimensional spaces is skewing of distance metrics like Euclidean or Manhattan as the ratio of distances to nearest and furthest neighbours approaches 1, making cosine similarity a more reliable metric in the current semantic space [1]. Furthermore, semantic similarities are better described by similarities between the direction of vectors, rather than their magnitude, which are captured more aptly by cosine distance [74].

3.2 Correlation

Correlation coefficients are one of the primary tools used to quantify relationships between variables and have been a popular method for studying the form-meaning relationship. Though each coefficient has different characteristics, all metrics vary between [-1 and 1] – the strongest disagreement and agreement respectively, where 0 indicates no relationship.

In our experiments, the variables of interest are the sets of semantic and form-based similarities between word pairs. Much of the previous research reports correlation scores using the Pearson coefficient, however characteristics of our dataset detailed below provided strong evidence against its applicability in this domain (see Discussion and General Conclusions section (3.5)). The Spearman coefficient was found to be more reasonable and thus was also computed. This section provides details of both coefficients.

3.2.1 Pearson

Given 2 variables X and Y, the Pearson correlation coefficient measures variable association based on the following formula:

$$r = \frac{\sum_{i=1}^{n}(u_i - \bar{u})(v_i - \bar{v})}{\sqrt{\sum_{i=1}^{n}(u_i - \bar{u})^2}\sqrt{\sum_{i=1}^{n}(v_i - \bar{v})^2}}$$

where $\bar{u} = \frac{1}{n}\sum_{i=1}^{n} u_i$ and $\bar{v} = \frac{1}{n}\sum_{i=1}^{n} v_i$

Though the more popular statistic, the Pearson coefficient is based on 2 relatively strong assumptions about the variables in question – both must be normally distributed, and linearly related to each other [64]. The results of the current implementation are reported in Pearson to allow comparison to previous studies, however neither assumption could be asserted in this experiment.

3.2.2 Spearman

Spearman's correlation coefficient makes no assumptions about the distribution of data and is applicable to non-linearly related variables. It is calculated using the same formula as the Pearson correlation metric where calculations are based on the rank of variable observations rather than the explicit observation values [64].

$$r = \frac{\sum_{i=1}^{n}(q_i - \bar{q})(r_i - \bar{r})}{\sqrt{\sum_{i=1}^{n}(q_i - \bar{q})^2}\sqrt{\sum_{i=1}^{n}(r_i - \bar{r})^2}}$$

where r_i, q_i denote the rankings of u, v respectively, $\bar{r} = \frac{1}{n}\sum_{i=1}^{n} r_i$ and $\bar{q} = \frac{1}{n}\sum_{i=1}^{n} q_i$

As neither the distribution of form-based distances, nor the quality of the relationship between them could be assumed to match the Pearson assumptions, Spearman's correlation measure is a better choice of statistics for the current experimentation.

3.3 Statistical Significance

Statistical significance was measured using the same method as Shillcock et al. [90]. After computing a correlation value, a tailored randomisation test was used to demonstrate how the true correlation compares to what would be expected if form-meaning mappings were entirely arbitrary (*the null hypothesis*).

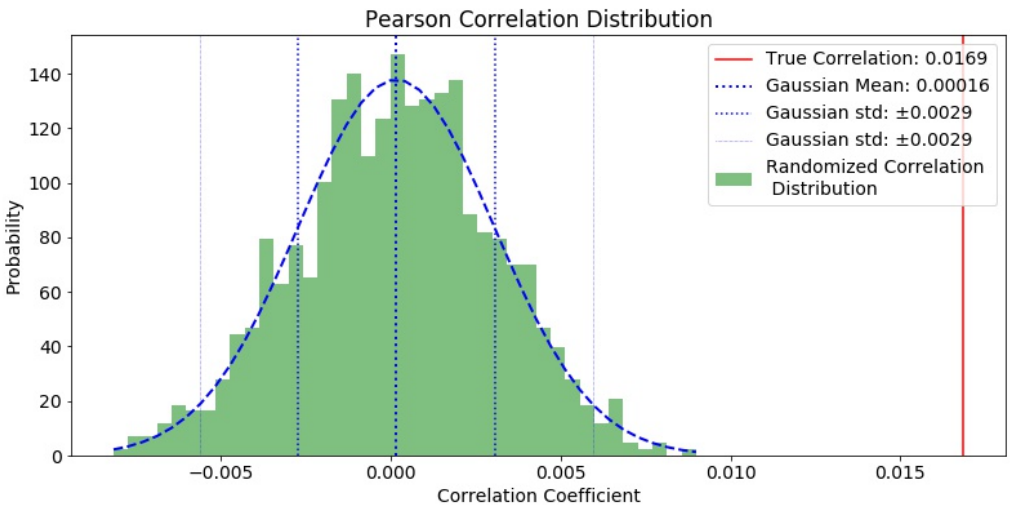


Figure 3.1: Significance testing: comparing the true correlation score to the distribution of correlation scores for randomised vocabularies. This is an example of Phonological MED.

To produce vocabularies under the null hypothesis, each word was randomly paired such that form-meaning mappings in the random vocabulary were based on the original word's phonological representation and the paired word's semantic representation. Phonosemantic correlations were calculated for 1000 randomised vocabularies to produce a distribution of correlation scores that would be expected if mappings were entirely arbitrary.

Figure 3.1 demonstrates the statistical significance testing for correlation obtained using Phonological MED; true correlation is 5.77 standard deviations away from the randomised distribution mean ($z = 5.77$) – strong evidence against the null hypothesis.

3.4 Results

Current results highlight some issues that were noted during experimental design, namely the validity of the form-based distance metrics described in 3.1.2 and the choice of correlation metric detailed in 3.2, however the results also shed light on some unforeseen deviations from previously reported findings which are explored in this section.

3.4.1 Pearson Correlation & Statistical Significance

Table 3.4: Pearson Correlation & Statistical Significance

Distance Metric	**Pearson Correlation** r	z **score**	p **score**
Orthographic MED	0.0060	1.15	0.06
Phonological Feature MED	0.0169	5.70	<0.0001
Phonological Feature MSD	0.0237	8.70	<0.0001
Mean Feature Vector	0.0416	6.97	<0.0001
Concat. of Feature Vectors	0.0204	4.25	<0.0001

As can be seen from the p and z scores in table 3.4, the correlation between semantic similarity and form similarity for all the form distance metrics replicate the statistical significance reported by Shillcock, Monaghan, and Gutierrez, except for orthographic MED which is still considered 'marginally significant' by convention [87].

Even so, actual correlation values are not as consistent as those reported by papers in section 2.4 *Previous Works* and displayed in table 4.2. Some divergence is to be expected, as both phonological and semantic distances are computed differently across all papers. Other implementational differences were tested as potential explanations, including different conversions between distance and similarity, and lexicon variations.

Pearson correlation assumes a linear relationship between variables and therefore would be affected by the conversion function from distance to similarity. Another similarity conversion based on Sanders and Chin's phonological distance algorithm was tested, which involved normalising phonological distances between $[0,1]$, before subtracting them from 1 to produce similarity scores [89]:

$$similarity(u,v) = 1 - \left(\frac{phonologicalEditDistance(u,v)}{maxPhonologicalEditDistance}\right)$$

This produced marginally smaller Pearson correlation scores and thus did not account for divergence between our correlation scores and those previously reported. Interestingly, current correlation scores match comparable previous results much more closely when the duplicate removal step (detailed in section 3.1.1.3) is ignored, as can be seen in table 3.5.

Table 3.5: Pearson Correlation Comparison. The Pearson correlation scores for the *Cleaned* and *Raw* datasets (duplicates removed and retained) compared to the most relevant correlation scores reported in previous research.

Distance Metric	**Clean** r	**Raw** r	**Shillcock** r	**Monaghan** r	**Gutierrez** r
Orthographic MED	0.0060	0.0167	–	0.016	0.019
Phonological Feature MED	0.0169	0.0464	0.061	0.034	0.046
Phonological Feature MSD	0.0237	0.0539			
Mean Feature Vector	0.0416	0.0427	–	–	–
Concat. of Feature Vectors	0.0204	0.0220	–	0.031	–

Figure 3.2 demonstrates the changes to significance tests when duplicates are retained. Such filtering wasn't discussed in previous papers, all of which also extracted lexicons from CELEX database, and could explain differences in correlation values.

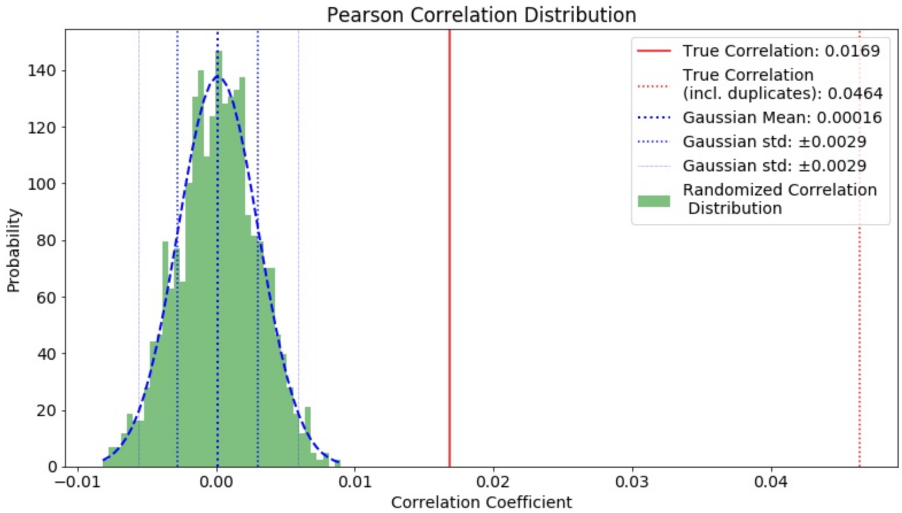
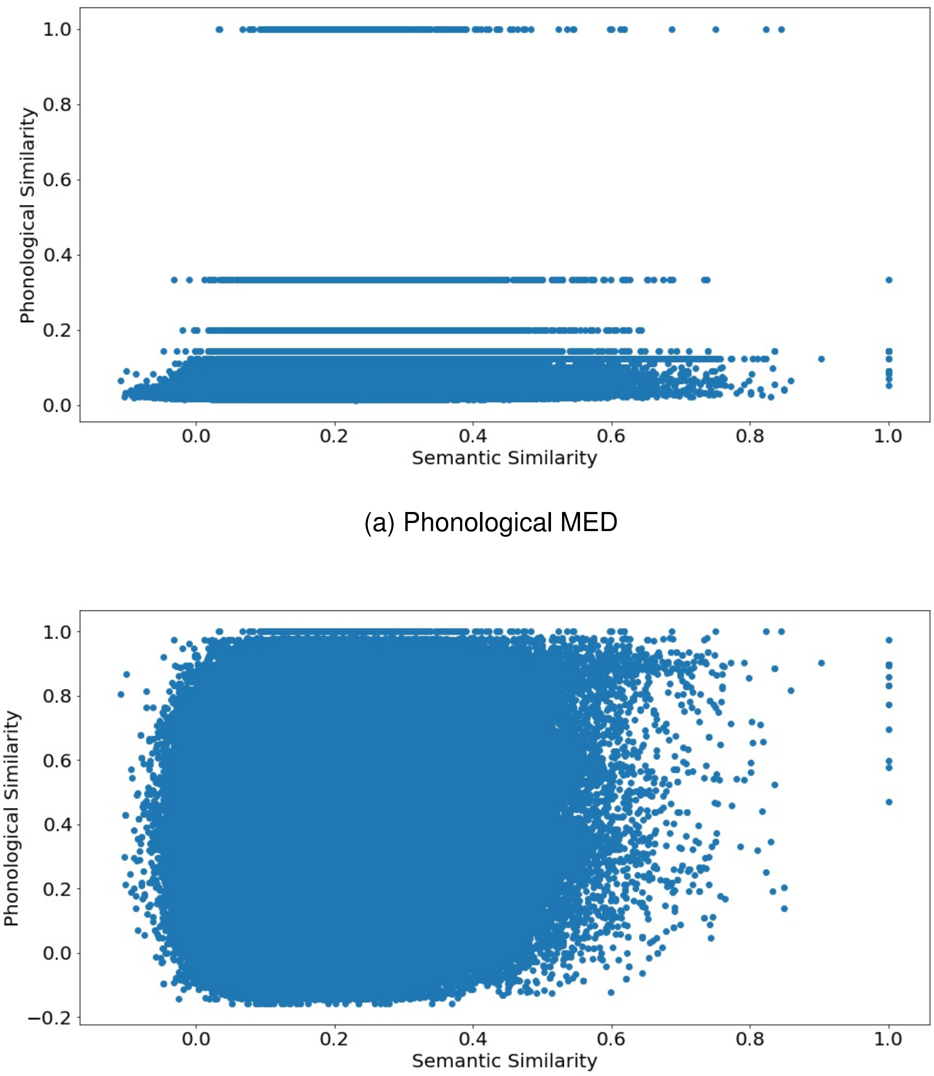


Figure 3.2: Significance testing - comparing the true correlation of Phonological MED obtained with and without duplicate pairs

3.4.2 Spearman Correlation & Statistical Significance

Figures 3.3(a) and 3.3(b) display similarity scores obtained from a direct distance metric and a vector-based distance metric, respectively.



