# Polylogues on The Mental Lexicon

Edited by
Gary Libben, Gonia Jarema and Victor Kuperman

John Benjamins Publishing Company

Polylogues on The Mental Lexicon

# Polylogues on The Mental Lexicon

## An exploration of fundamental issues and directions

*Edited by*

Gary Libben
Brock University

Gonia Jarema
Université de Montréal

Victor Kuperman
McMaster University

John Benjamins Publishing Company

Amsterdam / Philadelphia

The paper used in this publication meets the minimum requirements of the American National Standard for Information Sciences – Permanence of Paper for Printed Library Materials, ANSI Z39.48-1984.

DOI 10.1075/z.238

**Cataloging-in-Publication Data available from Library of Congress:**
**LCCN 2021021472 (PRINT) / 2021021473 (E-BOOK)**

ISBN 978 90 272 0925 2 (HB)
ISBN 978 90 272 5961 5 (E-BOOK)



John Benjamins Publishing Company · https://benjamins.com

# Table of contents

# Preface

In developing this volume, our goal was to create a book whose contents match its title as closely as possible. Thus, this is a volume of *Polylogues on The Mental Lexicon* that, together, constitute *an exploration of fundamental issues and directions*.

The book is built upon the foundation of Issue 11.3 of *The Mental Lexicon*, which brought together researchers who have articulated a vision of what is important in the development of new mental lexicon research and have engaged in productive debate with colleagues across subfields, generations, and theoretical perspectives.

It seems to us that from its beginnings, mental lexicon research has been at the crossroads of research and scholarship. It is this that makes it most exciting and most important. In this volume, we have endeavoured to capture that excitement and to contribute to a deeper understanding of key issues and debates in the field.

Here, in this book, authors invite you to join a polylogue – a textual conversation of many voices. The first chapter, *The Mental Lexicon as Polylogue* by Kuperman, Jarema and Libben sets the stage by presenting an examination of how the notion of a polylogue developed from antiquity and how it has particular relevance to the mental lexicon. The chapter also creates a context for the subsequent polylogue chapters by presenting an analysis using the technique of 'structural topic analysis'. This analysis examines what topics have been at the forefront of research as it is represented in the journal literature and, in particular, in the John Benjamins journal *The Mental Lexicon*, over the past 15 years. For students, in particular, the initial chapter provides a snapshot of key themes and topics and therefore also a way to identify gaps, areas of developing prominence, and opportunities to shape the future of the field.

Each of the subsequent six chapters has the following structure: (a) the author thesis, (b) commentaries on the thesis, and (c) the article. The thesis is a brief presentation of the central claims. Each thesis addresses a core theme in mental lexicon research. These are: the value of cross-linguistic megastudies (Myers), the nature of meaning (Westbury), how to capture truly natural language (Tucker & Ernestus), what can be learned from lexical acquisition (Ravid et al.), the advantages of a functionalist perspective (Richie), and the role of schemas in understanding morphology and the lexicon (Jackendoff & Audring). Each thesis is followed by five commentaries. These commentaries are opportunities for the other authors to contribute insights on the central themes in the thesis. The final section of each chapter is the article upon which the polylogue is based (from The Mental Lexicon, vol.11:3). In this way, the organization of each chapter is designed to enable the reader to quickly appreciate the core claims made by the authors and to engage in the discussion of consequences and implications. It is our hope that readers will find that this volume and its interactive structure captures the excitement of the field and also provides important new perspectives.

We wish to extend our deepest gratitude to the extraordinarily competent and supportive team at John Benjamins Publishing Company. In particular, we thank Esther Roth, who has guided us through this project, Susan Hendriks, who has been a constant and indispensable support in our work on *The Mental Lexicon*, and to Kees Vaes, whose encouragement and insight were critical to getting the whole enterprise started.

We are also extremely grateful to the Social Sciences and Humanities Research Council of Canada (SSHRC), which has supported our Mental Lexicon Research over a number of projects. Our work on this book has received direct support from the SSHRC Partnership Grant, "Words in the World".

CHAPTER 1

# The mental lexicon as polylogue

Victor Kuperman[1], Gonia Jarema[2,3] and Gary Libben[4]
[1]McMaster University / [2]Université de Montréal / [3]IUGM /
[4]Brock University

In many ways, and on many levels, the notion of a polylogue captures core features of the mental lexicon. This is the case when we consider the construct of the mental lexicon as a psycholinguistic construct – the dynamic cognitive system that constitutes lexical ability. It is also the case when we consider how words themselves, as the shared possessions of a community, are indispensable to the opportunities for polylogues within that community.

In the sections below, we discuss the nature of the polylogue as a form of knowledge exchange, as it has developed from antiquity. We consider, in some depth, a particular community, and a particular contemporary polylogue – that is the community of mental lexicon researchers and their continuing scientific polylogue over the past thirty years in the texts of academic journals such as The Mental Lexicon. Finally, we present the ways in which dynamic co-activation within the mental lexicon and the multidimensional connections among lexical representations may also be seen to form a cognitive polylogue.

## Roots of the polylogue

The Oxford dictionary defines the word *polylogue* as "a discussion involving more than two people" and traces its origin to the late nineteenth century. The term is modeled after *dialogu*e and formed of the Greek prefix *poly-* (many) and the bound stem *logue-*, from Ancient Greek λόγος, 'discourse or argument', in its Aristotelian sense. It is important to note that the word *dialogue* does not decompose into *di(duo)logue*, as *διά* means 'through', across', or 'inter', rather than 'two'. Thus, a *dialogue* can, in fact, involve two or more people. This leads us, then, to the question of what distinguishes a dialogue from a polylogue. Tracing their usage historically might give us an answer.

https://doi.org/10.1075/z.238.01kup

In Europe, the dialogue has served as a key literary genre since Antiquity. The most widely known example of this is Plato's *Republic*, which features the Socratic dialogue. But as a genre, the dialogue has been documented to have existed several millennia prior to Greece's Golden Age: In the third millennium BCE, Sumerians devised 'disputations', or literary debates written in poetic form, in which two protagonists (e.g., a sheep and a fish, winter and summer) debate their usefulness to humankind. At the end of the debate, the Sumerian deities Enlil and Enki would settle the dispute. In India, the hymns of the *Rigveda* and later the *Mahābhārata* are examples of early usage of dialogue as a literary genre. The latter contains the Bhagavad Gita, which features the famous dialogues between Lord Krishna, disguised as a charioteer, and the Pandava prince Arjuna. These dialogues are shown in visual artistic form in Figure 1.

