Quantifying the Psychological Properties of Words

by

Tomas Engelthaler

A thesis submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy in Psychology

University of Warwick, Department of Psychology

September 2018

Contents

List of Figures

List of Tables

Acknowledgements

My studies were funded by a studentship from the Engineering and Physical Sciences Research Council (EPSRC). I appreciate the council for seeing potential in my work and letting me carry it out successfully.

Reflecting on my time at Warwick, it has been an influential and exciting period of life. I arrived seven years ago, filled with hopes and dreams. I am leaving with the same hopes and dreams, and with an ability to make my ideas a reality. I had the time to explore my passions and to grow both academically and personally.

Thomas Hills had an overwhelmingly positive impact on both my studies and my personal growth. I was enormously lucky to meet him in my undergraduate degree and to later have him be my supervisor on the PhD. His guidance, mentorship, and friendly approach was the single most important factor making Warwick a fruitful place of growth. I have been thankful and mindful of his contributions every day. I hope one day I can have half the impact on a student that he had on me.

Sotaro Kita's brilliant leadership of the Language Group made networking and learning from others a breeze. He has never hesitated to offer advice and support me in my career search. He helped me see my degree on a longer trajectory, appreciating the broader context education has in life.

Claudie Fox provided personal support whenever things were challenging. I will always be thankful for her making sure I am comfortable and confident in my studies. I am leaving Warwick a happy person because of her proactive, caring guidance in both my undergraduate and postgraduate degrees.

Elisabeth Blagrove has been a member of staff I look up to, as well as a friend. She helped me make a seamless transition from my undergraduate course to my postgraduate degree. Her passions for outreach and student engagement encouraged me to pursue the same path, never forgetting to look after my environment as much as I look after myself.

Friederike Schlaghecken encouraged my teaching, providing me with the opportunities and feedback to grow as a mentor to others. She guided me in developing successful frameworks for my students, being directly responsible for the positive results I had when teaching.

The Department of Psychology at Warwick and all its staff have been both my university and my home. They will forever hold a special place in my heart.

Suzanne Aussems has been a flawless role model and a supportive friend, always nudging me in the right direction.

Li Ying and Cynthia Siew have been amazing colleagues and friends. They made my studies at Warwick fun while producing stellar work that motivated me to reach higher in my research.

I appreciate the support from my own family – my sister, parents, and grandparents, powering my studies both emotionally and financially. It would not be possible without them.

Thank you, Winston, for being the scientist I aspire to be.

Declaration

This thesis is submitted to the University of Warwick in support of my application for the degree of Doctor of Philosophy. It has been composed by myself (except for the Macroscope, read below) and has not been submitted in any previous application for any degree.

The work presented (including data generated and data analysis) was carried out by the author except in the cases outlined below:

An exception should be noted for the last chapter of the body, named "The Macroscope". This is a project conducted in the final year of my PhD studies. The Macroscope is a collaborative effort, for which I am only partly responsible. Li Ying, my PhD colleague, is the primary author of the publication that stemmed from this research. My contributions were on the applied side, designing a software platform to demonstrate and deliver the results of my colleague's work. As the software platform was a substantial part of my final year, I include the publication in my thesis. The chapter represents a collaborative effort and I do not take majority credit for the writing – an identical chapter will appear in Li Ying's thesis. A detailed explanation of my contribution to the Macroscope is provided as a preface to the chapter.

All the thesis chapters have been published during my PhD:

- Chapter one: In print as a book chapter in the *Frontiers of Cognitive Psychology* (Network Science in Cognitive Psychology) handbook.
- Chapter two as: Engelthaler, T., & Hills, T. T. (2017). Feature biases in early word learning: network distinctiveness predicts age of acquisition. *Cognitive science*, *41*, 120-140.
- Chapter three as: Engelthaler, T., & Hills, T. T. (2018). Humor norms for 4,997 English words. *Behavior research methods*, *50*(3), 1116-1124.
- Chapter four as: Ying, L., Engelthaler, T., Siew, C. S. Q., & Hills, T. T. (*in review*) The Macroscope: A Tool to Examining the Historical Structure of Language. *Behavior research methods*

My supervisor, Thomas Hills, supported my research throughout and provided feedback on the writing.

Abstract

This thesis explores the psychological properties of words – the idea that words carry links to additional information beyond their dictionary meaning. It does so by presenting three distinct publications and an applied project, the Macroscope. The published research respectively covers: the modelling of language networks to explain lexical growth; the use of high dimensional vector representations of words to discuss language learning; and the collection of a normative dataset of single word humour ratings. The first publication outlines the use of network science in psycholinguistics. The methodology is discussed, providing clear guidelines on the application of networks when answering psychologically motivated questions. A selection of psychological studies is presented as a demonstration of use cases for networks in cognitive psychology. The second publication uses referent feature norms to represent words in a high dimensional vector space. A correlative link between referent distinctiveness and age of acquisition is proposed. The shape bias literature (the idea that children only pay attention to the shape of objects early on) is evaluated in relation to the findings. The third publication collects and shares a normative dataset of single word humour ratings. Descriptive properties of the dataset are outlined and the potential future use in the field of humour is discussed. Finally, the thesis presents the Macroscope, a collaborative project put together with Li Ying. The Macroscope is an online platform, allowing for easy analysis of the psychological properties of target words. The platform is showcased, and its full functionality is presented, including visualisation examples. Overall, the thesis aims to give researchers all that's necessary to start working with psychological properties of words – the understanding of network science in psycholinguistics, high dimensional vector spaces, normative datasets and the applied use of all the above through the Macroscope.

Preface

My PhD started as a fascination with the psychology behind language. The studies during my degree were supported by an overarching theme: “What psychological information can be attached to a word?” It is a strangely broad question. The sheer amount of psycholinguistic research produced daily, in both the private and academic sectors, makes it near impossible to provide a comprehensive answer. Despite the naïve background, my question resonates with a genuine interest. Not necessarily an interest in the psychological states of individual writers, but in the expressive power of language. The fact that language lets us share our inherently unique and personal experiences with others is a captivating premise. The information we exchange goes well beyond simple facts, often containing attitudes, psychological states or even abstract feelings. These properties are usually preserved through written language, demonstrated by a reader’s ability to empathise or emotionally align with the writer.

The goal of my PhD was getting insight into operationalising the psychological information attached to words. The psychological components of words are of interest to researchers in multiple domains. Mohammad (2016) outlines several applied use cases, including: the detection of depression (Cherry, Mohammad, & De Bruijn, 2012), inference of political attitudes (Conover, Ratkiewicz et al., 2011), prediction of movements on the stock market (Yu, Wu, Chang, & Chu, 2013), or optimisation of learning environments (Suero Montero & Suhonen, 2014). The diverse range of applications poses an interesting challenge when studying the methodology being psychological world properties. Insights that are applicable to the field as a whole need to be general enough to cross domains, but also offer theory and application that isn’t already established by the individual fields.

My research proposal was awarded a studentship from the Engineering and Physical Sciences Research Council (EPSRC), with the aim of developing a “Text Analysis and Adaption Toolkit”. This was outlined at an easy-to-use tool that would graphically represent the psychological properties of target words. A project like this first requires a sound understanding of the properties attached to words. In turn, my PhD and thesis reflect this approach. The thesis consists of diverse, theoretically driven studies that utilise the psychological properties paradigm in one way or another. Ultimately, the final chapter presents the Macroscope, which is a more robust version of the software platform originally proposed.

The thesis is organised into five sections. A theoretical introduction to meaning and four content chapters, where each corresponds to a project carried out during my PhD and a successful publication. The first section was published as a chapter in an introductory text book and therefore resembles the structure of a literature review. Sections two and three have been published as journal articles and follow the structure of a published research submission. Finally, the fourth section is an outline of an

applied project, the Macroscope. The Macroscope was a collaborative effort, with Li Ying being the main author and me carrying out the software engineering of the platform.

All sections are true to their published version. They underwent minor changes mainly due to the unification of the formatting. Where relevant, prefaces have been added to introduce the context of each work.

Thematically, the sections are the following:

> The first section is a literature review of using network science to infer semantic relationships between words.
>
> The second section is an analysis of normative datasets to discuss a referent feature distinctiveness effect in early word learning.
>
> The third section is a crowd sourcing project, creating and publishing a novel database of single word humour ratings.
>
> The fourth section is an outline of creating a software platform that shows how psychological properties of target words change over time.

All projects were carried out with the same motives – they allowed me to further my understanding of psychological properties, there was a lack of insight into the field that could benefit from my study, and there was an opportunity for me to carry out the research. This led to a thesis that is thematically diverse. At the same time, all my research is tied together nicely in the Macroscope. The applied platform would not be possible without first studying and understanding all three areas beforehand – that is the mapping of semantic spaces, quantifying psychological differences between words, and creating a database of psychological properties.

It is important to note that regardless of the Macroscope, the individual chapters are informative and contributing to their respective fields. The link to the applied platform is an added benefit, not necessarily the key merit of the publications. Ideally, it makes most sense to evaluate each publication on its own merits. And ultimately discuss the broader implications when reading the Macroscope.

Each chapter is preceded by a brief description. These short prefaces encompass the initial aims of each project were and how it was carried out. The project description is then followed by the published text.

Introduction

The present work explores how to best use linguistic data sets when making inferences about one's psychology. This approach is strongly pragmatic, in the sense that it balances theoretical validity with data availability. A statistical, data-driven view of language comes with a disclaimer at its core. The resulting models are tied to, and framed by, the underlying raw data. While the individual chapters are motivated by producing tangible outputs (public data sets, code or software), the very use of a statistical approach to language makes theoretical assumptions regardless. For example, Chapter 2 establishes a link between the age of acquisition of words (how early in life a word is learned) and the distinctiveness of the object (how unique the referent, i.e. the represented object, is). Such claim comes with the assumption that the process of learning words is fine-grained enough to appreciate the distinctiveness of real-world objects along different dimensions. Similarly, Chapter 3 creates a dataset of single-word humour. That is, a list of words, sorted by how humorous they are. This also poses a strong theoretical assumption. If construction of a humour list is possible, it implies that single words either have a humour dimension attached to them or at the very least link to a mental representation that includes (or can access) an assessment of humour.

These assumptions are not universally accepted. Over the thousands of years of scientific inquiry, academics have proposed a vast range of linguistic theories. The purpose of this introduction is to provide a general overview of relevant linguistic theory. This helps establish where the assumptions of this thesis stand in the broader theoretical field. As a result, both the limitations and the merits of the present work should be apparent. Additionally, framing the thesis in relation to past theory aims to stimulate discussion.

There are three relevant theoretical fields covered – semantics (the discussion of what word meaning entails), semantic change (the study of how meaning changes over time), and lexical development (the outline of factors that shape our word learning). Humour is discussed briefly in relation to semantic theory, and a more thorough theoretical introduction to the field of humour is provided preceding the respective Chapter (3).

The Study of Word Meaning

Ancient Views of Meaning

The idea of conceptualising the link between words and their meanings is not a recent one. In fact, the first study on semantics can be traced to Plato. In Cratylus (383a-d), Plato publishes dialogues with Socrates on the topic of a name (i.e. word) meaning. Word meaning is defined in a denotative sense – the purpose of the word is to define (and link to) the absolute properties of the referent (object the word represents). For example, the word 'runner' would map on to the physical object of a person who moves quickly. Socrates also establishes two additional concepts. He strongly believed that words are built by smaller, meaningful elements in a systematic way. This is what we now call *morphology*. In the above example, the word 'runner' has a base of 'run', which carries the core of its meaning. Unlike most current scholars, however, Socrates also held an absolute, metaphysical naturalistic view of the world, which also applied to his view of language (Cratylus 385e–390e). Naturalistic, because the universe was, at its essence, governed by natural laws (as opposed to arbitrary rules imposed by the society). And metaphysical, meaning these laws were not enforced by the nature around us, but by 'nature' in an abstract sense – an overarching force that transcends the physical world. Applied to language, this results in a unique perspective where words are linked to real-world objects through the powers of a metaphysical force. This view, in turn, implies there is one correct way of building a language system. Similarly, there is one true way of mapping words onto their referents. A variation in the system, such as a synonym (a meaning that has two possible forms), would be a deviation from the intended metaphysical order. Socrates suggested it is the scholars' role to educate themselves, understand the mappings and help form new words that best follow the rules of nature. The current term for this approach is *speculative etymology*, the process of establishing the definitions of words by studying their surface linguistic features (e.g. their form).

The *rhetorical tradition* is in stark contrast to the view of absolute natural meaning. Pioneered by Aristotle, the study of classical rhetoric assumes that meaning is given to language by the one producing it. Speakers express their intentions, and have a dialogue with the audience, by consciously arranging the individual words into patterns that take on meaning as a larger unit (Kennedy, 1994, p. 4). According to rhetoric, attributing meaning to single words is not worth studying. In fact, the 'invention' (a rhetoric term for meaning) cannot be untangled from 'arrangement', the multi-word linguistic unit (Kennedy, 1994, pp. 4-5). A variation in language, called 'ornamentation', is not a crime against the metaphysical nature, but instead a tool that helps facilitate an exchange between the speaker and their audience. Ornamentation "includes substitutions of one term for another as in metaphor; figures of speech, or changes in the sound or arrangement of a sequence of words, such as anaphora or asyndeton; and figures of thought, in which a statement is recast to stress it or achieve

audience contact" (Kennedy, 1994, p. 6). As seen in the above quote, the rhetorical tradition blends the lines between word meaning and the intention of the communicator – they are interdependent.

Both Socrates' and Aristotle's ideas were preceded by *lexicography* – the discipline of creating dictionaries of words. Ancient dictionaries were exclusively bilingual, the first of which trace back to Mesopotamian clay tablets, 2300BC (Encarta, 2009). While they don't discuss meaning directly, they must, by definition, appreciate that words in different languages map on to each other in terms of meaning. The first monolingual English dictionary was published in 1755 and represents an effort to create a lasting record of the denotative meanings in language (Mitkov, 2004, p. 50). This marks a turning point in how scholars thought about meaning. The need for a monolingual record of meaning is a response to the realisation that meaning is not permanent. It is transient, changing over time.

Meaning as a Changing Concept

The field of *historical-philological semantics* is mainly concerned with studying word meaning through observing its change (Allan, 2013, pp. 559). This approach, prominent in the late 19th and early 20th century, documents, analyses and defines the features of semantic change. It proposes three important theoretical pillars (Allan, 2013, pp. 560 - 562). First, word meaning is seen as a psychological concept. This is in clear contrast to the denotative definitions mentioned above. In a predominantly denotative theory, meaning is nothing but the link between a token (an instance of a word) and the referent (object it represents). Much like a label on a jar of cookies. Historical-philological semantics sees words as links to psychological representations – tokens are not directly linked to the referent at all. People mediate the relationships between words and referents. This idea was later defined as the *representational theory* of meaning. Second, if we assume that meanings are mediated by our psychology, semantic change is a psychological process. It is an outcome of a shift in our internal representations. As such it can be studied using empirical, cognitive experiments. And thirdly, if we agree that environment (at least partially) shapes human psychology, language is indirectly influenced by both the society (external environment) and other psychological processes (internal environment).

Meaning as a Structure

Saussure disagreed with the view of language as a psychological process. In particular, he saw linguistics as a bespoke academic field, comprised of an array of interconnected systems, largely independent of cognition (Hawkes, 2003, p. 8). His theory of *structuralism* proposes that language consists of two separate, but connected areas – and abstract framework of rules (*langue*) and the observed speech that follows these rules (*parole*). He noted that our discourse is rarely representative of the whole language. A piece of text is a small cross-section of all that language has to offer. Yet we

perceive and understand the conventions and patterns that govern the text we produce. Consequently, individual word meaning must be an outcome of an overarching, abstract rule set. Because the conventions of *langue* give meaning to all of the words simultaneously, the meanings of individual words are always relative to each other. A change in the meaning of one word would be accompanied by a change in the meaning of all the other words (Hawkes, 2003, pp. 9-13). The modern legacy of structuralism is seeing and subsequently studying language as a series of compartmentalised frameworks.

Structuralism paved the way to seeing lexical semantics, the study of word meaning, as an independent branch of science. The view of language as a system that can be studied in and of itself led to the development of prominent theories, exclusively tackling the definition of word meaning. There are hundreds of contemporary ideas contributing to the discussions in lexical semantics, addressing the whole spectrum of semantic views mentioned so far. The large majority of current theories assumes a representational structure – the fact that words map on to a lexicon (language store) in our mind. The theories then differ in describing how the lexicon mapping is structured and how the target referent is accessed based on an input word.

Meaning as a Category

Instead of seeing words as individual elements of meaning, the *semantic field theory* (Trier, 1931) proposes a *categorical representation*. Lexemes (units of meaning) are organised into thematical categories. To obtain the meaning of a word, a person must know two groups of information – category level information and relational information. Category level knowledge would explain the lexeme-category assignment (e.g. "dog belongs to the animal category") and the properties of the category as a whole (e.g. "animals need food" or "animals reproduce"). Relational information would state where the target word (e.g. "dog") sits in relation to other members of the category (e.g. "cat", "alligator"). Together, this allows the language user to navigate the whole semantic field to find the appropriate category, and narrow down the meaning based on relationships within the category. A notable feature of this theory is an enforced single-category membership. Words can not be members of multiple categories at once. Most of the methodologies in this thesis would be incompatible with a single-category representation, as a relational assessment of objects across categories is assumed in most statistical models of meaning.

The binary (yes / no) word membership of only a single category was refuted by the *prototype theory* (Rosch, 1973). Instead, words are representative of a category to a degree. The word "soup" may be a marginal member of the "appetizer" or "drink" category, while words like "olives" and "tea" have a much stronger category membership respectively. The theory claims that the process of extracting meaning is not about meeting the criteria of category membership (as the semantic field theory would

suggest), but instead is a similarity assessment. When determining category membership of a target (e.g. “snake”), people first compare it to the strongest (*prototypical)* members of the relevant categories (e.g. animals – “dog”, “cat”, “cow”). If prototypical members are not known, the comparison is made with other known exemplars of the category. The outcome of such semantic search is a probabilistic membership mapping – a graded list of most and least relevant categories to the target word. This is a considerable advantage of the prototype theory, as it allows for partial and multi-category membership. Equally, it does not prevent the assessment of similarity between lexemes across categories. A similarity matrix that is not category sensitive will be a recurring theme throughout the thesis.

An important aspect of the *prototype theory* is its experientially obtained category structure. This was first commented on by Chafe (1972), who established the term *experiential knowledge*. For instance, the fact that a ‘shirt’ is a more prototypical member of the ‘clothes’ category than a ‘scarf’ is unlikely to be an outcome of specific, direct learning. People are rarely put in a scenario where the prototypicality criteria of a category is explained or taught directly. Instead, the appreciation of category structure and the ability to assess the prototypicality of members is likely an outcome of a large number of observations.

The idea of experiential learning is a key foundation of *frame semantics* (Fillmore, 1977). As individuals go through life, they are exposed to a constant stream of complex scenarios. Even something as simple as looking at the sun requires the understanding of heat, light, colour, the day-night cycle, planets, and so on. While people may vary in their mastery (and depth) of the knowledge, at least a basic understanding of the related concepts (many of which are abstract) is required to understand what a sun represents. This pool of related knowledge is called a *frame*. Applied to language, Fillmore argues that a frame must first be established in order to later be supplied with a referring word label. The implication is that words refer not only to a referent object, but that the mental representations of words also consist of the related knowledge.

Meaning as a Psychological State

The concept of a *frame* mainly focuses on the encyclopaedic (factual) information needed to successfully interpret a word. However, one could easily extend this idea to include non-factual factors. During the rise of behaviourism in the early 20th century, psychologists started to think about signs (i.e. words) as cues for observable actions. The *disposition theory* (Morris, 1938) proposed a word is learned when exposure to it produces a state where the original response to the referent may happen. Words would represent the disposition to produce an action. Interestingly, the theory does not claim the response is always equivalent to that of experiencing the referent, as would be common with Pavlovian classical conditioning (Pavlov & Thompson, 1902), but that the set of possible responses

when experiencing a learned word would include the original response to the referent. Whether the original response is selected from the possible set is subject to environmental variables (Morris, 1946). In sum, words form direct relationships with objects, mediated by their shared behavioural response.

Osgood (1952) further reframed the idea, pointing out a shared response is not always necessary. In his example, seeing an apple often results in the response of eating, whereas seeing the word 'apple' may never result in the same response. He instead claims that words and their referents share a psychological state that accompanies the original disposition to act. When seeing an apple, we may feel hungry, refreshed or have a particular sense of smell. An appropriately learned word for 'apple' would be one that evokes similar psychological states. Osgood calls this the *mediation hypothesis*. Words and their referents do not share the dispositions towards a behaviour, but the mediators that accompany such dispositions in the referent. Because the mediators of behaviour are often psychological factors, this idea is also labelled as *psychological meaning* – a set of psychological states that link a word to its referent.

This view is at the core of the present thesis. The primary assumption, and the basis of the studies presented, is that words have access to the psychological state of their user. It is no doubt that different items in the world lead to different emotional responses from the people experiencing them. It is our belief that the labels (words) for the items also carry emotional responses. In turn, the act of writing (or more broadly communicating) is an act of the communicator sharing psychological frames with their audience. On the receiving end, as listeners or readers, we automatically process the language into our own associated psychological frames. We define the term *psychological frame* as: a set of psychological states that are related to a word. In a sense, this view takes Osgood's theory of psychological meaning and subsets it. Instead of trying to understand the process of linking words to their referents through measuring the mediating psychological states, it focuses on the first half exclusively – trying to retrieve the psychological states that are related to a word.

The psychological frames likely form as part of the process of word learning. We make no assumptions on whether the psychological frames of words are equivalent to a set of psychological states of directly experiencing the referents. Our studies indirectly suggest, however, that the psychological frames are often overlapping between communicators, allowing for effective communication. Provided the researcher knows the patterns of overlap (e.g. how the word 'sunshine' usually makes a person feel), it should be possible to take a piece of text and imply (with varying accuracy), what psychological states the communicator was experiencing at the point of writing. To do that, it is first necessary to establish that something as abstract as a psychological frame (or a state) is measurable and quantifiable.

Measuring Meaning

The original theoretical framework of psychological meaning was introduced with hopes of reliable measurement. This was to contrast the solely theoretical approaches to meaning dominating academia before the 20th century, while also addressing the rigidity of a purely behavioural approach (i.e. the need for a shared behavioural response of the word and the referent). Osgood, Suci, & Tannenbaum (1957) conducted a series of exploratory studies, trying to capture the psychological dimensions of meaning. This introduced a high-dimensional perspective of words, where each word may be visualised as a vector in a high-dimensional space. A dimension is a possible emotional state of a dichotomous nature (e.g. tense/relaxed, strong/weak). The meaning of a word is, therefore, a specific combination of values across all dimensions. Interestingly, the perception of these properties for a given word is generally consistent among participants (Figure 1a)

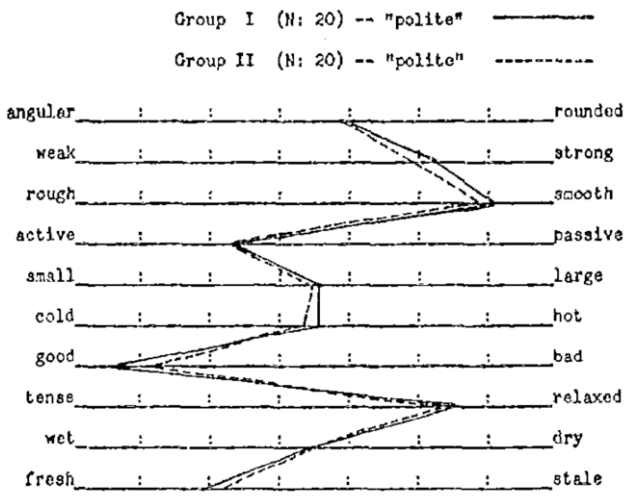


Figure 1a. Visualisation of the word polite in a high dimensional space. The values are means as rated by two independent groups of participants (n = 20). *Reprinted from: Osgood, C., Suci, G., & Tannenbaum, P. (1957). The measurement of meaning. Urbana, IL: University of Illinois.*

Osgood's work was highly influential, leading to the development of *Affective norms for English words* (ANEW; Bradley, & Lang, 1999). This collection of 1,034 words was rated on valence, arousal and dominance, allowing researchers to quantitatively analyse the psychological frames of words for the first time. Given the recent advances in crowdsourcing, obtaining normative datasets on a large scale has become a straight forward task. ANEW have been used as the basis for Warriner, Kuperman and Brysbaert's (2013) norms of valence, arousal and dominance, which include 13,915 words. Recently, Mohammad (2018) published reliable ratings of valence, arousal and dominance for 20,000 English words. The quantification of psychological variables is not restricted to valence, arousal and dominance. In Chapter 3, we demonstrate it is possible to apply this methodology to measure the humour appraisal of individual words.

The above studies demonstrate the ability to quantify the psychological meaning of words. Quantifying the semantic and encyclopaedic meaning has been equally as successful. Following Fillmore's theory on frame semantics (1977), the FrameNet project (Baker, Fillmore, & Lowe, 1998) uses manual annotation to establish the semantic frames of individual words. This results in an extensive database, listing frame descriptions (essentially semantic and relational definitions) of individual words. FrameNet also includes valence assessments of lexical units (Ruppenhofer, 2006). This poses a question of whether valence is a psychological state or a lexical property of a word. The common theme of the present thesis is that there is no clear distinction between psychology, a pattern in the observed language data, and word meaning. In fact, they are viewed as interdependent and interchangeable. Word meaning is a psychological frame, observable as a pattern in produced language output.

Meaning as a Data Pattern

The last paragraph claims word meaning can be observed as a data pattern in language. This is not far from *Firthian linguistics* (Firth, 1957) – also called *contextual meaning theory*. The core idea of this view is that words only have their meaning within the cultural and lexical context they appear in. From a cultural perspective, observed language is a representation of a cultural environment. From a statistical perspective, this implies the meaning of a word is deducible from the words (and broader lexical structures – i.e. sentences and documents) surrounding it.

The field of *computational linguistics* provides a diverse array of methods helping to quantify meaning, essentially allowing for a mathematical expression of the Firthian paradigm. The term computational linguistics is incredibly general, covering anything from the processing of phonology (sounds) to automatic syntax parsing (decomposition of sentence structure). We focus on the two most relevant subgroups – statistical semantics and connectionist modelling.

Statistical Semantics

Statistical semantics supposes that the meaning of words is extractable from text using statistical methods. More specifically, the *distributional hypothesis* claims that a words co-occurrence distribution (i.e. the pattern of words surrounding the target) holds its meaning (Harris, 1954). Logically, if two words share the same co-occurrence distribution, they should also share the same meaning. This was applied in the field of translation shortly after (Weaver, 1955). The meaning of an unknown word could be estimated by finding a word with a similar co-occurrence pattern for which the meaning is known.

In tandem with the assumptions of *historical-philological semantics*, which claims the meanings of words change over time (as discussed at the beginning of the introduction), this results in an

interesting semantic paradigm. Knowing the co-occurrence of a word at time points X_1 and X_2, we can accurately estimate the shift in meaning, simply by calculating the distance between those two co-occurrence patterns (Kulkarni, 2015). The term *distance* is used in a vector sense. A word's co-occurrence distribution is often quantified as a vector, that is, a list of words and the number of instances they co-occur with the target word (see Table 1a for an example).

Table 1a

Example of a co-occurrence distribution for words 'dog' and 'cat'

Target Word	Co-occurring words with the Target, and the number of co-occurrences			
-	bone	mouse	pet	animal
dog	23	1	62	132
cat	6	35	71	154

Table 1a. This table is a mock example of a co-occurrence table. The rows indicate two analysed target words (dog, cat), the columns indicate the co-occurrence values between the target and one of four words (bone, mouse, pet, animal). In this example, the word 'dog' appears with 'bone' 23 times, whereas 'cat' appears with 'bone' only 6 times. In practice, the number of columns is usually exhaustive – having one column for each English word in the data set. Each row is then referred to as a vector.

Co-occurrence is often measured in a *window* – a set range around a target. For example, in the sentence "I love to play with my dog, they are a lovely companion.", a window size of 3, with the target of 'dog', would calculate co-occurrence between 'dog' and 'play, with, my' (the three preceding words) and 'they are a' (the three following words). Higher window sizes appreciate a broader co-occurrence context, but may also include too much irrelevant information. Studies often explore a range of window sizes, and then either publish results across the whole range, or choose an optimal window size from the range (based on a pre-defined, objective criterion). The distance between two co-occurrence vectors can be expressed using any available vector distance measure, for which there are a range of choices. Three vector distance measures are explained and applied in Chapter 2.

We mention calculating a vector distance of two time points X_1 and X_2, measuring semantic change over time. However, this is a recent research question (Kulkarni, 2015). It is much more common to be calculating the distance between the vectors of two different words W_1 and W_2, obtaining a semantic similarity measure between the two (Turney & Pantel, 2010). This thesis does both – a distance between two words (in Chapters 2 and 4), and a distance between two time points (in Chapter 4).

Connectionist Models

Several of the semantic theories propose *frames*, pools of information linked to each word. This may be represented as a table (or a vector) – listing each piece of information for each word, or having words as rows and possible information as columns, recording '1' for a link and '0' for no relationship. Such a representation is not easily interpretable, especially for bigger sets. The need for a more easily interpretable (and more intuitive) data solution led to the development of *semantic networks*. Originally developed as a knowledge representation for machine translation (Richens, 1956), a semantic network consists of *nodes* (pieces of information), usually represented as dots, and *edges* (relationships), usually represented as links between the nodes.

Figure 1b. An example of a network representation, based on real co-occurrence data from Chapter 4. Nodes (dots) represent words of various emotions; edges (links) are formed between nodes if their vector co-occurrence distance is below a set threshold (i.e. if they have similar co-occurrence patterns).

Chapter 1 is a review of semantic network approaches and explains their creation in detail. For now, two important features of networks are worth noting. First, a network graph does not have meaningful axes. In Figure 1b (above), the absolute positions of the nodes are not representative – the only meaningful information is whether two nodes have a link. A network representation, consequently, does not support vector distance measures (it is not meaningful to comment on the distance of two nodes on a graph). This is not true for vector measures, where the distance between two words has an interpretation. However, the benefit of a network representation is its ability to readily represent a process. For example, it can show how information would spread across the nodes, travelling from one word to another using the edges provided (Siew, 2019). The abstraction of absolute position in a network makes the relationship structure, by definition, the core of the analysis. This is much more

difficult to model on a vector table, as vectors usually have more complex relationship definitions. The edges of a network make process modelling straightforward. Chapter 1 examines another use of a network approach – modelling lexicon growth during early word learning.

Data Sets for Computational Semantics

The data relevant to the reviewed theories broadly fall to three categories – a) language corpora collected specifically with academic research in mind, b) large scale language corpora adapted for semantic research, and c) data sets build specifically for quantifying semantic relationships. This thesis uses all of these types, as each has its strengths.

Language corpora collected specifically with academic research in mind

To make statistical judgements about the properties of language and word learning, a body of text is usually required. As with any research, there is the option to create a bespoke experimental design to first collect the data. This is useful for novel learning paradigms of specific research questions that require materials tailored to the premise of the study. The alternative is to use an already published corpus of language data. There are several prominent corpora built by academics, specifically with computational linguistics in mind.

For child directed speech and child production data, CHILDES (MacWhinney, 2014) is the most influential data set produces. At the time of writing, the associated publication it has 6887 citations, used in a wide spectrum of research. The database includes dialogues between parents and their children, with transcripts in 26 languages. It contains data on both monolingual and bilingual children, along with an age spectrum ranging from those barely producing any words to older school-attending children. Some of the transcripts represent conversations between siblings or between a child and a non-parent adult. It also includes children with language learning disabilities, allowing for the analysis of atypical language development patterns.

For corpora representing adult language, the British National Corpus – BNC (Leech, 1992) is a *balanced corpus* of 100 million words, including both written and spoken speech. The term ‘balanced corpus’ refers to a conscious effort of basing the data set on a variety of sources with the intention of high ecological validity of the research results. Including a range of types of publications in the data set tries to negate any systematic language patterns caused by the publication type itself. The Corpus of Contemporary American English – COCA (Davies, 2009) is the American English equivalent, including 560 million words in, also using balanced sourcing of materials. Both of these corpora aim to be a representation of recent language use. For studies into historical language patterns, the Corpus of Historical American English – COHA (Davies, 2012) is the historically oriented sibling of COCA,

including 400 million words, with the publication date range of 1800 to 2000. This data set is particularly useful for diachronic analyses (comparing vectors between two time points).

Large scale language corpora adapted for semantic research

The present day society generates data at an unprecedented pace. From a linguistic perspective, there is no shortage of online data to analyse. The disadvantage of using 'raw' corpora not intended specifically for research is that the data tends to be highly biased, over representing the user base that generated it. This is a far departure from the carefully balanced data sets such as BNC or COCA. However, there are two considerable advantages. First, the size of naturally generated online data sets exceeds that of any manually compiled corpus. Twitter users produce 500 million tweets per day (Sayce, 2018). Considering a tweet consists of up to 280 characters, Twitter users generate more content in a day than BNC, COHA and COCA combined. The second benefit is the recency of the data – online data is representative of the language today, as opposed to a corpus that may be decades old. Twitter data has been used for detecting sentiment (Agarwal, Xie, Vovsha, Rambow, & Passonneau, 2011), sarcasm (González-Ibánez, Muresan, Wacholder, 2011), gender (Burger, Henderson, Kim, & Zarrella, 2011), syntax structure (Derczynski, Ritter, Clark, & Bontcheva, 2013) or even crime rates (Gerber, 2014). For diachronic data, the Google Books nGram corpus (Lin et al., 2012) contains summary tables of approximately 8.1 million books, that is about 6% of the books ever published. These range from 1800 to 2009, allowing for a comparison of language patterns across time. Chapter 4 of this thesis uses the Google nGram corpus to visualise diachronic language patterns over time.

Factors Influencing Word Learning

It is not feasible to measure a child's lexicon directly. To obtain the data for modelling early word learning, researchers often resort to using a parent reported *inventory*. This is a research design where the parent lists all the words a child either understands or produces. Asking for the words that are understood results in a *receptive vocabulary*, while a list of the words a child produces represents the *productive vocabulary.* The receptive vocabulary is always larger than the productive one, as it is more difficult to produce words in the correct contexts compared to simply understanding them from someone else's speech. The most widely used inventory is the MacArthur-Bates Communicative Development Inventory – MCDI (Dale & Fenson, 1996), measuring both the receptive and productive vocabularies. An output is an estimated pool of words a child knows – i.e. a vocabulary size. This measure can be correlated with any other quantitative variable, getting a sense of what factors most influence early word learning.