(a) Phonological MED

(b) Concatenation of Phonological features

Figure 3.3: Form-meaning similarity scores for different Phonological distance metrics

As noted in section 3.2, the Pearson correlation metric involves strong assumptions about the variables being measured: that they are both continuously distributed, and linearly related. Throughout the study, it became clear that although semantic and vector-based form similarity scores met the first criteria, the direct distance metrics for computing form-based similarity did not, as displayed in figure 3.3. Table 3.6 further supports the discrete nature of phonological similarity scores obtained through direct distance metrics.

For reference, there were 54049 unique semantic similarity scores.

Aside from the variable distributions, there is no reason to expect a linear relationship between form and meaning. Both assumptions provide support for discarding Pearson. Tamariz reached a similar conclusion about the use of Pearson correlation, opting in-

Table 3.6: Direct Distance Metrics produce discrete similarity score distributions

Form-Based Distance Metric	Number of Unique Values
Orthographic MED	9
Phonological Feature MED	328
Phonological Feature MSD	111
Mean Feature Vector	66459
Concat. of Feature Vectors	45139

stead for a domain-specific variant of Fisher Information to measure distance between languages [93]. The standard Spearman correlation coefficient avoids both Pearson assumptions and was selected for its simplicity. It was also noted that although ignoring duplicate removal had a large impact on Pearson correlation, the effects on Spearman were negligible, further supporting the robustness of the Spearman metric. Scores are recorded in table 3.7.

Table 3.7: Spearman Correlation & Statistical Significance

Distance Metric	Spearman Correlation r	z score	p score
Orthographic MED	-0.0402	9.97	<0.0001
Phonological Feature MED	-0.0251	4.98	<0.0001
Phonological Feature MSD	-0.0073	1.33	0.09
Mean Feature Vector	0.0354	6.15	<0.0001
Concat. of Feature Vectors	0.0113	2.13	0.02

As can be seen in table 3.7, the correlation values for all direct distance metrics are negative while all vector representations are positive; 2 inferences can be drawn from these results. Firstly, given the assessment of the direct and vector-based distance measures in the Form Space section (3.1.2) which asserted the theoretical weaknesses of our vector-based representations, the differences in score sets provide further evidence to reject both the vector-based representations. Second, the negative correlation scores obtained by all direct distance metrics would imply that similar-sounding words are more likely to have distinct meanings. Given that all 3 previous studies claim a positive correlation between form and meaning, the current results were thoroughly checked.

A possible explanation for negative values is the Spearman coefficient's inability to handle 'ties'. The discrete nature of phonological edit distances resulted in many tied phonological similarities scores (displayed in figure 3.3(a)) which could have skewed Spearman's r. The Kendall τ coefficient is another standard rank-based correlation coefficient which includes tie-handling and was computed across all distance metrics [64]:

$$\tau = \frac{n_c - n_d}{n_t}$$

where $n_t = \frac{1}{2}n(n-1)$ and n_c, n_d are the numbers of concordant, discordant pairs.

Kendall correlation scores, as well as Spearman scores, are displayed in table 4.5. Though all Kendall scores are slightly smaller than Spearman scores, the similarity between both correlation scores across all form-based measures implies that ties do not provide an explanation for negative correlations.

Table 3.8: Kendall & Spearman Correlation Scores

Distance Metric	Spearman Correlation r	Kendall Correlation τ
Orthographic MED	-0.0402	-0.0310
Phonological Feature MED	-0.0251	-0.0170
Phonological Feature MSD	-0.0073	-0.0050
Mean Feature Vector	0.0354	0.0233
Concat. of Feature Vectors	0.0113	0.0072

Though significant negative correlation scores are surprising given that previous research agues a converse relationship, there is nothing inherently incomprehensible about such findings, especially given some of the issues with previous research discussed above. Tamariz also finds significant correlation for some datasets using an alternative to Pearson, however as her metric is positive definite and unrelated to Spearman correlation, it is comparing resulting values is irrelevant [93].

One of the motivating factors behind studying the relationship between form and meaning, as noted in this chapter's Introduction (Chapter 3), is understanding how the opposing pressures exerted by language use and language acquisition affect the structure of the vocabulary; though phonosemantic systematicity increases the ease of acquiring language, it also increases ambiguity in communication. Tamariz presents research suggesting that vocabularies contain specific phonological parameters responsible for systematicity and discriminability to reflect these conflicting forces [93]. A plausible explanation for the statistically significant, negative correlation between form and meaning reported above is that our measure may capture more information about phonological features that respond to pressures favouring the discriminability of words and opposing systematicy.

3.5 General Discussion & Conclusions

Over a century ago, De Saussure put forward the Arbitrariness of the Sign notion, implying that the relationship between word for and meaning was arbitrary [31]. Though local sets of exceptions have long been accepted, Shillcock et al. was the first study to examine the global characteristics of the form-meaning relationship [90]. The paper introduced 3 novel methods to study the overall phonosemantic systematicity, its statistical significance, and it's distributed across the lexicon, all of which have been reimplemented in proceeding studies.

The current experimental design drew inspiration from Shillcock et al.'s study among others to quantify the form-meaning relationship, aiming to achieve more reliable results by modifying the dataset, and the form and semantic-based word representations as detailed in section 3.1. Shillcock et al. drew 2 conclusions; first, that the monomorphemic, monosyllabic subset of English displays weak, global phonosemantic systematicity, and second, that it is distributed unevenly such that mass is concentrated over communicatively important words [90].

Though the current study replicated the significance of phonosemantic systematicity measured with Pearson correlation reported in previous studies, questions arose throughout the study about both the validity of our own implementation, as well as the general methodology which cast doubt on the validity of previous conclusions, specifically the reported correlation scores, and the claims that such correlations imply global phonosemantic systematicity.

3.5.1 Correlation Coefficients

As discussed in the Correlation section, though Pearson coefficients are widely reported in previous research, neither of its foundational assumptions, namely continuous variable distributions, can be asserted. This issue initially seemed purely a matter of statistical definitions, however it's importance grew significantly when Spearman correlation scores indicated drastically different conclusions about the form-meaning relationship.

It can be argued from table 4.3 that vector-based representations currently implemented meet the Pearson assumption of being continuously distributed, but the same is not possible for the direct distance metrics. Although direct distance implementations vary across studies, the general nature of direct distance metrics indicates that this may have been the case in previous studies as well – Tamariz also noted that Pearson was not applicable in her experiment for this reason. This is strong support for rejecting the Pearson metric.

3.5.2 Global Phonosemantic Systematicity

All previous research, including the current Spearman correlation results, have provided *some* evidence that the correlation between form and meaning measured using Shillcock et al.'s general methodology is statistically significant [90] [73] [47] [93]. However, it's difficult to demonstrate whether this correlation is simply a side-effect of highly systematic pockets in the lexicon, like phonaesthemes and onomatopoeic words, or whether it is a global characteristic of the lexicon. All papers cited above also demonstrated that systematicity is unevenly distributed across the lexicon; Monaghan et al. used characteristics of their systematicity distribution to support Shillcock et al.'s claim that systematicity is a global property across the lexicon [73], however the argument is not entirely convincing. Some questions about their reasoning and methodology, as well as some counter arguments, are discussed below.

Monaghan et al. base per-word systematicity on the effect of removing the target word on the lexicon-wide correlation score, such that words whose removal leads to greater decreases in overall correlation are more systematic. To test the global feature claim, Monaghan et al. used the Wilcoxon signed-rank test to compare the rankings of the systematicity distribution across the true lexicon to the systematicity rankings across distributions lexicons with randomly shuffled form-meaning mappings [101]. Results were further controlled with Bonferroni correction to avoid the 'multiple comparison problem' [13]. Finding that *the null hypothesis* (that the systematicity rankings of the true distribution could be expected under completely arbitrary conditions) could not be rejected, Monaghan et al. claim that observed systematicity is not a consequence of local highly-correlated word sets, but rather a global feature across the lexicon.

This rank-based test compared the topology of the systematicy distribution over the true lexicon to those of completely arbitrary vocabularies. If localised pockets of systematicity were at the heart of the overall systematicity, the topology of per-word systematicity over the true vocabulary would be distinct from those of randomised vocabularies. Though Monaghan's results are interesting, they are based on Pearson correlation; reasons why this metric is inappropriate have already been discussed. Perhaps more importantly though: from a purely logical perspective, failure to reject the null hypothesis is not sufficient evidence for accepting it.

Gutierrez et al.'s also studied effect of local sets of highly correlated words on the overall form-meaning correlation. Though Gutierrez employed Shillcock et al.'s general methodology of measuring form-meaning correlation, they proceed to measure overall systematicity based on the quality of semantic distance predictions using 2 innovations which provide greater model flexibility and ensure that effects local sets of phonosemantic systematicity are reliably captured in a global analysis [47]:

- *Kernel Regression*: This framework avoids making assumptions about the preditor-target variable relationship by basing predictions on the target values of nearby points in predictor space. This provides a drastic increase in modelling flexibility when compared to the Pearson correlation which assumes relationship linearity.
- *SMLKR*: This algorithm learns optimal weightings for Levenshtein edits. String MED assumes equal weighting across substitution edits, however previous research demonstrates that phonological and orthographic attributes play different roles in phonosemantic systematicity - for example, shared consonants drive semantic similarity while shared vowels promote distinct semantics [93] [25].

Gutierrez et al. find substantially higher systematicity using their methodology, hypothesising that effects of localised phonosemantic systematicity are underestimated in direct form-meaning correlation experiments. Using their own per-word systematicity measure based on the SMLKR model regression error and the systematicity distribution over word sets sharing 2-letter beginnings, they test this hypothesis by studying the likelihood that each word set exhibits the same mean regression error if systematicity was randomly distributed over the lexicon. p-values are used to represent this likelihood and are defined as the percentage of randomly selected word sets with a lower mean regression error than the target set. If systematicity were randomly distributed across the lexicon, each word-beginning set would have similar mean regression errors therefore one would expect a uniform distribution of p-values across word-beginning sets, however Gutierrez et al. findings do not conform to such a distribution. What's more, word beginnings with statistically significant p-scores correspond to phonaesthemes [77].

3.5.3 Conclusions

A number of conclusions can be drawn from the analysis of the current results. First, although previous research provides evidence for a positive correlation between form and meaning in a representative subset of English, when measured with an appropriate correlation metric, the relationship appears to be slightly negative. This would imply that words with similar form are more likely to convey distinct meanings, a conclusion which is not unfathomable.

Second (and most important), the current experiment demonstrated that drawing firm conclusions about the form-meaning relationship is more nuanced that previously reported, specifically regarding the effects of local pockets of highly-systematic words. Correlation scores and their statistical significance measuring using the Shillcock et al. method are highly dependent on the experimental design - from the algorithms used to measure semantic and form-based distance, to the conversion between distance and similarity, and to the choice of dataset. Though such variability seems to be diminished by the use of Spearman correlation metric which avoids the strong Pearson assumptions, it is an undesirable characteristic for scientific conclusions and implies that systematicity may be less robust than previously claimed.

Third, along with raising questions about the robustness of phonosemantic systematicity and the Shillcock et al. method for examining it, this study also cast doubt on claims that such systematicity is a global feature of the lexicon. Results from Gutierrez et al. and Tamariz provide convincing evidence that the effects of highly systematic local pockets were underestimated by assuming a linear relationship between form and meaning, and an equal weighting of phonological attributes (ie. Levenshtein edits) with respect to systematicity [47] [93].

3.5.4 Future Work

The current experiments make a strong case for rejecting the Pearson correlation coefficient to measure phonosemantic systematicity due to the nature of the distribution of phonological similarity scores when measured with direct distance metrics, but also to avoid making unsupported assumptions about the nature of the form-meaning relationship. Future work should take care when selecting their metrics to quantify the relationship.

There are some clear improvements that could be made to the current implementation, especially regarding vector-based phonological representations, the results of which were for the most part ignored in this study; an improvement that was considered involved the using the Levenshtein algorithm to guide padding which would better account for ordering and thus produce more reliable representations. Better phonological word representations could have implications for speech recognition or synthesis systems. However, as noted in the discussion, measuring correlation directly may not be a suitable method for quantifying the form-meaning relationship; further study into the weightings produced by Gutierrez et al. concerning the different roles of phonological and orthographic attributes may lead to more interesting findings about the specific roles of phonological features in language acquisition or use [47] [93] [25].