**Figure 1.** Painting of the Bhagavad Gita's dialogue between Krishna and Arjuna

Credit: Gift of Dr. and Mrs. Edwin Binney. 3rd Copyright: San Diego Museum of Art. Downloaded from Wikimedia Commons. https://commons.wikimedia.org/wiki/File:Dialogue_between_Krishna_and_Arjuna_on_the_battlefield_of_Kurukshetra_(6124590103).jpg

## The polylogue as conversation and text

In her article "Introducing polylogue", Kerbrat-Orecchioni (2004) restricts the category of dialogues (or "dilogues" as the author prefers to name them in an effort to rectify the etymological confusion brought about by the prefix dia-) to dyadic interactions viewed as conversations between two individuals. Multi-party conversations, on the other hand, termed by the author as "multi-participant interactions" among "real live individuals", belong to the category of polylogues and include trilogues (a conversational triad), tetralogues, etc. Kerbrat-Orecchioni (2004) points out that polylogues can vary greatly not only in number, but also in nature. They are structurally highly complex, as well as unstable and unpredictable in their organization. This, in her view, makes polylogues extremely difficult to formalize.

It seems to us that it is precisely the complexity and apparent unpredictability of polylogues that make them so valuable to culture and, as we will discuss below, to the mental lexicon. In culture, and particularly in text, the polylogue enables us to span time and location and to maintain a heterogeneity of perspective that can be transmitted over generations and which allows distinct perspectives to be presented so that their significance, perhaps overlooked at a particular point in time, can play a key role at a later date.

Perhaps the best example of such a diachronic polylogue created through accretions of textual commentaries is the Jewish Talmud. This is the core text of Rabbinic Judaism and can be seen as a polylogue that spans well over a thousand years. The text of the Talmud is comprised of two components – the Mishna, which is a legal text committed to writing in Hebrew around 200 CE and the Gemara, which is a commentary written in Aramaic, completed around 500 CE. To this text, many other commentaries were added over the next thousand years. This diachronic polylogue was first printed around 1520 CE in Venice by the Catholic Belgian printer Daniel Bomberg. The Bomberg print layout placed the Talmud text in the center of the page (with the legal Mishna text above and the commentary Gemara text below. This text block was surrounded by a margin of white space which itself was surrounded by commentaries by the most prestigious scholars dating from the 11th to 15th centuries. Those commentaries, in turn, were surrounded by another margin of white space, and then by additional commentaries. The hierarchy among commentators corresponds to their proximity to the core text. This arrangement of a page of Talmud is shown in Figure 2.

Noteworthy also among the many Medieval Jewish scholars who produced textual commentaries is the work of 13th century Paula Dei Mansi of Verona, the first known Jewish female scholar and scribe, who translated the Jewish Bible from Hebrew into Italian – and annotated it with her own scholia (Taitz & Tallan, 2003).

**Figure 2.** A page of the Talmud. The core Talmud text is shown in the center. The text is surrounded by commentary, including that of Rabbi Sholomo Yitchaki (1040–1105) in the right inner margin. (Source: https://commons.wikimedia.org/w/index.php?curid=31718332)

Despite her 'lower' status in society, here is a scholarly woman who has left a mark in a domain considered solely the prerogative of men.

Scholarly textual polylogues were also prevalent in and across other traditions and thrived particularly in cities in which cultures interacted. In the Middle Ages, Eustathios of Tessalonike, a Greek 12th century scholar, commented on the Homeric Epics. He taught rhetoric in Constantinople and was in time appointed Archbishop of Thessalonike (today's Thessaloniki, or Thessalonica). An avid reader well-acquainted with classical scholarship, his Homeric scholia (plural of Ancient Greek *σχόλιον*, 'comment', interpretation') on both the Iliad and the Odyssey were characteristically commentaries reproducing previous commentaries, with the occasional addition of some original commentary. This exemplifies extraordinary transitional times: a time in which a high-ranking Christian cleric could devote his energies to commentary on a heathen epic.

Cordoba and Toledo in Muslim Spain of the middle ages were also cities of cultural interaction. In his 2003 book "Aristotle's Children", Richard Rubenstein examines the events and interactions that surrounded the re-introduction of Aristotle's writings to European thought. This re-introduction was, to a great extent, the result of translations of Arabic texts overseen by Francis Raymond de Sauvetât who was Archbishop of Toledo between from 1126 to 1151 CE. Crucially, the members of his Toledo School of Translation did not only translate the Arabic versions of Aristotle's writing into Latin – they also translated the textual polylogues of Muslim and Jewish scholarship. These included the work of the great scholars Avicenna, Averroës, and Maimonides. Thus, through this, Europe re-inherited not only the writings of a single great philosopher. It also inherited the tradition of philosophical scholarship.

In a very real way, the history of science and scholarship is the history of the textual polylogue. It is the polylogue that allows ideas to converge, diverge, lay dormant, and re-awake. In a historical perspective, then, polylogues appear to have taken the form of mainly textual commentaries, rather than that of a 'conversation' – oral or written – between simultaneously present protagonists. A type of interpretation, rather than disputation. In this way, the notion of polylogue is related to Julia Kristeva's notion of intertextuality (Kristeva, 1980). Intertextuality, as a construct, draws attention to the fact that the production and interpretation of a text depend on a language user's knowledge of other texts (de Beaugrand & Dressler, 1981).

But now, the internet revolution has given birth to a new and extraordinarily powerful form of polylogue. Social media platforms have become polylogue spaces that are unprecedented in their inclusiveness and impact. Polylogal web postings may be conceptually far removed from the traditional literary polylogal commentary found in the margins of manuscripts, incunabula (i.e., early typographical books) and even newer forms of (re)printed works. They are, nevertheless, polylogues – still written commentaries, but now genuinely conversational. These new polylogues have true interactive spontaneity. They involve almost instantaneous recording of response, with its often explosiveness and stancetaking. And, as is currently expected, they will remain available for a very long time.