A known correlate of vocabulary size is the frequency of child directed speech - CDS (Hart & Risley, 1995), the more parents speak to a child, the larger the vocabulary size. Hart & Risley also found that controlling for CDS, there are no differences in vocabulary size across different socioeconomic status

(SES) groups. This implies the effects of SES on vocabulary size are due to parents from higher social classes speaking to children more. The child's ability to properly discern the phonological makeup of a word is particularly important (Kuhl et al., 2015) during word learning. It has been proposed that child directed speech is exaggerated in terms of the intonation, allowing for better sound segmentation (Thiessen, Hill, & Saffran, 2005). Repetition of phonological information (which is common in CDS) also facilitates the identification of the auditory patterns, leading to faster word acquisition (Gathercole, 2006; Stokes & Klee, 2009). Intuitively, there is also an effect of age, showing that the lexicon consistently grows with time (Park et al., 2002).

Rather than examining vocabulary size, it is also possible to investigate which words we learn first. There is a well-established effect frequency, with the most common words in the written language learned earlier (Keuleers, Stevens, Mandera, & Brysbaert, 2015). The same is true for word prevalence – the use of a word across the adult population. The words with the most users are usually among those learned earliest (Keuleers et al., 2015). Words that consist of frequent phonological patterns are also positively correlated with early age of acquisition (Storkel, 2001).

Visual attention span is an important contributor to language acquisition. Children who hold their attention to words longer show better reading skills (Bosse & Valdois, 2009). When it comes to focusing on objects, not all types of information are attended to equally. Chapter 2 focuses on a set of experimental findings labelled *shape bias* (Landau, Smith, & Jones, 1988). That is the effect of children paying most attention to the shape of objects when learning new words. In frame semantics terms – the finding that children's semantic frames of words include shape features more often than any other type of feature (e.g. material or function). A specific review of factors influencing vocabulary growth is included in Chapter 1 from a network perspective, and in Chapter 2 from the perspective of a biased feature set of semantic frames.

Aims of the Thesis

Taking into account the reviewed literature, the thesis has one broad theoretical aim, and four individual aims for each chapter. The paradigm of *frame semantics* has been influential, contributing to our understanding of language representation. Combined with Osgood's idea of *psychological meaning*, this thesis proposes a framework labelled a *psychological frame*. That set of psychological states connected to the sign of a word. The publication outputs of this thesis aim to explore the notion of psychological frames, especially from a quantitative perspective. It employs and discusses perspectives that offer avenues into quantifying human psychology using language (primarily text) as a cue. By reading through the chapters, it should be clear that there is methodology to investigate psychological frames (Chapter 1); that vector representations can be used to make inferences about word learning (Chapter 2); that complex psychological concepts such as humour can be defined

quantitatively (Chapter 3); and that the outputs of rigorous analyses of psychological frames can be presented in a fashion understandable by a non-specialist (Chapter 4).

Individually, Chapter 1 aims to provide an entry-level knowledge into how best compose network representations of semantic relationships. Network science is not widely used in psychology, so an easily accessible publication aims to open the approach up to more academics. Chapter 2 aims to apply the idea of *frame semantics* to the field of *shape bias*. To our knowledge, this has never been done before. The methodology in this context represents a novel contribution to a long-standing debate. Chapter 3 creates a publicly available data set of single-word humour. This opens possibilities to analysing humour in a fine-grained, quantitative way. Such analysis has not been possible to date, as most research into humour was either theoretical or using jokes (i.e. large pieces of text). Finally, Chapter 4 aims to offer diachronic insights into language use patterns, while making the methodology available to non-specialists and even non-academics, in real time, with no prior understanding of the methodology. This is a considerable effort, reducing the barrier of entry into diachronic analyses of language. Macroscope, the tool developed in Chapter 4, makes access to quantifying *psychological frames* open to the public.

Chapter 1: The Influence of Semantic Structure in the Growth of the Early Lexicon

Project Overview

Having worked with linguistic networks throughout the first year of my doctorate, I decided to formalize the merits of the approach as a text book chapter. This was an excellent opportunity for me to review the network science literature. More importantly, it allowed me to share my take on how to construct network spaces out of language data. The chapter aims to be an easy to understand introduction into network approaches, focusing on the specific implementation for psychology. Despite the psychological focus, it is general enough to mention a variety of data sources and possible network types that can be constructed. It is a summary of the main approaches to network generation I came across during my other research.

The review was later published as part of a book compiling recent innovations in psychological network science, inviting me to highlight the "growth of the lexicon" – in other words, the process of learning new words. This means the methodology includes a growth component, explaining how to model time series data using networks. The chapter breaks down the process into steps, first building a static network and then expanding it using a growth model. This mimics the order in which a scientist would learn this methodology.

Thomas Hills published several specific implementations of this approach. The second half of the chapter follows some of his research, acting as a showcase of examples how the methodology can be used. Ultimately, the chapter hopes to introduce some of the concepts behind my work, but also to be an entry-point into network science for any interested psychologist.

In the greater scheme of things, network science is only one set of methodologies employed during my PhD. When building the Macroscope, Li Ying (my colleague) and I decided to use networks heavily. The results networks generate are incredibly visual, easily stimulating the user's interest in the presented data.

One potential downside of using networks is the lack of explicit dimensions. The nodes in a network are connected, but their position in the graph is not necessarily representative of anything. Dealing with psychological properties of words, there are times when using dimensions of the space is beneficial. Vector spaces have this property, which can be seen in further chapters of the thesis.

Introduction

Early word learning is characterized by a *vocabulary spurt.* The sudden acceleration of adopting new words results in a non-linear vocabulary growth rate (Goldfield & Reznick, 1990; Mayor & Plunkett, 2010). While there is discussion whether the accelerated pattern is due to systemic changes or simply due to a delay in acquiring complex words (McMurray, 2007), the adoption of new words in early childhood is, without doubt, a non-linear process. To study early word learning, it is important to consider a child's lexicon as a developing structure. On one hand, the process of adopting new words is facilitated by external factors, such as the exposure to repeated words (Gathercole, 2006), the properties of the parents' lexicon (Weizman & Snow, 2001) or the interests of the child (DeLoache, Simcock, & Macari, 2007). At the same time, the state of the child's lexicon itself influences the words to be learned. Applying the network science framework allows for the description of both the external factors of word learning, as well as the effects of the lexicon itself. In turn, applications of network and graph theory have greatly contributed to understanding word learning (Vitevitch, 2008; Beckage & Colunga, 2016).

This chapter will review approaches to modelling semantic lexicon growth through network science. Namely, it will focus on modelling the growth of a vocabulary, by studying the growth models of semantic network representations. The methodology consists of three steps - building semantic networks, incorporating longitudinal data, and examining the properties of networks to draw conclusions. The chapter provides all the information needed to start working with lexical networks, as well as shows examples of publications where the approach was successfully used.

First, we provide an overview of lexical network generation. In practice, a researcher has a variety of options on how to generate the networks. We describe three different resources for building semantic networks: feature norms, free association norms, and natural language corpora. Secondly, the theory behind measuring word production is explored. This allows the networks to be analyzed longitudinally, as a time series. Finally, applications of this methodology are shown. These include modelling child lexicon growth, predicting word learning order using semantic distinctiveness, predicting word learning order using contextual diversity, exploring lexical structure differences in atypical development, and exploring differences in the developing network structure of bilinguals.

Constructing Semantic Networks

Networks generally consist of *nodes*, connected by *edges*. In the case of semantic networks, a node represents a single word adopted by a child in their lexicon. This makes it easy to study lexicon growth. The growth of the network itself is, inferred to be, directly representative of the growth of the lexicon. In contrast, the definition of edges varies widely across the literature. For our purposes, edges

represent a relationship between words. The nature of the relationship may be quantified in different ways (see Table 1b).

TABLE 1b

Different edge types in semantic lexical networks used to predict word learning

Edge type	Description	Referenced Publication
Perceptual and functional features	Edges are based on shared semantic features.	(McRae et al., 2005; Vigliocco, 2008)
Free associations	Edges are based on cue-target relationships in the free-association task.	(Nelson, McEvoy, & Schreiber, 2004)
Natural language corpora	Edges are based on similar patterns of usage in natural language.	(Lin et al., 2012)

Table 1b. An overview of the edge types covered in the chapter. The referenced publication allows the reader to quickly look up the datasets as well as guidance on specific implementation.

Feature norms: Feature norms allow for the production of network edges based shared perceptual or functional features. Feature norms are datasets, generated by adult participants who produce lists of features for individual words. An example is the *McRae feature norms* (McRae et al., 2005). A dog may be described as "having fur" and "having four legs". A cat, described using similar features, would then be strongly linked to dog in a network based on shared features. Additionally, the features are categorized by type (e.g. visual features – "is large", encyclopedic – "is a vehicle", function – "used in cooking"). These divisions allow for the generation of networks focusing on different kinds of feature similarity. The *Vinson and Vigliocco feature norms* (Vinson & Vigliocco, 2008) offer an alternative dataset, including verbs as well as nouns.

While the literature reviewed predominantly focuses on perceptual feature norms, it is possible to extract lexical relationships from any data set that quantifies dimensions and provides dimensional values for each word. Provided we are interested in semantic relationships, the only requirement is that the quantified dimensions are of a semantic nature. For example, Lynott and Connell (2013) provide a list of modality norms for 400 nouns. These specify how words relate to sense modalities (hearing, taste, touch, smell, and vision). Consequently, such data set allows the researcher to construct a matrix (or a network) of relationships between the words themselves.

Free association norms: Free association norms are also suitable for network edge creation. These norms are a result of participants responding to cue words with the first word or word that comes to mind, the target work. For example, a participant may be presented with the cue "angry", responding

with the word “furious”, “friend” or “red”. Unlike feature norms, word associations establish a direct cognitive link between a cue and the generated word. Numerous free association norms are available for network construction (Postman & Keppel, 2014; Nelson, McEvoy, & Schreiber, 2004).

Natural language corpora: Word co-occurrence measures allow for the linking of nodes through context. These utilize text corpora, establishing an edge between words when the words appear in the vicinity of each other in the corpus. Their advantage is that they do not rely on participants’ conscious response and thereby attempt to derive psychological structure from the input that people are likely to experience over their lifetime (Jones, Hills, Todd, 2015). Co-occurrence datasets are much larger in size than production norms, resulting in fewer restrictions on the size of the networks. One such example is the Google Ngrams, a large collection of word co-occurrence data (Lin et al., 2012). Additional sources of co-occurrence edges can be derived from semantic space models, such as BEAGLE and LSA (Jones, Kintsch, & Mewhort, 2006). These use co-occurrence data to derive word similarity, but do not require that words appear together in order to have similar meaning. Instead, words need to share similar patterns of usage across language.

Measuring Word Production

Simulating network growth in semantic networks requires knowing which words a child knows, at different stages in their life. This data supports a time series network analysis – a key component of investigating semantic network growth. The Child Language Data Exchange System (CHILDES) is one example of such database (MacWhinney, 2000). It consists of direct transcriptions of child-parent speech, along with the age of the child. Tools like Childfreq (Bååth, 2010) allow for easy extraction of CHILDES frequency data, returning per-word frequency values. These can serve as a basis for implementing frequency thresholds in network growth models. For example, the input of a network growth model can be pre-filtered to include words above a certain CHILDES frequency. This would make the network less sensitive to spontaneous one time mentions of a word.

The MacArthur-Bates Communicative Development Inventory (MCDI) is a parent report measure of lexical development (Dale & Fenson, 1996). It is not transcript based, but instead includes a checklist that a parent fills out, at set points in time. The result is a table of words and the ages from which a child either comprehends or produces each word. Averaged over groups of participants, the MCDI provides an accessible way to compute population average (or normalized) lexical development.

The Kuperman age of acquisition norms (Kuperman, Stadthagen-Gonzalez, & Brysbaert, 2012) are the least direct measure. The procedure shows lists of words to adult participants, asking for the month at which the adult believes they had learned the word as a child. The advantage is the size of the dataset, containing age of acquisition estimates for 30.000 English words. The Kuperman norms are significantly correlated with the MCDI, suggesting the ratings are valid $r_s(107) = .64$, $p < .001$.

Modelling Network Growth

Computational models provide useful ways to formalize and compare different hypotheses. To understand early lexical acquisition in children, these models should include various influences, including the structure of the language children are exposed to, the structure of the lexicon that children already know, and the relationship between these two. Hills and colleagues investigated the performance of three network growth models that captured these relationships (Hills, Maouene, Maouene, Sheya, & Smith, 2009). The MCDI was used to produce a growth trajectory for a normalized child lexicon. Networks were generated for the ages of 16 to 30 months, resulting in 15 unique networks of up to 130 nodes (nouns), arranged developmentally from the youngest to oldest. The edges were formed using either semantic feature production norms or word association norms, leading to two different sets of networks (feature and associative). Note that this approach allows for testing hypotheses about both structure and process. That is, what are the most informative semantic relationships for network construction and what are the best processes for predicting network growth, respectively.

The McRae feature norms were used as a basis for the feature networks. Children are sensitive to feature categories (Keil & Batterman, 1984; Sheya & Smith, 2006), suggesting feature networks could perform well when predicting word learning.

The associative networks were formed using the University of South Florida Free Association Norms (Nelson, McEvoy, & Schreiber, 1998). Adults utilize free association when identifying learned objects (Nelson, Zhang, & McKinney, 2001) and previous work has shown that they predict the age of acquisition (Steyvers & Tenenbaum, 2005).

To model the network growth, three formal growth models were proposed (see Figure 2 for an overview). A *z*-score was calculated for each model, comparing the model-appropriate growth value of the learned word with the distribution of growth values for all words that could have been learned. The three models are described as follows:

Preferential attachment

Preferential attachment states that highly connected already known words are more likely to get new connections in the future. The model is therefore driven by the structure of the growing lexicon, not the information it is exposed to. Patterns consistent with preferential attachment have already been found in non-linguistic domains (Barabasi & Alberts, 1999) and Steyvers and Tenenbaum (2005) proposed such a model for lexical growth over cultural or developmental time-scales providing a useful hypothesis for early word learning. It is believed these patterns emerge due to the nature of how large scale networks grow in general – where

existing nodes are duplicated and the edges are then rearranged, in a process called duplication and divergence (Kumar et al., 2000; Wagner, 2001).

In the context of lexical development, preferential attachment would lead to networks where the longest known words have the most connections. In other words, the founding nodes of the network would have the highest degree. Steyvers and Tenebaum (2005) investigated this question and found patterns supporting the idea by looking at the structure of adult free association networks.

Preferential acquisition

Preferential acquisition is not lexicon driven, but environment driven. Preferential acquisition states that highly connected words in the language environment are more likely to be learned. Hills et al. (2009) suggest that this could be due to word salience in the environment. Highly connected words would be more apparent in the environment, would appear in more contexts, and would be more frequent in general. This model puts all of the emphasis on the external stimuli during word learning. As attention to stimuli plays a large part in learning (Olson & Sherman, 1983), it may be that children, whose cognitive resources are still limited, simply focus solely on the stimuli. This further suggests that earliest word learning may consist of word islands, which are not yet semantically related. It has the further implication that language learning is not sensitive to what children learn first, but is sensitive to the structure of the learning environment, which may make the learning process more robust.

Lure of the associates

The third model, lure of the associates, states that: “Unknown words may be highlighted by known words to which they are related and learned in proportion to those relations” (Hills et al., 2009). This effectively means children would be more likely to learn words with the most connections to known words overall. This model is less driven by prototypical words (central nodes) and instead driven by semantic relatedness with the constellation of previously learned words.

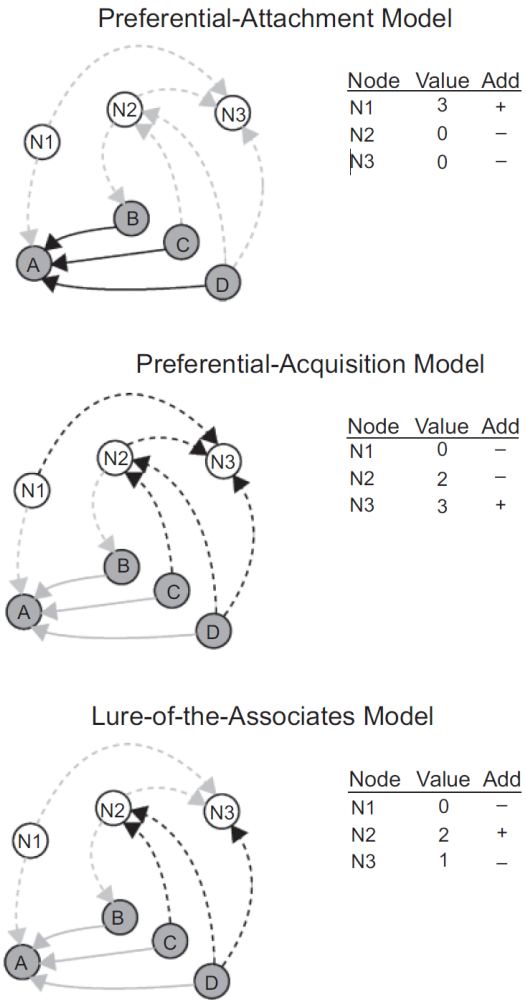


Figure 2. Overview of three different word learning models**.**

Three example networks show what words are predicted to be learned, according to the different growth models. Note that the already learned words (A-D), shown in gray, are identical across the networks. Black lines indicate relationships the growth model views as relevant, gray lines are relationships the growth model deems irrelevant. The relationships depicted here are mock examples, representing any possible type of a lexical relationship. The sum of important relationships, leading to possible new words (N1, N2, N3), is presented in the table on the right as "Value". The node with the highest Value is learned. Preferential attachment predicts learning of words that connect to learned words of high degree within the known network. Preferential acquisition predicts learning of words that are highly connected within the learning environment (i.e. the unknown set). Lure of the associates predicts learning of words that have high degree from the nodes in the known network.

Reprinted from **"**Longitudinal analysis of early semantic networks preferential attachment or preferential acquisition?" by Hills, T. T., Maouene, M., Maouene, J., Sheya, A., & Smith, L., 2009, *Psychological Science*, *20*(6), 729-739.

None of the growth models were well supported when the network was constructed based on the feature norms. However, the results were different for the network based on free association norms. The preferential acquisition and lure-of-the-associates models both described the observable growth of the associative networks, and the preferential acquisition model fits the associative growth data the

best out of the two. The preferential attachment model was not any better than a model based on random word learning.

These results suggest that the growth of associative networks is not due to central nodes in the learned network, but instead due to highly connective words present in the learning stimuli. Upon including additional parameters, the best performing model consisted of preferential acquisition, word frequency in CHILDES, and the number of phonological neighbors. This implies a shared responsibility of semantic, frequency, and phonological factors during word learning.

Semantic Feature Distinctiveness

Children are known to use feature cues during word learning. We know children are sensitive to shape (Landau, Smith, & Jones, 1988), texture and material (Samuelson & Horst, 2007; Soja, 1992; Soja, Carey, & Spelke, 1991). The failure of feature networks warranted further investigation. According to the reviewed literature, it seems unlikely that relationships between these features would not have an impact on the growth of the lexicon.

Engelthaler and Hills (2016) investigated the role of semantic features in early word learning. As in Hills et al. (2009), the McRae feature norms were used to create vector feature spaces. Each word was assigned a feature vector, based on the number of participants listing each feature per given word (as found in the McRae norms). The vector dimensions were all the possible features listed in the norms. Features not listed for a word at all were assigned the value of 0. Each word was then represented as a single point in a high dimensional vector space.

In a vector space, the distance between nodes represents their dimensional similarity. This can be in turn correlated with age of acquisition of the nodes, examining the relationship between feature similarity and word learning. To do this, three different distance measures were adopted (see Figure 3).

First, the non-shared feature distance represents the absolute difference between the listed number of features for two words. This counts up all features that two words have that are not shared between the two words. This is a dissimilarity measure, concerned about the feature differences between two words, not their similarity.

Second, the Manhattan distance[1] represents the difference between listed features for two words, weighted by the number of participants listing a feature. Features seen as prototypical (listed by many

[1] Manhattan distance here refers to a specific application of the metric relevant to multi-dimensional feature spaces. It could alternatively be seen as a "weighted non-shared feature distance". For clarity, and to help distinguish the measure from our non-shared feature distance, we will refer to the weighted measure as Manhattan distance. The measure is more specifically discussed in Chapter 2.

participants) would have a more profound impact on the distance if they were listed for one word by not the other. Like the measure above, the Manhattan distance is a dissimilarity measure, but in this case, it treats a feature as important to the difference in proportion to the number of times it is listed for one word but not the other.

Third, the Jaccard distance is one minus the size of the intersection of shared features divided by the size of the union. In our case, this is the number of shared features divided by the number of all features listed for the two words. Unlike the other measures, this measure reduces the distance in proportion to the number of shared features.

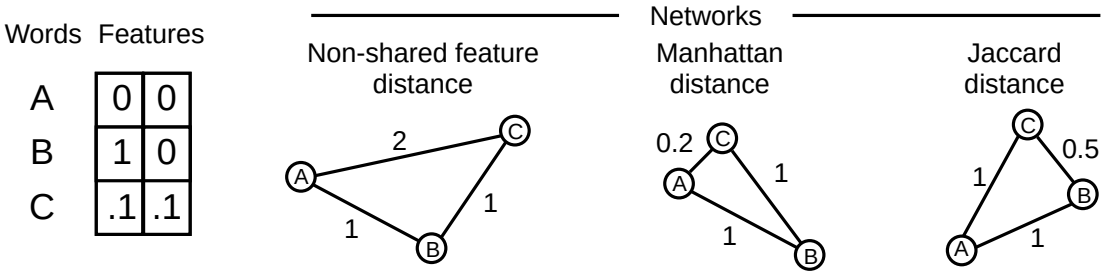


Figure 3. An example of the three different distance measures used to compute feature distinctiveness. Each of the three words (A, B, and C) each has a vector representing the proportion of individuals who reported each of two features; each feature is represented by a column in the feature matrix. To the right, networks are shown for each of the distance metrics described in the text. Reprinted from "Feature biases in early word learning: network distinctiveness predicts age of acquisition." by T. Engelthaler, & T. T. Hills 2016. In print at *Cognitive Science*.

Both the non-shared feature distance and the Manhattan distance were significantly correlated with age of acquisition $r_s(490) = -.24$, $p < .001$ and $r_s(490) = -.34$, $p < .001$, respectively. This implies that on average, more distinctive words are learned earlier. The Jaccard distance was not correlated with age of acquisition $r_s(490) = -.04$, $p > .05$[2]. These findings hint at the lack of predictive power in Hills et al. (2009). Feature similarity does not seem to be an important factor in word learning, while feature dissimilarity (distinctiveness) does.

A clustering analysis showed the effect to be local as well as global. This means dissimilarity is perceived both in the context of the whole environment (an object is unique in respect to all the other words), as well as the local environment (an object is unique within a group of similar objects – a category). Both aspects explain different parts of the variance of age of acquisition, meaning they may work in tandem.

[2] A detailed analysis of the semantic feature distinctiveness is shown in Chapter 2. This includes an analysis of multicollinearity, suggesting that non-shared feature distance and Manhattan distance are highly correlated with each other, while Jaccard distance isn't correlated with either. This is to be expected, as the first two measures explore similarity, while Jaccard distance captures dissimilarity – these are two fundamentally different relationships.

Contextual Diversity

The variety of contexts a word appears in facilitates its cognitive processing (Adelman, Brown, & Quesada, 2006). A diversity of context is also linked to better word segmentation (Hayes, & Clark, 1970) and easier word recognition (Newman, 2008). Some studies also suggest that contextual diversity in child-directed speech leads to a higher rate of word learning, as well as a more diverse vocabulary (Yu, & Smith, 2007; Gillette, Gleitman, Gleitman, & Lederer, 1999).

The link between context distribution and cognitive properties has further been investigated by Hills, Maouene, Riordan and Smith (2010). Contextual diversity was measured through a matrix of word co-occurrences, similar to the word co-occurrence detector method (Li, Farkas, & MacWhinney, 2004). A moving window co-occurrence was computed for the CHILDES corpus. An N x N matrix was formed, where N is the number of unique word types in the corpus. Each cell of the matrix *ij* was then populated with co-occurrence data, where +1 is added to the cell's initial value, whenever word *i* and *j* appear within a set range (*k)* of each other. Because the window range *k* is arbitrarily set, the process was repeated for *k* values between 2 and 100. A binary representation of the matrix then shows whether or not a word *i* ever appears with the word *j*. A sum of the binary values for word *i* is then a measure of the word's contextual diversity. This approach frames the context of a word as a bag of its immediate neighbours (within a sliding window of a size *k*). A highly contextually diverse word would be one that is often surrounded by unique words within the distance of *k*. This view is different to document-based contextual diversity measures (e.g. Adelman, Brown, & Quesada, 2006), which define a highly contextually diverse word as one that appears in many unique documents.

Contextual diversity was found to be negatively correlated with age of acquisition, where more diverse words are learned earlier. This effect was the strongest for a co-occurrence moving window of 5, but still significant for larger windows of up to 40, though this differed by word class (e.g., verbs and nouns).

Contextual diversity was a significant predictor of age of acquisition, even after controlling for frequency and free associations. The effect was broken down by word types, being strongest for nouns and weakest for all word types combined. This may suggest contextual diversity plays different roles in word learning, dependent on the word type. Comparing contextual diversity and associative network degree (associative networks from Hills, Maouene, Maouene, Sheya, & Smith, 2009), we see a model including contextual diversity outperforms a model using free associations.

Using the three network growth models described in Figure 2, the lure of the associates was the best fitting model for nouns and all word classes combined. Preferential acquisition was the best fitting model for verbs and function words. None of the models were a significant fit for adjectives. A possible theoretical explanation is that learning nouns relies of semantic connections between the

novel words and the already learned lexicon, while verbs and function words need to be prototypically salient in the learning environment. Such inferences are speculative, however, and are presently being investigated at the time of writing of this thesis. Making any cognitive and structural claims would need to be supported by experimental studies. It is also important to note that preferential attachment performed poorly for all word classes, supporting the findings reported above for free associations. Analysis of the three different models suggests that the learning of different word classes may be accomplished by inherently different learning mechanisms.

Structure in Abnormal Language Development

Adult lexical networks show a small-world structure (i Cancho, & Solé, 2001; Steyvers, Tenenbaum, 2005). The small-world structure is a network property, suggesting nodes are connected in clusters, with several groups of nodes with a high in-group degree (Watts, & Strogatz, 1998). Past research showed that vocabulary size at an early point in life strongly predicts lexical capabilities of an individual in the future (e.g. Thal, Bates, Goodman & Jahn-Samilo, 1997; Anderson, & Freebody, 1979; Moyle, Weismer, Evans, & Lindstrom, 2007). Children who show low vocabulary sizes relative to their peers, referred to as *late-talkers*, often show lexical difficulties in the future.

Beckage, Smith, & Hills (2011) investigated how individual differences in semantic network structure relate to word acquisition (network growth). The study compared the vocabularies of 66 children (aged 15 to 36 months). 39 children had typical vocabulary size for their age, while 27 children had relatively small vocabulary size. Lexical networks were created for each child, using parent reported checklists of the child's vocabulary. The network nodes were connected using CHILDES co-occurrence. Words that highly co-occur within the CHILDES database were assigned edge connections.

Network statistics were computed, averaged over all words in the network. These included in-degree, clustering coefficient and geodesic distance. The in-degree was computed as the number of words preceding the target word in a fixed window as it moved through the corpus (the number of edges pointing towards a node). The clustering coefficient of a node was computed as the connectedness of the node's nearest neighbors. Specifically, it is the fraction of observed neighbor connections, divided by all possible connections. The clustering coefficient is an established measure of network structure. Finally, the geodesic distance was computed as the average shortest path between two nodes. See Newman (2003), for a review of the various network measures.

Additionally, to investigate the importance of semantic structure in the environment, 300 random acquisition networks were constructed for each child. These were created by sampling an N number of random words from the pool of all the words on the vocabulary questionnaire, where N is the vocabulary size of the child. The nodes were then connected in the same way as above, using

CHILDES co-occurrence. All of the 300 networks were then averaged for each child. This results in networks that contain semantic relatedness information of the environment, but not the structure obtained through learning.

As a control, 100 completely random networks were generated using the Erdős-Rényi method (Erdös, & Rényi, 1960). These shared the number of nodes and edges, but did not share the structure. The edge structure was generated randomly, instead of using CHILDES semantic relatedness. This results in a network that contains properties generated by the size and the connections of the network itself, but not the semantics of the environment.

A comparison of the random acquisition networks and the completely random networks showed that random acquisition networks have significantly more local connections. The distribution of edges in the random acquisition graphs supported local clustering, unlike in the completely random networks. They also have higher median in-degree, suggesting a hub-like network structure. These findings suggest that the learning environment is highly structured in itself, to the point where random sampling results in structured, small-world networks. As such, in early word networks, the small-world structure does not depend on the growth process, but on the inherent structure of the learning environment.

Significant differences were found between typically developing children (TD) and late talkers (LT). TD showed similar patterns to the random acquisition network model, with local clustering. LT however, showed significantly less small-world structure.

Generally, networks that exhibit small-world structure trend towards smaller geodesic distances as they grow. This is in line with the results for typically developing children, where a relatively small geodesic distance was observed. In contrast, late talkers show a much higher geodesic distance, which could suggest the failure to pick up on the small world structure present in the learning environment.

This set of findings offers new insights into network growth. Namely, it highlights the importance of external environment structure when forming the structure of the growing network. Children who have normally developing lexicons may find it easier to notice the patterns in the stimuli, leading to the emergence of local clustering as the network grows. On the other hand, atypical developing children struggle to develop local network clustering, perhaps due to the inability to recognize structure in the "to be learned" network.

Network Structure of Bilinguals

While the majority of lexical research focuses on a single language, bilinguals are an intriguing focus group, showing how the development of network structure behaves across multiple languages. Both bilinguals and monolinguals perform similarly in terms of concept learning rate (Hoff et al, 2012; De

Houwer, Bornstein, & Putnick, 2014; but see Bilson et al., 2015). They also exhibit comparable ability in word-object mapping (Byers-Heinlein & Werker, 2013; Werker, Byers-Heinlein, & Fennell, 2009). At the same time, differences are found in the way bilinguals perceive words (Ramon-Casas, Swingley, Sebasti´an-Gall´es, & Bosch, 2009) and in their underlying processes of word disambiguation (Byers-Heinlein & Werker, 2009).

Bilson et al. (2015) investigated the differences in network structure of bilinguals and monolinguals. Vocabulary checklists (MCDI) were collected from 254 English speaking monolinguals. Bilinguals spoke English and: Spanish (n = 111), Mandarin (n = 25), Vietnamese (n = 19), Malayalam (n = 9), Japanese (n = 8), Arabic (n = 4), French (n = 4), and Russian (n = 1). The collected vocabularies were used as the nodes in the network, with the edges constructed using the University of South Florida Free Association Norms (FAN; Nelson, McEvoy, & Schreiber, 2004).

To assess the growth of the networks, a maximum log-likelihood binomial distribution model was proposed. The best fits and maximum likelihood estimators were then compared across networks. Monolingual children were found to learn English words at a faster rate. Both groups showed acceleration in the vocabulary growth, with monolinguals showing significantly higher acceleration rate than bilinguals. These results are consistent with the fact that bilinguals are exposed to lower word frequency in their environment (David & Wei, 2008; Pearson, Fernández, & Oller, 1995).

The indegree of words in the associative network was predictive of age of acquisition, for both monolinguals and bilinguals, with a stronger predictive power for bilinguals. While bilinguals do develop their English vocabulary slower, they are overall, more receptive to the associative structure of words.

When comparing two different growth models (preferential attachment and preferential acquisition), results were in support of previous research. Preferential acquisition better explained the growth of the network for both monolinguals and bilinguals. Associative structure of the learning environment seems to influence learning more, in comparison to the associative structure of the already adopted lexicon. This also means monolinguals and bilinguals do not process words in a qualitatively dissimilar way.

Analyzing the network structures of monolinguals and bilinguals, showed that they do not significantly differ in terms of network clustering coefficients or density. Additionally, both types of network show a similar level of small-worldness. However, the structure of an English network of a bilingual does differ to that of a monolingual in terms of indegree and average path length. The indegree and average path length are, on average, higher for bilinguals, a result observed over a range of lexical sizes. Bilingual English networks also tend to adopt more small-world structure overtime. This shows the bilingual's' early preference for words more dissimilar to what they already know. See Figure 4 for a comparison of the two networks.

Finally, Bilson et al. (2015) showed, using the computational framework provided by preferential acquisition, that bilingual first language learners learned more translational equivalents than would be expected for children who were learning two languages independently. For each child, Bilson et al. simulated 1000 learners using preferential acquisition to grow lexicons equivalent in size to those of the target child. This then provided a distribution of expected numbers of translational equivalents. Children were compared with their own distributions, and the majority had learned more translational equivalents than would be expected by chance. Several further analyses confirmed that bilingual first language learners showed semantic facilitation (not inhibition) in the learning of translational equivalents. Thus, if a bilingual English-Spanish child learned 'ball' in English, they would be more likely to learn the word 'pelota'.

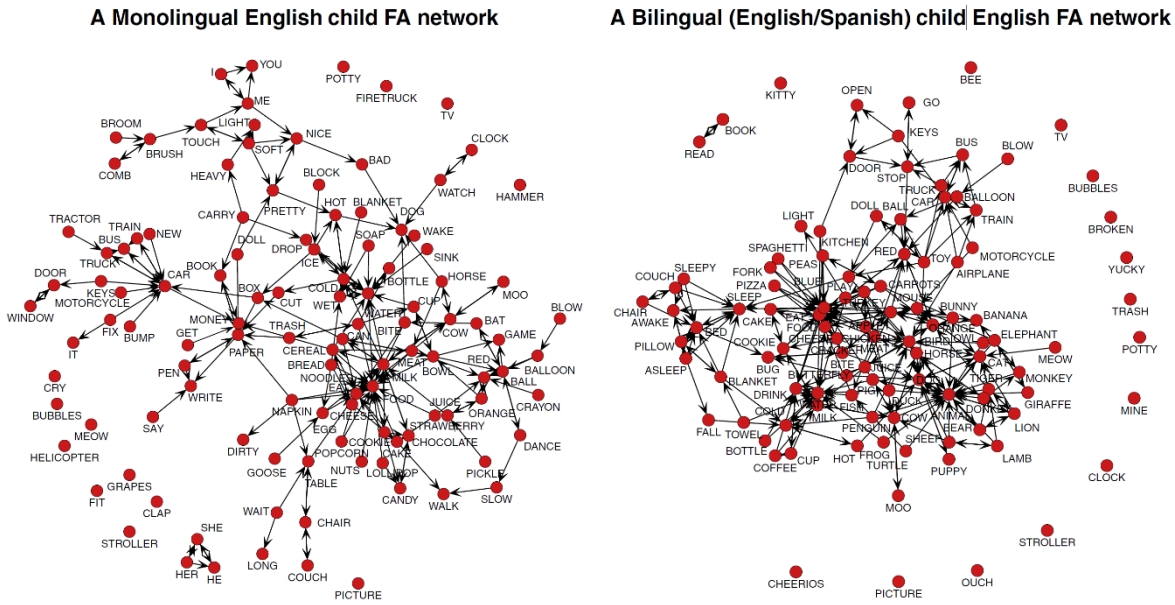


Figure 4. Representative English networks for a monolingual and Spanish–English bilingual based on free associations (FA). Reprinted from "Semantic facilitation in bilingual first language acquisition." by S. Bilson, H. Yoshida, C. D. Tran, E. A. Woods, & T. T. Hills, 2015. *Cognition*, *140*, 122-134.