As noted in section 2.1.2, the orthographic depth principle characterised how closely the orthography and phonology of a language are related. This would be an interesting property to consider in future work that investigates the roles of different orthographic and phonological symbols in driving systematicity or distinctiveness, especially how these roles differ between languages.

Chapter 4

Background: Language Design

Language has evolved to a precarious optimum, balancing between tensions caused by evolutionary and social forces, the mutually incompatible goals of minimising effort and maximising accuracy, and the characteristics of language systems required for acquisitions and use [65] [103] [73] [21]. The first strand of this study provided some insight into the relationship between form and meaning and the high degree of arbitrariness; the second strand explores how such a vocabulary structure interacts with cognitive processes, specifically language acquisition. This section details some causes and effects of the vocabulary structure.

4.1 Vocabulary Structure

Even under the De Saussurian assumption that the form-meaning relationship is arbitrary, language structure exhibits systematicity stemming from a number of widely-accepted sources. The most studied displays of systematicity are pockets of local systematicity discussed in sections 2.3 and 2.4 which can be language specific, taking the form of sound symbolism, onomatopoeic words, and phonaesthemes, but also occur at a general level. For example, expressives referring to size, distance, and shape across languages have consistently been found to contain different vowel classes; those denoting large size contain low vowels while those denoting small size contain high vowels [82] [52] [98]. There is growing evidence for the strength of such effects – the 'bouba-kiki' effect is the robust tendency to associate sounds requiring rounded or angular mouth shape with similarly-shaped objects and has been demonstrated to bias face-name mappings in the same way [10]. Another important source of systematicity is morphology which effectively reflects semantic features like grammatical category in word form and is a key component of the compositional nature of language construction [58] [30].

Though Experiment 1 did not produce definite conclusions for the degree of systematicity in the vocabulary, results provided evidence that the overall relationship between form and meaning is overwhelmingly arbitrary. Arbitrary form-meaning mappings introduce a high cost for language acquisition as the existing mental lexicon offers no assistance in determining correct semantic mappings for novel words. This has troubled researchers from as early as the 1600's [20]. One such example is Wilkins' 'Essay Towards a Real Character and a Philosophical Language'. Published in 1668, Wilkins attempted to create a language based entirely on scientific principles, where letters map systematically to meaning classes, that would be free from the ambiguity and irregularities that permeate natural language [102].

Wilkins demonstrated that it was possible to create a 'perfect', systematic language but why hasn't language evolved this way?

4.2 Counteractive Forces

Multiple features of language structure have been posited to arise from the need to balance counteractive forces, many of which stem from the evolutionary principle of least effort [103]; for example, Clark and Wilkes-Gibbs [22] assert that utterances in dialogue are designed to minimise the joint effort of the speaker and listener for production and comprehension, respectively. The form-meaning relationship is no different - a host of opposing pressures must be considered to understand the balance between systematicity and arbitrariness in the form-meaning relationship, a number of which are discussed below.

4.2.1 Cognitive Processes: Acquisition vs. Expression

The power of language lies in enabling communication – the transfer of complex information and ideas. Such expression first requires language to be acquired. These 2 processes depending on opposing characteristics of language; given that these are 2 most crucial language processing tasks, it is hard to imagine that language has not evolved to optimise both.

Language learning involves learning mappings between form and meaning, as well as how the specific language groups words into semantic categories through the use of language-specific features like morphology and grammar. Learning such categories is facilitated by simple, structured, *systematic* mappings; learning true and artificial categories is facilitated by systematic phonological coherence between forms belonging to each category [39] [17] [57]. Systematicity allows the existing mental lexicon to contribute to the acquisition of new words, thus reducing the amount of effort required for learning [73]. From an acquisition point of view, the level of arbitrariness in the vocabulary is especially strange.

Expression, on the other hand, is enhanced by arbitrary mappings. The goal of expression is to convey potentially novel meanings that will be understood precisely and reliably by listeners while reducing the effort required for language production [21]. From an Information Theory perspective, systematic mappings represented in the form-meaning space can be described using a small set of components. These can become crowded as the vocabulary increases, thus increasing the potential for interference between words [43]. Not only does this increase the probability for confusion in communication, but it also makes the expression of novel ideas more difficult as the semantics of every word will constantly be diluted with the connotations of neighbouring phonological words [73].

4.2.2 Evolution: Transmission vs. Expression

The competing pressures of cognitive processes can be reformulated from an Evolutionary point of view [16]. The fact that a system as complex as language can be acquired and used with very little effort suggests an intrinsic evolutionary relationship between the neural mechanisms responsible for language, and language structure itself. This could be interpreted as biological evolution – the brain has evolved to suit language – however the rate of change in language far exceeds the rate of genetic change, implying that language has evolved and adapted to human neural mechanisms in a process of cultural evolution [21]. This idea was put forward by Darwin who compared language to a complex organism of highly connected constraints which evolves under the antagonistic pressures of human cognition and determines the survival of linguistic items [28]; language must be expressive enough to be useful, while being simple to allow transmission between generations.

Transmission of language is closely linked to the ease of acquisition; it relies on language being compressible to minimise the cost of precisely conveying *any* meaning [58]. Compressibility is enhanced by systematicity, which allows languages to be described concisely. As expected, learners are naturally biased to such languages as compressed mental representations require less effort to learn [18].

Expressivity in the context of evolution is constrained by the same goals as described above – favouring arbitrary mappings which allow the reliable transfer of distinct, potentially novel meaning with minimal effort – however can also be interpreted in terms of compressibility. Though transmission requires high system compressibility, expressivity favours the compressibility of individual signals which is generally compromised by system compression [58]. Take for example the word 'run': a compressible system would most likely contain the past tense 'runned', however 'ran' is a more compressible individual signal.

4.3 Potential Explanations for Arbitrariness

These opposing pressures offer some explanation for the extreme degree of arbitrariness in form-meaning vocabulary mappings, especially the forces exerted by the need to enhance expressivity. An additional explanation arises from the fact that language doesn't occur in isolation. The context in which language is learned and used provides a multitude of cues which guide and constrain the semantic interpretation of a corresponding form [73].

Contextual information that accompanies language can stem from many sources - social cues like the model of participants, eye gaze, gesture, and tone of voice, environmental cues like the saliency of present features or objects, or cues embedded directly in language [3] [40] [15]. Computational models of language acquisition that include contextual information are better able to capture aspects of human language learning like mutual exclusivity and object individuation [40].

Context directly affects language structure by providing new solutions to optimise forces acting on language, such as those discussed above. For example, it has been posited that all efficient communication systems will contain some ambiguity as long as context encodes some information to guide semantic interpretation; ambiguity allows the reuse of sounds and words that require less effort to produce and understand, and if context encodes useful semantic information, unambiguous languages would be at least partially redundant [80].

In terms of the relationship between form and meaning, context has been hypothesised to play a role in explaining the balance between systematicity and arbitrariness. Monaghan et al. [73] posit that arbitrary mappings allow contextual information to have a maximal impact on determining the semantic representation of a corresponding form, thus satisfying both the constraints for expressivity by maintaining distinct referents, and for acquisition by relying on context to reduce the cost of learning.

Chapter 5

Experiment 2: Language Design

This section explores the interactions between vocabulary structure and cognitive processes, specifically language acquisition and use. Experiment 1 cast doubt on the existence of global phonosemantic systematicity; though language use is enhanced by arbitrary mappings which enable reliable communication of potentially novel ideas, such a high degree of arbitrariness is unexpected as language acquisition favours simple, systematic, and structured vocabularies that are compressible.

From a cultural evolutionary perspective, it's hard to imagine that language, especially one as wide-spread as English, hasn't evolved to optimise both learning and expression. To further explore the origins of arbitrariness in vocabulary structure, this chapter tests 2 hypotheses that attempt to explain such a high degree of arbitrariness by proposing potential advantages to language acquisition caused by arbitrary mappings. The first is based on the role of contextual information in guiding the semantic interpretation of phonological form and was put forward by Monaghan et al. [73], while the second proposes that arbitrary mappings become necessary for effective language acquisition as vocabulary size increases [43].

Using a simulation inspired by Monaghan et al. and drawing on 2 tasks vital to language acquisition, both hypothesis are tested. This chapter first presents the general simulation design and reimplementation of Monaghan et al.'s study. Secondly, the validity and robustness of Monaghan et al.'s results are tested. Finally, the same simulation is used to explore the vocabulary size hypothesis.

5.1 General Methodology

Monaghan et al. posit that the vocabulary structure's balance between systematicy and arbitrariness has been struck to satisfy the conflicting requirements of learning to individuate the specific meaning of a word and to categorise similar words based on their semantic features, both of which are vitally important to communication [73]. Categorisation is a complicate task as words will belong to multiple categories, however Monaghan et al. focus on the task of learning the grammatical category of a word.

Specifically, they hypothesise that systematic form-meaning mappings facilitates such category learning, while arbitrary mappings facilitates individuation when leaning occurs in the presence of additional contextual information that encodes semantic cues to guide comprehension and learning.

To test this hypothesis, they modelled how arbitrariness and systematicity in vocabulary mappings affect these key language acquisition tasks – individuation and categorisation. A feedforward neural network was trained on systematic and arbitrary sets of mappings between phonological and semantic vectors to predict semantic representations, and the quality of learning over time was assessed over the 2 tasks. 2 simulations were devised; the first using only phonological information as input, and the second including additional contextual information where context is modelled as an additional input feature directly indicating semantic category.

Monaghan et al.'s simulations were reimplemented as a basis for further experimentation into the interactions between vocabulary structure, language acquisition, and expression. These simulations consist of 4 key components:

1. *Training data*: arbitrary and systematic sets of mappings between phonological and semantic feature vectors
2. *Network*: a simple, fully connected network trained with backpropagation
3. *Tasks*: the accuracy scores for individuation and categorisation used to access the quality of learning
4. *Training & Testing*: measuring model performance across training epochs

Details of how each component was implemented in this study are presented below.

Training Data Training data used by Monaghan et al. consists of phonological and semantic representations used to build a small toy language of 12 mappings between 2 distinct phonological classes and 2 distinct semantic classes. The mappings between phonological and semantic vectors reflect vocabulary structure such that in the systematic condition, items from the same phonological class are mapped to the same semantic class, while in the arbitrary condition, half the items of each phonological class are mapped to each semantic class.

All word forms included 3 IPA symbols of the same $Consonant - Vowel - Consonant$ (CVC) form. Each symbol was associated with 11 phone features from a scheme of distinctive feature combinations devised by Harm and Seidenberg [48], resulting in 33-dimensional vectors. To create the 2 distinct phonological classes, 12 words were build using either fricative consonants and front vowels, or plosive consonants and back vowels. A similar language was designed for the current experiment and is displayed in table 5.1.

Semantic space consisted of 10 dimensions, where the semantic classes (A and B) were represented by centres at 0.25 and 0.75 in each dimension. Semantic representations associated with each phonological vector were generated by adding uniform noise in the range ± 0.25 to the corresponding category centre in semantic space.

Table 5.1: Toy Language. In the *systematic* condition, semantic class A contains fricative consonants and front vowels, while class B contains plosive consonants and back vowels. Half the instances of each phonological class are assigned to the semantic classes in the *arbitrary* condition.

Semantic Class	Systematic Condition		Arbitrary Condition	
	Class A	**Class B**	**Class A**	**Class B**
Phonological Form	fi:z	gOk	fi:z	fIz
	fIz	gu:k	zIf	zi:f
	zi:f	kOg	fIf	zi:z
	zIf	ku:g	ku:g	gOk
	zi:z	kOk	kOk	gu:k
	fIf	gu:g	gu:g	kOg

The Network The basic network architecture from Monaghan et al. included 33 input units, 10 hidden units, and 10 output units; weights for each unit were initialised to a uniform distribution [73]. Given a phonological representation of a word, the network was trained with backpropagation using stochastic gradient descent with a learning rate of 0.05 to predict the corresponding semantic representation. Given the size and simplicity of the dataset, this architecture seems well-suited to the study.

The Tasks (Accuracy Scores) After being trained and presented with phonological forms, semantic representation predictions were used to assess the performance on the individuation and generalisation tasks:

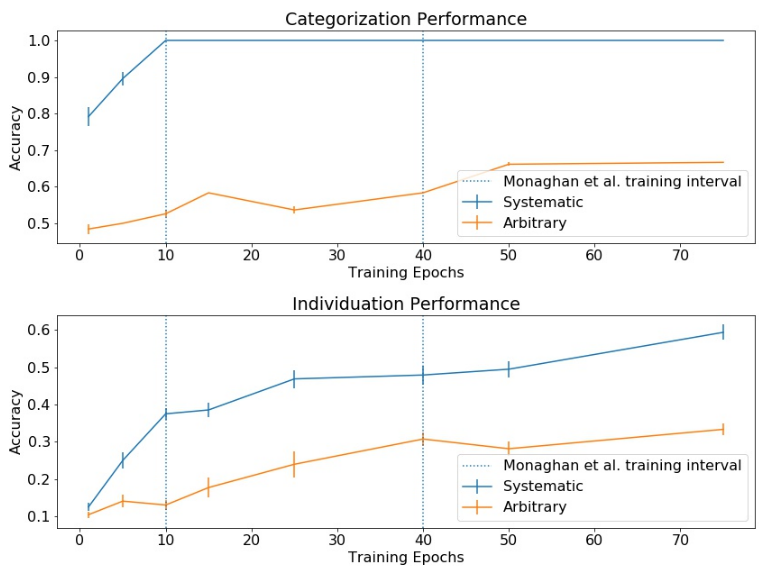
- *Generalisation*: a prediction was correct if the semantic category of the closest semantic pattern to the predicted output was that of the test items true category
- *Individuation*: a prediction was correct if the predicted semantic vector was closest to the true semantic vector associated with the test item.