It is difficult to fully appreciate the extent to which such developments have expanded the quantity of polylogue information that is created daily. Taking Twitter use as an example, at the time of writing, over 9,000 tweets are sent per second (Twitter stats, 2020). Using an average of 34 characters per Tweet and 6 characters per word, this would result in 51,000 words being tweeted every second. It would thus take Twitter 11.5 seconds to produce the number of words in Tolstoy's War and Peace.

This explosion of information has transformed the fields of natural language processing, text linguistics, and discourse analysis. It has also created the need for new computational tools to serve the needs of researchers of polylogues and other texts. In the section below, we explore the use of one such tool, Structural Topic Modeling, and apply it to the polylogue of mental lexicon research.

## Structural topic modeling applied to the mental lexicon as polylogue

As we have stated above, the entirety of the scientific enterprise may be seen as a specific polylogue in which common characteristics of style, evidence, and assumptions about truth and progress may be found. Polylogues that focus on content, however, tend to be at the level of research community – groups of scholars that have substantial overlap in the journals that they read, the conferences that they attend, and, very often, the people that they know. We have focused on one such group, the community of mental lexicon researchers.

Mental Lexicon research, as a domain of inquiry, may be said to have developed as a branch of psycholinguistics in the 1970s (but see Levelt, 2013 for evidence that psycholinguistics emerged much earlier). We have chosen the past 30 years of research (from 1990 to present) as a span of time within which to examine the topics that have emerged as dominant within the research domain. Thirty years is also a span of time that is sufficient in length to enable us to examine how the relative dominance of topics may have changed. This is a key component of our analysis because, in many important aspects, scientific production of knowledge is a time-series (Crispi & Geuna, 2008). It is a chronologically ordered, mostly incremental process, in which the future outcomes are contingent on the outcomes in the past and the present. Thus, when talking about the future of a scientific field, an overview of its track record is not only a homage to tradition, but also a practical necessity. In our analysis of the current state of research on the Mental Lexicon over three decades, we identify the most prominent topics within the research domain and their temporal dynamics. We do this by examining the statistical regularities of language use in scientific publications on the Mental Lexicon.

When humans engage in the task of categorizing documents into topics, there is a substantial danger of subjectivity and experimenter bias, especially if the categorization is conducted by adherents of the field under analysis. Fortunately, within the field of natural language processing, the machine-learning technique of *Topic Modeling* has been developed. This technique takes text documents as input and provides automated solutions to topic categorization problems without resorting to researchers' intuitions (Blei, 2012; Griffits, Steyvers, & Tenenbaum, 2007; Rosen-Zvi, Griffiths, Steyvrs, & Smyth, 2004).

The core of any type of a topic model is a probabilistic algorithm that examines patterns of co-occurrences of words within and across documents (Blei, Ng, & Jordan, 2013). It has long been known that semantically similar words tend to co-occur within a small text window (Firth, 1957). Thus, documents that share a larger number of identical or semantically similar words tend to be thematically related. In *Topic Modeling*, each topic is defined as a probability distribution over words (e.g., *book* will have a higher probability of belonging to the same topic as the words *journal, laptop* and *theatre* and a lower probability of belonging to the same topic as *engine, metal*, and *friction*). The algorithm assigns to each word a probability of being associated with each of the topics. A topic is defined as a collection of words that have a relatively high probability of co-occurring in the same documents, and a lower probability of co-occurring with words that form other topics. Typically, researchers name a topic based on the words that have the highest probability of association with this topic (e.g., possible names for word collections in our example above would be 'writing' vs 'engineering').

Topic modeling typically starts from assuming a random uniform distribution of topics over words and a uniform representation of topics in the documents. An iterative process updates probability distributions and assignments of words to topics and topics to documents by factoring in the statistics of word co-occurrence (for the underlying method, see Blei et al., 2003). The process continues until no further improvement can be achieved in topic assignments, or the maximum number of iterations is reached. While the actual algorithmic implementations of topic modeling vary, their common goal is to infer topics based on textual data and annotate documents as more or less pertinent to an inventory of topics. Each document represents a mixture of topics rather than just a single topic. For instance, it may contain 80% of book-related words, 10% of engine-related words and smaller percentages of other topics: this distribution would determine the semantic field of "writing" as a predominant topic of the document, with smaller traces of "engineering". Topic modeling is thus indispensable for automatically organizing and summarizing data archives, including the archives of scientific literature (e.g., Blei & Lafferty, 2007).

An additional analytical advantage comes from a specialized machine-learning technique called *Structural Topic Modeling*. This technique makes it possible to include covariates into the model, and thus study how these covariates influence either topic prevalence (popularity) in the corpus, or topic content (collection of words that constitute a topic). Multiple covariates are allowed, and they can be categorical (e.g., how topics differ in the prose written by male vs female authors), continuous variables (e.g., how does topic popularity change over time) or interactions between variables. Structural topic modeling differs from other solutions in that it allows topics to be correlated, assumes that "each document has its own prior distribution over topics, defined by covariate X rather than sharing a global mean"; and allows word use within a topic to vary as a function of a covariate (Roberts et al., 2014, page 1067).

In this chapter, we use structural topic modeling to identify topics that characterize scientific production of research into the Mental Lexicon and track their change over time (for precedents see Hall, Jurafsky, & Manning, 2008; Cohen-Priva & Austerweil, 2015). Thus, we are interested in the change in the prevalence of the topics rather than their content. Our materials consist of abstracts to all papers published in *The Mental Lexicon* journal since its inception in 2006 ($n$ = 199) as well as abstracts of all papers that were listed in the Web of Science database and contained the keyword "mental lexicon" ($n$ = 1104). The total pool of 1,303 abstracts covered the period from 1990 to 2017 and was supplied as a set of input documents to the structural topic model implemented in the stm package (Roberts, Stewart, Tingley, Lucas, Leder-Luis, Gadarian, Albertson, Albertson, & Rand, 2014) in the statistical software R 3.4.3 (R Core Team, 2017). As meta-data to each document, we provided the year in which it was produced and the data source (*The Mental Lexicon* journal or the *Web of Science*).