Future Applications

The words that children learn during early development contains an inherent semantic structure. The work described above indicates that this structure is important to early word learning, and can inform both typical and atypical learning trajectories, as well as differences between monolingual and bilingual first language learners. The network approach provides measures of structural differences and facilitates computational models of language learning over time, producing generative network growth models that differ from those initially described in other domains.

But there is still work to be done. One clear angle for future work is developing neural network models or semantic space models that experience text and learn language similar to that of children.

The network structure investigated here provides an additional variable against which such models can be tested. It is not merely about learning words at the right rate, but learning words in the appropriate pattern to match that observed in children.

Another open question is over what range is semantic structure important. That is, how sensitive are the computational models to broad structure. This review mentions window size when computing co-occurrence matrices from natural language corpora. The 'optimal' size, where semantic structure is the most informative in terms of language acquisition, differed for the various word types. Future work might look at why the sensitivity to structural breadth is not consistent across different word types. For example, why do nouns appear to gather structural information from a larger context than verbs or adverbs? Second, how might age impact the optimal window size?

Finally, it is possible to create frameworks that bridge the various edge types. So far, each of the applications only generated edges using one dataset. For example, words were related through semantic similarity exclusively. This allows us to isolate the effects of a specific relationship, but may not be representative of the real world. It is likely that children use a variety of relationships to link words together. Future work should investigate networks that use multiple edge types at once (Stella, Beckage, & Brede, 2016).

Conclusion

Network modelling provides a compelling way of exploring lexical development. They offer several benefits worth noting. Being a data-driven approach, the methodology requires little to no a priori assumptions in terms of structure. The relationships within the network are emergent. As such, the observed patterns may be, to an extent, reflective of lexical patterns in the real world. If the input words are accurate and the edge type is in fact perceived by children, the observed network properties may be attributed as a property of the actual lexicon.

Since the approach is not tightly construed by theory, its biggest advantage is the ability to bridge domains. As shown in this chapter, the same methodology is used to examine semantic features, free association, abnormal language development or bilingual lexicon structure. Beyond the present context, similar approaches are used in the fields of climate modelling (Havlin et al., 2012), epidemiology (Pastor-Satorras, & Vespignani, 2001), social sciences (Borgatti, Mehra, Brass, & Labianca, 2009) and energetics (Albert, Albert, & Nakarado, 2004). This allows for cross-domain comparison of processes related data. The network representation of an energetic power grid may be similar to that of a child's lexicon, while the underlying mechanisms may be completely different. With that in mind, it is interesting to observe how network representations blend with established theory in the various research fields.

On the other hand, network science should not be seen as independent of theory. As with any analytical framework, rigorous theoretical knowledge is often useful to explain the findings. In addition, model competition is absolutely vital to test the predictive power of one model over another. Though previous models of network growth provided preferential attachment as a proof of principle model, model comparisons showed that one could indeed do much better. Proof of principle is sufficient in many domains, but it is often not difficult to produce a model that beats random acquisition. The power of modeling is being able to test the necessity of certain assumptions against others. The network science approach offers an additional layer of complexity for these models, but also provides more information about the language acquisition process as one not simply of vocabulary size but vocabulary structure.

Chapter 2: Feature Biases in Early Word Learning, Network Distinctiveness Predicts Age of Acquisition

Project Overview

The hope of capturing the psychological differences between words is the main motivation behind the whole thesis. While network science offers a powerful methodology to capture the broad structure of language, it is limited when it comes to quantifying differences between individual words. High dimensional vector spaces are the perfect complement, able to precisely compare words along any number of pre-defined dimensions. This chapter explores how to form, transform and visualise linguistic high dimensional vector spaces.

Much like networks, vector spaces require the researcher to decide what psychological properties are represented by the connections (in this case dimensions) between words. In this case, we opted for referent feature norms, also known as semantic feature production norms or McRae feature norms (McRae, Cree, Seidenberg, & McNorgan, 2005). These quantify the properties of referents (e.g. both elephants and airplanes are large objects, so they would have a link). In turn, the constructed vector spaces capture the perceived real-world structure of the objects represented by words.

The discussion of "shape bias" was the perfect field to apply this in. There has been a long-standing dialogue between groups of researchers as to what exactly shape bias represents, and even raising questions whether it is a valid concept at all. The academic exchange on this topic usually references experimental studies and their results. I felt that exploring the discussion mathematically, from a very different perspective, would be both intriguing and valuable.

The methodology used operationalises distances between words in vector space. Different distance measures are proposed and contrasted with one another. This was a crucial piece of my PhD, as it allows for the quantification of psychological differences. The distance measures form the backbone of the Macroscope platform, powering some of the comparative figures we generate.

Introduction

When learning new words, children need to generalize word-object-mappings across different items that vary along numerous dimensions. For example, suppose a child learns the word "spoon" in relation to one spoon. In learning new words for other objects, they need to be able to distinguish the category of spoons (which do not need new labels) from other categories of objects that do, such as forks, bowls, and toothbrushes. If an item looks too much like a spoon (e.g., a spork), the inability to distinguish this item from other spoons may, in principle, make it more difficult to learn the new word. Children are proposed to overcome this apparently difficult task with the assistance of learning biases (Markman, 1990). In this article we will combine what we see as two complimentary biases that help children solve this problem, and we will use these to develop and investigate a new notion of feature distinctiveness in age of acquisition.

One bias known to influence word generalization is *shape bias*. When children are asked to learn the name of an unfamiliar object, they tend to generalize the word to other objects based on shape rather than other feature types, such as texture, color, or material (Landau, Smith, & Jones, 1988). For example, if a child is presented with a round, rubber object in association with the word "dax," it is more likely that the child will generalize the word "dax" to other objects that are round, rather than generalizing "dax" to other objects made of rubber. Shape bias is not always found when investigating early world learning (Cimpian, & Markman, 2005) and is clearly influenced by developmental trends (see Landau, Smith, & Jones, 1988). Questions about the origination and generality of shape bias are quite common in the literature (Markson, Diesendruck, & Bloom, 2008; Kemp, Perfors, & Tenenbaum, 2007; Booth, & Waxman, 2008), as well as how it may be influenced by more general learning or feature biases—the topic we take up here.

There are a number of additional biases that may be considered to fall into the broad class of *feature biases*, with the two most prominent being texture and material biases (e.g., Jones, Smith, & Landau, 1991; Samuelson & Horst, 2007; Soja, 1992; Soja, Carey, & Spelke, 1991). Though there are a number of additional word features associated with phonology and structure in the language (Hills, 2013; Morgan, 1996; Iversen, Patel, & Ohgushi, 2008; Thiessen, & Saffran, 2007), here we are concerned with the early influence of object feature biases by which children generalize word labels associated with objects to different degrees depending on their feature similarity. We define the term feature bias as a focus on a specific dimension of the referent object during word learning, relative to other available dimensions. Statistically, a feature bias may be represented as a correlation between successful word learning and a systematic feature pattern in the learned words (and their referents) along a specific dimension. Shape bias would then represent situations where the unequal generalization of word labels is correlated with a difference in shape, but not a difference in other feature dimensions.

A second bias known to influence word learning, which may work in tandem with feature biases, is *mutual exclusivity*. In early lexical development children prefer to assign novel words to objects that do not yet have names. This means that two different words are often interpreted to refer to two different objects (Markman, Wasow, & Hansen, 2003; but see Bilson, Yoshida, Tran, Woods, & Hills, 2015). Mutual exclusivity has been shown to influence word-object associations and this is related to a preference for labelling novel objects with novel words (e.g., Mather, & Plunkett, 2009; Mather & Plunkett, 2012). In turn, mutual exclusivity is likely to play a strong role in early word learning (Hills, 2013; Kachergis, Yu, & Shiffrin, 2012; Yurovsky, Yu, & Smith, 2013)

Mutual exclusivity may play a still broader role in learning if it extends beyond individual objects to categories. If children can assess whether a newly encountered object is a member of an already named category, then they can assign novel names to objects in novel categories; if an object belongs to an already named category, it should be less likely to receive a new label. How categories are defined should be influenced by children's perceptual and cognitive systems and what properties or features of objects children pay attention to (Colunga & Smith, 2008; Smith & Samuelson, 2006). For example, the more visually dissimilar objects are, the more inclined children are to show a shape bias (Tek, Jaffery, Swensen, Fein, & Naigles, 2012). Feature distributions over categories are also a central concept in the categorization literature, where the influence of feature types and shared versus distinctive features is well known to influence category learning (Love, Medin, & Gureckis, 2004; Sloutsky, & Fisher, 2004; Weitnauer, Carvalho, Goldstone, & Ritter, 2014). If children use mutual exclusivity at the category level, they must do so based on information provided in the shared and non-shared features between objects. This sensitivity to the feature similarity between objects, in turn, may facilitate feature biases in the process of lexical acquisition.[3]

The majority of research on shape bias has been experimental in nature (e.g., Booth, Waxman, & Huang, 2005; Collisson, Grela, Spaulding, Rueckl, & Magnuson, 2014; Graham & Diesendruck, 2010). This research strongly supports a link between shape bias and word learning. Specifically, in a study by Smith and colleagues, a group of children too young to systematically show shape bias during natural word learning were trained over a 9-week period by exposure to novel words for novel objects, presented in categories organised by shape. Children with exposure to this training learned nouns outside the laboratory at a faster rate than control children. This suggests some sensitivity to shape-based (and possibly other feature-based) categories during word learning (Smith, Jones, Landau, Gershkoff-Stowe, & Samuelson, 2002).

[3]Whether this feature similarity extends to the conceptual understanding of features (and categories) is the topic of an ongoing debate (for a review of the debate see Elman, 2008). We do not attempt to resolve this debate, but nonetheless aim to provide a novel investigation into the relationship between feature biases, mutual exclusivity, and age of acquisition.

If mutual exclusivity and feature biases work together to facilitate early lexical learning then structural relations between one object's features and the features of other objects a child is exposed to should provide information with respect to an object's age of acquisition. That is, feature biases may help a child assess the similarity between a novel object and the already learned lexicon (by restricting which dimensions are being considered). If a word for a similar referent already exists in the lexicon, mutual exclusivity would suggest the novel object does not need a new word label – in essence hindering word learning. On the contrary, if a novel object is encountered that is highly dissimilar to anything the lexicon has labels for, mutual exclusivity would not pose a restriction to word learning, indirectly facilitating the acquisition of a new word label. In this paper we look at some of the implications of combining feature biases and mutual exclusivity. Through feature network analysis we aim to offer insight into how these implications are reflected in age of acquisition.

To date, however, extensions of feature information to predict age of acquisition have been unsuccessful. Hills, Maouene, Maouene, Sheya, & Smith (2009a) investigated the role of feature similarity in age of acquisition by producing weighted networks of words based on shared features. In these networks, nodes in the network represented nouns and edge weights between nodes represented the number of shared features. Using several different modelling approaches, Hills et al. (2009a) found no evidence that feature similarity could predict age of acquisition. While subsequent work has investigated the role of semantic and associative factors in early word learning (Hills, Maouene, Riordan, & Smith, 2010; Hills, 2012), to our knowledge no further progress has been made with features. However, one potential problem with the approach in Hills et al. (2009a) is that it overlooks the inference from mutual exclusivity outlined above—that is, learning may be predicted not by shared-features but by feature distinctiveness, which we operationalize here as a measure of non-shared features between one or more objects.

Network analysis, or graph theory, allows investigations of structured information based on the relationships between objects, or nodes. In studies of language learning and language processing, the relationships (i.e., edges) between nodes have been based on, for example, phonetic similarity (e.g., Arbesman, Strogatz, & Vitevitch, 2010), free association norms (e.g., Hills, Maouene, Maouene, Sheya, & Smith, 2009a), and co-occurrence in language (e.g., Beckage, Smith, & Hills, 2011; Hills et al., 2010). In the present work, we use several quantitative measures of distinctiveness to produce networks of words for objects based on the features of those objects. In these networks, edges now represent how dissimilar two referents (objects in the real world) are. This then allows us to compute distinctiveness for objects in the network, as a function of their overall dissimilarity to other objects in the network. In addition, this also allows us to investigate additional structural properties of distinctiveness, such as to what extent this effect is driven by specific feature types or is a local (near neighbour) or global (across all words) property.

Our approach is based on the inference that feature biases and mutual exclusivity should lead to two distinct patterns in age of acquisition. First, words representing objects that are more distinctive should be learned earlier than words for objects that are less distinctive. That is, if the basis for two objects being considered the same is a function of their shared and non-shared feature distribution, then objects that share fewer features with other objects—which are therefore more distinctive—should be learned earlier. Mutual exclusivity could then be viewed as an example of a distinctiveness driven relationship between objects, rather than a stand-alone learning bias. Secondly, the classic account of shape bias (as a prominent subset of feature biases) should predict that the above finding will be more prominent with features relating to object's shape, rather than other non-shape related features of an object (e.g. function, sound, or material). As noted above, the precedence of shape over other feature categories (e.g. function) has been subject to conflicting experimental results (e.g., Diesendruck & Bloom, 2003). However, it may also be that shape bias is a component of feature biases, where each of the component biases (including texture and material) are each driven by a similar process of distinctiveness.

Finally, we also ask to what extent distinctiveness represents a global or local property by investigating how distinctiveness operates within categories (i.e., sub-networks) in comparison with all words simultaneously. We know of no prior work on the topic of local versus global distinctiveness. However, for distinctiveness to be effective for word learning it is particularly important that it function locally, allowing children to discern similar objects from one another. Should distinctiveness only work on a global level, it may be primarily driven by the fact that some superordinate categories of words may overall be more distinctive than other superordinate categories. Such an effect would be less useful when trying to distinguish two closely related objects. The contribution of local effects is investigated using network clustering analyses.

Our goal is to better understand how feature distinctiveness might contribute to early learning biases (e.g., shape bias). Specifically, if feature distinctiveness is driving the early influence of processes like mutual exclusivity, it may also provide a broader explanation for feature biases more generally, and offer an explanation for previous findings in the literature. Word learning is a complex process shown to involve a large variety of factors. The present work is not an attempt to explain how word learning works as a whole. Instead, it is aimed to be a novel approach to understanding how feature distinctiveness might affect certain feature-related learning biases.

Methods

Features

We used the McRae semantic feature production norms as the basis of our network generation (McRae, Cree, Seidenberg, & McNorgan, 2005). This is a collection of 541 living and non-living

concepts (nouns) with features collected from approximately 725 adult participants for each concept. These include 7259 unique features, which are labelled and categorised both according to the division proposed by Cree & McRae (2003) and also the taxonomy of Wu & Barsalou (2009). Each listed feature is assigned to one Cree & McRae category and one Wu & Barsalou category. See Table 2 for examples of the listed features and their Cree & McRae category assignment.

TABLE 2

Examples of listed features with their respective assigned categories

Assigned category (exhaustive list)	Three randomly sampled example features
visual form and surface	has legs, is big, made of wood
visual motion	runs, crawls, is fast
visual colour	is green, is dark, is colourful
taxonomic	is a fruit, is an animal, is a tool
encyclopaedic	used long ago, found in houses, made by bees
function	is eaten, used for building, requires slicing
sound	is loud, is buzzing, plays music
tactile	is rough, is sharp, is soft
taste	is sweet, tastes good, tastes hot
smell	is smelly, smells nice, smells bad

Table 2. Examples of listed features with their respective assigned categories. Each of the 7259 listed features is assigned to exactly one of the 10 Cree & McRae categories. The "assigned category" is an exhaustive list of all the possible categories. The three example features have been randomly sampled to illustrate what features these categories contain.

The feature norms allow us to see each object as a concept representing a list of features. Furthermore, the division of these features into categories allows us to quantify to what extent objects differ from one another in terms of these feature categories (e.g. *does one word represent a higher proportion of visual features than another?*).

Some developmental studies indicate that children do not simply learn words, but also perceive them as categories containing features (Sheya & Smith, 2006). While children may, or may not view the world through the fine-grained categories listed above, it has been established that children can and do

use object features as one of the main components of lexical development and that they do use broader categories that with age become similar to adult categories (Schyns, Goldstone, & Thibaut, 1998; Keil & Batterman, 1984). Infants also make links between words, their referents and the referents' functions to facilitate lexical learning, hinting at the ability to understand deeper connections behind words (Booth & Waxman, 2002). At the same time, the extent to which children understand deeper categories is a topic of continuing discussion within the literature. Some studies suggest children's learning of categories may not mean children understand the deeper structure of the category (Plunkett, Hu, & Cohen, 2008). The 'encyclopaedic' feature category may be an especially inaccurate representation of how a child views the world, as it relies on extensive general knowledge. For that reason, we removed encyclopaedic features from our analyses, reducing the dataset to 5842 features. Leaving them in, however, does not change the general conclusions of this work, with similar results throughout. With this in mind, we encourage the reader to not interpret the feature categories as direct representations of a child's view. Instead, we suggest the present study is an exploratory view into what adult-perceived word categories are most susceptible to distinctiveness effects in relation to age of acquisition.[4]

Age of acquisition

We obtained the age of acquisition from the Kuperman norms dataset (Kuperman, Stadthagen-Gonzalez, & Brysbaert, 2012). This is a database of over 30,000 words, ranked on age of acquisition by a large sample of online participants. These norms show the retrospectively estimated age of acquisition of a given word. This database also contains 492 words in the McRae feature dataset (out of 541), which we used in this study. The age of acquisition rating represents the year in which an adult participant estimated that they understood a given word. The data used in our study is the mean age of acquisition rating for a given word – this is the mean retrospective age of acquisition rating based on 20 participants' responses to each word. It is important to note that these are retrospective estimates of age of acquisition by an adult population. Adult age of acquisition ratings are generally much easier to collect than parent-observed (direct) age of acquisition indexes, resulting in bigger datasets. The bigger dataset is the primary reason for opting to focus on adult retrospective ratings in the present study. Using the Kuperman norms, our pool of analysable data grows considerably in contrast to using parent-indexed age of acquisition. One can get some idea of how well the Kuperman (adult retrospective) norms actually relate to children's lexical pool by correlating the ratings with a parent-indexed age of acquisition rating dataset. One such parent-indexed dataset is the MacArthur-

[4] It is important to keep in mind that the McRae feature norms are generated by adults. They do not offer direct insight into how infants view the world, but should be viewed as a plausible proxy representation of an infant's view. Our results then, rely on the assumption that the adult view of an object's features is to an extent representative of an infant's.

Bates Communicative Developmental Inventory, toddler version (MCDI; Dale, & Fenson, 1996). This is a set of 680 words along with production metrics for 1789 children, collected monthly from caregivers of children between the ages of 16 to 30 months. These show the proportion of toddlers using a given word at a specific month of age. For the words that were present in both norm databases, the reported age of acquisition in the Kuperman norms was positively correlated with the reported MCDI age of acquisition (first recorded month of age where at least 50% of the toddlers demonstrated the use of a given word) $r_s(107) = .64$, $p < .001$[5]. While the Kuperman norms correlate well with direct age of acquisition measures and also with a cross validation adult sample (see Kuperman et al., 2012), our main results reflect retrospective adult ratings. Any inferences in this paper to child learning are based on the assumption that retrospective adult age of acquisition ratings are representative of the size of a child's lexical pool. While we believe this to be true, additional possible explanations are also considered in the discussion.

Furthermore, the mean adult-rated estimate age of acquisition for our sample of 492 words was 4.3 years (SD = 0.83). The mean parent-observed (direct) age of acquisition for the same word sample, collected from the MacArthur-Bates Communicative Developmental Inventory, toddler version (MCDI; Dale, & Fenson, 1996), was 1.8 years (SD = 0.28). Combined with the mean rating correlation mentioned in the previous paragraph, this suggests that while adult retrospective ratings may be generally indicative of which words are learned earlier in relation to other words, the rating may be inaccurate with respect the exact time of learning. Specifically, all of our results should be viewed as based on relative age-of-acquisition relationships between words, not claims about how distinctiveness relates to absolute age of acquisition (e.g. specific months of learning).

For the above reasons (see footnote), the majority of our analyses are non-parametric. As a result, the reported correlations are based on rank measures rather than absolute values. This is not only because of normality assumptions, but also to ensure the effects are not misinterpreted due to the inconsistency in absolute age between adult age of acquisition ratings and parent-observed indexes.

[5] Spearman's rank correlation was used due to a possible violation of the normality assumptions. A Shapiro-Wilk test showed that both the Kuperman age of acquisition data ($W = 0.97$, $p < .001$) and the MCDI age of acquisition data ($W = .97$, $p < .05$) are significantly different from a normal distribution.

Distinctiveness Measures

Our measures of distinctiveness involve both counts of distinctive features as well as network representations. *Relative feature distinctiveness* is a measure of how rare a noun's individual features are with respect to all other words in the norms. This is defined as the sum of the distinctiveness of all the features reported for a word:

$$x = \sum_{i=1}^{m} \frac{1}{w_i} \quad (1)$$

where *x* is the overall relative feature distinctiveness of a word; *m* is the total number of features listed for a word; and w_i is the number of words listing feature i. Thus, the more objects that have a feature, the less distinctive the feature. Note that this measure is not computed at the pair-wise level between words, but computes rareness over features and then sums this for individual words.

Network measures. Our network distinctiveness measures involve assigning each node in our networks to a noun from the McRae feature dataset. The distance between nodes in the network then represents a measure of relative distinctiveness. The categorisation of feature types mentioned above allowed us to generate networks based on all features and also based on subsets of features associated with specific categories, where the distances between the nodes of a network then represent feature dissimilarity on only one feature dimension (e.g. 'visual colour').

We calculated the network distinctiveness using three different approaches, which each make different underlying assumptions about the features involved. These are the non-shared feature distance, the Jaccard distance and the Manhattan distance. Together, these three measures allow us to isolate the effects of shared features (comparing Jaccard distance with non-shared feature distance) and the role of salience (comparing Manhattan distance with non-shared feature distance). Salience, here, refers to a sensitivity to the prototypicality of features – i.e. the ability to distinguish frequently reported features for a word (e.g. "dog has legs"), from features reported by only a few participants (e.g. "dog smells").

The *non-shared feature distance* calculates the distance between two nouns as the sum of all the features the two respective words do not have in common. In set theory, this is known as the symmetric difference between the two feature sets:

$$d = (n_1 + n_2) - 2n_s \quad (2)$$

Where n_1 is the total number of features listed for the first word; n_2 the total number of features listed for the second word; and n_s the total number of features that were shared by word one and two. This measure focuses only on the dissimilarity between two concepts, because shared features are excluded.

It also represents a feature listed by one respondent with the same strength as a feature listed by ten respondents, and thus does not weight features in relation to the frequency with which they are produced.

The *Jaccard distance* calculates the distance between two nouns as the ratio of the symmetric distance between the intersection of the two nouns' feature sets and their union. Thus, the Jaccard distance normalizes the distance in relation to the total number of features available for comparison. Formally, the Jaccard distance between two nodes is

$$d = 1 - \frac{n_s}{n_1 + n_2 - n_s} \tag{3}$$

Where n_1 is the total number of features listed for the first word; n_2 the total number of features listed for the second word; and n_s the total number of features that were shared by word one and two.

In contrast to all our other measures we use, the Jaccard distance is sensitive to shared features. Nouns with the same number of non-shared features but with a larger set of shared features will have a smaller Jaccard distance than nouns with fewer shared features. Thus, Jaccard distance is reduced when shared features are added.

The *Manhattan distance* calculates the distance between two nouns as the sum of the absolute differences between the proportions of participants reporting each feature. Thus, the Manhattan distance measure takes into account feature salience, where salience is indicated by the number of participants who produce a given feature. The Manhattan distance is

$$d = \sum_{i=1}^{n} |p_i - q_i| \tag{4}$$

Where n is the number of features recalled for both words (i.e., the union of the two feature sets); p is the proportion of people listing feature i for the first word and q is the proportion of people listing feature i for the second word. Proportion data can be found in the McRae norms and is defined as the total number of people listing a feature for a given word divided by 30 (each word was annotated with features by 30 people). Unlike the other measures, Manhattan distance places emphasis not on the difference in number and types of features listed, but quantifies to what extent these differences in features are salient. Figure 5 shows an example of how these three measures, provided the same input data, generate different networks.

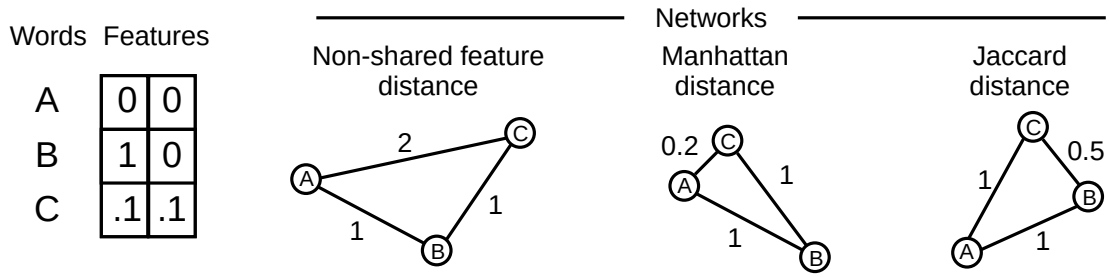


Figure 5. An example of the three different distance measures used to compute feature distinctiveness. Each of the three words (A, B, and C) each has a vector representing the proportion of individuals who reported each of two features; each feature is represented by a column in the feature matrix shown on the left. To the right, networks are shown for each of the distance metrics described in the text.[6]

Results

Relative feature distinctiveness predicts age of acquisition

Overall relative feature distinctiveness was negatively correlated with Kuperman's age of acquisition $r_s(490) = -.21$, $p < .001$, showing that words with more distinctive features are perceived to be learned earlier. This negative correlation was also found between overall distinctness and the MCDI (parent-reported) age of acquisition, $r_s(111) = -.19$, $p < .05$, showing that words with more distinctive features are learned earlier when using the parent-reported age of acquisition norms as the base of the analysis.[7] All of the correlation results mentioned in this paper are Spearman's rank correlations. This is both due to the fact that the age of acquisition data is likely not normally distributed and also due to an absolute difference between the two types of age of acquisition (parent-indexed and adult rated, as mentioned in the method's section). When using adult rated age of acquisition, the Spearman's rank correlation should produce results more relevant to child learning.

Relative feature distinctiveness of visual form and surface features best predicts age of acquisition

The division of features into feature types allow us to investigate whether features of different types predict age of acquisition to differing extents. Each feature was labelled with one of ten types as proposed by Cree & McRae (2003)—for example, 'visual form and surface,' 'visual motion', etc. For

[6] Data and code to reproduce the results found in this work can be found online: http://www2.warwick.ac.uk/fac/sci/psych/people/thills/thills/engelthalerhills-analysis_scripts.zip

[7] Limiting the analysis to 109 words in the MCDI in a regression predicting distinctiveness, neither MCDI (parent indexed age of acquisition) nor the Kuperman norms (adult retrospective age of acquisition) explain additional variance after controlling for the other. This suggests that the two norms account for a common variance in explaining feature distinctiveness.

each word, we computed the relative feature distinctiveness for each of the feature types for which it contained features. We then used the relative feature distinctiveness across feature types to predict age of acquisition. A multiple linear regression revealed that only two feature types were predictive of age of acquisition – 'visual form and surface' and 'visual motion' (Table 3). Visual form and surface was the best predictor among all the features.

When looking at parent-reported age of acquisition (MCDI), no significant effects were found. The main reason for using adult-rated age of acquisition is the fact that they are much larger in size, allowing us to cover a broader range of words (gaining more statistical power). The MCDI only contains 113 words for which there are McRae feature norms. In comparison, the Kuperman adult age of acquisition norms contain 492 words overlapping with the McRae feature norms. When looking at subsets of feature categories, this difference in sample size becomes even more apparent. When interpreting our results and in comparing it with earlier work (Hills et al., 2009a), it is important to keep this difference in mind.

TABLE 3

Variance for each feature category as an outcome of a linear regression model. (Relative feature distinctiveness predicting adult reported age of acquisition)

	Age of acquisition	
Variable	*B*	CI
Visual form and surface	-0.12*	[-0.20, -0.03]
Visual motion	-0.10*	[-0.19, -0.00]
Sound	-0.09	[-0.18, 0.00]
Tactile	-0.06	[-0.15, 0.03]
Function	-0.03	[-0.12, 0.06]
Taste	-0.02	[-0.11, 0.07]
Smell	-0.02	[-0.11, 0.07]
Visual colour	-0.00	[-0.09, 0.09]
Taxonomic	-0.00	[-0.09, 0.10]

Table 3. Variance for each feature category as an outcome of a linear regression model. *Note.* N = 492. Adjusted $R2$ = .02, F = 2.26*. B = standardized beta coefficient. *CI* = 95% confidence interval. $*p < .05$. $***p < .001$.

Network distinctiveness predicts age of acquisition

As shown in Table 4a, both the non-shared feature network and Manhattan distance network showed negative correlations between distinctiveness and age of acquisition—$r_s(490)$=-.24, $p < .001$ and $r_s(490) = -.34$, $p < .001$, respectively. The Jaccard distance network did not show a significant correlation ($r_s(490) = -.04$, $p > .05$). These correlations show that network measures of distinctiveness based on dissimilarity are more strongly correlated with age of acquisition than a measure of relative feature distinctiveness, and they are also consistent with our prediction based on feature biases and mutual exclusivity.

When the three distance measures were analysed as predictors in a linear regression model, only Manhattan distance explained additional variance after controlling for the other two distance measures (Table 4b).

As mentioned in the methods section, the Jaccard distance metric is sensitive to shared features. On the other hand, non-shared feature distance and Manhattan distance use dissimilarity in feature makeup to calculate distance between words. The absence of a significant correlation of Jaccard distance suggests that presence of shared features does not relate to age of acquisition the same way the presence of distinctive features does. Hence, feature dissimilarity (rather than feature similarity) may be more important when looking at age of acquisition effects. This may also explain the limited predictive power of the network similarity approach taken in Hills et al. (2009a), which was based exclusively on shared features.

TABLE 4a

Word distinctiveness and age of acquisition

Model	r_s	CI
Nonshared feature distance	-.19***	[-0.27, -0.11]
Jaccard distance	-.03	[-0.12, 0.06]
Manhattan distance	-.28***	[-0.36, -0.20]

Table 4a. Word distinctiveness and age of acquisition. *Note.* N = 492. Df =490. R_s = Spearman rank-order coefficient. CI = 95% confidence interval. *$p < .05$. ***p < .001.

TABLE 4b

Regression table for distinctiveness measures predicting Kuperman age of acquisition

Variable	Age of acquisition	
	B	CI
Nonshared feature distance	-0.03	[-0.20, 0.13]
Jaccard distance	-0.06	[-0.14, 0.03]
Manhattan distance	-0.17*	[-0.33, -0.01]

Table 4b. Regression table for distinctiveness measures predicting Kuperman age of acquisition. *Note*. N = 492. Adjusted $R2$ = .04, F = 7.07***. B = standardized beta coefficient. *CI* = 95% confidence interval. **p < .05. ***p < .001*.

TABLE 4c

Correlations between the distance measures

Variable	1	2
1 Nonshared feature distance		
2 Jaccard distance	0.01	
3 Manhattan distance	0.84	-0.10

Table 4c. Correlations between the distance measures. Values represent spearman rank correlation coefficient.

Due to potential multicollinearity, Table 4c shows correlations between the implemented distance measures. It is apparent that the one strong correlation is between Nonshared feature distance and Manhattan distance. This is to be expected, as Manhattan distance is essentially Nonshared feature distance weighted by the numbers of participants reporting a feature. Jaccard distance is not strongly correlated with either one (and even shows an inverse direction with Manhattan distance). This reinforces the claim that Jaccard distance is a measure of concept difference rather than concept similarity, and these two concepts capture fundamentally different aspects of semantic relationships between objects in the world. In addition, the calculated variance inflation factor are as follows: Non-shared feature distance (3.39), Manhattan distance (3.42), Jaccard distance (1.04). Similarly, our interpretation is that the first two are capturing the same effect, while Jaccard distance is a fundamentally different measure. A possible explanation for the fact that Non-shared feature distance (NFD) does not

correlate with Age of Acquisition in the regression table is that NFD is too crude of a measure, not sensitive to partial feature membership with associated words. Quantifying which features are prototypical for words is only possible using Manhattan distance. Theoretically, this suggests that estimating similarity of words is important during word learning, and that this estimation is sensitive to the prototypicality of features.

Distinctiveness within feature types

As noted above for relative feature distinctiveness, some feature types may be more salient than others and, in turn, more correlated with age of acquisition. The developmental progression of feature salience may be taken to indicate that non-perceptual features should be less predictive than perceptual features (e.g., Sloutsky, 2010). To address this, we used Manhattan distance—our most predictive distinctiveness measure—to construct networks for each feature category separately. This allowed us to calculate the overall per-category distance for each word. The per-category distance was computed by running our Manhattan distance analysis on only a subset of the features – in other words, a metric of how distinct of a feature makeup a word has only taking into account one feature category at a time. This measure then shows how distinctiveness by category correlates with age of acquisition, resulting in a correlation table of category-specific distinctiveness by age of acquisition. Table 5 shows that the strongest correlation was with 'visual form and surface' distinctiveness, showing a negative correlation with age of acquisition $r(484) = -.19$, $p < .001$. This finding suggests that words that are less similar to other words in a 'visual form and surface' network are learned earlier. On the other hand, the 'taxonomic' category is also significant, although in the opposite direction, $r(394) = .10$, $p < .05$. Here the inference may be that more taxonomically distinctive words are learned later, indicating they may belong to more uncommon taxonomic categories.[8] Additionally, comparing this analysis (Table 5) with the previous analysis of *relative feature distinctiveness* (Table 3 – i.e. an approach not sensitive to the prototypicality of features), Visual Motion feature distinctiveness is no longer significantly predictive of age of acquisition. This suggests referents that have highly unusual visual motion features may be learned earlier, but outside of the rarely reported features, the salience of all the visual motion features does not seem to be a significant predictor. This is in contrast to Visual form and surface, where having rare features is predictive of early age of acquisition (Table 3), and distinctiveness in terms of the prototypicality of features is also predictive (Table 5).