Training & Testing To measure the quality of learning with the 2 tasks, the network completed a varying number of training blocks where each block consisted of a presentation of the 12 mappings in a random order before being tested on the set of phonological representations. Monaghan et al. assessed model performance after intervals of $\{10,12,30,40\}$ training blocks, however the current implementation assessed performance after $\{1,5,10,15,25,40,50,75\}$ training blocks to maintain comparable results from all experiments. The model was tested over the training interval 16 times to mitigate the effects of random weight initialisations; the mean performance is reported.

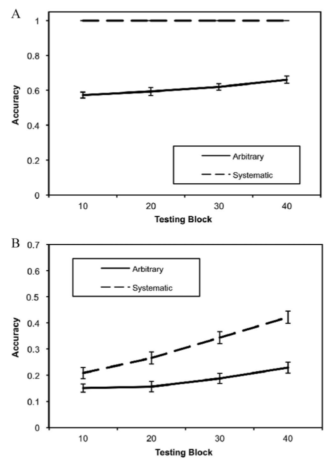
5.1.1 Simulation 1: Effects of Vocabulary Structure on Language Acquisition

The implementation described above models acquisition of systematic and arbitrary mappings without contextual information; results are displayed and discussed below.

Results & Discussion The Monaghan et al. hypothesis expects categorisation and individuation performance to be higher in the systematic condition than in the arbitrary one. Figure $5.1(a)$ demonstrate that this is the case, mirroring the results achieved by Monaghan et al., displayed in figure $5.1(b)$; the accuracy with which the model can categorise and individuate words is consistently higher under systematic conditions than arbitrary ones. Slight differences between Monaghan et al.'s results and the current ones were attributed to differences in network design such as the initialisation of weights which was not discussed by Monaghan et al..



(a) Current Results



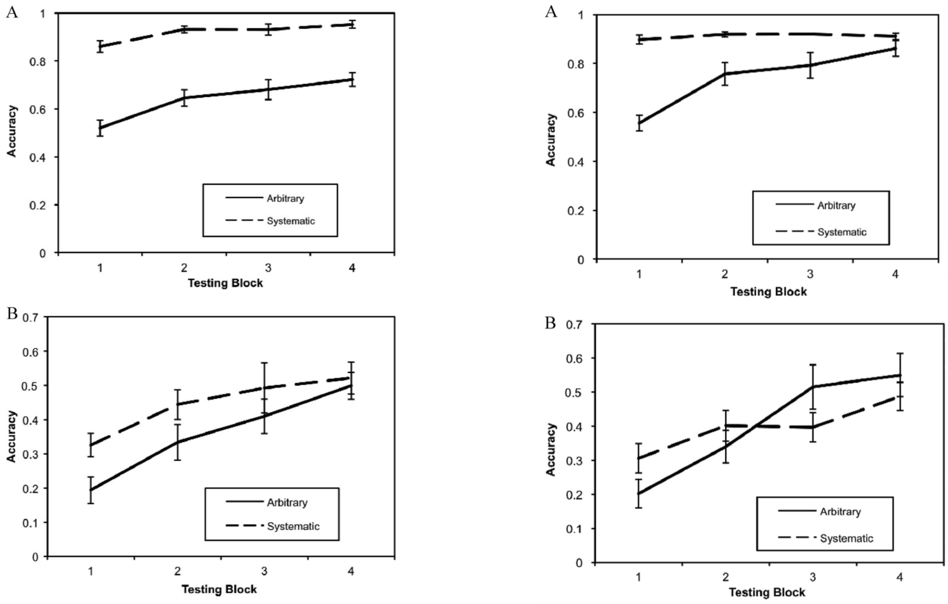
(b) Monaghan et al.'s Results for (A) Categorization and (B) Individuation [73]

Figure 5.1: Average Performance Variation with Training Epochs on Categorization and Individuation Tasks Without Context

There were 2 main concerns about general simulation design that cast doubt on the reliability and realistic credibility of these results; firstly, the structure of model training and testing, and secondly, the simplicity of the simulation and dataset.

The first issue is centred around a golden rule of Machine Learning - separation of training and test sets. Given that this simulation aims to model language acquisition which involved multiple exposures to new words [72], it seems justified to make an exception to this Machine Learning practice.

The second involves simulation design, specifically the extremely simplified representations of learning, form, and semantics. To test to robustness of their simulation results, Monaghan et al. completed parallel behavioural studies using the same toy words and pictures from 2 distinct categories (objects and actions) as semantic representations; results for both simulations are displayed in figure 5.2.



(a) Simulation 1: no context

(b) Simulation 2: context

Figure 5.2: Average Human Performance Variation with Training Epochs for Language from table 5.1. Figures from Monaghan et al. [73]

Results from the first behavioural study matched the patterns in results from simulation 1 – systematic mappings resulted in higher accuracies for both categorization and individuation. Together, these results to provide support for the credibility of simulation design, as well as the view that systematicity facilitates language acquisition while arbitrariness is a hinderance. However, natural language learning doesn't occur in isolation as it has been modelled in this simulation.

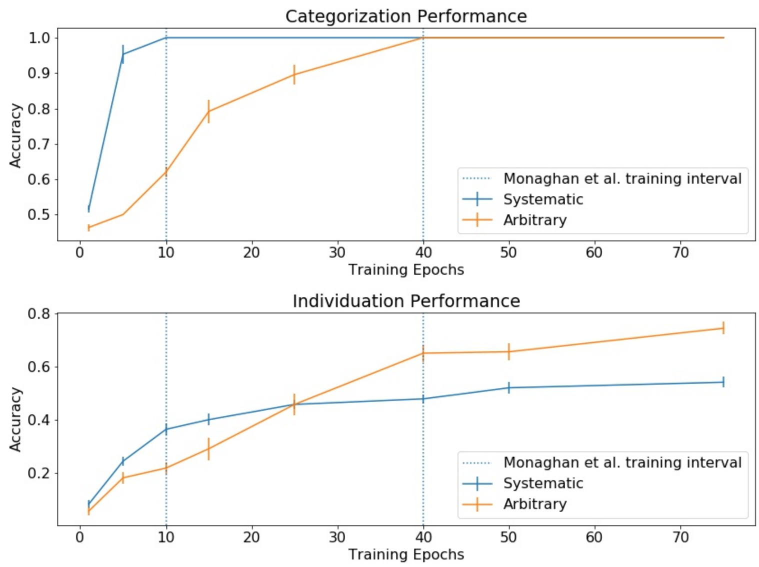
5.1.2 Simulation 2: Effects of Vocabulary Structure and Context on Language Acquisition

As noted in section 4.3, language learning is dependent on a host of contextual cues that constrain the potential meaning of a new word. Monaghan et al. hypothesis that the arbitrary form-meaning mappings maximise the effects of such contextual information in guiding semantic individuation, thus providing an advantage to acquisition and an explanation for the degree of arbitrariness in real vocabularies.

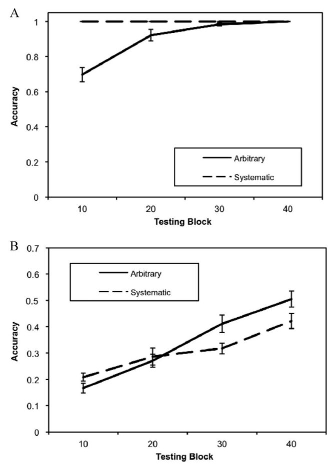
To test this, the experimental set up from *Simulation 1* was slightly modified. The new method, results, and discussion are detailed below.

Method Given that task categorisation in these simulations is constrained to the grammatical categories, Monaghan et al. integrated contextual information contained in language that denotes the grammatical category of a word, such as morphological inflections or surrounding function words. These were operationalised as an additional binary input feature for each phonological vector that corresponded perfectly to the true semantic category [73]. The current experimental set-up was identical to Simulation 1, with this addition.

Results Given that the additional contextual cue encodes semantic information, performance should improve across both conditions and tasks; the context cue is directly indicative of word category, and can help guide semantic predictions especially in the arbitrary condition. To support the hypothesis described above, individuation performance under arbitrary conditions should not only improve, but also exceed systematic performance. Figure 5.3 demonstrates that current results match those obtained by Monaghan et al.; not only does individuation performance under arbitrary conditions increase compared to Simulation 1 without context, but it surpasses systematic individuation performance.



(a) Current Results



(b) Monaghan et al.'s Results for (A) Categorization and (B) Individuation [73]

Figure 5.3: Average Performance Variation with Training Epochs on Categorization and Individuation Tasks Without Context

The issues identified in the results section for simulation 1 (5.1.1) are also applicable in simulation 2, as are the justifications. Monaghan et al. completed a second behavioural study in tandem to this simulation (see figure 5.2), the results of which match the patterns produced by the simulation [73].

These results imply that context improves performance for categorisation and individuation in the arbitrary conditions, however, because the results also demonstrate that context doesn't necessarily improve individuation performance under systematic conditions, they support the hypothesis that the effects contextual cues are maximised by arbitrary mappings.

5.2 Investigation 1 - Phonological Features

Results from the 2 simulations described in sections 5.1.2, 5.1.2 provide some interesting insights to how language acquisition and vocabulary structure interact, however the dataset of 12 mappings is extremely small. To ensure that results were robust to a different representation of phonology, and to allow construction of larger datasets, the same simulations were run using the phonological features derived in Experiment 1 (section 3.1.2).

This experiment involved generating new sets of mappings based on the phonological features described in Experiment 1.

5.2.1 Method

As the simulation structure remained unchanged, the bulk of this experiment involved automating the generation of training languages. Given that the network input consists of concatenated feature vectors, all phonological items were of a specified *Consonant – Vowel* form, from specified sets of consonants and vowels. Generating datasets involved the following steps:

1. *Grouping Sounds*: as described in section 2.1.1.3 and figure A.2, IPA consonants and vowels can be split into categories based on Manner and Place features. Such sets were constructed from the CELEX IPA symbols to enable generation of word forms from particular phonological classes.

Table 5.2: Phonological categories used to generate languages

	Class	CELEX IPA Symbols
Consonants	*Fricative*	`[f, v, T, D, s, z, x, h, Z]`
	Plosive	`[p, b, t, d, k, g]`
	Nasal	`[m, n, N]`
Vowels	*Front*	`[i, e, 3, &]`
	Back	`[u:, V, O:, A, O]`

2. *Generate Phonological Representations*: Given a specified form, and the sound sets from which consonants and vowels should be drawn, phonological representations were generated by randomly selecting symbols from the sound sets in table 5.2.

3. *Convert to Feature Sets*: similarly to Experiment 1, sequences of IPA symbols were parsed to produce vector concatenations of the corresponding phonological feature vectors.

The network input layer was also altered to account for the change in input length.

5.2.1.1 Test 1.a: Same Language

The first goal of this investigation was to ensure that results from simulations 1 and 2 were robust to different phonological feature representations. The same word forms displayed in table 5.1 were converted into feature representations using the phonological features from Experiment 1, resulting in $3*24=72$-dimensional vectors.

Simulations with the new phonological representations were run 16 times, reporting the performance mean and standard deviations.