The prepDocuments() function in the stm package implements basic text-processing utilities. Thus, words in all documents were converted to lower case, and punctuation, numbers and function words were removed. Words were then 'stemmed', i.e., wordforms like *walk, walked, walking, walkable*, and *walker* and their occurrences across documents were assigned to the same lemma *walk*. The implementation of stemming in stm consists of a relatively crude, morphologically-naïve method of truncating word ends until a common stem is found: it strips inflectional and most derivational morphemes and occasionally identifies as a lemma a truncated word form (e.g., *bilingu* for bilingual, bilingualism, etc). This procedure reduced the dataset to 8,913 word types and 93,076 tokens. A further trimming step was to remove all words that occurred in fewer than 10 documents: this step attenuates the spurious influence of low-frequency terms on topic selection and assignment. All 1,303 documents were retained, with 1,283 remaining word types and 78,912 word tokens.

A structural topic model was fitted to this dataset, with two covariates: source of publication and year of publication since 1990 (ranging from 1 to 27) and default control parameters (Spectral type of initializing the topic assignments, and an allowed maximum of 75 iterations for the model to converge). Given a relatively small number of input documents and recommendations of the stm package (Roberts et al., 2014), we set the desired number of topics to 10. Since analysis did not reveal any topical difference between the two sources of abstracts, we report a model with one covariate: year of publication.

Table 1 reports the topics identified by the structural topic model. The outcomes are presented as collections of words that are most diagnostic of each topic. Different metrics are used to quantify word association to topics. We report three of them because they provide complementary insights into the topic content: (a) words with the highest probability in the topic, (b) words that are both frequent and exclusively found in the given topic, and (c) words that are particularly frequently found within the topic given their empirical distribution. The top five words are reported for each metric. The order of topics is arbitrary and not informative. For the ease of reference, we provided a single-word label to each topic. These labels are nothing but short-cuts, and we warn against interpreting them as full representations of respective topics.

**Table 1.** The *Mental Lexicon* topics represented as collections of most diagnostic words

| Topic | Highest probability in topic | Frequent & exclusively in topic | Frequently in topic, given their distribution |
|---|---|---|---|
| 1. *Theory* | mental, lexicon, model, research, cognit | express, health, social, psycholog, approach | belief, feel, depress, health, care |
| 2. *Masked priming* | prime, morpholog, word, experi, effect | prime, root, mask, morpholog, hebrew | prefix, prime-target, mask, prime, soa |
| 3. *Bilingualism* | languag, speaker, english, bilingu, nativ | learner, bilingu, nativ, colloc, speaker | efl, learner, foreign, colloc, bilingu |
| 4. *Word recognition* | word, frequenc, effect, lexic, recognit | frequenc, neighborhood, competit, respons, recogni | deaf, neighborhood, densiti, neighbor, estim |
| 5. *Neuroscience* | process, lexic, activ, semant, brain | left, patient, tempor, brain, neural | lobe, gestur, magnet, gyrus, cortex |
| 6. *Children* | children, group, read, age, studi | children, age, abil, skill, score | peer, dyslexia, year-old, dyslex, month |
| 7. *Speech* | speech, phonolog, word, represent, model | speech, phonet, syllabl, sound, phonem | tone, tonal, acoust, voic, syllabl |
| 8. *Inflection* | form, verb, inflect, regular, morpholog | irregular, regular, verb, inflect, plural | irregular, participl, plural, tens, singular |
| 9. *Chinese/Japanese* | word, compound, read, mean, chines | compound, chines, sens, constitu, charact | colleg, compound, polysem, self-pac, chines |
| 10. *Syntax* | noun, semant, categori, name, context | gender, grammat, phrase, idiom, categori | bare, idiom, gender, block, phrase |

Table 1 reveals that the Mental Lexicon topics cover both the theoretical position of this research field in a broader scientific, social, clinical and cultural context (Topic 1) and more practice-grounded topics like language acquisition and processing in children and bilinguals (Topic 3 and Topic 6). Neuroscience, speech science, and linguistic research beyond the word level (e.g., syntax, idiomatic expressions) form their own topics (Topics 5, 7, and 10). Several specialized research questions and even particular experimental paradigms proved influential enough to warrant their own topic: this was the case with inflectional morphology, especially irregular forms (Topic 8); Chinese and Japanese study, especially on compounding (Topic 9); and the masked priming lexical decision task (Topic 2). Finally, the methodological effort of studying lexical variables and tasks is reflected in Topic 4. While division into a greater number of topics might have highlighted additional corners of the Mental Lexicon landscape, we believe that the present inventory of topics offers a fair representation.

A critical question of interest is how these topics evolved over the last 30 years and what the current thematic and methodological preferences in the field are. A partial answer to this question comes from the estimation of how time (measured here as year of publication since 1990) affects the prevalence of our topics. The expected proportion for any given topic (out of 10) in the total corpus is 10%: a deviation from this number over time indicates either an increased or decreased topic popularity. We used the estimateEffect() function to quantify the influence of time on topic prevalence both as a nonlinear effect (using basic splines) and a linear effect. Figure 3 illustrates the temporal dynamics of topic change.

Figure 3 points to a very substantial increase in the number of publications on theoretical aspects of the mental lexicon (Topic 1), with a current estimated prevalence much higher than expected (18 vs 10%). We also observed a substantial increase since 1990 in studies on bilingualism, and teaching and learning foreign languages (Topic 3): this topic is currently somewhat more prevalent than expected in the The Mental Lexicon corpus (12%). A drastic decrease characterizes research that uses masked priming (Topic 2), studies on inflectional morphology (Topic 8), research on speech (Topic 7), and – rather unexpectedly – neuroscience of the mental lexicon (Topic 5). The current state of these four topics is at or just below the expected level of prevalence. The decrease associated with masked priming can, quite likely, be related to a diversification of methodology within the field over the temporal period. Similarly, the decrease associated with inflectional morphology may be seen as capturing a trend toward diversification to other morphological processes (e.g., compounding, derivation) and to other non-morphological lexical properties. It is worth noting a sharp increase in neuroscientific work in the last few years, reflected in the non-linear trend: it is possible that the tide has already turned, but the linear approximation is too rudimentary to mirror it.