[8]Corrections for Type I errors in multiple correlations are not straightforward, especially in cases where the correlations are unlikely to be independent. Several of our analyses have fairly low sample sizes, meaning that adjusting for multiple tests could also result in Type II errors (rejecting a significant result when there in fact is one). Nevertheless, where appropriate, we include both unadjusted and Bonferroni adjusted *p* values. We encourage the reader to compare these two and make their own judgement on how robust our results are. For a discussion on the merits of not using corrections of *p* values, see Rothman (1990) or Nakagawa (2004).

TABLE 5

Relationship between feature-network word distance and age of acquisition

Model Feature Type	r_s	CI	n
Visual form and surface	-0.19*** [bbb]	[-0.28, -0.12]	486
Smell	-0.18	[-0.56, 0.25]	19
Taste	-0.14	[-0.39, 0.13]	51
Tactile	-0.11	[-0.28, 0.05]	165
Visual Colour	-0.08	[-0.20, 0.04]	264
Sound	-0.05	[-0.29, 0.17]	93
Visual Motion	-0.05	[-0.21, 0.11]	165
Function	0.05	[-0.05, 0.15]	418
Taxonomic	0.11*	[0.01, 0.21]	396

Table 5. Relationship between feature-network word distance and age of acquisition. *Note*. N = number of observations. Df = (n – 2). R_s = Spearman rank-order coefficient. CI = 95% confidence interval. *p < .05. ***p < .001. [b] = p < .05. (Bonferroni correction). [bbb] = p < .001. (Bonferroni correction).

Due to the differences in the category sample size, a possible interpretation is that the relatively strong correlation of 'Visual form and surface' features is due to the sample size. Smaller samples could mean noisier distance measures, and therefore a lower correlation coefficient. We can investigate this by analysing the relationship between the correlation coefficients (r_s) in Table 5 and the respective sample sizes of the categories (n). A spearman's rank correlation showed no relationship between the coefficients and the respective sample sizes $r(7) = -.06$, $p = 0.87$. This suggests larger sample sizes are not the primary contributors to the effects described in this chapter.

Distinctiveness networks within 'visual form and surface'

The 'visual form and surface' feature category is the largest feature category, covering 32% of all the reported features. This category can be further broken down into more specific subsets. To do this, we used the taxonomy proposed by Wu & Barsalou (2003) to investigate the distinctiveness within visual form and surface features. This resulted in subsetting visual form and surface features into 5 subcategories (omitting two additional categories that represented less than five features). We then constructed Manhattan distance networks for each word in each subcategory network and calculated the overall network distance for each word. Table 6 shows that only two groups displayed significant negative correlations with age of acquisition: 'external surface property' $r_s(344) = -.12$, $p < .05$; and 'material' $r_s(238) = -.25$, $p < .001$. These findings suggest that only two of the visual form and surface subgroups are indicative of distinctiveness associated with age of acquisition, where words for more

distinctive items are learned earlier. 'Made of' represents a sensitivity towards material. Whereas 'external surface property' represents a sensitivity towards general shape (e.g. 'is long', 'is round'). We address this further in the discussion in relation to shape bias.

TABLE 6

Correlation between Manhattan distance of words and Age of acquisition for various feature types within visual form and surface networks.

Model Feature Type	r_s	CI	n
Made of (material)	-0.25*** [bbb]	[-0.36, -0.13]	240
External surface property	-0.12*	[-0.22, -0.02]	346
Internal surface property	-0.10	[-0.50, 0.31]	29
External component	0.00	[-0.10, 0.11]	377
Internal component	0.12	[-0.06, 0.28]	113

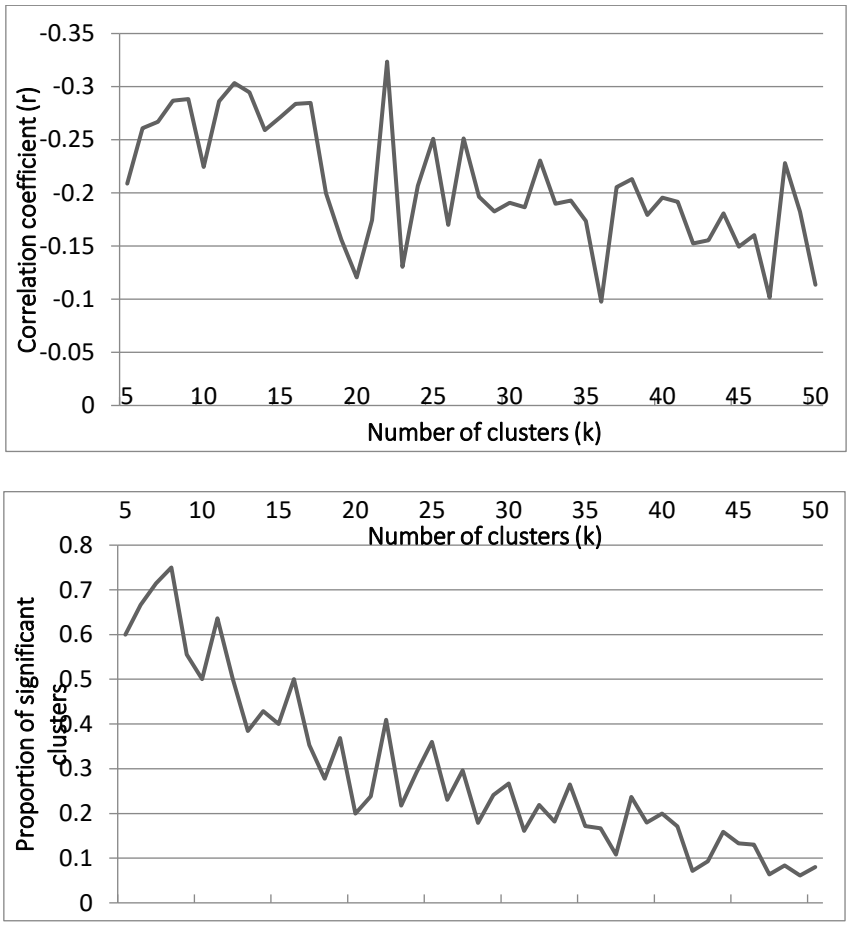
Table 6. Correlation between Manhattan distance of words and Age of acquisition for various feature types within visual form and surface networks. *Note.* n = number of observations. df = (n - 2). r_s = Spearman rank-order coefficient. CI = 95% confidence interval. *p < .05. ***p < .001. [b] = p < .05. (Bonferroni correction). [bbb] = p < .001. (Bonferroni correction).

Global versus local feature distinctiveness as a predictor

All of the above distinctiveness measures are global. That is, feature distinctiveness is measured as a function of relationships between *all* nouns in the data set. However, this does not allow us to distinguish between local and global distinctiveness, and may indicate that object names are learned either a) because they are in distinctive clusters (small clusters of objects that are fairly distinct from all other objects) or b) because they are fairly distinct within a cluster. In laboratory studies distinctiveness is implicitly taken to be a local measure among other objects in the study, though some objects may be more or less similar to other items the child is familiar with. In previous work, Hills et al. (2009b) found that objects clustered by features tended to form meaningful categories, such as clothes, food, and animals. Moreover, feature similarity strongly influences children's inductive generalizations (Sloutsky, 2010; Sloutsky & Fischer, 2011). For object distinctiveness to be useful for word learning, distinctiveness within categories should also be predictive of age of acquisition.

To test this, we used the spherical k-means network clustering algorithm to cluster nouns in relation to feature similarity. Spherical k-means finds an appropriate clustering given a number of clusters, *k*, such that objects are placed in clusters with the nearest mean similarity (Dhillon & Modha, 2001). The methodology finds the best centroid locations of each cluster, by minimising the angle between the vectors of the components. For a detailed explanation of the mathematical foundations of spherical k-means, see Buchta, Kober, Feinerer, & Hornik (2012). In the present case, words in the same cluster

are likely to share similar features, based on the cosine similarity of their shared feature vectors. Following clustering, we computed the mean Manhattan distance for each noun from its other category members (i.e., the distance to the cluster centroid). For each cluster, we then correlate age of acquisition with cluster centroid distance. We computed the analyses for a range of cluster numbers (5 to 50). Figure 6a shows the mean average correlation for all clusters within a given cluster breakdown. Similarly, Figure 6b shows the proportion of clusters showing a significant *p* value ($p < .05$). This allows us to see both what number of clusters shows the strongest mean effect, but also highlights the fact that our results are apparent over a range of assumptions about the number of clusters.



Figures 6a and 6b. Mean correlation coefficients and proportionate p value significance of within-cluster distinctiveness and age of acquisition. For each cluster number, *k*, words are divided into *k* clusters using spherical-k-means, then the words' Manhattan distance from all other members of its cluster were computed and correlated with age of acquisition. One correlation is calculated per cluster, with the mean across all clusters presented in Figure 6a. The proportion of clusters showing a significant *p* value in said correlation is presented in Figure 6b.

Table 7 shows the cluster breakdown for the cluster number $k = 7$, which had the largest proportion of significant clusters, showing a significant correlation between centroid distance and Age of acquisition (71%). The proportion of significant p-values is not a measure of effect size, but rather an indicator of the relationship between cluster size and effect size. Naturally, as the number of clusters increases,

their size of each cluster will decrease and finding a significant correlation between cluster members and age of acquisition will prove more difficult. Consequently, this analysis is not intending to find the best clustering setting, but rather find a setting of clusters that is interpretable using a rank correlation, staying consistent with the approach to previous analyses. As shown, a noun's feature distance from its category members is strongly negatively correlated with age of acquisition across the majority of clusters. Words most distinctive from their other category members are learned earliest. Distinctiveness therefore appears to function as a local measure of how dissimilar are the most similar objects, and would therefore be appropriate for word learning within these subcategories. However, due to the fact that the local measure is calculated as a subset of the global measure, it is not possible to put the two metrics into the same regression (the global measure explains the variance of the local measure by definition). To establish a clear insight into whether the effect is driven by a local or a global system, a study with independent local and global measures would be necessary.

TABLE 7

Correlation between words' distinctness within a cluster and age of acquisition

Cluster	*r*	CI	n	exemplars
1	-0.21	[-0.43, 0.05]	61	car, trolley, piano
2	-0.34*	[-0.55, -0.08]	55	skyscraper, door, box
3	-0.31**	[-0.48, -0.12]	98	fork, sink, hatchet
4	-0.25*	[-0.43, -0.05]	97	broccoli, honeydew, sardine
5	-0.36*	[-0.58, -0.09]	49	butterfly, swan, wasp
6	-0.35**	[-0.55, -0.13]	68	mouse, coyote, sheep
7	-0.29*	[-0.50, -0.05]	64	scarf, wand, necklace

Table 7. Correlation between words' distinctness within a cluster and age of acquisition. *Note.* n = number of observations. df = (n - 2). *r* = Pearson product-moment correlation coefficient. CI = 95% confidence interval. Exemplars show three example members of a given cluster. Spherical k-means clustering algorithm was used to categorise data into 7 clusters. *$p < .05$. **$p < .01$. ***$p < .001$.

Discussion

Object features influence how we experience the world and how we discriminate and name different objects. In the present work, we showed that it is viable to explore feature biases and mutual exclusivity through network analysis, and that this approach can elucidate potential factors influencing early word learning. The present work makes three contributions to this area of research. First, object distinctiveness was negatively correlated with age of acquisition, implying that words associated with more distinctive object features are learned earlier. Second, different feature types contribute to this effect to differing extents, with the principal feature types associated with visual form and surface

properties. Third, using cluster analysis we demonstrated that this effect was a local property of distinctiveness, indicating that these effects may come about via distinctiveness between near neighbours in the feature space as opposed to items belonging to distinctive categories. In what follows we briefly describe these results and their implications.

Throughout the literature, mutual exclusivity has been defined as the tendency to pair one object with only one word label (Markman, & Wachtel, 1988; Jaswal, & Hansen, 2006). This concept is often framed as a word learning principle. However, if word learning is in fact an example of a categorisation process, then mutual exclusivity may represent an example of a principle based on categorisation more generally. As such, mutual exclusivity may be framed as a property of categorisation with the implication that the more dissimilar two objects are, the less likely they are to be assigned to the same category. In the context of word learning, this would mean that dissimilar objects are more likely to take on a new label – a finding demonstrated by the present paper. This view – of mutual exclusivity as a categorisation facilitator – is a novel addition to how the field has framed mutual exclusivity so far, and one we feel is provocatively supported by the present research.

All of our analyses showed that distinctiveness negatively correlates with age of acquisition – meaning that more distinct words are learned earlier. This finding offers additional support to the mutual exclusivity principle, which we interpreted to suggest that children should learn names for more dissimilar objects more easily. The observation that Manhattan distance and non-shared feature networks are correlated with age of acquisition, while the Jaccard distance is not, suggests that mutual exclusivity may be based more on distinguishing features of objects than on the number of shared features. The role played by shared and distinct features in mutual exclusivity is an important line of future research.

The importance of distinctive versus shared features was demonstrated by the relative performance of Manhattan distance versus Jaccard networks. The accuracy of the Jaccard metric relies on comparing the proportion of non-shared features to overall features in the union of the two features sets for two objects. On the other hand, the Manhattan distance calculation is built around contrasting two concepts on differences in feature salience; the addition of equally salient shared features does not influence this calculation. This suggests, perhaps paradoxically, that the kinds of information that make items similar (e.g., in relation to inductive inference) may be quite different from the kinds of information that make objects distinctive (e.g., in relation to the mutual exclusivity principle). This suggests a cognitive model that places different weights on shared versus distinctive information – a finding that may further inform the underlying assumptions behind shape bias (Gentner & Imai, 1995; Samuelson & Smith, 1999).

Investigating different dimension weighting in more detail, our regression showed that visual form and surface features were the best predictor of age of acquisition among features. Looking at visual form

and surface more specifically, we found that external surface property and material are particularly important. External surface property covers features like ‘is big’ and ‘is round’ – features that define the general shape of an object. Material defines the material of an object such as, for example, ‘water’ or ‘rubber.’ Additional features also explained variance in age of acquisition in Table 3. Visual motion tells us how something moves, ‘it flies,’ and tactile represents how it feels to the touch, ‘is soft.’ These findings imply that the notion of shape bias is not as simple as comparing the contours of two objects and claiming they are similar if they match and dissimilar if they differ. Instead, it seems plausible that children make an assessment based on general shape in combination with other features (see Samuelson & Smith, 1999). The classic shape bias experiments that juxtapose contour and texture comparison may be looking at an important subsection of a higher dimensional comparison process.

The low predictive power of external components (features such as ‘has legs’ and ‘has buttons’) has several potential explanations. Firstly, it is possible that the nature of the feature coding process results in features that are overly specific or, alternatively, overly generalised. Our results depend on the processes used by adults to generate features, and this is very likely to contain biases that influence our results. For examples, defining a spider as ‘has six legs’ may not be particularly useful if all other objects have numerous distinctive external components as well. On the other hand, defining a human as ‘has legs’ may be too general, as many animals have legs. Thus, it may be the case that external components are not sufficiently detailed in relation to shape to predict age of acquisition—representing a kind of ceiling effect created by obtuseness of the data. Moreover, if each object is discriminable by a specific variation of external components, then the distinctiveness of external components may be uninformative, as they are all distinct. This explanation is supported by research on visual perception, as most of the visual perception studies agree that correct recognition of an object relies on successfully identifying sub-elements of the whole picture (Biederman, 1987; Logothetis, & Sheinberd, 1996; Tarr, & Bülthoff, 1988). Thus, if people identify objects primarily based on unique external components, then computing the distinctiveness of these is unlikely to predict age of acquisition. One way to test this explanation is to collect feature norms on novel object categories, for which names are learned in a second experiment.

As noted above, it is important to note that majority of our results are based on adult retrospective age of acquisition ratings. One possible explanation for some of our results may be that when asked to judge the age of acquisition of a word, adults use feature distinctiveness as a cue to rate the word. However, the fact that the MCDI ratings also show a correlation with distinctiveness suggests that our findings are not an artifact of adults using distinctiveness as an age of acquisition rating cue. Nevertheless, our ratings are correlative in nature, meaning they show a relationship between distinctiveness and age of acquisition, but do not imply causality in one direction or another.

Our results may also be interpreted as a general effect of distinctiveness on age of acquisition, most apparent through some feature categories (e.g. external form and surface). If feature distinctiveness of an object is related to the process of word learning, then it is crucial to control for distinctiveness when designing experimental stimuli. For example, when experimentally investigating whether shape is more "important" a cue than another feature category (such as color), it is important to establish that the stimuli is similarly distinctive in terms of shape as it is in terms of color. It may be the case that it is easier to design stimuli that stand out in terms of shape (some of the shapes used in early shape bias studies are quite unusual), while objects may not stand out in terms of color. Presenting this stimuli to a child may result in a preference for "shape driven learning", as that is the category that is most distinctive in context. In this sense, the inconsistencies in the shape-bias literature may be influenced by stimuli not being evenly distinctive across feature categories. However, it may also be that shape is a feature category that naturally allows for distinctiveness, because it has more dimensions along which it can vary, in ways that other feature categories may not. A literature review comparing the per-category distinctiveness of stimuli used in previous shape bias studies might offer an interesting view into this problem.

It is also important to note that our investigation suggests that features predict a fairly small amount of the variance in age of acquisition. Features should not be viewed as the single driving force behind acquiring early words, but a contributor working alongside a range of other, well established correlates of age of acquisition – some of which are imageability (Gilhooly, & Logie, 1980), frequency (Carroll, & White, 1973), phonemic and letter word length (Whaley, 1978), and the semantic organisation of language (Hills et al., 2010; Hills, 2012). Taken together, these suggest word learning depends not only on the properties of words and how those words are used, but also on the features of the objects to which the words refer.

Chapter 3: Humor norms for 4,997 English words

Project Overview

Constructing a psychological network or a vector space requires a dataset assigning words to psychological values of a given property. As mentioned in the previous chapters, these can be perceived valence values (how positive is a word?), a sum list of referent features (how many participants think an apple is red?) or any other word value pairing. Collecting large scale norms has become more accessible this century, due to the ease of accessing large pools of online participants.

Discussing the topic of humour with my supervisor, we realised there are no systematic, large-scale norms of word humour. If a researcher wanted to know whether something is funny, they would either have to carry out a semantic analysis or use a database of jokes. This was fascinating. Humour is supported by a long history of philosophical research. Classical thinkers proposed a plethora of theories on what humour is, while modern psychologists often run experimental studies on our responses to jokes. Considering the established humour research, it was curious that normative datasets were available for a range of psychological dimensions, but not for humour.

At first, I was sceptical. Perhaps the reason for having no humour dataset is simply that it doesn't work. Most studies focus on jokes or phrases. People have mentioned to me stories of comedians reciting lists of funny words, though that is far from a controlled research condition. In spite of my scepticism, a pilot study with our university students showed a consensus among participants and *turd* was king (among humorous words).

I committed to polishing the data collection process. Creating a custom website offered the flexibility needed to collect data in a bespoke fashion. Following several rounds of testing and optimising the website, the study was deployed to 1000 participants. With the data in place, I consciously decided to publish my findings in a minimalistic way. This allows for other researchers to apply the norms in any direction they see fit.

The first two chapters of the thesis covered the analysis and modelling of psychological properties of words. In contrast, this chapter outlines how to generate new normative datasets of psychological values. It does so while providing the literature on humour, demonstrating how a normative dataset can be collected in a theoretically driven way.

Theories of Humour[9]

Much like the concept of meaning, humour has been discussed since the ancient times. Every person implicitly understands when something is funny. We smile and laugh with no effort. The view of this thesis would pitch humour as a part of a word's psychological frame. Still, intuitively, humour seems a much more elusive topic than for example valence. It is used in social contexts, often times with very specific goals and attitudes. Constructing an effective joke is a craft that often fails when we try the hardest, but works when we least expect it. The complexity of humour is reflected in the theory surrounding it. They are diverse, disconnected and tackling very different views of what humour is. Broadly, the humour theories fall into five categories: evolutionary, play, superiority, energy release, and incongruity. Hurley, Dennett, Adams, & Adams (2011, pp. 37-56) published a brief overview of these categories.

Evolutionary Theories

The fact that smiling and laughing is innate has led evolutionary researchers believe humour must have an intrinsic biological value (Foot, & Chapman, 1976). Laughing and an appropriate facial expression, as a response to a stimulus, is seen not only in very young infants, but also in congenitally blind babies (Thompson, 1941). Additionally, humour and humour-like interactions are seen across a range of human cultures, from highly complex ones to primitive tribal groups (Haig, 1988). All of these findings provide evidence for the claim that humour is not learned, but rather a biological, almost instinctive response to patterns in the environment. Research in this area focuses less on explaining the foundations of behaviour, and more on finding and documenting just how prevalent humour is for humans, even in cases where social learning is unlikely. Consequently, humour can be seen as an in-built part of human existence, a biological response not too different from fear or anger.

Play Theories

While humour may or may not be an innate trait, there is a pool of evidence suggesting it has a role when growing up. Specifically, it promotes play and curiosity – or as Darwin (1872) put it, humour "tickles the mind". The play theories claim that humour is a vehicle for harmless social interaction that promotes healthy organisation of social structures and facilitates individual growth through improved exploration (Cohen, 2008). As a contrast to the evolutionary theory, Spinka, Newberry and Bekoff (2001) point out that both humans and non-human primates participate in play, and that this play often takes on humour-like elements using exaggerated nonverbal communication. According to this view, it

[9] Note that this introductory chapter is written in addition to the humour introduction printed in the original publication of the journal article. This means some of the theories and references are overlapping with the section titled 'Introduction', immediately following 'Theories of Humour'.

is not that humour is an innate biological mechanism, but that the need for play is universal in primates, and humour is an outcome (or a variation) of play. In fact, play is often referred to as a universal counter balancing mechanism to aggression (Panksepp, 1993).

Superiority Theories

Superiority theories consider the motives of humour users. They see humour as an attitude – an expression of an imbalance of power between two parties. This was theorised from the perspective of a general victory – humouring their positive standing as an outcome of successfully conquering a challenge (Hobbes, 1840); and in more sinister light, using humour as a tool for ridiculing others, denouncing the already inferior social standing of those we don't like (Scruton, 1987). In both cases, humour is an attitude used in a complex social environment. The superiority theories assume humour only happens in a social environment that is sophisticated enough to a) have clear discrepancies of social power, and b) to have individuals consciously exploiting the imbalance. These criteria may be too restrictive however, inconsistent with the observations of humour in relatively trivial settings.

Energy Release Theories

According to the psychodynamic approach, humour is a means of releasing energy from the body. In this paradigm, the energy in the body is seen as a negative state, so any release of it outward from the body is a net positive for the organism. For Spencer (1860), every human emotion is a way of transferring energy from within the body to the environment. However, humour held a special role. While other emotions, such as anger, fear or passion accompany the release of bodily energy with a strong action (e.g. fighting an enemy), humour, according to Spencer, does not have an associated action. This makes humour a harmless dissipation mechanism, an invaluable piece in the emotional repertoire of a person. According to Freud (1928), humour is an outcome of a situation where we were plannig to express an emotion, but the environment stopped us from doing so. Humour would therefore be a confused placeholder – an emotion to fill in the space when the intended emotion is unavailable. For example, when facing a life or death scenario (e.g. meeting a wolf in the woods), the initial response might be one of immense fear. The person may be too overwhelmed to feel fear, or feeling it would be highly counter productive to the organism. As such, the body expresses humour instead, laughing at the dire situation. Energy release theories were primarily developer at the turn of the 19th and 20th century, lacking scientific support today.

Incongruity Theories

Humour, and jokes especially, often consist of a violation. Putting two mismatching concepts next to each other makes us laugh. This is the core idea behind incongruity theories – explaining humour as a

measure of stimuli incongruity. First propsed by Beattie (1779), humour was seen as a reaction to an incongruous set. When exposed to an environment that consists of poorly organised objects, people would respond with laughter. This definition is elegant in its simplicity, but also unreliable. There are many examples of incongruous sets from everyday life (such as a misorganised bookshelf) that cause frustration instead of laughter. Kant (1790) framed humour as an absurd property of an object or a situation. This moves away from set theory and into the philosophical consideration of what consists an absurd quality. The current view of humour as an incongruity was introduces by Schopenhauer [1883] (1969), who stated that humour requires a contrast between perception and representation. This implies a violation of expectation – our a priory formed representation contains features that are misaligned with what we experience. The recent evolution of this view is the *benign violation theory* (McGraw & Warren, 2010). Similar to Schopenhauer's view, a stimulus must violate our expectation. However, an additional criterion is added, requiring the context of the violation to be benign (non-harmful). If a person visits a bank, expecting an assistant, but being greeted with a clown instead, it may trigger a humorous response (violates an expectation in a benign way). In a similar situation happens, but instead of a clown, the person is greeted with a bank robber, it may be less of a laughing matter. McGraw adds that situations may be benign for multiple reasons, not only lesser severity. One way of making a context benign is adding time distance. This may be why we are able to laugh at dire situations that happened early in our lives – the fact that they have passed makes them benign enough.

Quantifying Humour

While not a theoretical paradigm per se, it is worth mentioning recent efforts in quantifying humour. The present thesis puts forth the idea that psychological states are quantifiable through observing language use. Even though the above theories of humour differ widely, all of them include a mental component of humour – be it through a complex attitude position of the superiority theories, or through a violation appraisal process of an incongruity theory. It is not too bold to suggest the perception of humour is therefore a psychological state. In effect, humour should be quantifiable through statistical methods applied to text.

Taking jokes as input, the JESTER algorithm (Gupta, Digiovanni, Narita, & Goldberg, 1999) represents a systematic approach to quantifying humour. The methodology uses principal component analysis, a clustering framework, to identify meaningful groups of related jokes. Then, using participant ratings of humour as a predicted variable, the algorithm learns to optimise which jokes a should be served to a participant, based on which jokes they found funny in previous trials. This results in a trained, humour recommendation system.

An alternative to a trained algorithm is an experimental paradigm. Mickes, Walker, Parris, Mankoff, & Christenfeld (2012), studied gender differences by having participants create humorous captions for

images. A primary analysis showed males tend to produce slightly funnier captions, as judged by both males and females. In a second study, participants were presented with the captions again, this time also showing the author's gender. The effect was even stronger this time, suggesting a bias towards relatively associating humour with males. This experimental design shows the possibility of generating humour data by having participants provide ratings for text. Similarly, it is possible to have participants rate a list of non-words on their humour, resulting with a list of how funny non-words are (Westbury, Shaoul, Moroschan, & Ramscar, 2016).

In this Chapter, the thesis develops a single-word database of word humour. It is important to note that unlike most experimental studies, there is little discussion on humour in single words. The non-word database by Westbury et al. (2016) is a rare exception, rather than an indication of a larger body of research. The reviewed theories have a similar property, mainly focusing on social contexts, events or situations, not on individual words. The recent *benign violation theory* assumes a violation of expectation – which begs the question whether an expectation can be violated by a single word. These ideas are unpacked in the following sections.

Introduction

The appreciation of humor is a fundamental, albeit mysterious, part of human cognition. We laugh at things like *Monty Python* and the work of Douglas Adams, but find topics like mass shootings and the holocaust off limits. Other topics, like sunsets and freedom, may lie somewhere in between. What makes one thing funnier than another? And what makes some topics inviolable in relation to humor? To help develop this research, we provide the first set of humor norms for a large collection of 4997 common words. The aim of providing this data is to help enrich the resources available for understanding the cognitive, developmental, and applied aspects of humor.

Humor has a long history of theoretical investigation. Darwin (1872) called humor "tickling the mind." Thomas Hobbes (1840) referred to it as a feeling of "sudden glory." These represent a selection from a long list of efforts to provide a theory of humor (reviewed in Keith-Spiegel, 1972; Hurley, Dennett, & Adams, 2011; Wyer & Collins 1992). These include biological theories—such as the Darwin-Hecker hypothesis that humor is a cognitive analogue of physical tickling (Fridlund & Loftis, 1990; Harris & Christenfeld, 1997); superiority theories, such as Hobbes notion of "sudden glory" over another individual or one's previous self (Hobbes, 1840); Release theories, such as that proposed by Spencer (1860) and later Freud (1928), that humor is a means of reducing excessive arousal; incongruity-resolution theories (Suls, 1972; Shults, 1976), perhaps first noted by Kant (1790), in his observation that "In everything that is to excite a lively convulsive laugh there must be something absurd," and later developed by Schopenhauer [1883] (1969), who suggested the "ludicrous" required a "contrast…between representation of perception and abstract representations." Still further theories have focused on the adaptive value of humor as an error correction mechanism and faulty logic detection system (Minsky, 1981), most recently and thoroughly developed by Hurley, Dennett, and Adams (2011). A similar version of this theory has been called the benign violation theory (McGraw & Warren, 2010) which suggests a person must realize the stimuli is incongruous with their expectations (violation), but also that this incongruity is not harmful given the context (benign).

The onslaught of theories aimed at understanding humor reflects our common experience that humor is a key ingredient in what it means to be a healthy human. It may even be uniquely human and, continuing the noble history validating intuition with Latin, Koestler (1964) referred to humans as *Homo ridens*, "laughing man" (see also Milner, 1972). Whether or not it is unique to humans, humor has well-documented influences on well-being and health, including self-concept, coping with stress, positive affect (Cann & Collete, 2014; Galloway & Cropley, 1999; Martin et al., 1993; Mora-Ripoll, 2011). Humor research also contains a wide body of literature concerned with understanding adult and child personality development (Martin, 1998; McGhee, 1971).

A large number of researchers also investigate the links between humour and gender (Abel, & Flick, 2012; Hay, 1995; Mickes, Walker, Parris, Mankoff, & Christenfeld, 2012). One possible evolutionary

hypothesis is that humour is that humor plays a role in male mating displays (McGee & Shevlin, 2009), and which is further supported by gender differences in response to humor in the brain (Azim, Mobbs, Jo, Menon, & Reiss, 2005; see also Goel, & Dolan, 2001).

In addition, cracking the riddle of what makes things funny has also been the motivation for a number computational algorithms designed to create humor, such as JAPE (Binsted, Pain, & Ritchie, 1997), STANDUP (Manurung et al., 2008), WISCRAIC (McKay, 2002), and HAHAcronym (Stock & Strapparava, 2003), as well as algorithms to detect and classify humor (Davidov, Tsur, & Rappoport, 2010; Mihalcea & Strapparava, 2005).

Much of the theory and empirical work briefly outlined above focuses on complete multi-word jokes, such as this zinger by Steven Wright: "I couldn't repair your brakes, so I made your horn louder." To this end, a number of studies have taken to rating and creating databases of jokes in an effort to allow researcher to disaggregate the various mechanisms that make them work (e.g., Goldberg, Roeder, Gupta, & Perkins, 2001; Wicker, Thorelli, Barron III, & Willis, 1981). A few studies have looked at single non-words (Westbury, Shaoul, Moroschan, & Ramscar, 2016), suggesting the absurdness of a non-word results in associated humor. None, to our knowledge, have focused on single English words.

The database we present here offers a basis for studying humor in perhaps a highly rudimentary "fruit fly" version, at the level of a single word. If single words have reliable humor ratings, they provide humor in miniature, allowing us to investigate humor in relation to the many existing lexical norms. These include some which are directly related to past theories—such as Freud's (1928) arousal theory—and others which offer at least some insight into processing and expectation, such as reaction times and frequency.

The collection of the humor norms follows on previous work demonstrating the advantage of crowdsourcing in psychological norm development: for example, Warriner, Kuperman, & Brysbaert (2013) have collected valence, arousal and dominance ratings for 13,915 English words; Brysbaert, Warriner, and Kuperman (2014) collected concreteness ratings for nearly 40,000 English words have been collected; and Kuperman, Stadthagen-Gonzalez, and Brysbasert (2012), collected age of acquisition ratings for 30,000 English words (Kuperman, Stadthagen-Gonzalez, & Brysbaert, 2012). These were in turn based on the value of previous norms, such as the Affective Norms for English, provided by Bradley & Lang (1999

).[10] Still other ratings norms have investigated different word properties, which have provided the basis for further investigating their influence on cognition, such as imageability and familiarity

[10] Dutch (Moors et al., 2013), Finnish (Söderholm, Häyry, Laine, & Karrasch, 2013), French (Monnier, & Syssau, 2014), German (Kanske, & Kotz, 2010), Italian (Montefinese, Ambrosini, Fairfield, & Mammarella, 2014), Portuguese (Soares, Comesaña, Pinheiro, Simões, & Frade, 2012) and Spanish (Redondo, Fraga, Padrón, & Comesaña, 2007).

(Stadthagen-Gonzalez & Davis, 2006), pleasantness (Bellezza, Greenwald, & Banaji, 1986) and meaningfulness (Paivio, Yuille, & Madigan, 1968).

These normative datasets have proven highly fruitful. For illustration, Dodds et al. (2015) used valence ratings to assess a universal positivity bias. Alhothali & Hoey (2015) used valence ratings to predicts readers responses to news articles. And Hills and colleagues (Hills & Adelman, 2015; Hills, Adelman, & Noguchi, 2016) used concreteness, age of acquisition, and lexical reaction times to evaluate the changing history of American English over the last two hundred years.

Here, we provide a large dataset of single-word humor ratings along with the demographics of the raters. The list of rated words was formed from the intersection of overlapping previous non-humor word norms, allowing us to provide an analysis of how word-level humor relates to valence, arousal, word length, concreteness, word processing time and word frequency. Secondly, breaking down our dataset by demographics, we provide a separation of humor by gender.

Methods

Stimuli

The words in the norms are chosen from the intersection of the valence, arousal, and dominance norms (Warriner, Kuperman, & Brysbaert, 2013), age of acquisition norms (Kuperman, Stadthagen-Gonzalez, & Brysbaert, 2012), lexical decision norms (Keuleers, Lacey, Rastle, & Brysbaert, 2012) and frequency norms (Van Heuven, Mandera, Keuleers, & Brysbaert, 2014). This resulted in 7775 words, from which the final word list of 5000 words was randomly sampled. This reduction in list size increases the number of raters exposed to a single word, given a fixed number of participants.

Participants provide information in response to demographics questions (age, gender, language, country growing up, and education), the humor rating of calibrator words, and the humor rating of 200 words randomly sampled from the pool of 5000 words. The calibrator words are a list of 11 words that spanned the range of humor rating in a pilot study (with 150 participants and 500 randomly sampled words). The calibrator words are presented in Table 8. Following previous studies (e.g., Brysbaert et al., 2014; Warriner et al., 2013), participants saw the calibrator words first, with the aim to show the participant the range of the humor scale and increase the reliability of subsequent ratings. The calibrator words were followed by the random sample of 200 words. The word sample was different for each participant, generated in real time when the participant opened the online questionnaire.