5.2.1.2 Test 1.b: Generated Languages

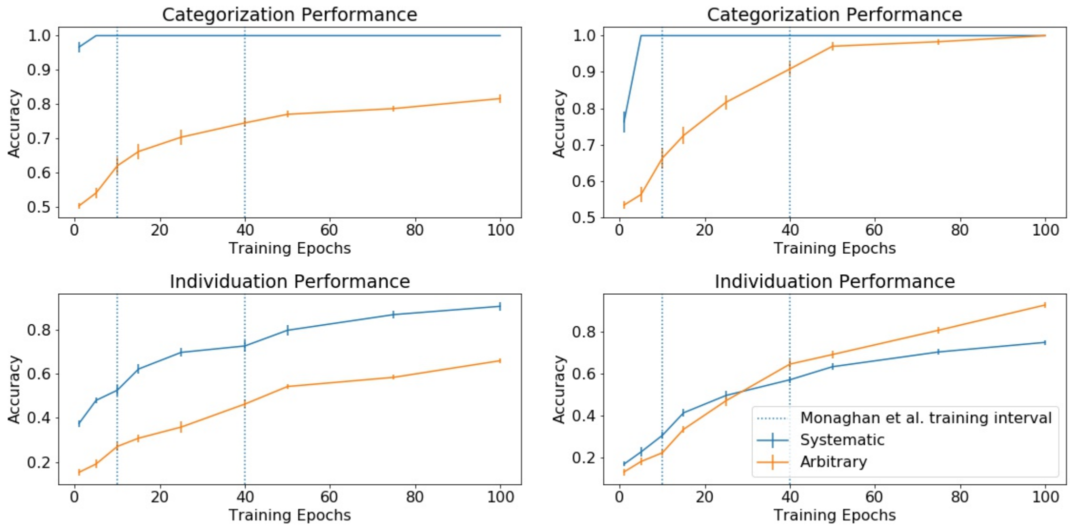
The automatic generation of phonological forms allowed different datasets to be generated, further testing the robustness of results from Simulation 1 and 2. Both simulations were run on 15 language sets of 12 mappings generated from the same phonological sound sets used by Monaghan et al. (fricative consonants and front vowels, or plosive consonants and back vowels). The performance of the model was measured on each language 16 times to reduce the effects of random weight and bias initialisation.

5.2.2 Results & Discussion

Test 1.a.: Same Language The results from Test 1.a. display the same trends as obtained when simulation 1 was run using phonological representations derived from Harm and Seidenberg's phonological features, and Monaghan et al.'s first behaviour study - systematic mappings result in better scores on on both tasks. Results, displayed in figure 5.4, demonstrate that learning trends of both simulations are robust to different phonological representations.

Though the performance trends are consistent when using both sets of phonological features, there is a notable difference between the accuracy scores achieved. Generally, scores improved when trained on mappings with form representations constructed from the Experiment 1 phonological features, however difference between the systematic and arbitrary conditions decreased. Given that all other factors remained constant, the performance differences must be due to how the network processed the different phonological feature representations.

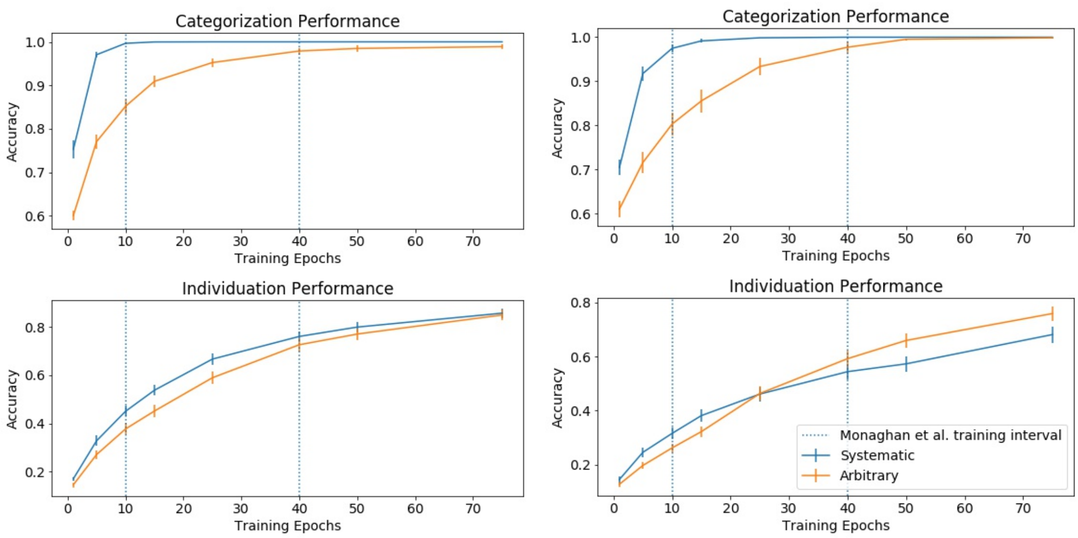
Harm and Seidenberg used 11 features to represent each IPA symbol while the current representations are 23-dimensional. From a Machine Learning perspective, the increase in general performance is likely due to extreme overfitting caused by the additional features while the decrease in performance difference between the arbitrary and systematic conditions could be due to the additional features making systematicity harder to detect and thus distinguish from arbitrary mappings. Though the additional phonological detail encoded in the extra features increases the neural network performance, it is difficult to relate these results to human learning and determine whether humans represent phonology with more or less detail.



(a) Simulation 1: no context (b) Simulation 2: context

Figure 5.4: Average Performance Variation with Training Epochs With Context on Language from table 5.1 (Experiment 1 Phonological Features)

Test 1.b.: Generated Languages Given that the toy language used throughout this experiment was extremely constrained, this experiment aimed to test if results were also robust to training languages. Simulations 1 and 2 were run on different sets of mappings based on phonological feature vectors randomly generated from the sound sets used by Monaghan et al. and of the same *CVC* form. Results are displayed in figure 5.5, displaying similar patterns found in all the previous experiments.



(a) Simulation 1: no context (b) Simulation 2: context

Figure 5.5: Average Performance Variation with Training Epochs With Context on Randomly Generated Language (Experiment 1 Phonological Features)

Though similar trends are present, they are less pronounced. Given that the semantic representations and phonological features used to construct phonological vectors were unchanged, these results imply that properties of the training language affect learning.

An important feature to consider in this case is the variability of the phonological sound sets used to generate languages; the language inspired by Monaghan et al.'s contained form presentations drawn from sound sets of 2 symbols while the magnitudes of the current sound sets are variable and all larger than 2 (see table 5.2). Both generated languages (`(fricative, back)` and `(plosive, front)`) are thus more variable than the Monaghan et al.-inspired languages. Such variation reduced the 'systematicity' across each training language, which could explain reduced differences in performance between the systematic and arbitrary conditions. Addition experimentation demonstrated that differences in performance were amplified to the levels found in Test 1.a. as the variability of phonological sound sets was reduced.

An unexpected feature of these results is that though arbitrary mappings produce a higher individuation accuracy than systematic mappings when context is present, the arbitrary condition still produce worse results than when no additional contextual information is provided. Monaghan et al.'s results, as well as the results from Test 1.a., demonstrate that contextual information reduced individuation performance under systematic conditions, but enhances performance in arbitrary conditions. Though individuation performance under systematic conditions decreases when context is provided, the final individuation performance under arbitrary conditions is slightly lower when context is provided.

Conclusions - Investigation 1 The performance trends reported by Monaghan et al. seem to be robust to different phonological feature representations, however are slightly sensitive to characteristics of the training language. Given that the language used in their original simulations was so constrained in terms of variability and size, this is to be expected. However, discovering such sensitivity encourages more rigorous testing in further work, specifically regarding the complexity of the training language by altering features such as variability and vocabulary size.

5.3 Investigation 2 - Vocabulary Size

This investigation explores another potential explanation for the prevalence of arbitrariness in vocabulary structure through a hypothesis put forward by Gasser: the sheer size of the vocabulary [43]. Gasser agrees that systematic form-meaning mappings should facilitate learning and comprehension in terms of increased learning speeds, reduced memory requirements, and added constraints on potential semantic interpretations of form, however he posits that as the vocabulary size grows, systematicity will eventually impede both acquisition and communication.

Consider the form-meaning space in which mappings can be represented as (x, y) coordinates; as the vocabulary grows, the average distance between mappings will decrease and eventually produce interference on both dimensions as representations overlap.

Form interference introduces synonymy while meaning interference produces ambiguity, both of which reduce the efficacy of acquisition and communication. Because systematic mappings can be represented on a single component, interference is much more likely to occur than if the relationship between form and meaning was arbitrary.

To test this hypothesis, Gasser simulates language acquisition by training a neural network to predict form representation from semantic representations. Though his simulation is more rudimentary than those implemented above, they involve similar components:

- *Training data:* training data involves of pairs of 'form' and 'meaning' representations. Each representation consists of 3 dimensions, which can each take a value between 0 – 9. Dimension values are represented by Gaussian patterns spread across 10 input units. Systematic mappings require semantic and form representations to match on 2 dimensions, while arbitrary mappings simply involve mappings between representations of randomly-selected dimension values. 2 vocabulary sizes were tested – 15 and 100 mappings.

- *Network:* Gasser also makes use of a connectionist network trained with back propagation. The network structure includes 30 input units, 64 hidden units, and 30 output units.

- *Training & Testing:* In each epoch of training, the network was presented with 5 instances of each mapping - 1 was the canonical mapping, while the other 4 included additional noise (each dimension value in the form and meaning representations was changed by ± 1 with a probability of 0.2). The network was trained on intervals of $[10, 20, ..., 80, 90]$ epochs. Performance was measured as the mean squared error after each training interval.

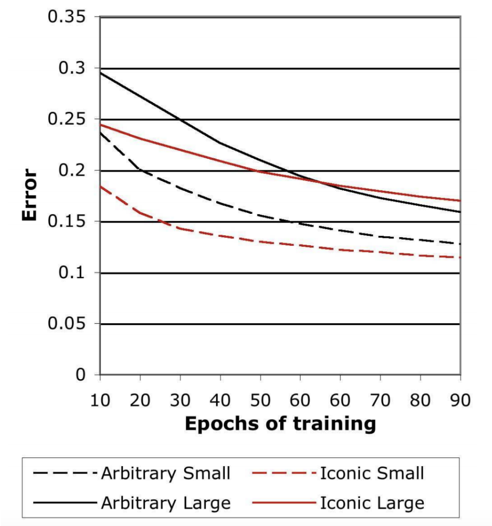


Figure 5.6: Gasser's Results: variation of average mean squared training loss with training for small (15) and large (100) vocabularies [43]

Gassers results, displayed in figure 5.6, confirm the expectation that systematic mappings provide an advantage for small vocabulary sizes, possibly due to the explicit correlations that back propagation can quickly discover. For large vocabularies, however, the network trained on arbitrary form-meaning mappings eventually out-performs the model trained on systematic mappings. He posits that the advantage occurs due increased confusion due to the increased proximity of mappings, and the presence of noise in the vocabulary.

This investigation involves testing Gasser's hypothesis with a more realistic simulation. The subsequent sections will describe the current methodology used to explore the effects of vocabulary size and structure on language acquisition, corresponding results, and discussion.

5.3.1 Method

Simulation 1 and 2 (sections 5.1.1, 5.1.2) were used as a basis for the current experimental design to test the robustness of Gasser's hypothesis as they provide some improvements with respect to Gasser's simulation.

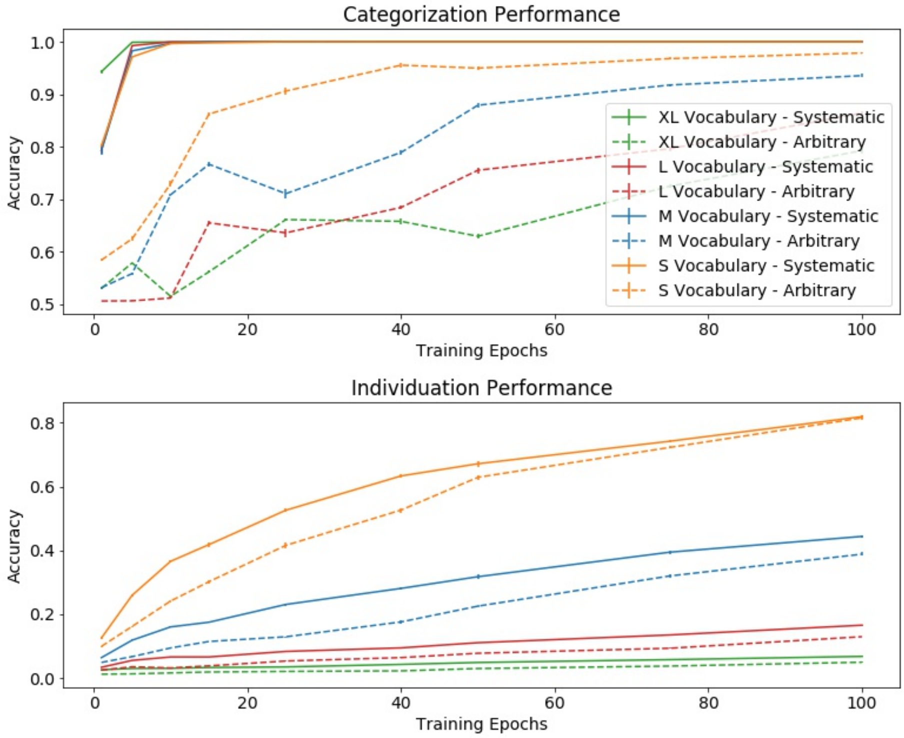
One of the key shortcomings of Gasser's experimental design was the extreme degree of abstraction. This issue was discussed with respect to simulations 1 and 2 above; though the semantic space employed in simulations 1 and 2 was also very symbolic, the form representations encode much more realistic phonological information. Another deviation from Gasser's design involves how performance is measured – the exemplar-based accuracy scores used in simulation 1 and 2 are a more realistic representation of how human learning occurs than mean squared training error as used in the Gasser experiment [76] [85].