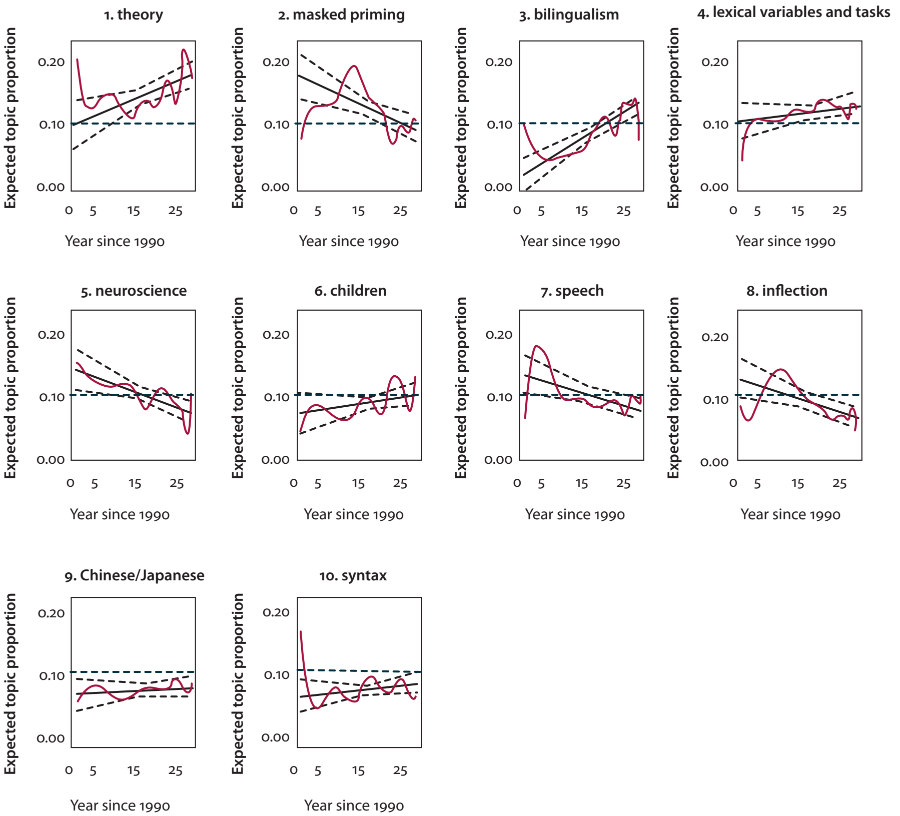



**Figure 3.** Change in topic prevalence since 1990. The nonlinear fit is shown in red, the linear fit in black (dotted lines mark the 95% confidence interval). The horizontal line in green shows the default 10% of the topic

Most other topics show a smaller amount of change. A slight nonsignificant increase is seen in the investigation of word recognition (Topic 4), language processing in children (Topic 6), as well as work on Asian languages (Topic 9) and syntax (Topic 10). Except for the word recognition topic, which is slightly more popular than expected, all other topics remain less prevalent than expected.

## The value of structural topic modeling in capturing polylogues and creating opportunities

Structural topic modeling such as that presented in Figure 3 produces a bird's eye view of the research field and its temporal dynamics as it emerges in the scientific publications that constitute this field. It offers a polylogue analysis. We believe that this analysis and the knowledge it produces is vital in at least three ways. First, the present-day state of the field can be taken to reflect its achievements and failures. A decrease in a topic's popularity may mean that the topic has been undeservedly abandoned, or it may mean that the key questions that engendered that topic have been successfully answered and research effort can be fruitfully reapplied to other endeavors. We leave such considerations regarding each topic to the reader.

Second, knowing where the field has moved so far enables us to predict its future. It appears that the themes and methods that boosted the Mental Lexicon as a field for many years – irregular inflection, or lexical decision with masked priming – are not on the forefront any longer. Looking toward the future, the trends suggest that one might expect to see more work that integrates the mental lexicon with broader research questions of linguistics and psychology, as well as a move away from conventional pools of self-selected, highly proficient, young undergraduate students and towards cohorts of children, L2 learners, and bilinguals.

Finally, the knowledge of the state-of-the-art gives researchers an opportunity not only to predict the future, but also to take concrete steps in shaping it. Boosts can be given to less popular topics by organizing special issues of journals or thematic workshops and symposia. Student training in under-utilized methods can be reinforced both in the classroom and through specialized learning opportunities (workshops, summer schools, online courses, etc.). Grant applications can benefit from these data to prove their novelty or demonstrate the need for additional examination of the topic. Understanding the landscape of topics and the relative effort invested in each, research may increasingly open itself to its subjects – children, clinical and aging populations, immigrants and language learners, as well as individuals highly variable in their language proficiency and demands for language use.

In all these ways, the polylogue of researchers may transcend the boundaries of learned discourse and embody itself as an evidence-based "doing of science". It is our hope that this polylogue continues and is joined by ever more voices.

## From the observers to the observed: The mental lexicon as polylogue

It is interesting to note that the tools that are used in structural topic modelling bear a striking similarity to the tools that are used within the field of mental lexicon research itself. The documentation and explanation of lexical frequency effects have been central to the field from its earliest stages. Similarly, the understanding of how words co-occur in language production and are co-activated in lexical recognition have been core desiderata in the field and have a served as foundations for the development of mental lexicon theory.

Seen in structural terms, it is perhaps not that surprising that the tools that can be used to gain insight into textual polylogues are tools that are also applied to the study of the mental lexicon. In both cases, of course, we are dealing with the analysis of words. But there is perhaps a deeper lesson that can be extracted from the comparison: Perhaps the mental lexicon itself is a special form of polylogue – one that is essentially cognitive in nature.

It has become clear that early conceptualizations of the mental lexicon as a dictionary in the mind (Aitchison, 2012) provided both a valuable metaphor and a liability for theory development. The dictionary metaphor suggested that the mental lexicon is a store of mental representations. On their own, however, representations do not actually do anything. Sitting on a shelf, a dictionary does not grow, it does not enrich the connections among its representations, it does not self-organize. In contrast, there is good reason to expect that the human mental lexicon does all of those things (Libben, 2020).