TABLE 8

Calibrator words presented to participants

Word	Mean Humor Rating (Pilot)
drought	1.13
deathbed	1.55
cleaver	1.69
oxide	1.8
rainstorm	1.91
lurch	2
maroon	2.08
driftwood	2.23
cleat	2.4
walnut	2.67
turd	3.78

Table 8. *Calibrator words presented to participants.*

Data Collection and Participants

Participants were recruited using Amazon Mechanical Turk. Any registered member of Amazon Mechanical Turk was allowed to participate, with the requirement of fully completing the study (partial data was not recorded), and only doing the study once. Upon accepting the study, the participant was redirected to a website which delivered the instructions and words for rating. The introduction read as follows:

> *You will rate how you felt while reading each word. There will be approximately 200 words. The rating scale ranges from 1 (humorless = not funny at all) to 5 (humorous = most funny). At one extreme of the scale, you find the word dull or unfunny; in that case, you should give the word a rating of 1. At the other extreme of the scale, you feel the word is amusing or likely to be associated with humorous thought or language (for example, it is absurd, amusing, hilarious, playful, silly, whimsical, or laughable); in this case, you should give the word a rating of 5. The scale also allows you to describe intermediate of humor; if you feel the word is neutral (neither humorous nor humorless), select the middle of the scale (rating 3).*
>
> *After you fill out some basic information about yourself, a word list will appear. Simply click the most accurate humor rating for each word. Once you finish rating the words, we will ask*

you a couple of questions about the way you use humor. Please work at a rapid pace and don't spend too much time thinking about each word. Rather, make your ratings based on your first and immediate reaction as you read each word.

The introduction was followed by the list of 211 words, each word having five buttons presented just below it, numbered from 1 to 5, with the extremes labeled “humorless” (1) and “humorous” (5). The first 11 words were the calibrator words. The combination of the remaining 200 words was different across participants. After selecting a rating for a word, the word disappeared from the list. Upon rating all words, the participant could press the “Submit” button. The participant was then presented with a debrief page and redirected back to Amazon. Each participant was paid $1. The study took approximately 15 minutes to complete, including reading the instructions and the debrief page.

Results

Data Trimming

The data was presented to 950 participants. 102 participants were removed due to incomplete submissions, errors in the data and improperly submitting their responses. 5 participants were removed due to low variability of their responses (the standard deviation of their humor ratings, on a 1-5 scale, was smaller than 0.2, indicating they chose roughly the same value for all words). 22 participants were removed because they indicated their primary language was not English. The final data consisted of 821 participants. The raw data had 173231 individual data points, referring to a single rating of a single word. Ratings were collected for 4997 words, with each word rated by at least 15 participants. The average number of participants rating a word was 33 (M = 32.93, SD = 5.64, n = 4986). The 11 calibrators were rated by all 821 participants.

Demographics

Participants identified as female in 478 cases (58%), as male in 341 cases (42%), and 2 participants chose not to answer (<1%). The mean age of participants was 35 years (M = 35.37, SD = 11.74, n = 821), ranging from 18 to 78 years old. Table 9 presents the education demographics.

TABLE 9

Education distribution of the participants

Education Type	Number of Participants	% of Participants
Elementary School	5	<1%
Some High School	5	<1%
High School Diploma	235	29%
Undergraduate Degree	434	53%
Postgraduate Degree	126	15%
Higher than Postgraduate Degree	16	2%

Table 9. *Education distribution of the participants*

Humor Ratings

For each word, all of the humor ratings were summed and divided by the number of participants rating the word. This resulted in a Mean Humor Rating (MHR) of each word. The split-half reliability of the individual ratings was 0.64, similar to that of previously collected for arousal ratings (0.69 in Warriner, Kuperman, & Brysbaert, 2013). The MHR for each word is provided in the supplementary material. MHR were also computed for each gender separately. Table 10 shows the descriptive statistics of MHR across all participants.

TABLE 10

Descriptive statistics of mean humor ratings (MHR)

Statistic	Value
Mean	2.41
Standard deviation	0.44
Median	2.34
Minimum	1.18
Maximum	4.32
Skew	0.78
Kurtosis	0.87

Table 10. *Descriptive statistics of mean humor ratings (MHR)*

The MHR distribution was positively skewed, indicating that more words are rated as *humorless* than *humorous*. This is in contrast to previously collected valence norms, which tend to be negatively skewed. People have an intrinsic positive bias for valence, interpreting most words as positive (Warinner et al., 2013; Dodds et al., 2015). For humor, the opposite is true – most words are rated closer to humorless than humorous. The shape of the MHR distribution is shown in Figure 7.

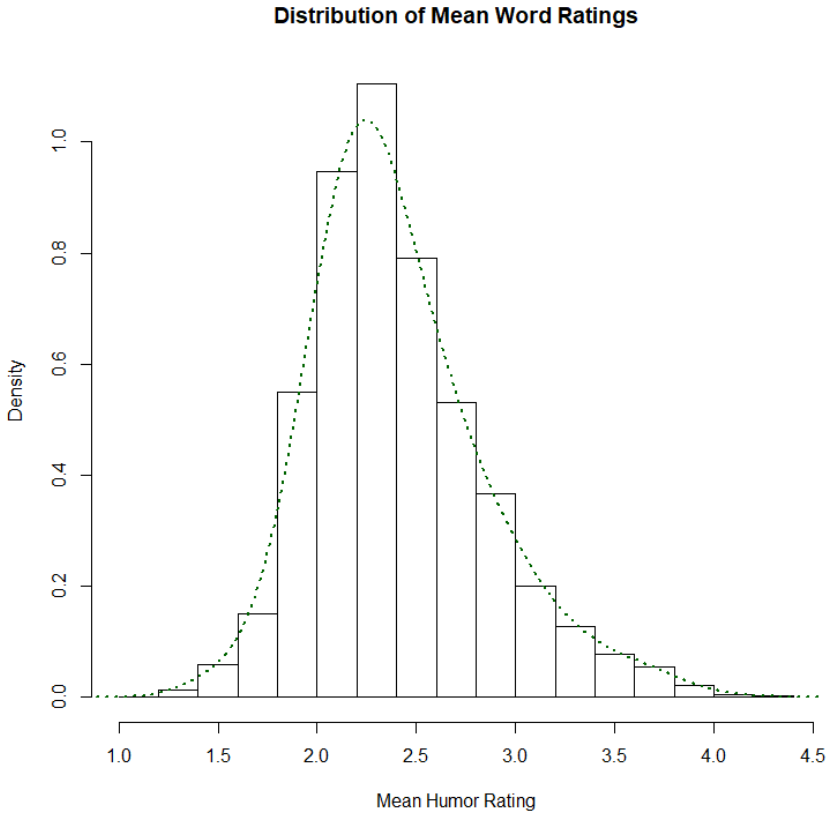


Figure 7. Distribution of mean humor ratings (MHR) across 4997 English words.

TABLE 11

Words with the most extreme mean humor ratings

Positive Extreme	Negative Extreme
booty (4.32)	rape (1.18)
tit (4.25)	torture (1.26)
booby (4.13)	torment (1.3)
hooter (4.13)	gunshot (1.31)
nitwit (4.03)	death (1.32)
twit (4)	nightmare (1.33)
waddle (4)	war (1.33)
tinkle (3.94)	trauma (1.35)
bebop (3.93)	rapist (1.37)
egghead (3.92)	distrust (1.38)
ass (3.92)	deathbed (1.39)
twerp (3.92)	pain (1.39)

Table 11. *Words with the most extreme mean humor ratings.*

The distribution of MHR covers a range of 3.14 units. The most humorless word in the norms is "rape" (1.18) and the most humorous word is "booty" (4.32). Table 11 lists the twelve most extreme words at end of the distribution.

The calibrator words were presented to all 821 participants. Their distributions were calculated individually. To provide an indication of how words across the distribution are rated by all of the participants, Figure 8 presents the distributions for each of the calibrator words separately.

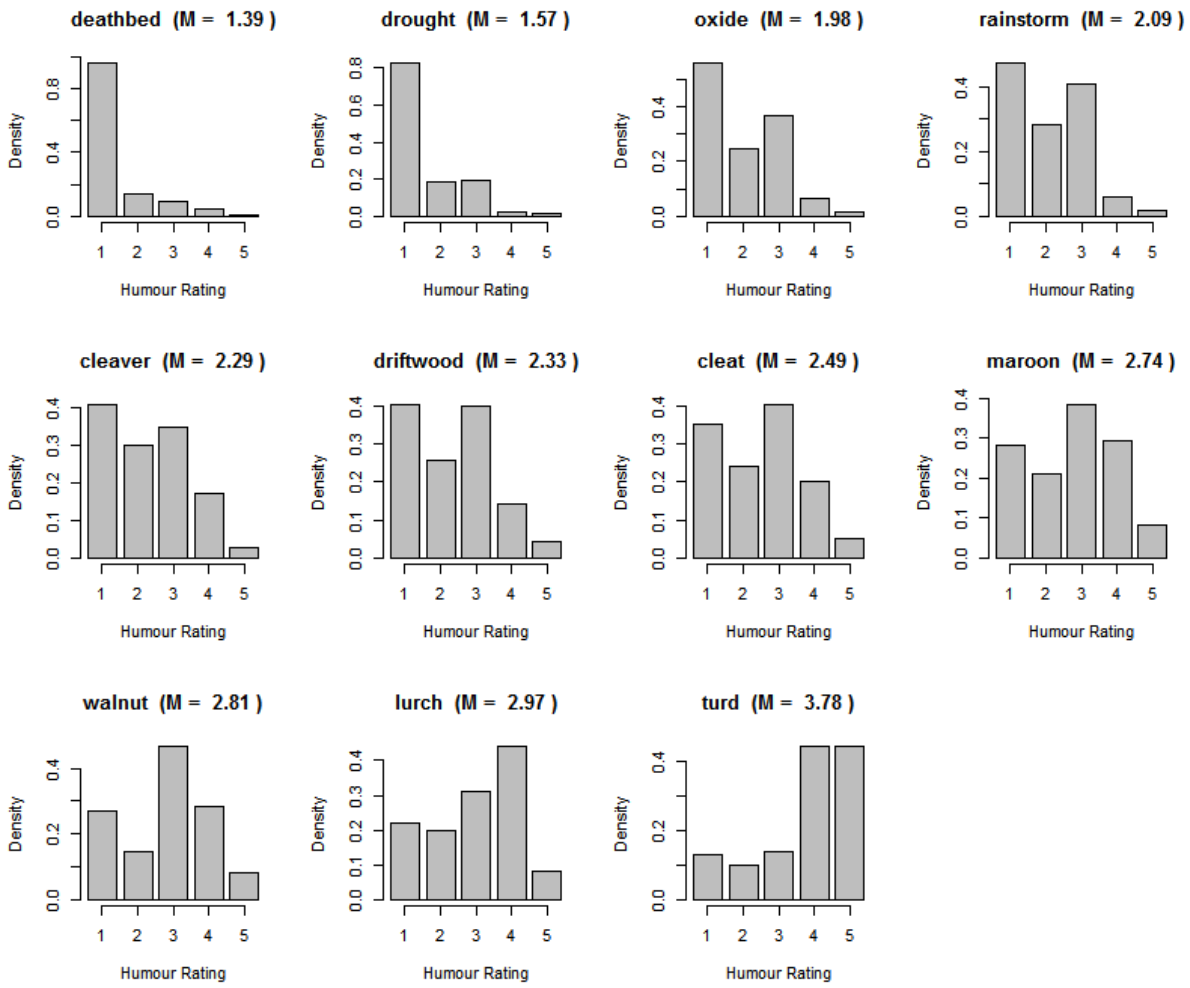


Figure 8. Distribution of ratings over all participants for each of the 11 calibrator words.

Correlations

Table 12 shows the correlations of the MHR with other linguistic metrics available from existing norms. The strongest correlation is with frequency (British National Corpus), with less frequent words rated as more humorous. Words less frequent in SUBTLEX (movie subtitles) were also rated as more humorous. Words that are associated with longer reaction times in lexical decision tasks were also rated as more humorous.

TABLE 12

Correlations between eleven lexical measures.

Variable	1	2	3	4	5	6	7	8	9	10	11
1 Mean Humor Rating											
2 Age of Acquisition	0.08										
3 Word Length	-0.06	0.26									
4 Frequency (BNC)	-0.42	-0.40	-0.26								
5 Frequency (SUBTLEX)	-0.30	-0.57	-0.33	0.78							
6 Lexicality RT	0.27	0.56	0.30	-0.71	-0.73						
7 Valence	0.09	-0.29	0.03	0.23	0.19	-0.22					
8 Arousal	0.05	0.07	0.05	-0.06	0.07	-0.04	-0.16				
9 Dominance	0.01	-0.22	0.00	0.23	0.18	-0.20	0.61	-0.15			
10 Concreteness	0.12	-0.35	-0.05	-0.11	0.00	-0.05	0.11	-0.18	0.05		
11 Frequency (ANC)	-0.40	-0.38	-0.27	0.88	0.78	-0.68	0.22	0.00	0.22	-0.15	

Table 12. Correlations between eleven lexical measures.

Gender Differences

The mean ratings for both genders were identical (M_M = 2.41, SD_M = 0.51; M_F = 2.41, SD_F= 0.48; males and females rate the same number of words, n=4997). There are, however, gender differences in the ratings of individual words. Table 13 shows words with the biggest disagreement between genders.

TABLE 13

Words with the largest differences between male and female ratings

Words Rated More Humorous by Males	Words Rated More Humorous by Females
bondage (1.55)	giggle (-1.92)
birthmark (1.47)	beast (-1.61)
orgy (1.47)	circus (-1.6)
brand (1.46)	grand (-1.5)
chauffeur (1.35)	juju (-1.45)
doze (1.34)	humbug (-1.38)
buzzard (1.34)	slicker (-1.38)
czar (1.30)	sweat (-1.38)
weld (1.29)	ennui (-1.36)
prod (1.27)	holder (-1.35)
corn (1.27)	momma (-1.35)
raccoon (1.26)	sod (-1.35)

Table 13. Words with the largest differences between male and female ratings. *Note*. Numbers in brackets are the difference in ratings between genders. They are computed as $MHR_M - MHR_F$: a positive value means the word is rated as more humorous by males, a negative values means it was rated as more humorous by females.

The words of biggest disagreement are in essence the outliers of an $MHR_M - MHR_F$ plot, where MHR_M is the mean humor rating of male participants and MHR_F is the mean humor rating of female participants. This relationship is shown in Figure 9. On a surface level, these seem to support the hypothesis that men use more humour including profanity and sexual themes (Mickes et al., 2013).

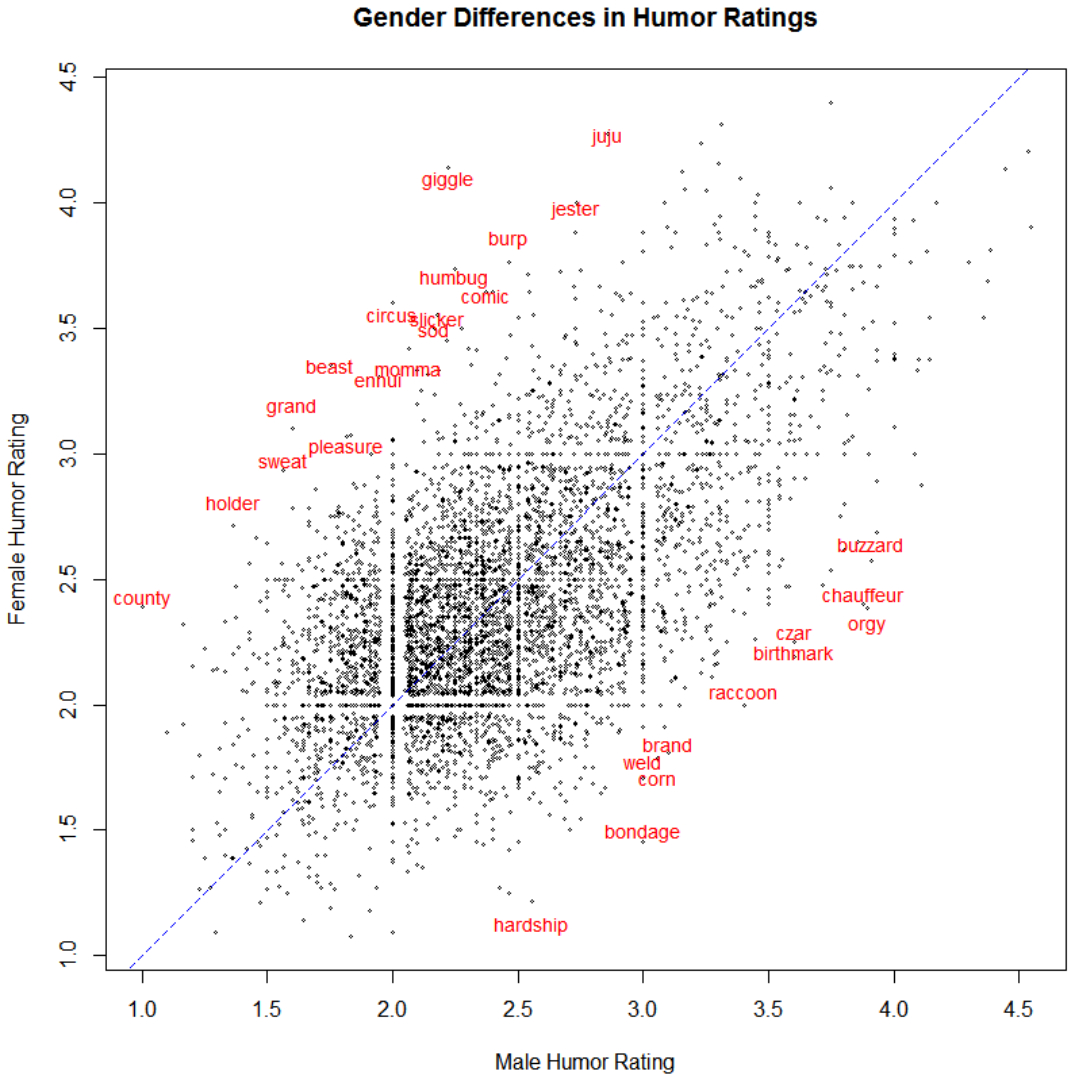


Figure 9. A plot of male and female MHR for each of the 4997 words. Words having an absolute gender difference larger than 1.25 are labelled in red. The blue line has an equation of y=x. Slight jittering was applied to the word labels to improve readability.

Discussion

Using the ready availability of large online data collection, the present study has created a database of single-word humor ratings. The statistical analyses show that people view words as humorous to a varying extent, with a skew towards seeing the majority of words as humorless. The appraisal of single-word humor can be reliably measured across participants, similarly to that of arousal.

The present study shows examples of analyses that can be carried out with the humor dataset. Specifically, it is possible to show correlational relationships between humor rating and other variables (i.e. frequency and lexical reaction times). This approach may in turn inform us on how the underlying

mechanisms of humor work, or at the very least, where to look in the future. Additionally, it is possible to investigate gender differences in humor appraisal.

In relation to the humour theory reviewed, the correlation analyses show no relationship between humour and arousal, which would be expected if the energy release theories were valid. Highly arousing words should cause more energy generation within the participants, and therefore higher ratings of humour. The inverse correlation with frequency may support a surprise or a violation theory. Specifically, low frequency words are more surprising to the participants, and simultaneously violating the distributional expectations. Participants should implicitly expect to see common words – seeing a low frequency word may violate that expectation. Such violation would also be benign (however we do not have situations of non-benign violations for comparison).

Besides the above mentioned examples, we identify three fields of interest for future research. First, using existing databases of jokes (e.g., Goldberg, Roeder, Gupta, & Perkins, 2001), the humor ratings make it possible to explore the relationship between the appraisal of humor on the joke level and on the single-word level. Second, the humor norms provide a resource for machine learning methods to establish the best predictors of word level humor, which can later be evaluated in psychological experiments. Third, individual ratings of words in relation to the norms can provide a basis for understanding individual differences in humor styles (e.g., Martin, Puhlik-Doris, Larsen, Gray, & Weir, 2003). Finally, like previous ratings, the humor norms may offer new insights into text analysis and the creation of psychological stimuli.

Chapter 4: The Macroscope, A Tool to Examining the Historical Structure of Language

Declaration and Statement of Contribution

I do not take majority credit for the writing in this chapter. The Macroscope was a collaborative effort between Li Ying and me, who is a PhD student at the University of Warwick. We worked together extensively to discuss the approach to the problem, come up with solutions and ultimately produce an analytical platform. The project is far more ambitious than we initially planned. Each of us contributed a substantial amount of work.

Our joint responsibilities included overseeing the project, discussing the best method of delivery, planning how to house our data structures, discussing how our figures would look, and generally learning from one another to make this a reality. We met regularly and collaborated freely, flexibly sharing the tasks as they come along. This naturally turned into an arrangement where I was tasked with creating the backend of the software platform (essentially coding the ecosystem we work in) and Li was tasked with analysing the data, producing figures and establishing the theory.

Li Ying was responsible (*I either have a minor contribution or no contribution*) for: getting the raw data, processing the big-data structures into tangible matrixes, coming up with analyses, doing the literature review, writing the main analytical code in Python, producing results and generating result figures.

I was responsible (*I have a majority contribution and Li advised on aspects of these tasks*) for: creating the server environment to house our data, coding a client-facing website to deliver our results, client-server communication, server operation and hosting, migrating Li's code and analyses into an online environment, and developing a cluster processing paradigm to enable fast processing of data.

The project represents different academic contributions to both of us. Li is very much interested in the evolution of language, the ability to capture shifts in meaning over time. The Macroscope makes substantial contributions in this area. To me, the Macroscope is a tool for quantifying psychological properties of words. It is able to visualise the psychological properties associated with a provided query using networks, vector spaces and normative datasets. In this sense, the collaborative project is a key component of both of our theses, albeit for different reasons.

The chapter was written collaboratively. During the time of its writing, I was coding the server platform to be released cooccurring with the journal submission. We agreed I'd only have a minority input on the text so that the software could be ready in time. I am closely familiar with the content, but do not take credit for the writing. The majority credit for writing the chapter should go to Li Ying.

Project Overview

As part of the EPSRC studentship funding my degree, I committed to creating a text analysis platform. The proposal was flexible – the platform should be easy to use and have academic impact, but the implementation was not set in stone. Inventing a platform like this from scratch requires a good amount of foresight. I decided to focus on learning as much as possible first, taking on and completing other projects before thinking about the platform. That way it could be robust and theoretically sound.

Over the course of my studies, the idea for the platform went through many iterations. It started as an R package, taking a list of words as an input and returning their valence properties. While interesting, it was not accessible to the general public – one would still have to understand R to use this package. The promise of moving this to an R-free environment was intriguing. I created a simple website that had a cooccurrence database and the valence/arousal norms. It took a single word as input and returned three figures: the most cooccurring words with the target, and the evolution of the target's valence and arousal over history. The website had great responses from people who tested it, reassuring me this an exciting direction for the text analysis toolkit.

During this time, Li Ying was working on a separate project, mapping the historical perception of risk. His work was impressive both theoretically and visually, generating eye-catching figures on the shift of meaning of risk. My website wasn't nearly as intricate as the figures generated for his publication, but it worked in real time. We realised our goals are near identical – we both wanted to have a tool for visualising the psychological properties of words.

The Macroscope was born out of my simple website and Li's understanding of the historical evolution of language. Over the final year of my PhD studies, we collaborated closely to make it as robust as possible. The outcome was exciting for both of us, as it exceeded our expectations. The functionality of the Macroscope is well beyond the proposed text analysis toolkit.

The Macroscope is an interactive client-server solution to providing language analyses. It takes queries from the client, sends them to our server, the query server redistributes the analytical duties to our workers (analytical servers), the queries get analysed and the data propagates back to the client's website. The website then displays the figures interactively to the user.

It is not a one-time product, but rather a continually improving service. At the time of writing this thesis, the Macroscope is available online in its first stable version. This corresponds to the functionality outlined in our publication (i.e. the chapter below).

It delivers on the promise of providing language insights to both researchers and the general public. The interface is sufficiently easy to remove a barrier to entry, while the analyses are sophisticated enough to be of use to language researchers.

Introduction

Lowenthal (2015) once wrote that "The past is a foreign country: They do things differently there" (p. 5). Understanding why they did those things and what they were thinking when they did them is partly about history, but it is also falls under the umbrella of historical psychology. A number of recent accounts have documented apparent historical changes in the way people thought in the past. These accounts follow in the footsteps of well-documented historical changes that have taken place even in the last several centuries, for example, in the diffusion of print materials and the industrial revolution's disarming of the Malthusian trap, releasing large parts of the world's population from hand-to-mouth economies (Clark, 2008; Eisenstein, 1980). These changes have led to numerous claims explaining the rising spectre of risk in society (Beck, 1992), the whittling away of violent behavior by the civilizing process (Pinker, 2011), urbanization's empowering of individuality and materialism (Greenfield, 2013), and the evolution of American English in response to information crowding (Hills & Adelman, 2015). The growing consensus appears to be that historical data represents a fertile ground for rolling our contemporary understanding of psychology back into the past.

The most common approach to studying historical beliefs and attitudes is what historians and literary critiques call *close reading*. A close read involves a human reader, who reads over original texts, attending to individual words and sentences. Scaling this approach to the volume of historical text currently available to make broad quantitative generalizations at the scale of hundreds of years is effectively impossible. A person reading 50,000 words a day would require 22,000 years to close read the text currently available in Google Ngrams book corpus. Over the past several decades, however, cognitive and language scientists have developed computational tools for *distant reading*, where researchers use algorithms to extract meaning from billions of words of text. These have been used to study properties of word recognition (Jones & Mewhort, 2007), the structure of memory (Hills, Jones, & Todd, 2012), the relationship between natural language production and individual differences (Pennebaker & Stone, 2003), changing frequencies of word usage across individual lifespans (Le, Lancashire, Hirst, & Jokel, 2011), and changes in word use over hundreds of years (Michel et al., 2011). In doing so, this progression has moved language analysis from synchronic investigation of single words to diachronic investigations of texts across cultural time, all of which can take place in the lifetime of a single researcher (or even in an afternoon).

The goal of the present work is to introduce a tool that adds an additional layer of structural depth to quantitative historical analysis, allowing researchers to zoom in and out on words--specifically, their semantics, and the associations they maintained in historical language. We call this tool the Macroscope, after the device in Piers Anthony's (1974) book by the same name which could zoom in and out on the cultural history of other alien civilizations. The key conceptual assumption upon which the Macroscope stands is that words provide information about the past and we can infer the meanings

of those words through the relations they keep with other words (e.g., Firth, 1957). Thus, meaning is derived through historical context, providing a new way of looking at semantic history. In what follows we describe the underlying computational machinery of the Macroscope and provide several case studies that demonstrate the Macroscope's utility for understanding historical language.

Method

The Macroscope

The Macroscope is a user interface consisting of a client-server interaction. The server, built in Node.js, handles user queries and analyses them in real time using Python. The data is then visualised on the client's website. It can be found at: http://www.macroscope.tech.

The Macroscope takes as input specific words of interest from the user, examines these in relation to a language corpus provided by the Macroscope, and outputs a range historical indicators about changing semantics over time. Here we take semantics in the broadest possible sense (see below). Data for each historical indicator can be downloaded in .csv format to the user's computer. A representation of the online interface for the Macroscope is shown in Figure 10.

The details of the language corpora and computational algorithms are provided below.

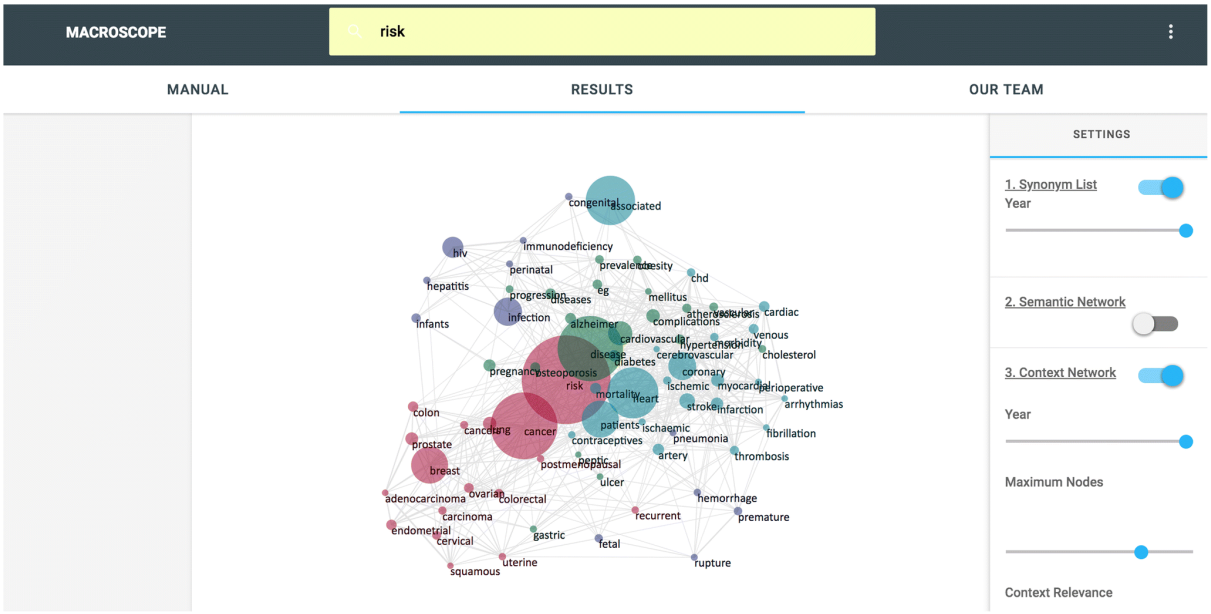


Figure 10. Screenshot of the Macroscope website. The search bar is on the top where users can input word of interest (state in the figure). The control panel on the right allows selecting specific analysis and manipulating parameters.

The language corpora

The first iteration of the Macroscope uses text from the English Google Ngram Book corpus (5-grams) (Michel et al., 2013). This will be supplemented with additional corpora (such as the Financial Times corpus and the Corpus of Historical American English) in forthcoming iterations, allowing users to compare data across multiple corpora. The Google Ngram Book corpus represents ~4% of all books published over the last several hundred years (Michel et al., 2013). Because the data representation is fairly sparse prior to 1800, we present data from 1800 to 2009 which contains approximately 155 billion words.

Frequency

Usage frequency is computed by dividing the number of instances of a word in a given year by the total number of words in the corpus in that year. For instance, in 1861, the word *slavery* appeared in the corpus 21,460 times, in 11,687 pages of 1,208 books. The corpus contains 386,434,758 words from 1861; thus the usage frequency of *slavery* in 1861 is 5.5x10-5. Users can input a search term into the search field and adjust various settings to capture and visualize the data of interest.

Co-occurrence matrix

To compute word properties from the words that a given word co-occurs with, the Macroscope relies on a co-occurrence. The Google Ngram data consists of a matrix using 5-gram data. The matrix records the number of times any two words co-occurred within a 5-gram over 209 years from 1800 to 2009. We include the top 50,000 most frequently used words across the 209 years, resulting in a 50,000 x 50,000 x 209 matrix. Each word in the co-occurrence matrix is represented as a vector of dimension 50,000 that stores its contextual information.

Sentiment and concreteness

Using the co-occurrence matrix, the Macroscope computes contextual sentiment (valence), arousal, and concreteness by taking the mean of the relevant ratings of all the words that co-occurred with a given word in a given year. We used the Warriner, Kuperman, and Brysbaert's (2013) norms to

retrieve contemporary valence and arousal ratings for each word, and the Brysbaert, Warriner and Kuperman's (2014) norm to retrieve contemporary concreteness ratings for each word. While the rating values may have likely changed over time, using contemporary ratings is the only viable option. Additionally, the user may easier interpret contemporary ratings – as they represent a valence space the user is familiar with.

Diachronic word embeddings

To find out which words are most semantically similar to each other and quantify their degree of similarity, we used distributional semantics, in which words are embedded in vector space according to their co-occurrence relationships (Bullinaria & Levy, 2007; Turney & Pantel, 2010). We constructed diachronic word embeddings for each year to allow comparisons across different years. This approach has been effectively demonstrated in a number of studies (Sagi et al., 2011; Xu & Kemp, 2015; Hamilton et al., 2016)**.** In our study, we constructed word embeddings as follows. First, vectors containing the number of times a given word co-occurred with all other words were directly obtained from the co-occurrence matrix described above. Second, we computed Positive Pointwise Mutual Information (PPMI) for each pair of words and constructed a PPMI matrix with entries given by

$$\mathrm{PPMI}(v_i, v_j) = \max(0, \log(\frac{P(v_i, v_j)}{P(v_i) \times P(v_j)}))$$

where vi, vj represents a pair of words from the corpus, and P(v) corresponds to the empirical probabilities of word co-occurrences within a sliding window size of 5 over the original text. As compared to a simple co-occurrence count, PPMI penalizes high-frequency words (i.e., of, the, and) that are used in the same context with a wide range of words, and favours words that frequently appear together but not with others (i.e., hong and kong). This is due to dividing the mutual co-occurrence by the individual frequencies of each word, in essence controlling for a possible inflated mutual-co-occurrence due to a high absolute frequency of one of the words. Forcing PPMI values to be above zero ensures that they remain finite and this has been shown to improve results (Bullinaria & Levy,

2007; Levy, Goldberg & Dagan, 2015). Lastly, we reduced the dimension of word embeddings to 300 using Singular Value Decomposition (SVD), comparing word similarities by computing cosine similarity of word embeddings. Using a lower dimensional space comes with multiple benefits. The dimensionality reduction acts as a form of regularization, reducing the influence of individual outlier words in the subsequent calculations. It prevents sparse matrices – the fact that most words never appear with majority of other words. And it reduces the data size considerably, allowing for real time data processing.

To validate that the word embeddings we trained on Google Ngram corpus accurately capture semantic relationships among words, we tested it on 200 multiple-choice synonym questions collected by Levy, Bullinaria, and McCormick (2017). Each question corresponds to a set of five words: the test word, followed by the correct synonym, followed by three incorrect choices. Due to some of the low frequency words (such as consommé and treacle) are not included in our analysis, we effectively tested on 183 synonym questions using word embeddings trained on aggregated data from 2000 to 2008. Our performance (89.5% correct) is comparable to word embeddings trained using 5 different algorithms by Levy and his colleagues (correct rate ranging from 86.5% to 92.0%).