Simulations 1 and 2 were were run as described in section 5.2 on different vocabulary sizes built from the phonological features from Experiment 1. This involved generating differently-sized vocabularies with 16, 32, 50, and 100 respective mappings. The network was trained on each vocabulary 20 times to account for network weight initialisations; the mean performance across 5 languages of each size was measured to mitigate the effects of random language generation.

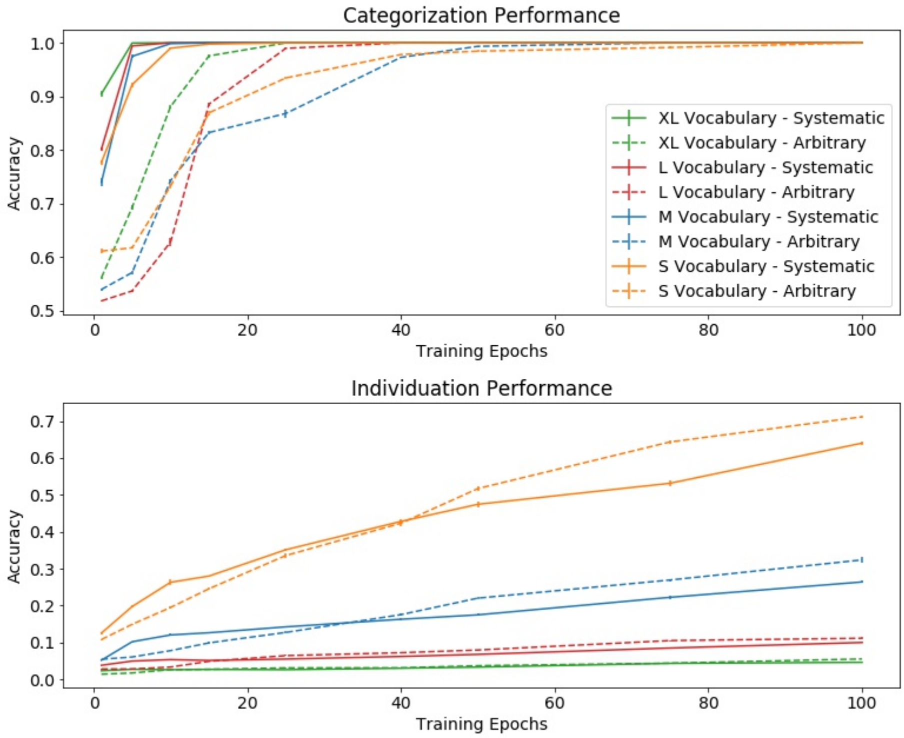
5.3.2 Results & Discussion

The results for both simulation 1 and 2 with varying vocabulary sizes are displayed in figures 5.7. Though differences in simulation design, network structure, and information representation make comparing resulting values to Gasser's results irrelevant, performance trends between Gasser's experiment and the 2 current simulations were explored with respect to Gasser's hypothesis that arbitrary mappings facilitate language acquisition as the vocabulary grows.

Gasser's design was more similar to the individuation task in Simulation 1 as it did not include contextual information and measured mean squared error (a variant of the individuation task), however current results show no advantage from arbitrary mappings for any vocabulary size. 2 potential explanations for such deviation stem from



(a) Simulation 1: no context



(b) Simulation 2: context

Figure 5.7: Average Performance Variation with Vocabulary Size Language. *For more detailed values, see table B.1*

the notion of interference and are explored below. Though interference makes acquisition and comprehension less efficient, it is a crucial component of real language – synonymy and ambiguity are features of many, if not all, human languages [35] and just as language processes like acquisition and expression occur in the presence of useful contextual information, they also occur amongst a host of environmental and psychological distractions.

A potential explanation is Gasser's addition of noise to both form and meaning representations. Gasser exposed the network to 5 potentially noisy variants of each mapping, forcing them to inhabit a larger portion of the total form-meaning space and thus increasing the probability of overlap (interference). Such noise has not been added to these simulations. Individuation performance is measured by accuracy where a semantic prediction is labelled as correct if it is *nearest* the true semantic mapping. The semantic space is therefore essentially a Voronoi partitioning where each true semantic representation occupies a partition of semantic space with no overlap.

An important component of Gasser's experiment that could also explain the absence of an advantage from arbitrary mappings in simulation 1 was noted in Gasser's paper – the number of dimensions used to represent form and meaning are key parameters to demonstrating the arbitrary advantage. Producing interference in form or meaning spaces with only a few more dimensions would require exponentially larger vocabularies. Given that Gasser's experiment involved 30 input units (3 dimensions) while the current experiment makes use of 72, this is a plausible explanation. The current form representations are more realistic than Gasser's as they are based on phonological features, however this could hint that humans represent phonology differently.

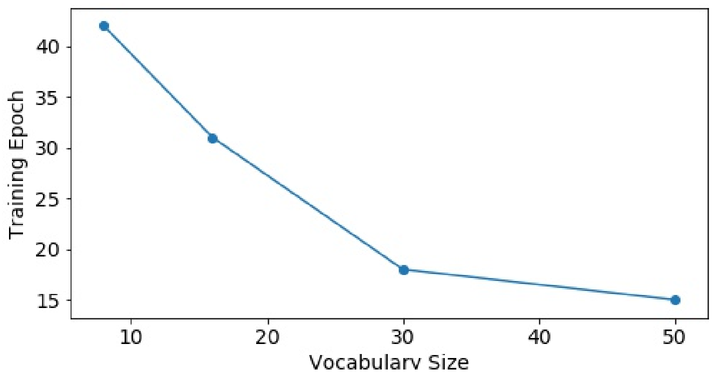


Figure 5.8: Arbitrary Advantage: demonstrates at which training epoch the network trained on arbitrary mappings begins to outperform one trained on systematic mappings for different vocabulary sizes.

On the other hand, results from Simulation 2 (figure 5.7) provide empirical evidence that vocabulary size could affect the balance of arbitrariness and systematicity in the vocabulary. Though systematic mappings provide an early advantage for individuation, eventually performance of the network trained on arbitrary mappings surpasses its systematic counterpart for all tested vocabulary sizes. What's most interesting is when this advantage occurs during training for different vocabulary sizes. Figure 5.8

demonstrates a relationship between the training epoch at which the advantage occurs and the vocabulary size: the advantage occurs after less training as vocabulary size increases. Though the difference in performance between the systematic and arbitrary conditions decreases as vocabulary size grows (see figure 5.7), these results still provide some support for such a relationship. The decrease in performance is to be expected given the task of individuation – the semantic space becomes crowded as the vocabulary grows, making it much less likely that the network's semantic prediction is nearest the true semantic representation. The different in performance under systematic and arbitrary conditions is affected by how semantic patterns are generated, the learning rate, and cost function.

Conclusions - Investigation 2 Though the advantage of arbitrary mappings for individuation is only present in simulation 2, simulation 2 is a more realistic model of acquisition as language never occurs in isolation. Results from both simulations provide further insight into Gasser's hypothesis that arbitrary mappings provide an advantage for language acquisition as vocabulary size increases [43].

Though results from Simulation 1 don't directly support Gasser's hypothesis that 'arbitrariness becomes necessary as the number of words [in the vocabulary] increases' [43], analysis of the results highlights the importance of interference for demonstrating the arbitrary advantage displayed in Gasser's results and suggests that vocabulary size affects acquisition indirectly. The results of Simulation 1 in conjunction with Gasser's findings demonstrate that the dimensionality of form-meaning spaces, the vocabulary size, and the degree of noise in mappings all interact to determine the degree of interference in a vocabulary. Arbitrary mappings can mitigate such interference and thus enhance acquisition.

Simulation 2 integrates a crucial component of language acquisition that was not modelled in Gasser's experiment – context. The results from simulation 2 imply that context may interact with vocabulary size in producing the arbitrary advantage. It is theoretically plausible that the strengths of pressures from systematicity and distinctiveness change according to vocabulary size, especially when considering developmental language acquisition. Psycholinguistics posit that distinctiveness is less important in early stages of acquisition than the promotion of acquisition itself which is facilitated by systematic mappings [91]; this theory is supported by the fact that words which are acquired early are highly systematic compared to words acquired later [74].

Both simulations demonstrate important elements in modelling language processes – noise and context. Similarly to conclusions from simulation 2 demonstrating context an vocabulary size factors interacting to produce the arbitrary advantage, previous research posits that noise and context components interact to maximise efficiency. Piantadosi et al. state that ambiguity arises in all efficient communication systems when useful contextual information is present. Given that semantic interference causes ambiguity, integration of context in language acquisition and comprehension could allow noisy semantic representations.

5.4 General Discussion & Conclusions

The language design experiment involved testing 2 separate hypotheses for how arbitrary form-meaning mappings could be advantageous to language acquisition and thus explain the extreme degree of arbitrariness in form-meaning mappings – the first posited that arbitrary mappings maximised the effects of useful contextual information on acquisition, and the second proposed that arbitrary mappings increase the effectiveness of learning as vocabulary size increases [73] [43]. Both were tested using a simulation design inspired by Monaghan et al. [73].

Though the simulation design was simple, it was selected as Monaghan et al. had completed parallel behavioural studies to support the validity of simulation results. On reflection, the simulation design managed to capture relatively realistic representations of form and meaning while maintaining a high degree of abstraction and control however improvements are possible. Though the form representations are much more uniform than real language as they all have the same *CV* form and are generated only from phonological symbols drawn from 3 sound classes, they encode some true phonological information. The semantic space is also relatively realistic – as discussed in section 2.2.1.2, vector space can provide realistic representations of meaning. However, additional consideration should be focused on more realistic generation of semantic vectors rather than simply adding gaussian noise to predetermined class centres. Other future work could focus on making representations less uniform, or making additional classes.

An interesting aspect of the design that has not been mentioned thus far is how the accuracy metrics of acquisition tasks – categorisation and individuation – were defined. Though mean-squared error as used by Gasser provides some information about how well the network learns to predict the specific corresponding mapping representation, the individuation accuracy score used throughout this study seems a much more realistic model of learning.

On the other hand, the current categorisation tasks could be viewed as a form of exemplar learning. Exemplar learning is a popular class of categorisation methods which involve comparing novel stimuli to known category instances and performing categorisation using similarity metrics, such as the Generalised Context Model [76]. Such models are often compared to Prototype learning which involves constructing prototypes of each category to which novel stimuli are compared to during categorisation, or rule-based exemplar categorisation such as General Recognition Theory which partitions the feature space using decision boundaries [5]. A prototype-based categorisation accuracy metric was implemented and run on simulations, however recent research proposes that humans use a combination of rule-based and similarity-based categorisations methods such as Tenenbaum's Bayesian model of Concept Learning [95].

Future work could modify the task accuracy metrics based on such models; the Bayesian model would be particularly interesting for testing Gasser's vocabulary size hypothesis as it includes a size principle mechanism for guiding which categorisation strategy is used based on category size.

5.4.1 Conclusions

Investigation 1 found that task accuracy trends matched Monaghan's results as well as human performance, generally supporting the hypothesis that arbitrary mappings facilitate language acquisition by maximising the effects of useful contextual information. Such trends were robust to different phonological feature representations but sensitive to complexity of the language driven by characteristics like phonological class variation [74].

On the other hand, though Investigation 2 provided some support for Gasser's hypothesis that large vocabulary size produce an advantage for acquisition of arbitrary mappings, Simulations 1 highlights that vocabulary size only indirectly drives this advantage [43]. Interference, which is affected by factors like vocabulary size, the dimensionality of representations and the degree of noise, is a more directly linked to the advantage of arbitrary mappings on acquisition.

The results of both Investigations together provide more general insights. Firstly, they demonstrate that acquisition may in fact benefit from arbitrary mappings which explains the high degree of arbitrariness in vocabulary mappings. Secondly, they highlight the interconnected nature of language; it is highly probably that many factors provide an arbitrary advantage to acquisition and that these effects combine in complex ways.

Chapter 6

Conclusions

Though language is acquired and used by nearly all humans with very little conscious effort, it is riddled with subtleties and nuances on every level. Cultural evolution posits that language structure is a result of balancing the counteractive pressures applied from the cognitive mechanisms for processing language [58]. This study has provided experimental exploration of the form-meaning relationship of language, as well as of potential explanations for why such a relationship may exist from a cognitive and cultural evolutionary perspective.

Experiment 1 highlighted potential flaws in previously-used methods for studying the form-meaning relationship, specifically the statistical analysis applied to the relationship, and the claims made about the global nature of phonosemantic systematicity [90] [74]. Current results report a negative relationship between form-meaning correlation when measuring with a more appropriate correlation coefficient (Spearman); though not in line with previous research, it could be partially explained by related research into how different phonological features may be reacting to pressures for systematicity and discriminability [93]. More importantly though, analysis of current results in conjunction with previous findings demonstrated that phonosemantic systematicity is more difficult to study that previously imagined, especially regarding how local phonosemantic systematicity affects the global relationship between form and meaning. The current study reveals that such local pockets were potentially oversimplified and underestimated with the Shillcock et al. method. Future studies should take care to avoid methods that make assumptions about the quality of the form-meaning relationship, and ensure that local systematicity is sufficiently accounted for.