Our linguistic development is characterized by an extraordinary rate of vocabulary development as children and teenagers. And lexical acquisition continues throughout the lifespan as we learn domain-specific vocabulary and often entirely new languages. Key to this dynamicity is the polylogue itself. Recent research has highlighted the ways in which the mental lexicon functions as an integrated system. Changes that seem to be localized at the level of individual lexical entries become changes to the system as a whole. Thus, it may be the case that the characteristics of our lexical systems result from the encoding of our experiences with words – and therefore our experience with people.

Jarema and Libben (2007) defined the mental lexicon as *"the cognitive system that constitutes the capacity for conscious and unconscious lexical activity"*. Lexical activity includes, of course, the activities of reading, writing, listening, speaking, and signing. Importantly, lexical activity also includes changes in the cognitive system that results from updating, re-analysis, and the integration of new experience. Under this view, the core activity of the mental lexicon can be considered to be polylogal activity.

The discussion above suggests that the notion of a polylogue may have to play a special role in bridging levels of lexical activity within the field of psycholinguistics. Importantly, polylogues may offer a means by which we can bring together the study of words as both possessions of groups and possessions of individuals.

## The thematic polylogue

We began our discussion of polylogues as textual forms by reaching back to Antiquity. Polylogues appear to have had their origins in the stories that are the core identities of cultures. These form the narratives around which people cluster and form (ideally) cohesive communities. The stories were transmitted from generation to generation at first orally, and in due course, with the advent of writing systems and the development of writing tools (tablets, parchment scrolls, etc.) in written form. With curious and creative scribes adding their own thoughts and interpretation of events, polylogues were born.

As we saw in our discussion of the Talmud, the polylogue can also be constructed as what we might call a *thematic polylogue*. In this type of polylogue, there is typically a core text or statement or topic, and the polylogue is focused around that core. Currently, there are academic journals that are organized as thematic polylogues. The creation of thematic polylogues is also often the explicit goal of academic conferences, workshops, and panel discussions. This book is certainly an example of a *thematic polylogue*. It aims precisely at joining the long literary tradition of the polylogue genre construed as written debates that involve critical commentaries and thoughtful interpretations that, although inevitably tinted by one's Weltanschauung and theoretical stance, are constrained by their very format to serve 'the greater good' of knowledge advancement. The volume's scholia or commentaries cannot be inscribed in the margins of chapters in the manner in which ancient manuscripts were glossed. Rather, the six chapters of Polylogues on the Mental Lexicon are briefly commented upon by its contributors in turn, at the outset of each chapter. Together, these commentaries are intended to enrich, albeit also challenge, the viewpoint presented by the chapter's author(s) and to thus offer readers a more balanced perspective on current and future core issues in mental lexicon research discussed in the volume.

## References

Aitchison, J. (2012). Words in the Mind: An Introduction to the Mental Lexicon. 4th edition, Oxford: Wiley-Blackwell.

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, 55(4), 77–84. https://doi.org/10.1145/2133806.2133826

Blei, D. M., & Lafferty, J. D. (2007). A correlated topic model of science. *The Annals of Applied Statistics*, 1(1), 17–35. https://doi.org/10.1214/07-AOAS114

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3, 993–1022.

Cohen-Priva, U., & Austerweil, J. L. (2015). Analyzing the history of Cognition using topic models. *Cognition*, 135, 4–9. https://doi.org/10.1016/j.cognition.2014.11.006

Crespi, G. A., & Geuna, A. (2008). An empirical study of scientific production: A cross country analysis, 1981–2002. *Research Policy*, 37(4), 565–579. https://doi.org/10.1016/j.respol.2007.12.007

De Beaugrande, R. A., & Dressler, W. U. (1981). *Introduction to text linguistics* (Vol. 1). London: Longman. https://doi.org/10.4324/9781315835839

Firth, J. R. (1957). A synopsis of linguistic theory, 1930–1955. *Studies in linguistic analysis*.

Griffiths, T. L., Steyvers, M., & Tenenbaum, J. B. (2007). Topics in Semantic Representation. *Psychological Review*, 114(2), 211–244. https://doi.org/10.1037/0033-295X.114.2.211

Hall, D., Jurafsky, D. & Manning, C. D. (2008). Studying the history of ideas using topic models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (pp. 363–371). Association for Computational Linguistics. https://doi.org/10.3115/1613715.1613763

Jarema, G. & Libben, G. (2007). Matters of Definitions. In G. Jarema & G. Libben (Eds.) *Core Perspectives on the Mental Lexicon*. Oxford: Elsevier, pp. 1–12). https://doi.org/10.1163/9780080548692_002

Kerbrat-Orecchioni, C. (2004). Introducing polylogue. *Journal of Pragmatics* 36(1), 1–24. https://doi.org/10.1016/S0378-2166(03)00034-1

Kristeva, J. (1980). *Desire in language: A semiotic approach to literature and art*. Columbia University Press.

Levelt, W. J. M. (2013). *A history of psycholinguistics: The pre-Chomskyan era. Oxford: Oxford University Press*

Libben, G. (2020) What can we learn from novel compounds? In S. Schulte im Walde & E. Smolka, (Eds.) *The role of constituents in multi-word expressions: An interdisciplinary, cross-lingual perspective*. Berlin: Language Science Press (pp. 107–128).

R Core Team. (2017). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/

Roberts, M. E., Stewart, B. M., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S. K., … & Rand, D. G. (2014). Structural topic models for open-ended survey responses. *American Journal of Political Science*, 58(4), 1064–1082. https://doi.org/10.1111/ajps.12103

Rosen-Zvi, M., Griffiths, T., Steyvers, M., & Smyth, P. (2004, July). The author-topic model for authors and documents. In *Proceedings of the 20th conference on Uncertainty in artificial intelligence* (pp. 487–494). AUAI Press.

Taitz, E., Sondra Henry, S., & Tallan, C. (2003). The JPS Guide to Jewish Women: 600 B.C.E.to 1900 C.E. Philadelphia: The Jewish Publication Society.

Tweets per second (2020, July 3). Retrieved from https://www.internetlivestats.com/one-second/#tweets-band

Wei, W. W. (2006). Time series analysis. In *The Oxford Handbook of Quantitative Methods in Psychology: Vol. 2*.