Results

Quantifying Semantic and Contextual Change

The Macroscope provides researchers with the ability to examine two distinct but related aspects of linguistic change in individual words over historical time as shown in Figure 11 below. First, diachronic word embeddings computed from the co-occurrence matrix enable us to discover words that are semantically similar to a given word for a given year (i.e., the semantic or synonym structure surrounding a word). These semantically related words are referred to as *synonyms* for the remainder of this paper (top half of Figure 11). Second, the co-occurrence matrix provides information regarding the context of a given word at a given year. Words that co-occur with the target word are referred to as *context words* for the remainder of this paper (bottom half of Figure 11).

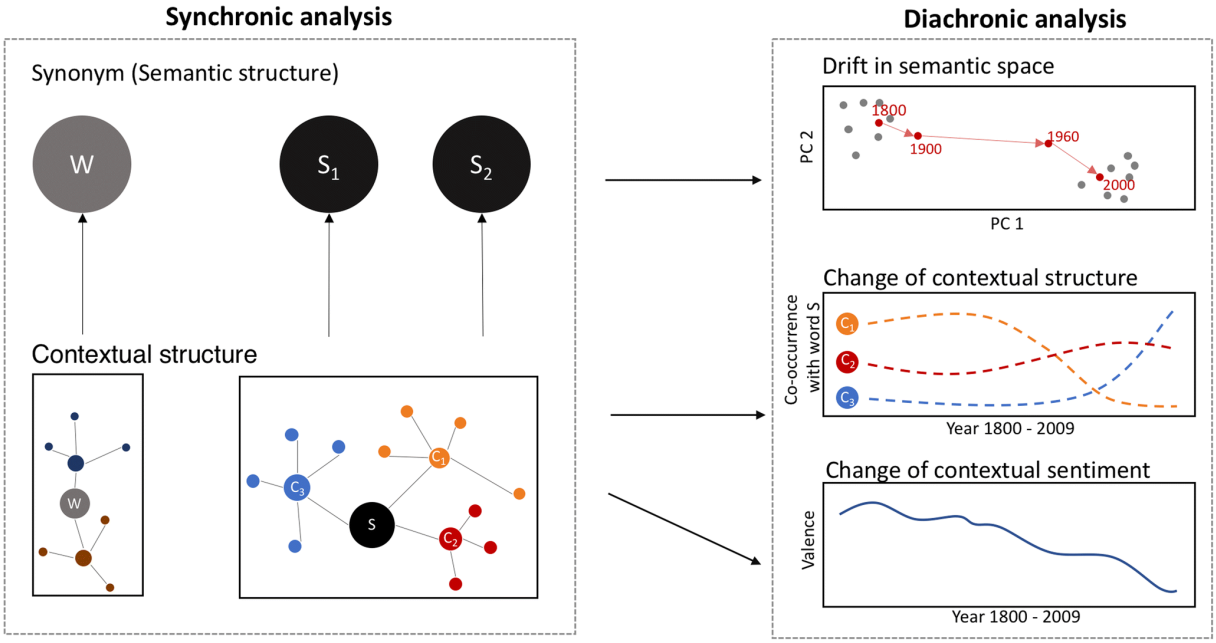


Figure 11. Conceptual framework summarizing the key features of the Macroscope. The Macroscope permits synchronic (left side) and diachronic (right side) analysis of the semantic/synonym (top) and contextual/co-occurrence (bottom) structure of words. *W* is the input word, S_1 and S_2 are identified synonyms. This is done by comparing the co-occurrence patterns of the proposed synonym *S*, looking at the most frequently co-occurring words C_1, C_2, C_3. In the semantic drift analysis, we visualise the best fitting position of the target word ($Word_i$), for their co-occurrence pattern using data in ten year increments. The best fitting positions (red dots) are then plotted into a single figure, in two dimensional space, using principal component analysis (PCA) – note that the axes (PC1, PC2) are not meaningfully interpretable.

On top of being able to "focus" the Macroscope on the semantics and contextual structure of an individual word in a particular year, the true power of the Macroscope is harnessed when the researcher "zooms" out to obtain a bird's eye view of changes in the semantic and contextual structure of words over historical time. Below we describe how the Macroscope can be used to examine the semantic (synonym) and contextual (co-occurrence) structure of individual words for a specific year (i.e., zooming in) and over historical time (i.e., zooming out). In the analyses described below, techniques from network analysis are employed to help with the interpretation and visualization of the synonym and co-occurrence structure of words. All analyses can be easily replicated using the Macroscope and the user can download the network graphs along with the data used to construct the graphs.

Synchronic semantic structure of words: Historical synonyms

How do we know what a word meant in the past? Using diachronic word embeddings, the Macroscope can quantify semantic similarity by computing the cosine distance of word embeddings for any pair of words. Therefore, a word's historical meaning can be inferred by finding its most semantically similar words in a given time period (i.e., *synonyms*).

Anxiety and depression are conceptualized as two distinct emotions by psychologists, yet often experienced by the general population as the same feeling (Barret, 2011). To examine how these concepts are represented in the written language and produced and read by people who do not necessarily have a psychology background, we used the Macroscope to identify synonyms of *anxiety*, *depression*, and *fear* using co-occurrence data from the year 2000 (see Table 14). *Anxiety* and *depression* share many synonyms that are associated with mental disorders. In contrast, *fear*, another commonly experienced negative emotion, appears to have different synonyms from *anxiety* and *depression*.

To better capture how these three emotion concepts are related to each other, the Macroscope provides a network graph representing the semantic similarity structure of their synonyms. The nodes shown in the network represent the top 5 synonyms to *fear*, *depression*, and *anxiety* as identified above, and the words *fear*, *depression*, and *anxiety* themselves. The edges between nodes are weighted by the strength of semantic similarity between word pairs (that is, cosine similarity between word embeddings). Edges that are greater than a threshold of 0.7 are shown in the network. The threshold of 0.7 is arbitrary and can be set by the user. Optimal visualisation settings vary on a word by word basis, influenced by the frequency of a word, its co-occurrence patterns, but also the intention of the researcher in how they intend to use the visualisation. If the synonyms of two words share a high degree of semantic similarity (i.e., they are connected to each other in the semantic network), this indicates that the two words are likely to be used in similar contexts and are semantically "close" to each other. Higher semantic similarity among the synonyms of two words offers an additional layer of depth to investigate how similar are the meanings of two words, even if the synonyms of two words were not necessarily the same. Though previous tools have provided quantitative information about word similarity (e.g., LSA from Landauer, Foltz, and Laham, 1998; BEAGLE from Jones & Mewhort, 2007), the present example demonstrates how the Macroscope provides and visualizes additional information about the broader semantic similarity structure of words via their synonyms. Figure 12a shows that the synonyms of *anxiety* and *depression* are also synonyms of each other but distinct from those of *fear*. Although psychologists treat anxiety and depression as two separate constructs, they appear to be used in semantically similar contexts in written language.

The same network approach used to represent concepts and their synonyms can also provide insights into the overlapping and distinctive components of two concepts. A similar analysis was conducted for the emotion words *fear*, *disgust*, and *anger*, 3 of the 6 basic emotions proposed to exist universally across cultures (Ekman, 1992). Results indicate that all three negative emotions intersect with some of each other's synonyms (see Table 14). Figure 12b shows that the concepts of *anger*, *fear*, and *disgust* share similar connections to words such as *dislike*, *resentment*, *indignation*, and *loathing*. However, each of these emotion concepts are also marked by their unique components that make them

distinctive from each other: *disgust* with *contempt*, *anger* with *rage*, and *fear* with *dread* and *apprehension*.

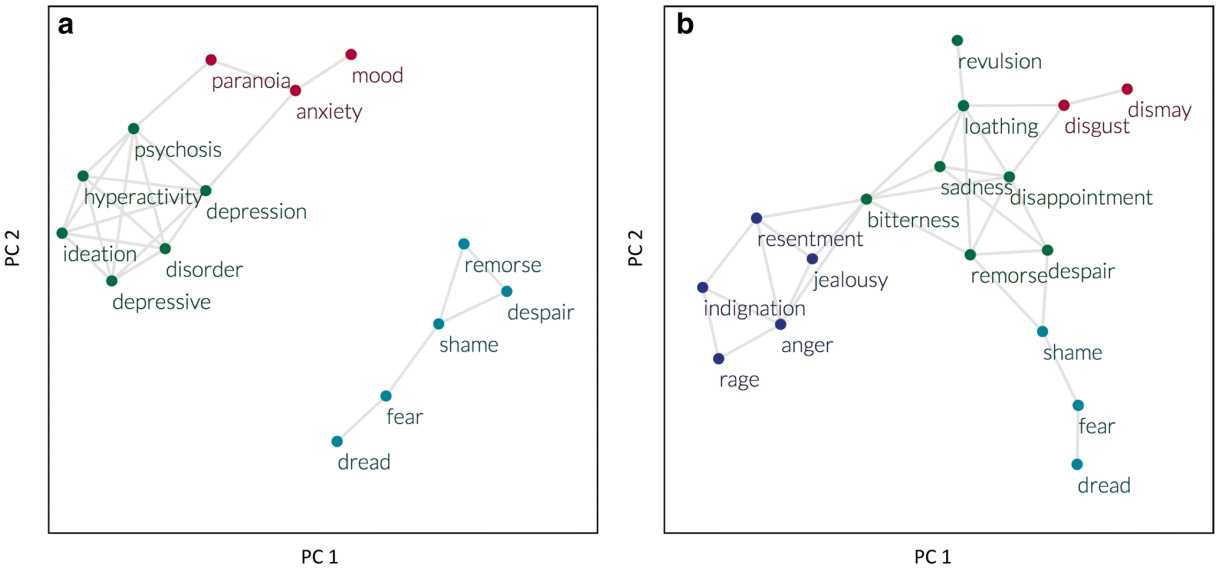


Figure 12. (a) Left: Synonym structure of anxiety, depression, and fear. (b) Right: Synonym structure of disgust, fear, and anger. The size of nodes is proportional to their usage frequency in the year 2000. The nodes represent the emotion concepts of interest and the top 5 most similar synonyms for each of the emotion concepts. The colors represent the community structure of nodes in the network and each community is represented with a different color. Community structure was detected by algorithm proposed by Blondel, Guillaume, Guillaume and Lefebvre (2008).

TABLE 14
The top 5 closest synonyms of depression, anxiety, fear, disgust, and anger provided by the Macroscope.

Target Word	Closest Synonym
Depression	Anxiety, Psychosis, Depressive, Hyperactivity, Disorder
Anxiety	Depression, Mood, Paranoia, Panic, Ideation
Fear	Dread, Shame, Anger, Remorse, Despair
Disgust	Loathing, Dismay, Disappointment, Revulsion, Sadness
Anger	Resentment, Bitterness, Jealousy, Rage, Indignation

Table 14. The top 5 closest synonyms of depression, anxiety, fear, disgust, and anger provided by the Macroscope.

Diachronic semantic structure of words: Semantic drift analysis

With large diachronic language data, the Macroscope is able to track how the semantics of individual words change over time. In the following examples we show how several words "move" along a path in a semantic space defined by their historical synonyms. A longer path moving from one point in the semantic space to another indicates significant changes in a word's semantic meaning over time. In contrast, a path that stays within a confined semantic space suggests that the word has retained its meaning over the time window examined.

Using the Macroscope the user can conduct a semantic drift analysis by inputting the word of interest, beginning and end time points (e.g., year 1850 and 2000), and intervening intervals (e.g., spaced by every 50 years). A semantic space was constructed for a target word by searching for its historical synonyms at the beginning time point (1850) and its modern synonyms at the end time point (2000). All synonyms' word embeddings are taken in their modern sense (2000). We also retrieved historical word embeddings of the target word for each time point of interest (i.e., 1900, 1950) and align their historical embeddings to its modern embedding using orthogonal procrustes (Schönemann, 1966), an algorithm to map one matrix to another of same shape. Lastly, these word embeddings were visualized on a two-dimensional space using principal component analysis (PCA). The PCA acts as a visualisation tool, allowing for the representation of a 300 dimension dataset (initially reduced by SVD) in a legible, two-dimensional image. All synonyms in this two-dimensional space are represented in their modern sense. Although in reality all word meanings fluctuate over time, we elected to adopt this approach in order to provide a clearer understanding of how changes in a word's historical meaning occur over time as benchmarked against its modern sense.

We used the Macroscope to examine the semantic change of three words that have been previously documented in historical linguistics (Jeffers & Lehiste, 1979). Figures 13a to 13c shows the semantic drift analysis of *broadcast, cell,* and *car* from the year 1850 to 2000 (with 50 year intervals). The idea to visualise word meaning change in this way was first pioneered by Kulkarni et al. (2015), who investigated these very same words. In 1850, the word *broadcast* referred to 'disperse upon ground by hand' and was closely associated with agricultural activity. In 2000, the word *broadcast* referred to radio and other media-related concepts. Our analysis shows that this change primarily took place between 1900 and 1950, a time period during which radio and television were invented (Figure 13a). *Cell* changed its dominant meaning from "a chamber in a prison" to a biological term and this change predominantly took place between 1850 and 1900 (Figure 13b). In 1850 the word *car* referred to a horse-driven wagon, but after the automobile was invented in 1885, it quickly acquired its modern sense. The semantic drift analysis shows that by 1900, *car* was no longer associated with a wagon (Figure 13c), but with modern transportation vehicles like *bus* and *truck*. In addition, we conducted a similar analysis for a word that was likely to be semantically stable over time: *happy*. The semantic drift analysis confirmed our intuitions: The word *happy* remained within the same semantic space over the past 150 years.

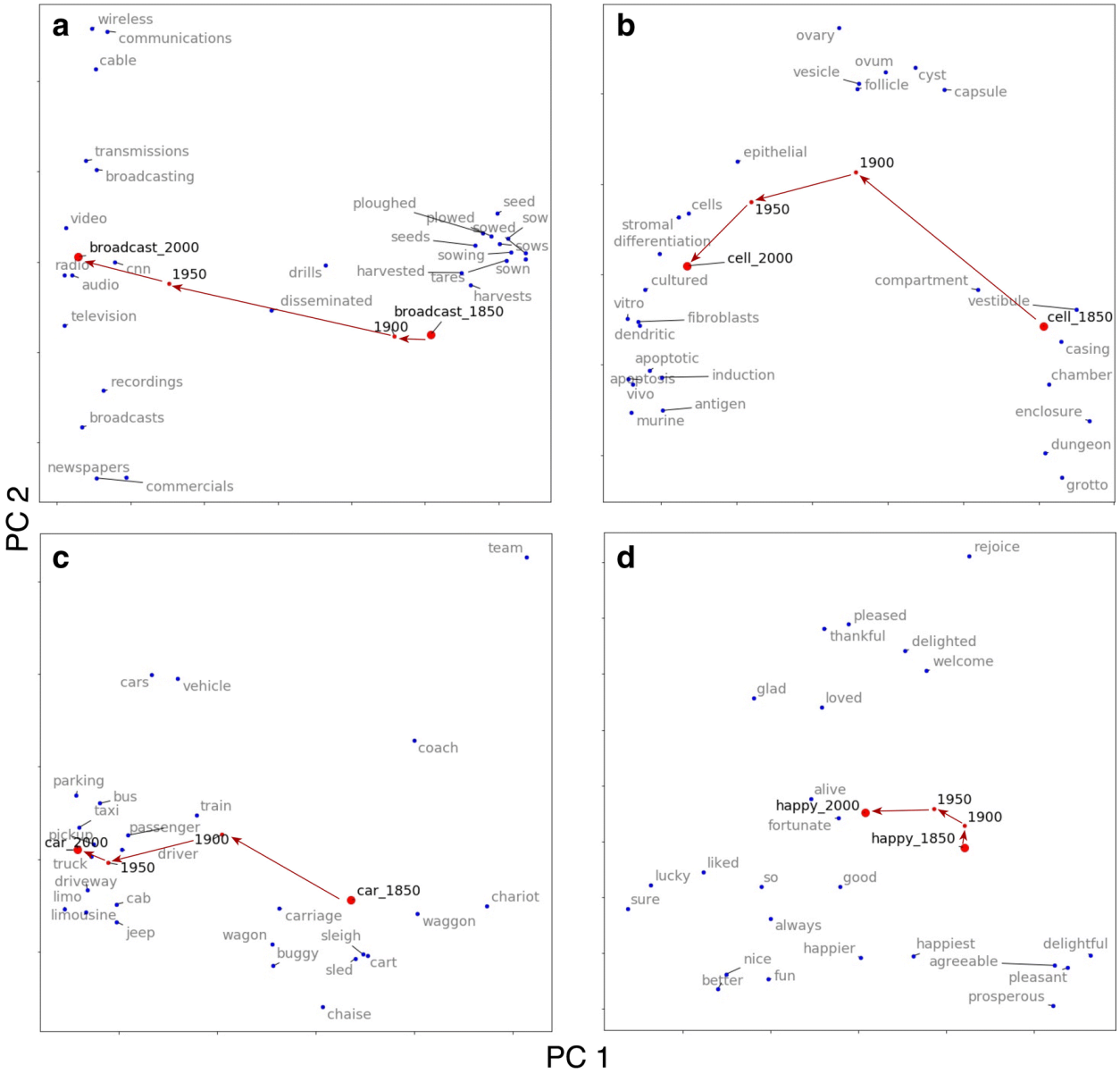


Figure 13. Semantic drift analysis for a) broadcast, b) cell, c) car, and d) happy from 1850 to 2000 with 50 year intervals. The blue dots indicate words that are semantically related to the target word of interest (i.e., its synonyms at the first and last time points). The path taken by the red dots indicate the "drift" in semantics of the target word from 1850 to 1900, from 1900 to 1950, and from 1950 to 2000. A comparison figure, using words from Hamilton et al. (2016).

Semantic drift analysis shown in Figure 13 offers a qualitative visualization on how word meanings changed over history, but it is not easy to quantitatively compare semantic stability between words (i.e. the semantic path travelled by *happy* relative to path travelled by *broadcast* from 1850 to 2000). Previous work has examined the properties of words that appear to show the highest degree of stability over historical time (e.g., Pagel, Atkinson, & Meade, 2007; Monaghan, 2014, Hamilton et al, 2016). Since the Macroscope provides information on diachonic changes in semantics, it can be used to quantify semantic stability of words as shown above in Figure 13,

$$\text{Stability}(w_i, t) = \text{cos\_sim}(w_i(T), w_i(T+t))$$

where wi(t) refers to the word embedding of word wi in year t. Semantic similarity ranges from 0 to 1. For example, the similarity of happy between year 1850 and 2000 is .74, much higher than the values for words that underwent greater semantic change, such as broadcast (.08), cell (.17), and car (.47). This allows researchers to examine potential forces that may have influenced semantic change. As a baseline for further examination, the Macroscope provides the semantic stability of a word in relation to its modern and historical word embeddings. Using this method, we retrieved the ten most stable words from 1800 to 2000. They are and, the, when, his, he, they, him, in, them, and a. A complete list of word stability between these two time points can be downloaded from the Macroscope. Pagel, Atkinson, & Meade (2007) first proposed the fact that the most stable words are all high frequency words, and our findings directly support this claim.

Synchronic contextual structure of words

Synonym analysis provides an accessible way to examine the semantic structure of words based on the conceptual assumption that words that are used in similar contexts are also semantically related to each other (e.g., Jones & Mewhort, 2007). On the other hand, identifying the particular context(s) in which a word was used can help us understand how polysemous words are used in their different senses across varying contexts, furthering our understanding of the relationship between the semantic and co-occurrence structure of words. For instance, it is possible for words to have a stable semantic/synonym structure but a varying co-occurrence structure over time. A concrete example can be seen in the word *woman*. Although the semantic meaning of the word *woman* has not changed much over past 200 years, in recent decades the word *woman* has been increasingly used it in the context of social issues surrounding feminism, gender discrimination, and abortion--contexts that were not commonly discussed during the 1800s.

The following co-occurrence networks of the words *monitor*, *option* and *gay* shows how the Macroscope can be used to understand the contextual structure of words. All networks were centred at the target word of interest. The context words, represented as nodes in the network, were selected based on their Positive Pointwise Mutual Information (PPMI) value with the target word. The edges were weighted by the PPMI values between each word pair. Next, nodes with low co-occurrence frequency with the target word and edges signalling low PPMI values were removed. Lastly, nodes with no edges (i.e., isolates) are removed. During the procedure, arbitrary thresholds for parameters must be specified in order to produce meaningful network graphs. The networks presented below were constructed using a PMI threshold of 3, and a minimum co-occurrence frequency of 200 times per 10 billion words. Communities are sub-groupings of nodes that are more likely to be connected to each other than to other nodes within the network. Community structures of the network are detected using a Louvain Method for community detection introduced by Blondel et al. (2008) based on a modularity optimization. Modularity is a scale value between -1 and 1 that measures the density of edges inside

communities to edges outside communities. This then uses an iterative process which redefines the community membership, merging communities until the modularity metric (a measure of the strength of the communities) is optimized.

Figure 14a shows the contextual network structure of *monitor* in the year 2000. Community detection analysis of the contextual network showed approximately 3 distinct contexts in which the word was used: as a computer device, in healthcare related settings, and a group of verbs that it often accompanies. From the contextual network structure of *monitor*, one can infer that it was used as a noun or a verb. As a noun, *monitor* is often referred to as a computer device; as a verb, *monitor* is often used in medical settings.

Figure 14b shows the contextual network structure of *nuclear* in the year 2000, which shows that the word *nuclear* is used in a number of distinct contexts: It can refer to a power source, physics phenomena, a technology known as nuclear magnetic resonance (NMR), or a weapon associated with some countries (*Soviet, Cuba, Korea*) but not other nuclear-armed states.

Figure 14e is an example of what the contextual structure of a polysemous word such as *option* looks like. Other than the conventional meaning of choosing among various possibilities, *option* also refers to a financial instrument. As Figure 14e shows, its contextual structure in the year 2000 is divided into two components. One involves its traditional sense, which incorporates the use of the option button on a keyboard. The other component consists of finance-related terms. It is important to note that such information would not be available if one only analysed the synonyms of *option* in the year 2000 (which are *options, cancel, default, item,* and *choose*), further highlighting how an analysis of a word's contextual structure can complement the analysis of a word's semantic structure.

As mentioned earlier, understanding the contextual usage of a concept can be useful to infer changes in the sociocultural environment. Figure 14c shows the context in which the word *gay* was used in the year 2000. It was not only associated with homosexuality, but also with a political movement associated with issues that extended beyond gay rights, such as feminism and abortion. Sexually transmitted diseases such as HIV (and it possible result AIDS) also appeared in this context, reflecting a social awareness of the association between homosexuality and the way that these diseases were transmitted among communities of gay men during the AIDS epidemic in the 1980s and 1990s. In contrast, 150 years ago, not only did all these associations not exist, the word *gay* simply did not refer to homosexuality. The contextual structure analysis suggests the word *gay* in 1850 was used in contexts involving fashionable clothes, cheerful mood, and pleasant colours (Figure 14d). These words were chosen specifically to follow Kulkarni et al. (2015), replicating their findings.

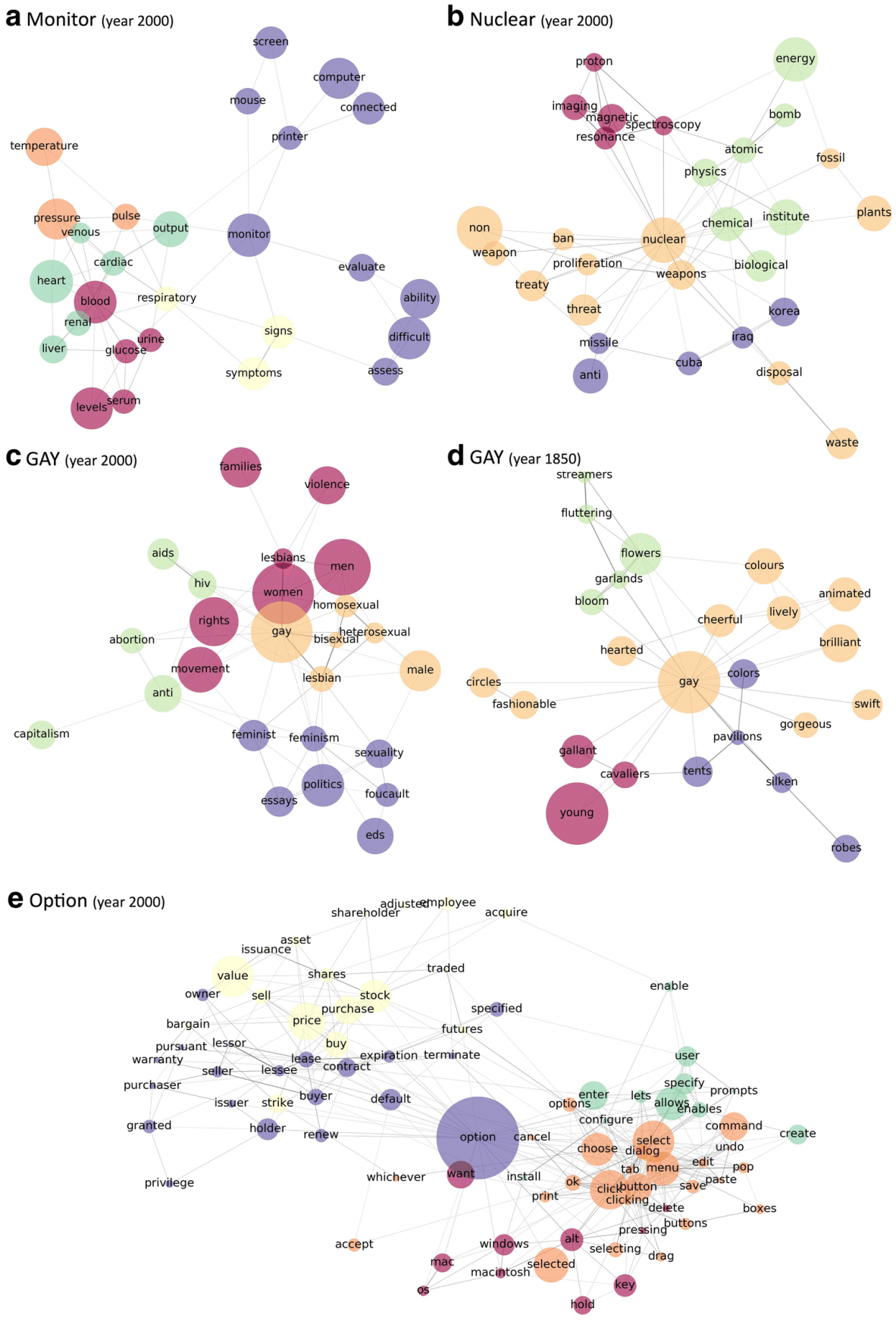


Figure 14. The contextual network structure of a) monitor, b) nuclear, c) gay in year 2000, d) gay in year 1850, and e) option. The nodes represent the context words that co-occurred with the target word in a given year. The

size of nodes is proportional to their usage frequency in a given year. The nodes were included in the networks if they had a PMI threshold greater than 3 with other words, and a minimum co-occurrence frequency of 200 times out of 1 billion words with the target word. The colors represent the community structure of nodes in the network and each community is represented with a different color.

Diachronic contextual structure of words

In addition to quantifying the contextual structure of words at a static point in time, the Macroscope allows users to quantify changes in the contextual structure of words diachronically. Figure 15 below shows how the frequency of co-occurrence of words co-occurring with *gay* and *nuclear* have changed between the years 1950 and 2000. Words with larger blue bars to the right (top of the y-axis) are words whose frequency of co-occurrence with the given word has increased the most from 1950 to 2000, whereas words with larger red bars to the left (bottom of the y-axis) are words whose frequency of co-occurrence with the given word has declined the most from 1950 to 2000. For instance, for the word *gay*, *lesbian* and *bisexual* increased the most in their frequency of co-occurrence whereas *happy* and *hearted* decreased the most in their frequency of co-occurrence. For the word *nuclear*, *weapons* and *magnetic* increased the most in their frequency of co-occurrence whereas *molecule* and *spin* decreased the most in their frequency of co-occurrence, reflecting the increased usage of nuclear as a weapon of destruction in recent years as compared to its scientific sense in the 1950s.

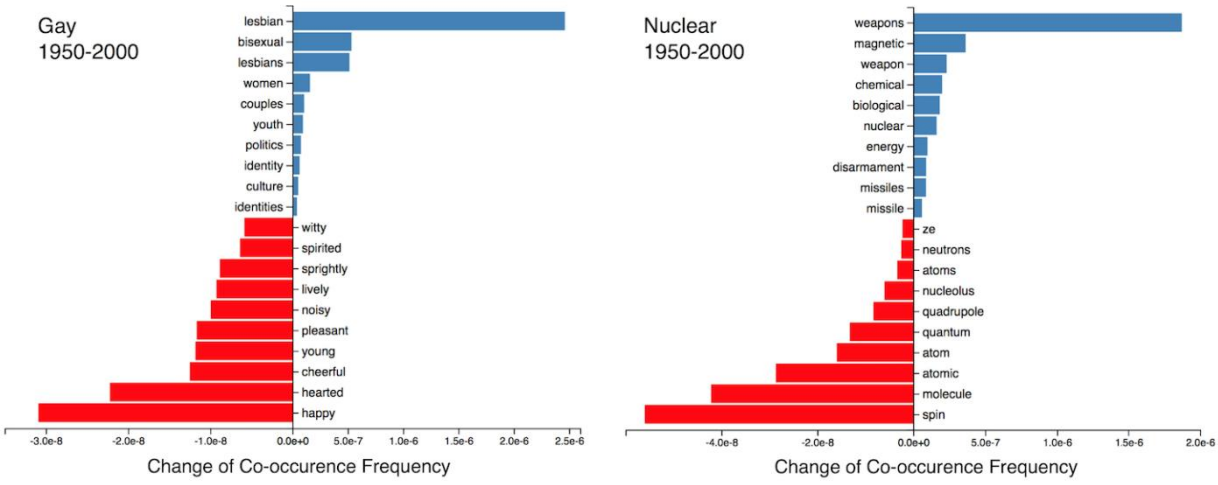


Figure 15. Words whose frequency of co-occurrence with gay and nuclear changed the most from 1950 to 2000. Words that increased the most in their frequency of co-occurrence with the target word from 1950 to 2000 are shown in blue near the top and words that decreased the most are shown in red near the bottom. The x-axes on the left and right side of the y-axis are scaled differently so that the y-axis is centered in the middle of the graph.

Although the previous analysis shows the largest changes in the frequency of co-occurring words between two time points, it is not completely clear to what extent a word has "lost" its old meaning. For instance it is possible for a word's old meaning to still be in use, albeit not as commonly used as

before. In addition, the previous analysis does not contain information regarding fine-grained changes in the frequency of co-occurring words during the time period in between the two specified time points.

One way to address these questions is to examine the extent to which a given word co-occurred with words found in its historical context. These context words can be obtained from the synchronic contextual structure analysis described earlier (see Figure 14). Users of the Macroscope can also enter words of particular interest to their research. The co-occurrence value in Figure 16 below (on the y-axis) was computed by summing the number of times the target word co-occurred with each word of interest (in this case, from its historical context identified in the contextual structure analysis in Figure 14) in each consecutive year after the historical reference year.

For instance, *gay* in 1850 co-occurred with words associated with cheerfulness, bright colors, and fashion (Figure 14c) and in 2000 co-occurred with words associated with homosexuality and sexually transmitted diseases (Figure 14d). The Macroscope can take these two lists of context words and compute their respective co-occurrence frequencies with the target word *gay* to capture how frequently its meaning in 1850 and its meaning in 2000 were used over the entire corpus (i.e., from 1800 to 2009). Figure 16 (left side) shows that how overall usage frequency of *gay* can be largely decomposed into two trends, with each corresponding to a different sense of *gay*. The co-occurrence between *gay* and its context words in the year 1850 declined quickly after 1900, whereas the co-occurrence between *gay* and its context words in the year 2000 emerged in the mid-1960s and increased dramatically after the 1980s. The pattern suggests that the old meaning of *gay* has been largely overwritten by its new emerging meaning.

Another example is the word *option* (shown on the right side of Figure 16). When looking at the contemporary contextual structure of *option* (Figure 14e), one can easily see that the word *option* refers to economic instruments: A *stock option* refers to stock warranted from a company to their employees as part of a remuneration package and a *lease option* refers to a real estate contract that gives the lessor an option to buy the property. A visual inspection of Figures 16d and 16f shows that a lease option probably existed in some form before the 19th century whereas a stock option was first introduced in the 1920s and the usage of this sense continued to grow in the 1980s.

By combining the synchronic contextual structure analysis of words with a diachronic analysis of co-occurrence frequency of context words with the target word, the Macroscope provides an accessible quantitative approach to track the association strength between a word and its various contextual structures over history, which could be used to investigate the evolution of word meanings or cultural change over time.

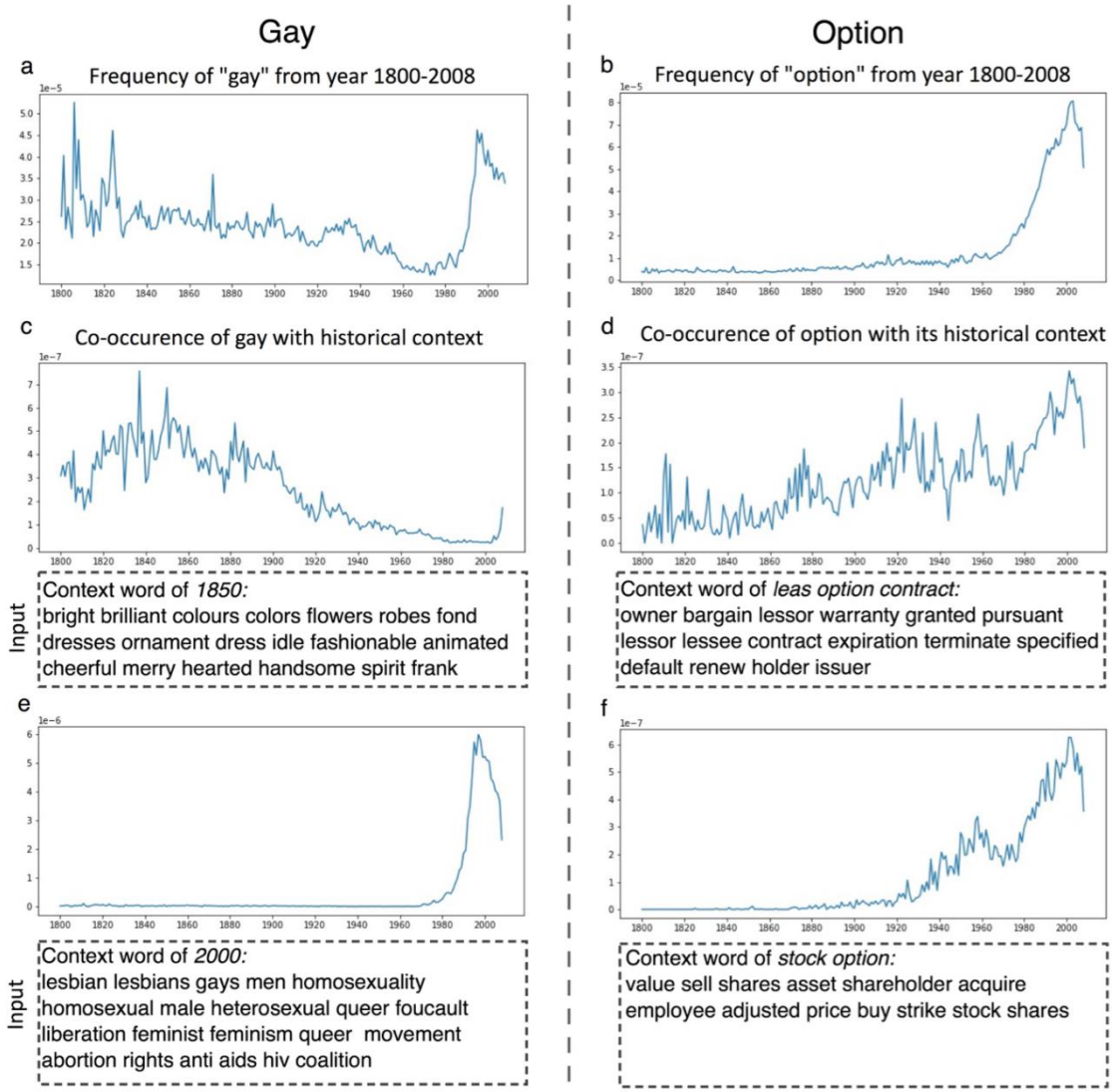


Figure 16. Co-occurrence frequency between the target word and its context words from 1850 and 2000. The context words were derived from the synchronic contextual structure analysis described earlier (see Figure 14 for examples). The co-occurrence frequency was computed by summing the number of times the target word co-occurred with each single word in the list of context words.