Regardless of how phonosemantic systematicity is studied, the current results along with all previous research demonstrate that vocabulary structure is highly arbitrary [74] [93] [90]. Experiment 2 aimed to explore how such structure interacts with cognitive processes required to acquire language. Simulations of language acquisition were used to explore 2 hypothesis for factors that could explain why the form-meaning relationship in terms of their effects on language acquisition – the integration of context, and vocabulary size [73] [43]. Investigations demonstrated that including useful contextual information produced an advantage of arbitrariness for acquisition, and that interference between form-meaning mappings, driven by factors including vocabulary size, impacted how quickly the advantage occurred; performance trends were robust to phonological representations but sensitive to the language complexity such as the variability of phonological classes. The current experimental design successfully captured some aspects of acquisition but highlighted the difficulty of modelling realistic language processes. Obtaining controlled conditions requires the use of simplifying assumptions, however current results demonstrate that interactions between complex components which may be difficult to model, like contextual information and noisy mappings, are vital to producing meaningful results.

Along with their individual conclusions, these experiments both suggest that taking an interdisciplinary approach to the study of language is the best way to uncover meaningful findings.

Bibliography

[1] Charu C Aggarwal, Alexander Hinneburg, and Daniel A Keim. On the Surprising Behavior of Distance Metrics in High Dimensional Space. *ICDT*, 1:420–434, 2001.

[2] Alexandra Y Aikhenvald. Typological distinctions in word-formation. In *Language Typology and Syntactic Description*, pages 1–65. 2007.

[3] Rami Al-Rfou, Marc Pickett, Javier Snaider, Yun-hsuan Sung, Brian Strope, and Ray Kurzweil. Conversational Contextual Cues: The Case of Personalization and History for Response Ranking. jun 2016.

[4] Stephen R Anderson, Anderson, and Stephen R. Morphology. In *Encyclopedia of Cognitive Science*. John Wiley & Sons, Ltd, Chichester, jan 2006.

[5] F Gregory Ashby and James T Townsend. Varieties of Perceptual Independence. *Psychological Review*, 93(2):154–179, apr 1986.

[6] The International Phonetic Association. *Handbook of the International Phonetic Association: A Guide to the Use of the International Phonetic Alphabet*. Cambridge University Press, 1999.

[7] R Baayen, R Piepenbrock, and L Gulikers. CELEX English Linguistics Guide, 1995.

[8] Marco Baroni, Georgiana Dinu, and Germán Kruszewski. Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 238–247, 2014.

[9] Marco Baroni and Alessandro Lenci. Distributional Memory: A General Framework for Corpus-Based Semantics. *Computational Linguistics*, 36(4):673–721, 2010.

[10] David N. Barton and Jamin Halberstadt. A social Bouba/Kiki effect: A bias for people whose names match their faces. *Psychonomic Bulletin and Review*, pages 1–8, jun 2017.

[11] Sergey Bartunov, Dmitry Kondrashkin, Anton Osokin, and Dmitry Vetrov. Breaking Sticks and Ambiguities with Adaptive Skip-gram. *Artificial Intelligence and Statistics*, pages 130–138, 2016.

[12] Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Janvin. A Neural Probabilistic Language Model. *The Journal of Machine Learning Research*, 3:1137–1155, 2003.

[13] Yoav Benjamini. Simultaneous and selective inference: Current successes and future challenges. *Biometrical Journal*, 52(6):708–721, dec 2010.

[14] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching Word Vectors with Subword Information. jul 2016.

[15] Amanda C. Brandone, Khara L. Pence, Roberta Michnick Golinkoff, and Kathy Hirsh-Pasek. Action speaks louder than words: Young children differentially weight perceptual, social, and linguistic cues to learn verbs. *Child Development*, 78(4):1322–1342, jul 2007.

[16] Henry Brighton, Simon Kirby, and Kenny Smith. Cultural selection for learnability: Three principles underlying the view that language adapts to be learnable. *Language origins: Perspectives on evolution, ed. M. Tallerman*, pages 291–309, 2005.

[17] Patricia J. Brooks, Martin D.S. Braine, Lisa Catalano, Ruth E. Brody, and Vicki Sudhalter. Acquisition of Gender-like Noun Subclasses in an Artificial Language: The Contribution of Phonological Markers to Learning. *Journal of Memory and Language*, 32(1):76–95, feb 1993.

[18] Nick Chater and Paul Vitányi. Simplicity: A unifying principle in cognitive science? *Trends in Cognitive Sciences*, 7(1):19–22, jan 2003.

[19] Noam Chomsky and Morris Halle. *The Sound Pattern of English*. Harper & Row, New York, 1968.

[20] Francis Christensen. Wilkins and Royal Society Reform. In Joseph L. Subbiondo, editor, *John Wilkins and 17th-century British Linguistics*. John Bengamins Publishing, 1992.

[21] Morten H. Christiansen and N Chater. Langauge as shaped by the brain. *Behavioral and Brain Sciences*, 31(05):79–89, 2008.

[22] Herbert H. Clark and Deanna Wilkes-Gibbs. Referring as a collaborative process. *Cognition*, 22(1):1–39, feb 1986.

[23] John Clark, Colin Yallop, and Janet Fletcher. *An Introduction to Phonetics and Phonology*. Blackwell Publishing, Oxford, UK, 3 edition, 2007.

[24] G. N. Clements. The geometry of phonological features. *Phonology*, 2(1):225–252, may 1985.

[25] Ronald a Cole, Yonghong Yan, Brian Mak, Mark Fanty, and Troy Bailey. The contribution of consonants versus vowels to word recognition in fluent speech. *Proc. ICASSP*, 2(4):853–856, 1996.

[26] Wikipedia Contributors. Simple English Wikipedia, 2017.

[27] Alan Cruse. *Meaning in Language: An Introduction to Semantics and Pragmatics*. 2011.

[28] Charles Darwin. *On the Origin of the Species*, volume 5. 1859.

[29] Mark Davies. Google Book Corpus (American English), 2011.

[30] Bart de Boer, Wendy Sandler, and Simon Kirby. New perspectives on duality of patterning: Introduction to the special issue. *Language and Cognition*, 4(04):251–259, dec 2012.

[31] Ferdinand de Saussure. *Course in General Linguistics*. NY: McGraw & Hill, New York, 1916.

[32] John DeFrancis. *Visible speech: the diverse oneness of writing systems*. University of Hawaii Press, 1989.

[33] Mark Dingemanse, Damián E. Blasi, Gary Lupyan, Morten H. Christiansen, and Padraic Monaghan. Arbitrariness, Iconicity, and Systematicity in Language, 2015.

[34] Editors of Encyclopaedia Britannica. Semantics, 2017.

[35] Christiane Fellbaum. WordNet. In *The Encyclopedia of Applied Linguistics*. John Wiley & Sons, Inc., Hoboken, NJ, USA, nov 2012.

[36] Edward Finegan. English. In Bernard Comrie, editor, *The World's Major Languages*, chapter 3. Routledge, 2009.

[37] J. R. Firth. Applications of General Linguistics. *Transactions of the Philological Society*, 56(1):1–14, 1957.

[38] John Firth. *The Tongues of Men, and Speech*. Oxford University Press, London, 1964.

[39] A. Fitneva, Stanka, Morten H. Christiansen, and Padraic Monaghan. From sound to syntax: phonological constraints on children's lexical categorization of new words. *Journal of Child Language*, 36(05):967, nov 2009.

[40] Michael C. Frank, Noah D. Goodman, and Joshua B. Tenenbaum. Using speakers' referential intentions to model early cross-situational word learning: Research article. *Psychological Science*, 20(5):578–585, may 2009.

[41] Angela D. Friederici. Neurophysiological markers of early language acquisition: From syllables to sentences. *Trends in Cognitive Sciences*, 9(10):481–488, oct 2005.

[42] Ram Frost and Marian Katz. The Reading Process is Different for Different Orthographies: the Orthographic Depth Hypothesis. In Ram Frost and Marian Matz, editors, *Orthography, Phonology, Morphology, and Meaning*, chapter 4. Elsevier, 1992.

[43] Michael Gasser. The origins of arbitrariness in language. *Proceedings of the 26th Annual Conference of the Cognitive Science Society*, pages 4–7, 2004.

[44] Sahar Ghannay, Benoit Favre, Yannick Estève, and Nathalie Camelin. Word Embedding Evaluation and Combination. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, 2016.

[45] John Goldsmith and Juliette Blevins. The Syllable in Phonological Theory. In John Goldsmith, Jason Riggle, and Alan C. L. Yu, editors, *The Handbook of Phonological Theory: Second Edition*, pages 206 – 244. Wiley-Blackwell, Oxford, UK, sep 1995.

[46] Stefan Th. Gries. Shouldn't it be breakfunch? A quantitative analysis of blend structure in English. *Linguistics*, 42(3):639–667, jan 2004.

[47] Dario Gutierrez, Roger Levy, and Benjamin K Bergen. Finding Non-Arbitrary Form-Meaning Systematicity Using String-Metric Learning for Kernel Regression. *ACL*, pages 2379–2388, 2016.

[48] Michael W Harm and Mark S Seidenberg. Phonology, reading acquisition, and dyslexia: Insights from connectionist models. *Psychological Review*, 106(3):491–528, 1999.

[49] Zellig S Harris. Distributional Structure. *WORD*, 10(2-3):146–162, 1954.

[50] Nigel Hewlett and Janet Mackenzie Beck. Basic Principles of Consonant Description. In *An Introduction to the Science of Phonetics*, chapter 6. Routledge, 2013.

[51] Nigel Hewlett and Janet Mackenzie Beck. Basic Principles of Vowel Description. In *An Introduction to the Science of Phonetics*, chapter 5. Routledge, 2013.

[52] Leanne Hinton, Johanna Nichols, and John J. Ohala, editors. *Sound symbolism*. Cambridge University Press, Cambridge, 1995.

[53] Charles Hockett. The Origin of Speech. *Scientific American*, 203(3):88–97, 1960.

[54] Laura Janda. Inflectional Morphology. In *The Oxford Handbook of Cognitive Linguistics*, page 308. 2012.

[55] Daniel Jurafsky and James Martin. *Speech and Language Processing*. 2016.

[56] Daniel Jurafsky and James H. Martin. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. *Computational Linguistics*, 26(4):638–641, 2007.

[57] Michael H Kelly. Using sound to solve syntactic problems: The Role of phonology in grammatical category assignments. *Psychological Review*, 99(2):349–364, apr 1992.

[58] Simon Kirby, Monica Tamariz, Hannah Cornish, and Kenny Smith. Compression and communication in the cultural evolution of linguistic structure. *Cognition*, 141:87–102, 2015.

[59] Hubert Kowalewski. Against arbitrariness: An alternative approach towards motivation of the sign. *Public Journal of Semiotics*, 6(2):14–31, 2015.

[60] Peter. Ladefoged and Sandra Ferrari. Disner. *Vowels and consonants*. Wiley-Blackwell, 2012.

[61] Roger. Lass. *Phonology: An introduction to basic concepts*. Cambridge University Press, 1984.

[62] Vladimir I. Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, 10(8):707–710, 1966.

[63] Isabelle Y Liberman, Donald Shankweiler, F. William Fischer, and Bonnie Carter. Explicit syllable and phoneme segmentation in the young child. *Journal of Experimental Child Psychology*, 18(2):201–212, oct 1974.

[64] Jinyuan Liu, Wan Tang, Guanqin Chen, Yin Lu, Changyong Feng, and Xin M Tu. Correlation and agreement: overview and clarification of competing concepts and measures. *Shanghai Archives of Psychiatry*, 28(2):115–120, apr 2016.

[65] Gary Lupyan and Rick Dale. Language Structure Is Partly Determined by Social Structure. *PloS one*, 5(1):e8559, jan 2010.

[66] N. Mantel. The detection of disease clustering and a generalized regression approach. *Cancer Research*, 27(2):209–220, 1967.

[67] Ignatius Mattingly. Linguistic Awareness and Orthographic Form. In Ram Frost and Marian Katz, editors, *Orthography, Phonology, Morphology, and Meaning*, chapter 1. Elsevier, 1992.

[68] Jean Baptiste Michel, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K. Gray, Joseph P. Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, Steven Pinker, Martin A. Nowak, and Erez Lieberman Aiden. Quantitative analysis of culture using millions of digitized books. *Science*, 331(6014):176–182, 2011.

[69] T Mikolov, K Chen, G Corrado, and J Dean. Efficient Estimation of Word Representations in Vector Space. 2013.