CHAPTER 2

# Meta-megastudies

James Myers
National Chung Cheng University

## Thesis

Cross-linguistic psycholinguistics has traditionally been conducted on small convenience samples, making it difficult to disentangle the effects of typological features that the sampled languages happen to share or that are cross-linguistically correlated. Extending the megastudy approach, which addresses within-language lexical confounding by treating lexical items as a random variable in regression analyses, meta-megastudies test a sufficiently large and varied cross-linguistic sample to allow language to be treated as a random variable as well. Meta-megastudies thus complement techniques like artificial lexicon learning in establishing which typological characteristics are associated with which lexical processing strategies, as revealed by statistical interactions between language type and lexical variables. Meta-megastudy language sampling is simpler than in linguistic typology, since historical relatedness and geographical proximity are unlikely to affect processing independently of the typological features themselves. While meta-megastudies raise other logistical challenges, their feasibility is already implied by the expanding scope of traditional cross-linguistic studies, an expansion made possible by the Web-enabled streamlining of international collaboration. True meta-megastudies will depend on decentralizing this Web-based approach via tools that encourage otherwise independent research groups not only to follow consistent experimental protocols, but also to share their results with the larger research community.

## Commentaries on Myers thesis

### Ray Jackendoff commentary on Myers thesis

It is by all means valuable to construct a database of crosslinguistic findings on language processing, as Myers suggests. It is also important to devise a platform for creating and running parallel experiments on a wide variety of languages. On

https://doi.org/10.1075/z.238.02mye

the other hand, it is also important to remember where the questions come from that motivate good experimental work: ideally, detailed and insightful theories of linguistic structure and of language processing. Such theories do not come (exclusively) from statistics run over a huge corpus of text, of speech, or of experimental outcomes. They depend on well-trained researchers thinking hard about particular problems raised by particular data in particular languages (an issue that comes up again in my commentary on Tucker and Ernestus).

For instance, consider the well-trodden issue of storage vs. computation of morphologically complex word forms. How does this issue play out in languages with exuberant productive morphology? Is it different in languages like Turkish, with its relentlessly suffixal morphology, and languages like Navajo, with templatic morphology? Similarly, how might VSO languages be processed differently from SVO, SOV, or free word order languages? How does the processing of right-headed relative clauses differ from that of left-headed relative clauses? Psycholinguistic questions like these and myriad others emerge from typological research within linguistic theory. But they cannot be tackled meaningfully without researchers (a) who are intimately familiar with the languages at issue, (b) who are sufficiently grounded in linguistic theory and linguistic typology so as to be able to formulate insightful questions, (c) who therefore are capable of fashioning appropriately sensitive stimuli for experimental research, and (d) who are furthermore capable of interpreting experimental results and bringing them to bear on general issues. It is not clear that minimally trained "citizen scientists" can be capable of such demanding tasks.

## Dorit Ravid commentary on Myers thesis

Grammars, under any theory currently held by linguists and psycholinguists, are not exactly "learned social conventions" but rather cognitive entities, constituting the cognitive organization of one's experience with of language. There is no doubt that grammars exist within speech communities, but relegating a grammar to the social rather than to the socio-cognitive realm misses the point about the essence of grammar. According to Bybee (2006), the general cognitive capabilities of the human brain, which allow it to categorize and sort for identity, similarity, and difference, go to work on the language events a person encounters, categorizing and entering in memory these experiences. The result is a cognitive representation that can be called a grammar. The fact that the mind responds to frequencies of patterns in the speech community does not mean that grammatical properties and their learning are solely the result of distributional features in that community. This assumption disregards the nature of grammar as a systematic framework conveying thought that always emerges in humans, with general features indifferent to the specific speech community. Thus, the need to take into account the deictic configuration of the discourse event – speaker, addressee and absentee roles, proximal and distal location, and temporal relativity – determines the basic structure of grammars, with lexical-grammatical features that reflect this configuration. For example, the grammars of many languages encode (non-biological) gender features as information-prominent markers formally organizing the message. While the details of the application of grammatical features to specific words or categories are definitely affected by distributions in the speech community (e.g., *dived* or *dove*), these details do not constitute the essence of "grammar".

### *Reference*

Bybee, J. (2006). From usage to grammar: The mind's response to repetition. *Language*, 82, 711–733. https://doi.org/10.1353/lan.2006.0186

## Russel Richie commentary on Myers thesis

The author argues that continued improvement in psycholinguistic theory – currently aided by development of megastudies on lexical and other linguistic items in a small number of languages – will require *meta-megastudies*, development of large datasets across *many and diverse languages*. My main comment on this thesis is to draw parallels with similar problems related to data sparsity in natural language processing, and to suggest how psycholinguists and computational linguists/NLP engineers can work together to improve their respective fields.

First, cross-linguistic data sparsity – among other causes – leads to problems in a number of NLP domains. For example, certain source-target language pairs (e.g., English->French) enable better machine translation than others (e.g., English->Xhosa), as anyone who has used Google Translate can attest. And even in a supposedly "high-resource" language like English, automated speech recognition on YouTube performs better on standard dialects (Californian English) than on nonstandard dialects (Scottish English), and even, in at least one case, better on male speech than on female speech (Tatman, 2017). In both cases (MT and ASR), the differences in task performance across language/dialect communities are commonly attributed to differences in quantity or quality of training data – pairs of English-French translations with which to train MT algorithms are more common than English-Xhosa, and groups with less power (women and linguistic minorities) are undersampled in common speech corpora. Simply put, socially less-powerful groups' language tends to be undersampled, and thus automatic systems do not serve them as effectively as more powerful groups.

Thus, just as psycholinguistic theory stands to gain from larger and more diverse datasets of linguistic data, so, too, do natural language processing tools. And so psycholinguists and computational linguists/NLP developers can achieve synergy by sharing resources and collaborating. To an extent, this already happens. For example, lexical norms developed for psycholinguistics (e.g., Wilson, 1998; Warriner, Kuperman, & Brysbaert, 2013) are being used for a range of NLP tasks, from sentiment analysis (e.g., Buechel and Hahn, 2017) to automated assessment of sentence specificity (Li and Nenkova, 2015). However, this cross-fertilization could happen more, and particularly in the other direction, by psycholinguists using NLP developers' datasets and tools. For example, Paetzold and Specia (2016) have recently shown that psycholinguistic properties of words including familiarity, age of acquisition, concreteness, and imageability can be predicted from word embeddings, with correlations between predicted and actual scores approaching .9. When word embeddings models already exist (as they do in several European languages; for example, see the SpaCy NLP toolkit) or the corpora necessary for such construction can be obtained, this approach may use fewer resources than manual/crowdsourced construction of psycholinguistic datasets in currently less-resourced languages.