Diachronic changes in word sentiment

So far we have demonstrated how the Macroscope can be used to investigate the semantic and contextual structure of words at a specific point of time and across historical time. Below we show how the Macroscope can also be used to examine diachronic changes in word *sentiment* and how that information can be used to infer cultural changes due to urbanization and understanding the changing social perceptions of risk.

Example 1: Cultural changes due to urbanization. Greenfield (2013) analyzed the changing psychology of culture in the US as a consequence of urbanization by selecting two lists of words associated with urban and rural cultural values respectively and tracking their usage frequency over time. She found that words signaling urban values have proliferated in the US over the past century, along with a declining trend among words signaling rural values. Rural values are built around tighter social communities that expect cooperation, value tradition and place social obligation on individuals. Urban societies tend to be more individualistic, prioritising materialistic progression, a greater degree of free choice and focus on the immediate family unit. As a result, words like "give" and "oblige" are implied to be associated with rural environments, whereas words like "get" and "choose" appear in urban settings. The Macroscope can not only track the usage frequencies of these words over time, but also track the sentiment change of words over time. Here we use the Macroscope to extend Greenfield's results by analyzing the sentiment of words that co-occurred with words associated with urban and rural values over historical time.

The results reproduce Greenfield's analysis (see left side of Figure 17) showing that the frequency of *give* and *obliged* (rural values; in blue) decreased over time and the frequency of *get* and *choose* (urban values; in orange) increased over time. The Macroscope adds additional information by showing that the sentiment of *get* and *choose* increased at a faster rate as compared to the sentiment of *give* and *obliged* (see right side of Figure 17). The increasingly positive sentiment of urban value words compliments and extends Greenfield's argument because increasing usage of a word such as *get* and *choose* does not necessarily imply that urban values are viewed positively and are increasingly adopted by people. To provide a counterexample, if a word is used more frequently but has an increasingly negative sentiment (such as the word *gay* in the 1980s during the AIDS epidemic), this concept may instead be viewed as dangerous and unfavourable (see Figure 18).

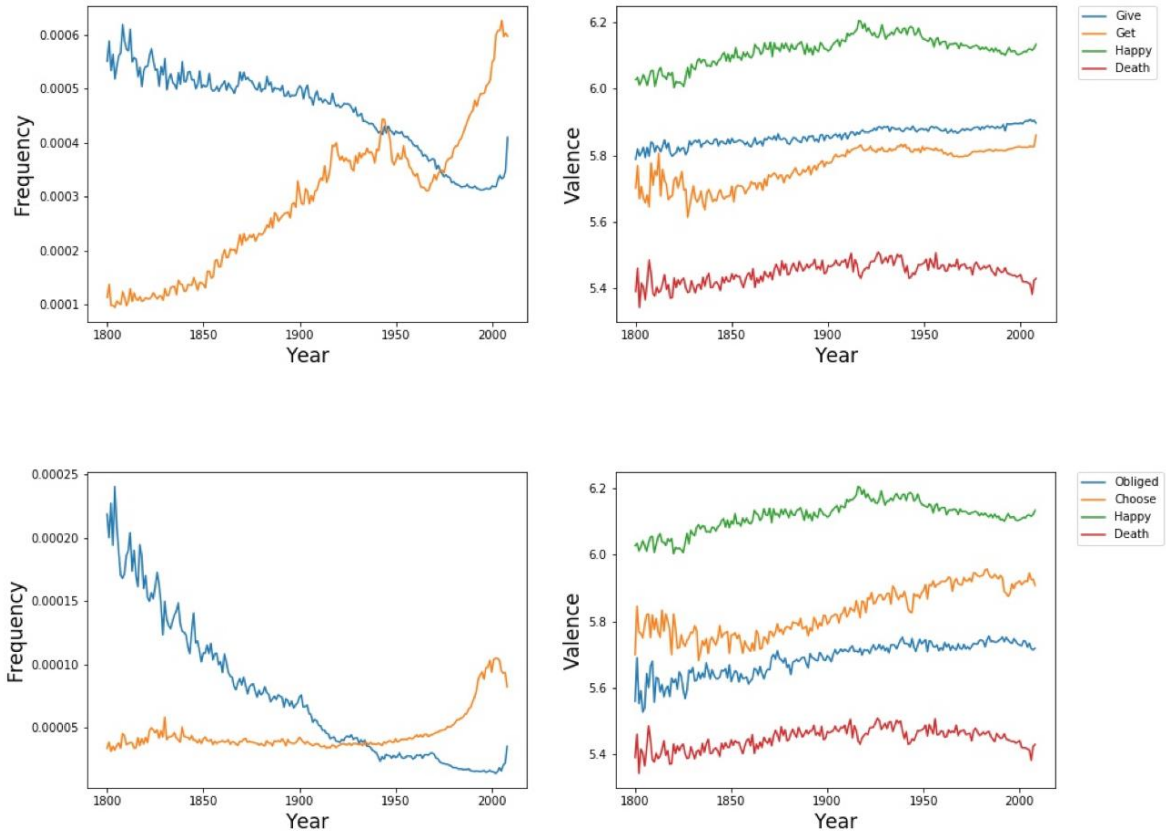


Figure 17. Frequency and valence from the Macroscope. The left side shows the usage frequencies for words associated with urban values (*get* and *choose* in orange) and words associated with rural values (*give* and *obliged* in blue) over historical time. The right graphs show the change in sentiment for the same words along with the change in sentiment for words such as *happy* and *death*, a high and a low valenced word whose sentiment is stable over time.

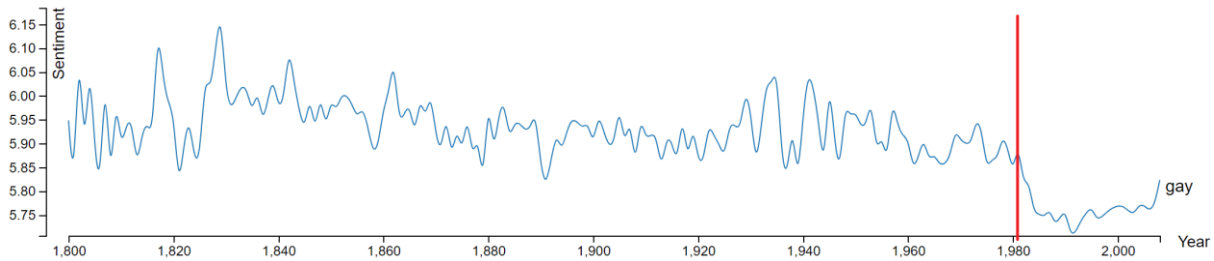


Figure 18. Valence of the word 'gay', taken from the Macroscope. The red line indicates the year of 1981, the sudden raise of public awareness of the HIV/AIDS pandemic in the United States.

Example 2: Changing social perceptions of risk. *Risk,* as defined by the Oxford English Dictionary, is synonym for *danger*, *hazard* and *fear*. However, sociologists and anthropologists have argued that *risk* represents more than just objective dangers or hazards in the real world. Instead, the notion of *risk* has been used to motivate social regulation and control or acts as a surrogate for other ideological concerns (Berk, 1992). In this example, we used the Macroscope to examine the relationships between *risk* and its synonyms over the past 200 years. Our results show that *risk* usage has experienced a rapid proliferation after 1950s compared to a stable usage of *hazard* and a declining usage of *danger* (Figure

19a). Correspondingly, the contextual sentiment of *danger* and *hazard* remained stable over time whereas the sentiment of *risk* became increasingly negative (Figure 19b). Output from the Macroscope (Figure 19c) shows how *risk* and its synonyms (i.e., *danger* and *hazard*) drift in semantic space between 1800 and 2000: *danger* and *hazard* have fairly limited semantic drift as compared to risk, which in the year 2000 was primarily associated with words related to medicine and health.

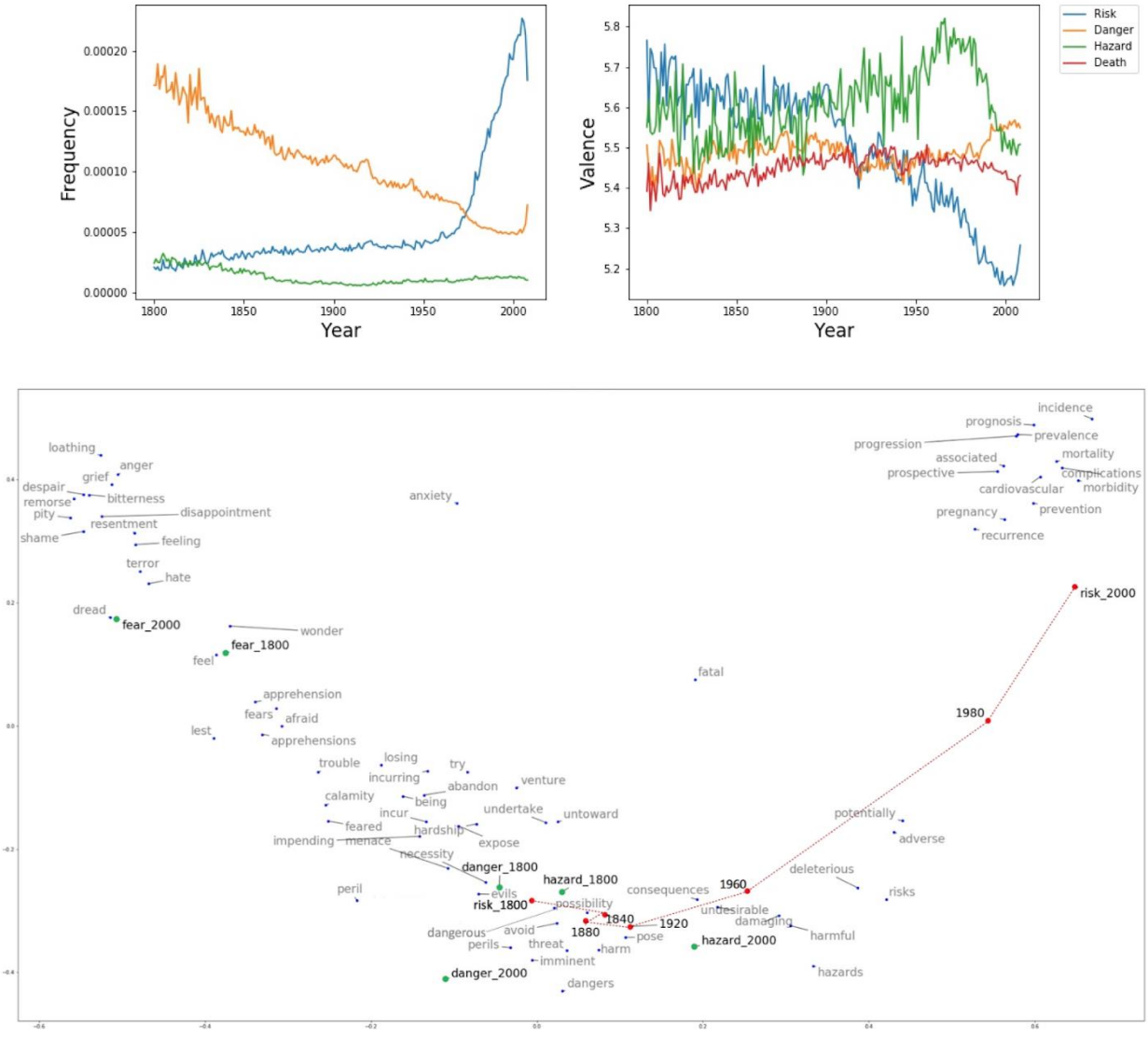


Figure 19. Frequencies and semantic drift. (a) Top left: Usage frequencies of *danger*, *hazard*, and *risk* over historical time. (b) Top right: Changes in the contextual sentiment of *risk*, *danger*, *hazard*, and *death* (*death* was selected as a benchmark) over historical time. (c) Bottom: Semantic drift of *danger*, *hazard*, and *risk* from 1800 to 2000. All figures were generated using the Macroscope.

General Discussion

Language has changed over historical time and that change is reflective of the kinds of things that people experienced and believed. The goal of the present paper is to introduce the features of the

Macroscope, an online algorithmic tool for zooming in and out on the semantic and contextual structure of words across historical time. A key conceptual assumption that the Macroscope neatly capitalizes is that words provide information about the past and we can infer the meanings of those words through the relations they keep with other words. To summarize, the Macroscope can provide (i) synchronic and diachronic analysis of a word's semantic structure (based on word embeddings derived from the co-occurrence matrix), (ii) synchronic and diachronic analysis of a word's contextual structure (based on word co-occurrences), and (iii) diachronic analysis of a word's sentiment.

In the numerous examples presented above, we provide evidence that the meanings of words can be derived through its historical context in language, and this provides researchers with a new way of looking at semantic history through historical language. Importantly, these analyses can be easily conducted by anyone via the Macroscope, which can accessed online.

The Macroscope offers numerous inroads to investigating many contemporary problems in psychology and historical linguistics. For example, what properties of words influence semantic shift (e.g., Zalizniak et al., 2012). How do word senses changed over time in relation to other word properties such as frequency, concreteness, and age of acquisition (e.g., Ferrer-i-Cancho & Vitevitch, 2017; Monaghan, 2014; Zipf, 1946)? Can we use nowcasting methods to 'backcast', examining how word usage reflects the influence of historical events (Lampos & Cristianini, 2012; Hills, Proto, & Sgroi, 2015)? What are additional structural properties of language that are associated the birth and death processes of words (Pagel et al., 2007; Vejdemo & Hörberg, 2016)? To what extent have words used in studies of age-related cognitive decline changed during the lifetime of individuals under study, for example in studies of memory and association (Hills, Mata, Wilke, & Samanez-Larkin, 2013; Ramscar, Hendrix, Shaoul, Milin, & Baayen, 2014)? We feel this is the tip of a large iceberg of potential questions.

Historical studies of any kind are limited in their generality by the artifacts that survive, who originally produced them, and who they were produced for. Studies of historical language are no different (see Hills & Adelman, 2015). Thus, the Macroscope is naturally limited in what it can see. As far as we know, there are no spoken language corpora, which means that individuals who could not write will not be reflected (probably ever) in historical language analysis. Historical texts may have also focused on different topics over time and therefore may not offer usage patterns that reflect common topical environments. Better understanding these patterns and their consequences for language is part of the question we hope the Macroscope can answer. For example, Dubossarsky, De Deyne, and Hills (2017) showed that free association networks changed non-linearly across the lifespan, between the ages of 8 and 80. This is mostly likely due to both developmental changes associated with factors underlying human cognition and changes in the lexical environment since roughly the 1920s. What language corpora best reflects this changing population? It is difficult to say. But studies of historical

language corpora nonetheless offer inroads into understanding what language structure can explain in the absence of additional assumptions. In forthcoming iterations of the Macroscope, additional corpora will be included to allow researchers to address specific question about generality.

To conclude, the language people use over historical time has been a primary source of understanding people's past beliefs and attitudes. The Macroscope brings quantitative approaches to a broader range of researchers interested in understanding historical psychology through the lens of language, enabling them to test and develop hypotheses about specific patterns of word usage and its semantics across history. In other words, the Macroscope is a passport to visit the foreign country of the past.

Conclusion

Summary

The idea to use high dimensional representations of words goes back to Osgood et al. (1957). Since then, the psychological properties paradigm has been used in a broad range of applied and theoretical fields. Quantifying the links between words and psychological concepts is not an unusual research design. This thesis highlighted two easy-to-use approaches to quantifying the psychological frames of words – networks science and vector spaces. In doing so, it demonstrates the feasibility of the quantitative approach to word meaning.

The thesis also provides novel insights into the fields of shape bias research, humour and diachronic semantic change. For shape bias, the work proposes a new distinctiveness-based paradigm. The advantage of this approach (over exclusively theoretical accounts) is that it offers a clear and testable hypothesis – that more distinctive words are learned earlier. This may serve as a path for revisiting the feature biases in early word learning, assessing future experimental studies using new criteria. For humour, the thesis offers equally as impactful additions. Until now, it was not possible to assess the humour of single words. If we consider larger pieces of text as collections of words, the ability to quantify the humour of individual elements paves the way to a mathematical model of humour of larger lexical units. The study on humour also shows that it indeed is a quantifiable concept even at the single word level – a novel finding contributing to the literature. Lastly, the Macroscope offers a platform for studying diachronic changes of word meaning. Analyses of semantic distributions generally require a reasonable effort in preparing the datasets. Running the analysis may be equally as time consuming, especially for large samples. The Macroscope abstracts this away from the end user, reducing the research process that used to take our lab months to an analysis that takes seconds. This should hopefully enable academics to spend more time thinking about the implications, and less time deploying their analyses.

Areas of Improvement

In spite of the thesis being successfully in addressing the goals it set out to investigate, there are areas of future improvement. These may either be pursued as follow up studies, or should at least be considered and discussed.

Google nGram Validity

Even though the Google nGrams are frequently used in computational linguistics, more and more researchers are questioning their validity. Namely, there is concern regarding the lack of balancing of the types of included publications and the overall mystery surrounding what publications are exactly

included (Pechenick, Danforth, & Dodds, 2015). The paradigm of psychological frames employs produced text as an insight into the psychological states of authors. It is tempting to claim the Google Books corpus represents the United States population as a whole. This is unlikely to be the case – not only because the types of publications in it are not balanced, but also because it 'only' represents 6% of the printed literature. This is a substantial number, but it is not representative of all the published literature at all. A possible remedy is including other corpora in the analyses. We are already planning to implement COHA in the Macroscope. Because the Macroscope is coded in an abstract fashion, it is possible to seamlessly interchange the underlying data sets, as long as they take on the form of a time series co-occurrence matrix.

Access to Written Language

A similar concern is regarding access to written literature in historical times. The ability to write is commonplace today. Anyone is able to publish any type of text online with little to no effort and even print publication is easily accessible. This would definitely not be the case in the 19th century. Any large-scale corpus of language should be interpreted with care. It is misleading to assume a large corpus size ensures representativeness of a population. The analyses conducted are always representative of the population having access to written language. This results in a skew towards the intellectual class in history, and a skew towards tech savvy individuals in the recently collected corpora.

Semantic Norm Size

The McRae feature norms used in Chapter 2 only cover a small cross section of language. This limits the viability of possible analyses. It was apparent that our investigation of distinctiveness was running into power issues due to low sample sizes, especially once we start to subset and cluster the semantic frames. The results are valid and should still be reliable, but there were directions we could not follow and questions we could not ask. Chapter 4 uses a co-occurrence based algorithm of inflating norms, essentially 'guessing' what the rating would be based on an implied psychological frame through a similarity of co-occurrence patterns between two words. This is an interesting premise and one that can help mediate the issue of low norm sizes for the time being. The ideal solution would be simply collecting larger pools of norms. Our lab is currently experimenting with a new data collection method that uses ranks instead of precise values. This method seems to have a higher inter-rater reliability, but most importantly, it dramatically decreases the number of required participants. The number of participants is reduced approximately 4-fold. This may increase the future normative data set size by an incredible margin, providing rank data is sufficient.

Future Directions

Over the course of the four-year degree, the published chapters have been presented at multiple conferences, as well as internally to the peers and staff. These presentations have always been invaluable, leading to new ideas and directions for the projects. There must have been over fifty follow up projects suggested over the years, each worth consideration. Here I briefly cover the most interesting ones.

A meta-analysis of distinctivenss

The shape bias literature is riddled with inconsistensies. Opposing results from different laboratories are common, and there is no shortage of discussion and theories either supporting or refuting the effect. An interesting idea suggested to me is this: because the field has not been focusing on general distinctiveness, it may be the case that some stimuli is simply more distinctive on the shape category, resulting in shape bias. Laboratories that do not find a shape bias may be more normalised in terms of the feature distinctiveness across all the perceptual categories. A meta-analysis of the general distinctiveness would help answer this question. This always seemed an incredibly promising idea that was never pursued due to new projects cropping up.

A predictive model of scaled humour

Single-word humour is really only the foundation for humour as a whole. Scaling up the study to larger lexical structures would be both insightful from a theoretical perspective and helpful for broader application. Cynthia Siew, Thomas Hills and I are currently working on a predictive model of word pair humour. We have an optimised model, which has been even validated using an independent sample of participants. The is incredibly strong predictive power between the humour rank of word pairs assumed by the model and the actual ranks from the validation sample (around 0.8).We are hoping to publish the results shortly, moving the humour norms in a new direction.

Connectionist models with an explanation

Vector spaces are incredibly useful when it comes to precisely quantifying a difference between two concepts. This is due to the in-built dimensionality of the vector paradigm. However, connectionist models, such as semantic networks, do not share this property. They are dimension-free by definition. This lack of dimension is not an issue for most simple descriptive observations – we can easily count the number of nodes, connections and even precisely observe information flow. However, as soon as we start using large scale, complex networks for data prediction (as is currently popular with propagating neural networks), the lack of dimensionality starts being an issue. Scientists are able to create incredibly powerful predictive models and trained machines, far surpassing humans in terms of data classification, complex decision-making or task mastery where reaction times play a role.

However, due to the lack of a clear dimensionality of the network paradigm, we are generally unsure why the models are performing the way they are. The networks are trained, they perform incredibly well – their structure and the learned data flow patterns yield exceptional outputs. But it is incredibly difficult to explain what features of the input contribute to the output, especially in a way that is informative to a human. Rethinking the network science paradigm to mitigate this downside would be a huge leap for data science as a whole. I am aware that researchers in the private sector are actively working on this question, and the potential is intriguing.

Closing Remarks

The research into the psychological frames of words is fascinating to me. At its core, it represents a mathematical way of quantifying human psychology. Considering people are generally abysmal at introspection – as is proved by behavioural research into cognitive biases – the possibility of having an objective look into one's psyche is exciting. On a personal level, research has always been about enabling others, not just about advancing the knowledge for knowledge's sake. With psychological frames, I see a future where people use this measurement to gain better introspection into their own psychological states, empowering individuals to make better, healthier decisions that help them reach their goals effectively.

References

Abel, M. H., & Flick, J. (2012). Mediation and moderation in ratings of hostile jokes by men and women. *Humor, 25,* 41–58.

Adelman, J. S., Brown, G. D., & Quesada, J. F. (2006). Contextual diversity, not word frequency, determines word-naming and lexical decision times. *Psychological Science, 17*(9), 814-823.

Agarwal, A., Xie, B., Vovsha, I., Rambow, O., & Passonneau, R. (2011). Sentiment analysis of twitter data. In *Proceedings of the Workshop on Language in Social Media (LSM 2011)* (pp. 30-38).

Albert, R., Albert, I., & Nakarado, G. L. (2004). Structural vulnerability of the North American power grid. *Physical Review E, 69*(2), 025103.

Alhothali, A., & Hoey, J. (2015). Good news or bad news: Using affect control theory to analyze readers' reaction towards news articles. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 1548–1558). Denver, CO: The Association for Computational Linguistics.

Allan, K. (Ed.). (2013). *The Oxford handbook of the history of linguistics*. OUP Oxford.

Anderson, R. C., & Freebody, P. (1979). *Vocabulary Knowledge. Technical Report No. 136.* Cambridge, MA: Bolt, Beranek and Newman, Inc.

Anthony, P. (1974) *The macroscope*. Sphere.

Arbesman, S., Strogatz, S. H., & Vitevitch, M. S. (2010). The structure of phonological networks across multiple languages. *International Journal of Bifurcation and Chaos, 20,* 679–685.

Bååth, R. (2010). ChildFreq: An Online Tool to Explore Word Frequencies in Child Language. [Publisher information missing]. Retrieved from http://lup.lub.lu.se/search/record/1776715/file/1776717.pdf

Baker, C. F., Fillmore, C. J., & Lowe, J. B. (1998, August). The berkeley framenet project. In *Proceedings of the 17th international conference on Computational linguistics-Volume 1* (pp. 86-90). Association for Computational Linguistics.

Barabási, A. L., & Albert, R. (1999). Emergence of scaling in random networks. *Science, 286*(5439), 509-512.

Barrett, L. F. (2017). *How emotions are made: The secret life of the brain*. Houghton Mifflin Harcour.

Beattie, J. (1779). *Essays: on poetry and music, as they affect the mind: on laughter, and ludicrous composition; on the usefulness of classical learning*. London: Printed for E. and C. Dilly and W. Creech, Edinburgh.

Beck, U. (1992). *Risk society: Towards a new modernity* (Vol. 17). Sage.

Beckage, N. M., & Colunga, E. (2016). Language networks as models of cognition: Understanding cognition through language. In A. Mehler, A. Lücking, S. Banisch, P. Blanchard, & B. Job (Eds.), *Towards a Theoretical Framework for Analyzing Complex Linguistic Networks* (pp. 3-28). Springer-Verlag Berlin Heidelberg.

Beckage, N., Smith, L., & Hills, T. (2011). Small worlds and semantic network growth in typical and late talkers. *PloS ONE*, *6*(5), e19348.

Bellezza, F. S., Greenwald, A. G., & Banaji, M. R. (1986). Words high and low in pleasantness as rated by male and female college stu- dents. *Behavior Research Methods, Instruments, & Computers, 18,* 299–303.

Biederman, I. (1987). Recognition-by-components: A theory of human image understanding. *Psychological Review, 94*, 115.

Bilson, S., Yoshida, H., Tran, C. D., Woods, E. A., & Hills, T. T. (2015). Semantic facilitation in bilingual first language acquisition. *Cognition, 140,* 122-134.

Binsted, K., Pain, H., & Ritchie, G. (1997). Children's evaluation of computer-generated punning riddles. *Pragmatics and Cognition, 5*, 309–358.

Borgatti, S. P., Mehra, A., Brass, D. J., & Labianca, G. (2009). Network analysis in the social sciences. *Science*, *323*(5916), 892-895.

Bosse, M. L., & Valdois, S. (2009). Influence of the visual attention span on child reading performance: a cross-sectional study. *Journal of Research in Reading*, *32*(2), 230-253.

Bullinaria, J. A., & Levy, J. P. (2007). Extracting semantic representations from word co-occurrence statistics: A computational study. *Behavior research methods*, *39*(3), 510-526.

Blondel, V. D., Guillaume, J. L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, *2008*(10), P10008.

Booth, A. E., & Waxman, S. (2002). Object names and object functions serve as cues to categories for infants. *Developmental Psychology*, *38*, 948.

Booth, A. E., & Waxman, S. R. (2008). Taking stock as theories of word learning take shape. *Developmental Science*, *11*(2), 185–194.

Booth, A. E., Waxman, S. R., & Huang, Y. (2005). Conceptual information permeates word learning in infancy. *Developmental Psychology, 41*, 491–505.

Bradley, M. M., & Lang, P. J. (1999). *Affective norms for English words (ANEW): Instruction manual and affective ratings* (pp. 1–45). Technical Report C-1. The Center for Research in Psychophysiology, University of Florida.

Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior research methods*, 46(3), 904-911.

Buchta, C., Kober, M., Feinerer, I., & Hornik, K. (2012). Spherical k-means clustering. *Journal of Statistical Software*, *50*(10), 1-22.

Bullinaria, J. A., & Levy, J. P. (2007). Extracting semantic representations from word co-occurrence statistics: A computational study. *Behavior research methods*, *39*(3), 510-526.

Burger, J. D., Henderson, J., Kim, G., & Zarrella, G. (2011, July). Discriminating gender on Twitter. In *Proceedings of the conference on empirical methods in natural language processing* (pp. 1301-1309). Association for Computational Linguistics.

Byers-Heinlein, K., & Werker, J. F. (2009). Monolingual, bilingual, trilingual: infants' language experience influences the development of a word-learning heuristic. *Developmental Science*, *12*(5), 815-823.

Byers-Heinlein, K., & Werker, J. F. (2013). Lexicon structure and the disambiguation of novel words: Evidence from bilingual infants. *Cognition*, *128*(3), 407-416.

Cann, A., & Collette, C. (2014). Sense of humor, stable affect, and psychological well-being. *Europe*'*s Journal of Psychology*, *10*, 464– 479.

Carroll, J. B., & White, M. N. (1973). Age-of-acquisition norms for 220 picturable nouns. *Journal of Verbal Learning and Verbal Behavior*, *12*, 563–576.

Chafe, W. L. (1972). Discourse structure and human knowledge. *Language comprehension and the acquisition of knowledge*, 41-69.

Cherry, C., Mohammad, S. M., & De Bruijn, B. (2012). Binary classifiers and latent sequence models for emotion detection in suicide notes. *Biomedical informatics insights*, *5*, BII-S8933.

Cimpian, A., & Markman, E. M. (2005). The absence of a shape bias in children's word learning. *Developmental Psychology*, *41*(6), 1003.

Clark, G. (2008). *A farewell to alms: a brief economic history of the world*. Princeton University Press.

Cohen, T. (2008). *Jokes: Philosophical thoughts on joking matters*. University of Chicago Press.

Collisson, B. A., Grela, B., Spaulding, T., Rueckl, J. G., & Magnuson, J. S. (2015). Individual differences in the shape bias in preschool children with specific language impairment and typical language development: Theoretical and clinical implications. *Developmental Science, 18*(3), 373–388.

Colunga, E., & Smith, L. B. (2008). Knowledge embedded in process: The self-organization of skilled noun learning. *Developmental Science, 11*, 195–203.

Conover, M., Ratkiewicz, J., Francisco, M. R., Gonçalves, B., Menczer, F., & Flammini, A. (2011). Political polarization on twitter. *Icwsm*, *133*, 89-96.

Cree, G. S., & McRae, K. (2003). Analyzing the factors underlying the structure and computation of the meaning of chipmunk, cherry, chisel, cheese, and cello and many other such concrete nouns. *Journal of Experimental Psychology: General, 132*, 163–201.

Dale, P. S., & Fenson, L. (1996). Lexical development norms for young children. *Behavior Research Methods, Instruments, & Computers*, *28*, 125–127.

David, A., & Wei, L. (2008). Individual differences in the lexical development of French–English bilingual children. *International Journal of Bilingual Education and Bilingualism*, *11*(5), 598-618.

Davidov, D., Tsur, O., & Rappoport, A. (2010). Semi-supervised recognition of sarcastic sentences in twitter and amazon. In *Proceedings of the fourteenth conference on computational natural language learning* (pp. 107–116). Association for Computational Linguistics.

Davies, M. (2009). The 385+ million word Corpus of Contemporary American English (1990–2008+): Design, architecture, and linguistic insights. *International journal of corpus linguistics*, *14*(2), 159-190.

Davies, M. (2012). Expanding horizons in historical linguistics with the 400-million word Corpus of Historical American English. *Corpora*, *7*(2), 121-157.

Darwin, C. (1872). *The Expression of the Emotions in Man and Animals*. London: John Murray.

De Houwer, A., Bornstein, M. H., & Putnick, D. L. (2014). A bilingual–monolingual comparison of young children's vocabulary size: Evidence from comprehension and production. *Applied Psycholinguistics*, *35*(6), 1189-1211.

DeLoache, J. S., Simcock, G., & Macari, S. (2007). Planes, trains, automobiles--and tea sets: Extremely intense interests in very young children. *Developmental Psychology*, *43*(6), 1579.

Derczynski, L., Ritter, A., Clark, S., & Bontcheva, K. (2013). Twitter part-of-speech tagging for all: Overcoming sparse and noisy data. In *Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013* (pp. 198-206).

Dhillon, I. S., & Modha, D. S. (2001). Concept decompositions for large sparse text data using clustering. *Machine Learning, 42*, 143–175.

Diesendruck, G., & Bloom, P. (2003). How specific is the shape bias? *Child Development*, *74*(1), 168–178.

Dodds, P. S., Clark, E. M., Desu, S., Frank, M. R., Reagan, A. J., Williams, J. R., … Megerdoomian, K. (2015). Human language reveals a universal positivity bias. *Proceedings of the National Academy of Sciences*, *112*, 2389–2394.

Dubossarsky, H., De Deyne, S., & Hills, T. T. (2017). Quantifying the structure of free association networks across the life span. *Developmental psychology*, *53*(8), 1560.

Edmonds, P., & Hirst, G. (2002). Near-synonymy and lexical choice. *Computational linguistics*, *28*(2), 105-144.

Eisenstein, E. L. (1980). *The printing press as an agent of change* (Vol. 1). Cambridge University Press.

Ekman, P. (1992). An argument for basic emotions. Cognition & emotion, 6(3-4), 169-200.

Elman, J. L. (2008). The shape bias: An important piece in a bigger puzzle. *Developmental Science*, *11*(2), 219–222.

Encarta (2009). *Dictionary.* Archived at https://www.webcitation.org/5kwbLyr75

Engelthaler, T., & Hills, T. T. (2016). Feature biases in early word learning: network distinctiveness predicts age of acquisition. *Cognitive Science, 41*, 120-140.

Erdös, P., & Rényi, A. (1960). On the evolution of random graphs. *Publications of the Mathematical Institute of the Hungarian Academy of Sciences*, *5*, 17-61.