[70] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed Representations of Words and Phrases and Their-Compositionality. *Advances in Neural Information Proessing Systems*, pages 3111–3119, 2013.

[71] George A. Miller. WordNet: a lexical database for English. *Communications of the ACM*, 38(11):39–41, nov 1995.

[72] Padraic Monaghan and Mh Christiansen. Why form-meaning mappings are not entirely arbitrary in language. *Proceedings of the 28th annual conference of the cognitive science society*, (1848):1838–1843, 2006.

[73] Padraic Monaghan, Morten H. Christiansen, and Stanka A. Fitneva. The arbitrariness of the sign: Learning advantages from the structure of the vocabulary. *Journal of Experimental Psychology: General*, 140(3):325–347, aug 2011.

[74] Padraic Monaghan, Richard C Shillcock, Morten H Christiansen, and Simon Kirby. How arbitrary is language? *Philosophical Transactions of the Royal Society B: Biological Sciences*, 369(1651):20130299–20130299, sep 2014.

[75] John Nerbonne and Wilbert Heeringa. Measuring dialect distance phonetically. In *Workshop on Computational Phonology, Special Interest Group of the Association for Computational Linguistics*, number 1995, pages 11–18, 1997.

[76] Robert M Nosofsky. Attention, Similarity, and the Identification-Categorization Relationship. *Journal of Experimental Psychology: General*, 115(1):39–57, mar 1986.

[77] Katya Otis and Eyal Sagi. Phonaesthemes: A Corpus-Based Analysis. *Proceedings of the 30th Annual Meeting of the Cognitive Science Society*, pages 1–6, 2008.

[78] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014.

[79] Pamela Perniss, Robin L Thompson, and Gabriella Vigliocco. Iconicity as a general property of language: Evidence from spoken and signed languages. *Frontiers in Psychology*, 1(DEC):227, 2010.

[80] Steven T. Piantadosi, Harry Tily, and Edward Gibson. The communicative function of ambiguity in language. *Cognition*, 122(3):280–291, mar 2012.

[81] Steven T Piantadosi, Harry J Tily, and Edward Gibson. The communicative lexicon hypothesis. *The 31st annual meeting of the Cognitive Science Society (CogSci09)*, (i):2582–2587, 2009.

[82] V.S Ramachandran and E.H. Hubbard. Synaesthesia – A Window Into Perception, Thought and Language. *Journal of Consciousness Studies*, 8(12):3–34, 2001.

[83] Radim Rehurek and Petr Sojka. Software framework for topic modelling with large corpora. In *LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 46–50, 2010.

[84] Jason Riggle. Phonological Feature Chart, 2008.

[85] Jeffrey N Rouder and Roger Ratcliff. Comparing Categorization Models. *Journal of Experimental Psychology: General*, 133(1):63–82, mar 2004.

[86] Josef Ruppenhofer, Michael Ellsworth, and Mrl Petruck. FrameNet II: Extended theory and practice. *International*, page 118, 2016.

[87] Harold Sackrowitz and Ester Samuel-Cahn. P values as random variables-expected P values. *American Statistician*, 53(4):326–331, nov 1999.

[88] Gerard Salton. The Smart document retrieval project. In *Proceedings of the 14th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 356–358, 1991.

[89] Nathan C Sanders and Steven B Chin. Phonological distance measures. *Journal of Quantitative Linguistics*, 16(1):96–114, feb 2009.

[90] Richard C Shillcock and Simon Kirby. Filled pauses and their status in the mental lexicon. *ISCA Tutorial and Research Workship on Disfluency in Spontaneous Speech*, pages 1–4, 2001.

[91] Ferrinne Spector and Daphne Maurer. Synesthesia: A New Approach to Understanding the Development of Perception. *Developmental Psychology*, 45(1):175–189, jan 2009.

[92] Gregory T. Stump. *Inflectional Morphology. A Theory of Paradigm Structure*. 2003.

[93] Monica Tamariz. Exploring systematicity between phonological and context-cooccurrence representations of the mental lexicon. *The Mental Lexicon*, 3(2):259–278, 2008.

[94] Paul Taylor, Alan W Black, and Richard Caley. The Architecture of the Festival Speech Synthesis System. *Proc. 3rd ESCA Workshop on Speech Synthesis*, pages 147–151, 1998.

[95] Joshua B Tenenbaum. Rules and similarity in concept learning. *Advances in neural information processing systems*, 12:59–65, 2000.

[96] Rebecca Treiman. The structure of spoken syllables: Evidence from novel word games. *Cognition*, 15(1-3):49–74, dec 1983.

[97] Peter D Turney and Patrick Pantel. From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, 37:141–188, 2010.

[98] Russell Ultan. Size-sound symbolism. *Universals of Human Lanugage*, 2:535–568, 1978.

[99] Clara Vania and Adam Lopez. From Characters to Words to in Between: Do We Capture Morphology? 2017.

[100] Robert A Wagner and Michael J Fischer. The String-to-String Correction Problem. *Journal of the ACM*, 21(1):168–173, 1974.

[101] Frank Wilcoxon. Individual Comparisons of Grouped Data by Ranking Methods. *Journal of Economic Entomology*, 39(2):269–270, 1946.

[102] John Wilkins. *An Essay Towards a Real Character and a Philosophical Language*. Gillibrand, London, 1668.

[103] George Kingsley Zipf. *Human behavior and the principle of least effort. An introduction to human ecology.* 1949.

[104] George Kingsley Zipf. *The Psycho-Biology of Language: an Introduction to Dynamic Philology*. Routledge, 2013.

Appendices

Appendix A

Phonosemantic Systematicity

IPA	example	SAM-PA	CELEX	CPA	DISC
ɪ	pit	I	I	I	I
ɛ	pet	E	E	E	E
æ	pat	{	&	^/	{
ʌ	putt	V	V	^	V
ɒ	pot	Q	O	O	Q
ʊ	put	U	U	U	U
ə	another	@	@	@	@
iː	bean	i:	i:	i:	i
ɑː	barn	A:	A:	A:	#
ɔː	born	O:	O:	O:	$
uː	boon	u:	u:	u:	u
ɜː	burn	3:	3:	@:	3
eɪ	bay	eI	eI	e/	1
aɪ	buy	aI	aI	a/	2
ɔɪ	boy	OI	OI	o/	4
əʊ	no	@U	@U	O/	5
aʊ	brow	aU	aU	A/	6
ɪə	peer	I@	I@	I/	7
ɛə	pair	E@	E@	E/	8
ʊə	poor	U@	U@	U/	9
æ	timbre	{~	&~	^/~	c
ɑ̃ː	détente	A~:	A~:	A~:	q
æ̃ː	lingerie	{~:	&~:	^/~:	0
ɒ̃ː	bouillon	O~:	O~:	O~:	~

Table 4: Computer phonetic codes for English vowels and diphth

(a) Vowels

IPA	example	SAM-PA	CELEX	CPA	DISC
p	pat	p	p	p	p
b	bad	b	b	b	b
t	tack	t	t	t	t
d	dad	d	d	d	d
k	cad	k	k	k	k
g	game	g	g	g	g
ŋ	bang	N	N	N	N
m	mad	m	m	m	m
n	nat	n	n	n	n
l	lad	l	l	l	l
r	rat	r	r	r	r
f	fat	f	f	f	f
v	vat	v	v	v	v
θ	thin	T	T	T	T
ð	then	D	D	D	D
s	sap	s	s	s	s
z	zap	z	z	z	z
ʃ	sheep	S	S	S	S
ʒ	measure	Z	Z	Z	Z
j	yank	j	j	j	j
x	loch	x	x	x	x
h	had	h	h	h	h
w	why	w	w	w	w
ʧ	cheap	tS	tS	T/	J
ʤ	jeep	dZ	dZ	J/	_
ŋ̩	bacon	N,	N,	N,	C
m̩	idealism	m,	m,	m,	F
n̩	burden	n,	n,	n,	H
l̩	dangle	l,	l,	l,	P
*	father	r*	r*	r*	R
(possible linking 'r')					

Table 3: Computer phonetic codes for English consonants

(b) Consonants

Figure A.1: CELEX to IPA sybmbol Conversion Charts [7]

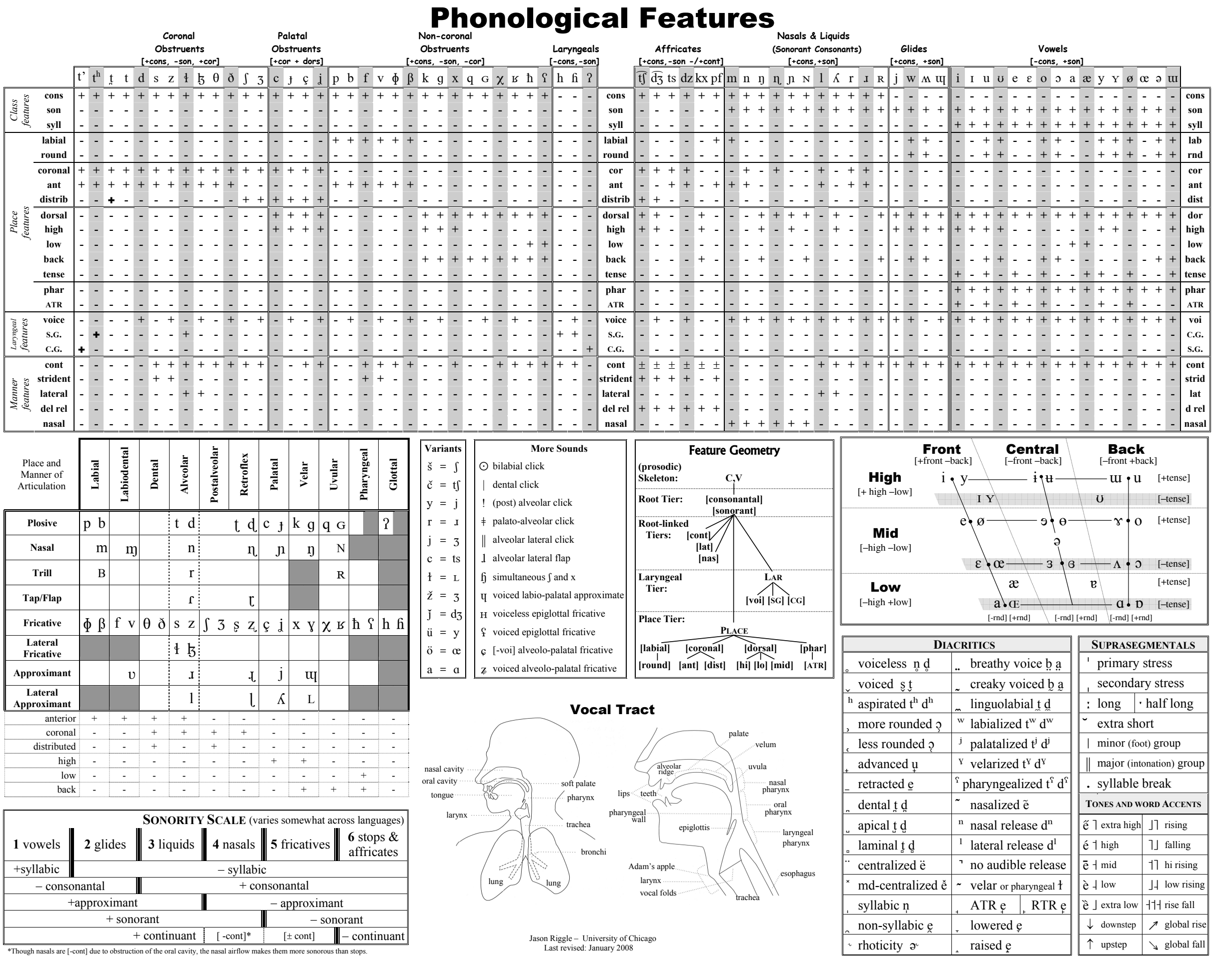


Figure A.2: Riggle's Phonological Feature Chart

Appendix B

Language Design

Table B.1: Final Accuracy Scores (after 100 epochs) of simulations 1 & 2 from Test 2.a.

Size	**Task**	**Condition**	**Final Accuracy Score** No Context	**Final Accuracy Score** Context
Small	*Categorization*	**Systematic**	1.0	1.0
		Arbitrary	0.979	1.0
	Individuation	**Systematic**	0.819	0.624
		Arbitrary	0.815	0.712
Medium	*Categorization*	**Systematic**	1.0	1.0
		Arbitrary	0.936	1.0
	Individuation	**Systematic**	0.444	0.265
		Arbitrary	0.388	0.324
Large	*Categorization*	**Systematic**	1.0	1.0
		Arbitrary	0.864	1.0
	Individuation	**Systematic**	0.166	0.100
		Arbitrary	0.130	0.112
XLarge	*Categorization*	**Systematic**	1.0	1.0
		Arbitrary	0.793	1.0
	Individuation	**Systematic**	0.068	0.047
		Arbitrary	0.050	0.056