Ferrer-i-Cancho, R., & Vitevitch, M. S. (2017). The origins of Zipf's meaning-frequency law. arXiv preprint arXiv:1801.00168.

Ferrer i Cancho, R., & Solé, R. V. (2001). The small world of human language. *Proceedings of the Royal Society of London B: Biological Sciences*, *268*(1482), 2261-2265.

Fillmore, C. J. (1977). Scenes-and-frames semantics. *Linguistic structures processing*, *59*, 55-88.

Firth, J. R. (1957). *Papers in linguistics 1934-1951*. Oxford University Press.

Foot, H. C., & Chapman, A. J. (1976). The social responsiveness of young children in humorous situations. *Humor and laughter: Theory, research, and applications*, 187-214.

Freud, S. (1928). Humour. *International Journal of Psychoanalysis, 9,* 1– 6.

Fridlund, A. J., & Loftis, J. M. (1990). Relations between tickling and humorous laughter: Preliminary support for the Darwin-Hecker hy- pothesis. *Biological Psychology, 30,* 141–150.

Galloway, G., & Cropley, A. (1999). Benefits of humor for mental health: Empirical findings and directions for further research. *Humor, 12,* 301–314.

Gathercole, S. E. (2006). Nonword repetition and word learning: The nature of the relationship. *Applied Psycholinguistics*, *27*(4), 513-543.

Gentner, D., & Imai, M. (1995). A further examination of the shape bias in early word learning. In E. V. Clark (Ed.), *Proceedings of the twenty-sixth annual child language research forum* (pp. 167–176). Stanford, CA: CSLI Publications.

Gerber, M. S. (2014). Predicting crime using Twitter and kernel density estimation. *Decision Support Systems*, *61*, 115-125.

Gilhooly, K. J., & Logie, R. H. (1980). Age-of-acquisition, imagery, concreteness, familiarity, and ambiguity measures for 1,944 words. *Behavior Research Methods & Instrumentation, 12*, 395–427.

Gillette, J., Gleitman, H., Gleitman, L., & Lederer, A. (1999). Human simulations of vocabulary learning. *Cognition*, *73*(2), 135-176.

Goel, V., & Dolan, R. J. (2001). The functional anatomy of humor: Segregating cognitive and affective components. *Nature Neuroscience, 4,* 237–238.

Goldberg, K., Roeder, T., Gupta, D., & Perkins, C. (2001). Eigentaste: A constant time collaborative filtering algorithm. *Information Retrieval, 4,* 133–151.

Goldfield, B. A., & Reznick, J. S. (1990). Early lexical acquisition: Rate, content, and the vocabulary spurt. *Journal of child language*, *17*(01), 171-183.

González-Ibánez, R., Muresan, S., & Wacholder, N. (2011, June). Identifying sarcasm in Twitter: a closer look. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers-Volume 2* (pp. 581-586). Association for Computational Linguistics.

Graham, S. A., & Diesendruck, G. (2010). Fifteen-month-old infants attend to shape over other perceptual properties in an induction task. *Cognitive Development, 25,* 111–123.

Greenfield, P. M. (2013). The changing psychology of culture from 1800 through 2000. *Psychological science*, *24*(9), 1722-1731.

Gupta, D., Digiovanni, M., Narita, H., & Goldberg, K. (1999, August). Jester 2.0 (poster abstract): evaluation of an new linear time collaborative filtering algorithm. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 291-292). ACM.

Haig, R. (1988). Some sociocultural aspects of humour. *Australian & New Zealand Journal of Psychiatry*, *22*(4), 418-422.

Hall, S. (1973). Encoding and decoding in the television discourse.

Hamilton, W. L., Leskovec, J., & Jurafsky, D. (2016). Diachronic word embeddings reveal statistical laws of semantic change. arXiv preprint arXiv:1605.09096.

Harris, Z. S. (1954). Distributional structure. *Word*, *10*(2-3), 146-162.

Harris, C. R., & Christenfeld, N. (1997). Humour, tickle, and the Darwin- Hecker hypothesis. *Cognition & Emotion*, *11*(1), 103–110.

Hart, B., & Risley, T. R. (1995). *Meaningful differences in the everyday experience of young American children*. Paul H Brookes Publishing.

Havlin, S., Kenett, D. Y., Ben-Jacob, E., Bunde, A., Cohen, R., Hermann, H., ... & Portugali, J. (2012). Challenges in network science: Applications to infrastructures, climate, social systems and economics. *The European Physical Journal Special Topics*, *214*(1), 273-293.

Hawkes, T. (2003). *Structuralism and semiotics*. Routledge.

Hay, J. (1995). *Gender and humour: Beyond a joke*. Wellington, New Zealand: MA thesis, Victoria University of Wellington.

Hayes, J. R., & Clark, H. H. (1970). Experiments in the segmentation of an artificial speech analog. *Cognition and the development of language*, 221-234.

Hills, T. (2013). The company that words keep: comparing the statistical structure of child-versus adult-directed language. *Journal of child language*, *40*(03), 586-604.

Hills, T. T., & Adelman, J. S. (2015). Recent evolution of learnability in American English from 1800 to 2000. *Cognition*, *143*, 87-92.

Hills, T. T., Adelman, J. S., & Noguchi, T. (2016). Attention economies, information crowding, and language change. In Jones, M. N. (Ed.), *Big Data in Cognitive Science*. Psychology Press.

Hills, T. T., Jones, M. N., & Todd, P. M. (2012). Optimal foraging in semantic memory. *Psychological review*, *119*(2), 431.

Hills, T. T., Maouene, M., Maouene, J., Sheya, A., & Smith, L. (2009a). Categorical structure among shared features in networks of early-learned nouns. *Cognition*, *112*(3), 381-396.

Hills, T. T., Maouene, M., Maouene, J., Sheya, A., & Smith, L. (2009b). Longitudinal analysis of early semantic networks preferential attachment or preferential acquisition? *Psychological Science*, *20*(6), 729-739.

Hills, T. T., Maouene, J., Riordan, B., & Smith, L. B. (2010). The associative structure of language: Contextual diversity in early word learning. *Journal of memory and language*, *63*(3), 259-273.

Hills, T. T., Mata, R., Wilke, A., & Samanez-Larkin, G. R. (2013). Mechanisms of age-related decline in memory search across the adult life span. *Developmental psychology*, *49*(12), 2396.

Hills, T., Proto, E., & Sgroi, D. (2015) Historical analysis of national subjective wellbeing using millions of digitized books. *IZA Discussion Paper No. 9195*.

Hobbes, T. (1840). Human Nature. In W. Molesworth (Ed.), *The English Works of Thomas Hobbes Of Malmesbury*, 4th ed. London: Bohn.

Hoff, E., Core, C., Place, S., Rumiche, R., Señor, M., & Parra, M. (2012). Dual language exposure and early bilingual development. *Journal of child language*, *39*(01), 1-27.

Hurley, M. M., Dennett, D. C., & Adams, R. B. (2011). *Inside jokes: Using humor to reverse-engineer the mind*. Cambridge: MIT Press.

Inkpen, D. (2007). A statistical model for near-synonym choice. *ACM Transactions on Speech and Language Processing (TSLP)*, *4*(1), 2.

Iversen, J. R., Patel, A. D., & Ohgushi, K. (2008). Perception of rhythmic grouping depends on auditory experiencea. *The Journal of the Acoustical Society of America*, *124*(4), 2263–2271.

Jaswal, V. K., & Hansen, M. B. (2006). Learning words: Children disregard some pragmatic information that conflicts with mutual exclusivity. *Developmental Science*, *9*, 158–165.

Jeffers, R. J., & Lehiste, I. (1979). *Principles and methods for historical linguistics*. MIT press.

Jones, M. N., Hills, T. T., & Todd, P. M. (2015). Hidden processes in structural representations: A reply to Abbott, Austerweil, and Griffiths (2015). *Psychological Review*, *122*(3), 570-574.

Jones, M. N., Kintsch, W., & Mewhort, D. J. (2006). High-dimensional semantic space accounts of priming. *Journal of memory and language*, *55*(4), 534-552.

Jones, M. N., & Mewhort, D. J. K. (2007). Representing word meaning and order information in a composite. *Psychological review*, *104*, *1-37*.

Jones, S. S., Smith, L. B., & Landau, B. (1991). Object properties and knowledge in early lexical learning. *Child development, 62*(3), 499–516.

Kachergis, G., Yu, C., & Shiffrin, R. M. (2012). An associative model of adaptive inference for learning word–referent mappings. *Psychonomic Bulletin & Review, 19*(2), 317–324.

Kanske, P., & Kotz, S. A. (2010). Leipzig affective norms for German: A reliability study. *Behavior Research Methods, 42,* 987–991.

Kant, I. (1914). *The Critique of Judgement* (J. H. Bernard, Trans.). London: Macmillian. (Original work published 1790).

Keil, F. C., & Batterman, N. (1984). A characteristic-to-defining shift in the development of word meaning. *Journal of verbal learning and verbal behavior*, *23*(2), 221-236.

Keuleers, E., Stevens, M., Mandera, P., & Brysbaert, M. (2015). Word knowledge in the crowd: Measuring vocabulary size and word prevalence in a massive online experiment. *The Quarterly Journal of Experimental Psychology*, *68*(8), 1665-1692.

Kemp, C., Perfors, A., & Tenenbaum, J. B. (2007). Learning overhypotheses with hierarchical Bayesian models. *Developmental Science, 10*(3), 307–321.

Keith-Spiegel, P. (1972). Early conceptions of humor: Varieties and is- sues. In J. H. Goldstein & P. E. McGhee (Eds.), *The Psychology of Humor: Theoretical Perspectives and Empirical Issues* (pp. 4–39). New York: Academic Press.

Kennedy, G. A. (1994). *A new history of classical rhetoric*. Princeton University Press.

Keuleers, E., Lacey, P., Rastle, K., & Brysbaert, M. (2012). The British lexicon project: Lexical decision data for 28,730 monosyllabic and disyllabic English words. *Behavior Research Methods, 44,* 287– 304.

Koestler, A. (1964). *The act of creation*. New York: Penguin Books.

Kuhl, P. K., Conboy, B. T., Padden, D., Nelson, T., & Pruitt, J. (2005). Early speech perception and later language development: Implications for the" critical period". *Language learning and development*, *1*(3-4), 237-264.

Kulkarni, V., Al-Rfou, R., Perozzi, B., & Skiena, S. (2015). Statistically significant detection of linguistic change. In *Proceedings of the 24th International Conference on World Wide Web* (pp. 625-635).

Kumar, R., Raghavan, P., Rajagopalan, S., Sivakumar, D., Tomkins, A., & Upfal, E. (2000). Stochastic models for the web graph. In D. C. Young (Ed.), *Proceedings 41st Annual Symposium on Foundations of Computer* Science (pp. 57-65). Redondo Beach, CA: IEEE.

Kuperman, V., Stadthagen-Gonzalez, H., & Brysbaert, M. (2012). Age-of-acquisition ratings for 30,000 English words. *Behavior Research Methods*, *44*(4), 978-990.

Ladd, D. R., Roberts, S. G., & Dediu, D. (2015) Correlational studies in typological and historical linguistics. Annual Review of Linguistics, 1 4.1–4.21.

Lampos, V., & Cristianini, N. (2012). Nowcasting events from the social web with statistical learning. *ACM Transactions on Intelligent Systems and Technology (TIST)*, *3*(4), 72.

Landau, B., Smith, L. B., & Jones, S. S. (1988). The importance of shape in early lexical learning. *Cognitive Development*, *3*, 299-321.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse processes*, *25*(2-3), 259-284.

Le, X., Lancashire, I., Hirst, G., & Jokel, R. (2011). Longitudinal detection of dementia through lexical and syntactic changes in writing: a case study of three British novelists. *Literary and Linguistic Computing*, *26*(4), 435-461.

Leech, G. N. (1992). 100 million words of English: the British National Corpus (BNC).

Levy, J., Bullinaria, J., & McCormick, S. (2017). Semantic vector evaluation and human performance on a new vocabulary MCQ test. In *CogSci*.

Levy, O., Goldberg, Y., & Dagan, I. (2015). Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics*, *3*, 211-225.

Li, P., Farkas, I., & MacWhinney, B. (2004). Early lexical development in a self-organizing neural network. *Neural networks*, *17*(8), 1345-1362.

Lin, Y., Michel, J. B., Aiden, E. L., Orwant, J., Brockman, W., & Petrov, S. (2012). Syntactic annotations for the google books ngram corpus. In *Proceedings of the ACL 2012 System Demonstrations* (pp. 169-174). Stroudsburg, PA: Association for Computational Linguistics.

Logothetis, N. K., & Sheinberg, D. L. (1996). Visual object recognition. *Annual Review of Neuroscience*, *19*, 577–621.

Love, B. C., Medin, D. L., & Gureckis, T. M. (2004). SUSTAIN: A network model of category learning. *Psychological Review*, *111*, 309–332.

Lowenthal, D. (2015). *The past is a foreign country-revisited*. Cambridge University Press.

Lynott, D., & Connell, L. (2013). Modality exclusivity norms for 400 nouns: The relationship between perceptual experience and surface word form. *Behavior research methods, 45*(2), 516-526.

MacWhinney, B. (2000). *The CHILDES Project: Tools for Analyzing Talk (Volume II: The Database)*. Mahwah, NJ: Lawrence Erlbaum Associates

MacWhinney, B. (2014). *The CHILDES project: Tools for analyzing talk, Volume II: The database*. Psychology Press.

Manurung, R., Ritchie, G., Pain, H., Waller, A., O'Mara, D., & Black, R. (2008). The construction of a pun generator for language skills de- velopment. *Applied Artificial Intelligence, 22,* 841–869.

Markman, E. M. (1990). Constraints children place on word meanings. *Cognitive Science, 14*, 57–77.

Markman, E. M., & Wachtel, G. F. (1988). Children's use of mutual exclusivity to constrain the meanings of words. *Cognitive Psychology, 20*, 121–157.

Markman, E. M., Wasow, J. L., & Hansen, M. B. (2003). Use of the mutual exclusivity assumption by young word learners. *Cognitive Psychology*, *47*, 241–275.

Markson, L., Diesendruck, G., & Bloom, P. (2008). The shape of thought. *Developmental Science*, *11*(2), 204–208.

Martin, R. A. (1998). Approaches to the sense of humor: A historical review. In W. Ruch (Ed.), *The sense of humor: Explorations of a personality characteristic* (pp. 15–60). Berlin: Walter de Crruyter.

Martin, R. A., Kuiper, N. A., Olinger, L. J., & Dance, K. A. (1993). Humor, coping with stress, self-concept, and psychological well- being. *Humor, 6*, 89–104.

Martin, R. A., Puhlik-Doris, P., Larsen, G., Gray, J., & Weir, K. (2003). Individual differences in uses of humor and their relation to psycho- logical well-being: Development of the Humor Styles Questionnaire. *Journal of Research in Personality, 37,* 48–75.

Mather, E., & Plunkett, K. (2009). Learning words over time: The role of stimulus repetition in mutual exclusivity. *Infancy, 14*, 60–76.

Mather, E., & Plunkett, K. (2012). The role of novelty in early word learning. *Cognitive Science, 36*, 1157– 1177.

Mayor, J., & Plunkett, K. (2010). Vocabulary Spurt: Are Infants full of Zipf?. In S. Ohlsson, & R. Catrambone (Eds.), *Proceedings of the 32nd Annual Conference of the Cognitive Science Society* (pp. 836-841). Austin, TX: Cognitive Science Society

McGee, E., & Shevlin, M. (2009). Effect of humor on interpersonal attraction and mate selection. *The Journal of Psychology, 143,* 67– 77.

McGhee, P. E. (1971). Development of the humor response: A review of the literature. *Psychological Bulletin, 76,* 328–348.

McGraw, A. P., & Warren, C. (2010). Benign violations: Making immoral behavior funny. *Psychological Science, 21,* 1141–1149.

McKay, J. (2002). Generation of idiom-based witticisms to aid second language learning. In *Proceedings of the Twente Workshop on Language Technology, 20*. The University of Twente.

McMurray, B. (2007). Defusing the childhood vocabulary explosion. *Science*, *317*(5838), 631-631.

McRae, K., Cree, G. S., Seidenberg, M. S., & McNorgan, C. (2005). Semantic feature production norms for a large set of living and nonliving things. *Behavior research methods*, *37*(4), 547-559.

Michel, J. B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., The Google Books Team, Pickett, J. P., Hoiberg, D., Clancy, D., Norvig, P, Orwant, J., Pinker, S., Nowak, M. A., & Aiden, E. L. (2011). Quantitative analysis of culture using millions of digitized books. *Science*, *331*(6014), 176-182.

Mickes, L., Walker, D. E., Parris, J. L., Mankoff, R., & Christenfeld, N. J. (2012). Who's funny: Gender stereotypes, humor production, and memory bias. *Psychonomic Bulletin & Review, 19*(1), 108–112.

Minsky, M. (1981). Jokes and their Relation to the Cognitive Unconscious. In Vaina, L., Hintikka, J. (Eds.) *Cognitive Constraints on Communication* (pp. 175-200). Boston: Reidel.

Mihalcea, R., & Strapparava, C. (2005). Making computers laugh: Investigations in automatic humor recognition. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing* (pp. 531–538).

Milner, G. B. (1972). Homo ridens: Towards a semiotic theory of humour and laughter. *Semiotica, 5,* 1–30

Mitkov, R. (Ed.). (2004). *The Oxford handbook of computational linguistics*. Oxford University Press.

Mohammad, S. M. (2016). Sentiment analysis: Detecting valence, emotions, and other affectual states from text. In *Emotion measurement* (pp. 201-237).

Mohammad, S. (2018). Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 English words. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 174-184).

Monaghan, P. (2014). Age of acquisition predicts rate of lexical evolution. Cognition, 133(3), 530-534.

Monnier, C., & Syssau, A. (2014). Affective norms for French words (FAN). *Behavior Research Methods, 46,* 1128–1137.

Montefinese, M., Ambrosini, E., Fairfield, B., & Mammarella, N. (2014). The adaptation of the Affective Norms for English words (ANEW) for Italian. *Behavior Research Methods, 46,* 887–903.

Moors, A., De Houwer, J., Hermans, D., Wanmaker, S., Van Schie, K., Van Harmelen, A. L., … Brysbaert, M. (2013). Norms of valence, arousal, dominance, and age of acquisition for 4,300 Dutch words. *Behavior Research Methods*, *45*, 169–177.

Mora-Ripoll, R. (2011). Potential health benefits of simulated laughter: A narrative review of the literature and recommendations for future research. *Complementary Therapies in Medicine, 19,* 170–177.

Morgan, J. L. (1996). A rhythmic bias in preverbal speech segmentation. *Journal of Memory and Language*, *35*(5), 666–688.

Morris, C. W. (1938). Foundations of the Theory of Signs. In *International encyclopedia of unified science* (pp. 1-59). Chicago University Press.

Morris, C. (1946). Signs, language and behavior. Oxford, England: Prentice-Hall.

Moyle, M. J., Weismer, S. E., Evans, J. L., & Lindstrom, M. J. (2007). Longitudinal relationships between lexical and grammatical development in typical and late-talking children. *Journal of Speech, Language, and Hearing Research*, *50*(2), 508-528.

Nakagawa, S. (2004). A farewell to Bonferroni: the problems of low statistical power and publication bias. *Behavioral ecology*, *15*(6), 1044-1045.

Nelson, D. L., McEvoy, C. L., & Schreiber, T. (1998). The University of South Florida word association, rhyme, and word fragment norms. Retrieved from http://www.usf.edu/FreeAssociation/.

Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (2004). The University of South Florida free association, rhyme, and word fragment norms. *Behavior Research Methods, Instruments, & Computers*, *36*(3), 402-407.

Nelson, D. L., Zhang, N., & McKinney, V. M. (2001). The ties that bind what is known to the recognition of what is new. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *27*(5), 1147.

Newman, M. E. (2003). The structure and function of complex networks. *SIAM Review*, *45*(2), 167-256.

Newman, R. S. (2008). The level of detail in infants' word learning. *Current directions in psychological science*, *17*(3), 229-232.

Ogden, C. K., & Richards, I. A. (1923). *The Meaning of Meaning: A Study of the Influence of Language upon Thought and of the Science of Symbolism* (Vol. 29). K. Paul, Trench, Trubner & Company, Limited.

Olson, G. M., & Sherman, T. (1983). Attention, learning, and memory in infants. In. M. Haith & J. Campos (Eds.), *Manual of child psychology: Vol. 2. Infancy and developmental psychobiology* (pp. 1001-1080). New York: Wiley.

Osgood, C. E. (1952). The nature and measurement of meaning. *Psychological bulletin*, *49*(3), 197.

Osgood, C., Suci, G., & Tannenbaum, P. (1957). The measurement of meaning. *Urbana, IL: University of Illinois.*

Pagel, M., Atkinson, Q. D., & Meade, A. (2007). Frequency of word-use predicts rates of lexical evolution throughout Indo-European history. Nature, 449(7163), 717.

Paivio, A., Yuille, J. C., & Madigan, S. A. (1968). Concreteness, imagery, and meaningfulness values for 925 nouns. *Journal of Experimental Psychology, 76,* 1–25.

Panksepp, J. (1993). Rough and tumble play: A fundamental brain process. In K. MacDonald (Ed.), *SUNY series, children's play in society. Parent–child play: Descriptions and implications* (pp. 147-184). Albany, NY, US: State University of New York Press.

Park, D. C., Lautenschlager, G., Hedden, T., Davidson, N. S., Smith, A. D., & Smith, P. K. (2002). Models of visuospatial and verbal memory across the adult life span. *Psychology and aging*, *17*(2), 299.

Pastor-Satorras, R., & Vespignani, A. (2001). Epidemic spreading in scale-free networks. *Physical review letters*, *86*(14), 3200.

Pavlov, I. P., & Thompson, W. H. (1902). *The work of the digestive glands*. Charles Griffin.

Pearson, B. Z., Fernández, S., & Oller, D. K. (1995). Cross-language synonyms in the lexicons of bilingual infants: One language or two?. *Journal of child language*, *22*, 345-345.

Pechenick, E. A., Danforth, C. M., & Dodds, P. S. (2015). Characterizing the Google Books corpus: Strong limits to inferences of socio-cultural and linguistic evolution. *PloS one*, *10*(10), e0137041.

Pennebaker, J. W., & Stone, L. D. (2003). Words of wisdom: language use over the life span. *Journal of personality and social psychology*, *85*(2), 291.

Piketty, T. (2014). Capital in the 21st century. Harvard University Press.

Pinker, S. (2011). *The better angels of our nature: The decline of violence in history and its causes*. Penguin UK.

Plunkett, K., Hu, J. F., & Cohen, L. B. (2008). Labels can override perceptual categories in early infancy. *Cognition*, *106*(2), 665–681.

Postman, L., & Keppel, G. (Eds.). (2014). *Norms of word association*. Academic Press.

Ramon-Casas, M., Swingley, D., Sebastián-Gallés, N., & Bosch, L. (2009). Vowel categorization during word recognition in bilingual toddlers. *Cognitive psychology*, *59*(1), 96-121.

Ramscar, M., Hendrix, P., Shaoul, C., Milin, P., & Baayen, H. (2014). The myth of cognitive decline: Non-linear dynamics of lifelong learning. *Topics in cognitive science*, *6*(1), 5-42.

Recchia, G., & Louwerse, M. M. (2015). Reproducing affective norms with lexical co-occurrence statistics: Predicting valence, arousal, and dominance. *The Quarterly* Journal of Experimental Psychology, 68(8), 1584-1598.

Redondo, J., Fraga, I., Padrón, I., & Comesaña, M. (2007). The Spanish adaptation of ANEW (Affective Norms for English Words). *Behavior Research Methods, 39,* 600–605.

Richens, R. H. (1956). Preprogramming for mechanical translation. *Mechanical Translation*, *3*(1), 20-25.

Roeckelein, J. (2006). *Elsevier's dictionary of psychological theories*. Amsterdam [Netherlands]: Elsevier.

Rosch, E. H. (1973). Natural categories. *Cognitive psychology*, *4*(3), 328-350.

Rothman, K. J. (1990). No adjustments are needed for multiple comparisons. *Epidemiology*, 43-46.

Ruppenhofer, J., Ellsworth, M., Schwarzer-Petruck, M., Johnson, C. R., & Scheffczyk, J. (2006). FrameNet II: Extended theory and practice.

Sagi, E., Kaufmann, S., & Clark, B. (2011). Tracing semantic change with latent semantic analysis. Current methods in historical semantics, 161-183.

Samuelson, L. K., & Horst, J. S. (2007). Dynamic noun generalization: moment-to-moment interactions shape children's naming biases. *Infancy*, *11*(1), 97-110.

Samuelson, L. K., & Smith, L. B. (1999). Early noun vocabularies: Do ontology, category structure and syntax correspond? *Cognition*, *73*, 1–33.

Sayce, D. (2018) *Number of tweets per day?* Retrieved from https://www.dsayce.com/social-media/tweets-day/

Schönemann, P. H. (1966). A generalized solution of the orthogonal procrustes problem. *Psychometrika*, *31*(1), 1-10.

Schyns, P. G., Goldstone, R. L., & Thibaut, J. P. (1998). The development of features in object concepts. *Behavioral and Brain Sciences*, *21*, 1–17.

Scruton, R. (1987). Laughter. In J. Morreall (Ed.), *The philosophy of laughter and humor* (pp. 156–171). Albany, NY: SUNY Press.

Sheya, A., & Smith, L. B. (2006). Perceptual features and the development of conceptual knowledge. *Journal of Cognition and Development*, *7*(4), 455-476.

Shultz, T. R. (1976). A cognitive-developmental analysis of humour. In J. Chapman & H. C. Foot (Eds.), *Humor and laughter: Theory, research, and applications* (pp. 11–36). London: John Wiley & Sons

Siew, C. S. (2019). spreadr: An R package to simulate spreading activation in a network. *Behavior research methods*, 1-20.

Sloutsky, V. M. (2010). From perceptual categories to concepts: What develops? *Cognitive Science*, *34*, 1244–1286.

Sloutsky, V. M., & Fisher, A. V. (2004). Induction and categorization in young children: A similarity-based model. *Journal of Experimental Psychology: General*, *133*(2), 166.

Sloutsky, V. M., & Fisher, A. V. (2011). The development of categorization. In B. H. Ross (Ed.), *The psychology of learning and motivation* (Vol. 54, pp. 141–166). San Diego, CA: Academic Press.

Smith, L. B., Jones, S. S., Landau, B., Gershkoff-Stowe, L., & Samuelson, L. (2002). Object name learning provides on-the-job training for attention. *Psychological Science*, *13*, 13–19.

Smith, L. B., & Samuelson, L. (2006). An attentional learning account of the shape bias: Reply to Cimpian and Markman (2005) and Booth, Waxman, and Huang (2005). *Developmental Psychology*, *42*, 1339–1343.

Spinka, M., Newberry, R. C., & Bekoff, M. (2001). Mammalian play: training for the unexpected. *The quarterly review of biology*, *76*(2), 141-168.

Soares, A. P., Comesaña, M., Pinheiro, A. P., Simões, A., & Frade, C. S. (2012). The adaptation of the Affective Norms for English Words (ANEW) for European Portuguese. *Behavior Research Methods, 44,* 256–269.

Soja, N. N. (1992). Inferences about the meanings of nouns: The relationship between perception and syntax. *Cognitive development*, *7*(1), 29-45.

Soja, N. N., Carey, S., & Spelke, E. S. (1991). Ontological categories guide young children's inductions of word meaning: Object terms and substance terms. *Cognition*, *38*(2), 179-211.

Söderholm, C., Häyry, E., Laine, M., & Karrasch, M. (2013). Valence and arousal ratings for 420 Finnish nouns by age and gender. *PloS One, 8,* e72859.

Spencer, H. (1860). The physiology of laughter. *Macmillan*'*s Magazine*, *1,* 395–402.

Stadthagen-Gonzalez, H., & Davis, C. J. (2006). The Bristol norms for age of acquisition, imageability, and familiarity. *Behavior Research Methods, 38,* 598–605.

Stella, M., Beckage, N. M., & Brede, M. (2016). Multiplex lexical networks reveal patterns in early word acquisition in children. Advance online publication. *arXiv:1609.03207*.

Steyvers, M., & Tenenbaum, J. B. (2005). The Large-scale structure of semantic networks: Statistical analyses and a model of semantic growth. *Cognitive Science*, *29*(1), 41-78.

Stock, O., & Strapparava, C. (2003). HAHAcronym: Humorous agents for humorous acronyms. *Humor, 16,* 297–314.

Stokes, S. F., & Klee, T. (2009). Factors that influence vocabulary development in two-year-old children. *Journal of Child Psychology and Psychiatry*, *50*(4), 498-505.

Storkel, H. L. (2001). Learning new words. *Journal of Speech, Language, and Hearing Research*.

Stuart, M. J. (1843). A system of logic. *Londres: Parker*.

Suero Montero, C., & Suhonen, J. (2014, November). Emotion analysis meets learning analytics: online learner profiling beyond numerical data. In *Proceedings of the 14th Koli calling international conference on computing education research* (pp. 165-169). ACM.

Suls, J. M. (1972). ATwo-Stage Model for the Appreciation of Jokes and Cartoons: An Information-Processing Analysis. In J. H. Goldstein & P. E. McGhee (Eds.), *The Psychology of Humor: Theoretical Perspectives and Empirical Issues* (pp. 81–100). New York: Academic Press

Tarr, M. J., & Bulthoff, H. H. (1998) Image-based object recognition in man, monkey and machine. *Cognition*, *67,* 1-20.

Tek, S., Jaffery, G., Swensen, L., Fein, D., & Naigles, L. R. (2012). The shape bias is affected by differing similarity among objects. *Cognitive Development*, *27*, 28–38.

Thal, D. J., Bates, E., Goodman, J., & Jahn-Samilo, J. (1997). Continuity of language abilities: An exploratory study of late and early talking toddlers. *Developmental Neuropsychology*, *13*(3), 239-273.

Thiessen, E. D., Hill, E. A., & Saffran, J. R. (2005). Infant-directed speech facilitates word segmentation. *Infancy*, *7*(1), 53-71.

Thiessen, E. D., & Saffran, J. R. (2007). Learning to learn: Infants' acquisition of stress-based strategies for word segmentation. *Language Learning and Development*, *3*(1), 73–100.

Thompson, J. (1941). Development of facial expression of emotion in blind and seeing children. *Archives of Psychology (Columbia University)*.

Trier, J. (1931). *Der deutsche wortschatz im sinnbezirk des verstandes: die geschichte eines sprachlichen feldes. 1. von den anfängen bis zum beginn des 13. jahrhunderts*. Winter.

Turney, P. D., & Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. *Journal of artificial intelligence research*, 37, 141-188.

Van Heuven, W. J., Mandera, P., Keuleers, E., & Brysbaert, M. (2014). SUBTLEX-UK: A new and improved word frequency database for British English. *The Quarterly Journal of Experimental Psychology*, *67*, 1176–1190.

Vejdemo, S., & Hörberg, T. (2016). Semantic factors predict the rate of lexical replacement of content words. *PloS one*, *11*(1), e0147924.

Vinson, D. P., & Vigliocco, G. (2008). Semantic feature production norms for a large set of objects and events. *Behavior Research Methods*, *40*(1), 183-190.

Vitevitch, M. S. (2008). What can graph theory tell us about word learning and lexical retrieval?. *Journal of Speech, Language, and Hearing Research*, *51*(2), 408-422.

Wagner, A. (2001). The yeast protein interaction network evolves rapidly and contains few redundant duplicate genes. *Molecular biology and evolution*, *18*(7), 1283-1292.

Warriner, A. B., Kuperman, V., & Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior research methods*, 45(4), 1191-1207.

Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature*, *393*(6684), 440-442.

Weaver, W. (1955). Translation. *Machine translation of languages*, *14*, 15-23.

Weitnauer, E., Carvalho, P. F., Goldstone, R. L., & Ritter, H. (2014). Similarity-based ordering of instances for efficient concept learning. In P. Bello, M. Guarini, M. McShane, & B. Scassellati (Eds.), *36th Annual Conference of the Cognitive Science Society* (pp. 1760–1765). Austin, TX: Cognitive Science Society.

Weizman, Z. O., & Snow, C. E. (2001). Lexical output as related to children's vocabulary acquisition: Effects of sophisticated exposure and support for meaning. *Developmental psychology*, *37*(2), 265.

Werker, J. F., Byers-Heinlein, K., & Fennell, C. T. (2009). Bilingual beginnings to learning words. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, *364*(1536), 3649-3663.

Westbury, C., Keith, J., Briesemeister, B. B., Hofmann, M. J., & Jacobs, A. M. (2015). Avoid violence, rioting, and outrage; approach celebration, delight, and strength: Using large text corpora to compute valence, arousal, and the basic emotions. *The Quarterly Journal of Experimental Psychology*, *68*(8), 1599-1622.

Westbury, C., Shaoul, C., Moroschan, G., & Ramscar, M. (2016). Telling the world's least funny jokes: On the quantification of humor as entropy. *Journal of Memory and Language, 86,* 141–156.

Whaley, C. P. (1978). Word—nonword classification time. *Journal of Verbal Learning and Verbal Behavior*, *17,* 143–154.

Wicker, F. W., Thorelli, I. M., Barron, W. L., III, & Willis, A. C. (1981). Studies of mood and humor appreciation. *Motivation and Emotion, 5,* 47–59.

Wu, L. L., & Barsalou, L. W. (2009). Perceptual simulation in conceptual combination: Evidence from property generation. *Acta Psychologica*, *132,* 173–189.

Wyer, R., & Collins, J. (1992). A theory of humor elicitation. *Psychological Review, 99*(4), 663–688.

Xu, Y and Kemp, C. 2015. A computational evaluation of two laws of semantic change. *Proc 37th Annu Conf Cogn Sci Soc 2015*.

Yu, C., & Smith, L. B. (2007). Rapid word learning under uncertainty via cross-situational statistics. *Psychological Science*, *18*(5), 414-420.

Yu, L. C., Wu, J. L., Chang, P. C., & Chu, H. S. (2013). Using a contextual entropy model to expand emotion words and their intensity for the sentiment classification of stock market news. *Knowledge-Based Systems*, *41*, 89-97.

Yurovsky, D., Yu, C., & Smith, L. B. (2013). Competitive processes in cross-situational word learning. *Cognitive Science, 37*(5), 891–921.

Zalizniak, A., Bulakh, M., Ganenkov, D., Gruntov, I., Maisak ,T., & Russo, M. (2012). The catalogue of semantic shifts as a database for lexical semantic typology. *Linguistics*, 50, 633–69.

Zipf, G. (1949). *Human behavior and the principle of least effort.* New York: Addison-Wesley.