Finally, as the author notes in the full paper, a challenge of conducting meta-megastudies is the massive human and financial cost. To meet these costs, in addition to volunteer 'citizen scientists', psycholinguists might take a page from computational linguists' and NLP developers' playbook and collaborate with – and apply for funding from – private firms. For example, it is in language teaching companies like Rosetta Stone and Duolingo's financial interest to develop larger collections of teaching materials across a diverse set of languages, and thus such firms may make for good partners in the effort to develop larger meta-megastudies.

## *References*


Buechel, S. & Hahn, U. (2017). A Flexible Mapping Scheme for Discrete and Dimensional Emotion Representations: Evidence from Textual Stimuli. In *CogSci 2017 – Proceedings of the 39th Annual Meeting of the Cognitive Science Society. London, UK, July 26–29, 2017*, pages 180–185.

Li, J. J., & Nenkova, A. (2015). Fast and Accurate Prediction of Sentence Specificity. *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2281–2287.

Tatman, R. (2017). Gender and Dialect Bias in YouTube's Automatic Captions. *Ethics in Natural Language Processing*. https://doi.org/10.18653/v1/W17-1606

Warriner, A. B., Kuperman, V., & Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13, 915 English lemmas, *Behavior Research Methods*, 1191–1207. https://doi.org/10.3758/s13428-012-0314-x

Wilson, M. (1988). MRC psycholinguistic database: Machine-usable dictionary, version 2.00. *Behavior Research Methods, Instruments, & Computers*, 20(1), 6–10. https://doi.org/10.3758/BF03202594

## Benjamin Tucker commentary on Myers thesis

Myers presents an intriguing new domain of research topics that can be generated by the proposed meta-megastudies. While large-scale meta-megastudies will be logistically challenging, they promise great theoretical benefits and are becoming ever more feasible via Web-based coordination. The described website in the paper offers a platform which can be applied to many possible tasks and also provides a way for diverse research groups to coordinate efforts among themselves and with independent researchers. The notion of a meta-megastudy is incredibly attractive for a researcher seeking to understand the processing of language. Ultimately though, we may never solve the "language-as-a-fixed-effect fallacy", which is likely due to our inability to randomly sample the set of languages available or to sample all living languages. Further, there will always be some bias in our language sampling, just as there is in our participant sampling. The sampling issue will continue to be a challenge even using web based solutions as noted by the author. For example, web based methods will cater to certain language families/languages better than others, due to factors often outside the control of the researcher like literacy or access to the internet. This is not to say that we should not pursue meta-megastudies and that there is no value in these proposed types of studies, but the limitations should be kept in mind.

Some other considerations with regard to web studies. The more accessible we make the studies the more we lose control over many parameters, like size of font or size of screen in a visual experiment. Also a question should be raised here, should these meta-megastudies also be done in the auditory domain? A quick overview of the megastudy literature shows that there is a strong bias toward visual experimental paradigms (Ernestus & Cutler, 2015; Jarema et al., 2015). Would a web-based experiment platform also perpetuate this bias? The excellent "WorldLikeness" website does allow for auditory presentation of stimuli but there is very little information about how this might be implemented on the website. This topic also raises the question of what visual only studies tell us about language processing and whether there are differences between auditory and visual language processing. The literature comparing visual to auditory domaines is relatively sparse (Bradley & Forster, 1987; Taft, 1986). Are the responses for word likeness different for orthographically presented words as opposed to auditory presentation? The author also discusses word naming studies and it seems in this case that the author is specific about typing in words. How would the authors go about the collection of actual vocal word naming latencies? Would the variability in microphones and headphones pose problems for data collection, particularly with peripheral hardware that largely cannot be controlled?

Further, these orthographically based experiments bias the sampling to large literate communities and leave out languages/dialects without an official writing system.

While I believe there are many challenges in pursuing the types of meta-mega-studies proposed by the author, I do believe that it is worth pursuing these topics. It is in the pursuit of general factors about psycholinguistics using a much wider array of languages that we will make progress and that solutions to many of the challenges identified and unidentified will present themselves. If researchers adopt the proposals in this paper, I have great hope for the coming decade of research and potential for new discoveries about the lexical representation of language.

*References*

Bradley, D. C., Forster, K. I. (1987). A Reader's View of Listening. *Cognition*, 25, 103–134. https://doi.org/10.1016/0010-0277(87)90006-0

Ernestus, M., & Cutler, A. (2015). BALDEY: A database of auditory lexical decisions. *The Quarterly Journal of Experimental Psychology*, 68(8), 1469–1488. https://doi.org/10.1080/17470218.2014.984730

Jarema, G., Libben, G., Tucker, B. V. (2015). The integration of phonological and phonetic processing: A matter of sound judgment. In: Jarema, G., Libben, G. (Eds.), Benjamins Current Topics. Vol. 80. John Benjamins Publishing Company, Amsterdam, pp. 1–14. https://doi.org/10.1075/bct.80.002int

Taft, M. (1986). Lexical access codes in visual and auditory word recognition. *Language and Cognitive Processes*, 1 (4), 297–308. https://doi.org/10.1080/01690968608404679

## Chris Westbury commentary on Myers

In 2001, Pauelsu, Démonet, Fazio et al. published an article in *Science* showing Italian dyslexics showed less reading impairments than French and English dyslexics. The reason they offered was straightforward: as anyone who has ever been to Italy knows, Italian has a very transparent orthography. It is just a lot easier to read Italian than it is to read English or French. The reading difficulty that characterizes dyslexia is partially a reflection of aspects of the language that the dyslexic is dealing with.

As well as orthographic transparency, languages differ on a myriad of other factors: average word length, morphological productivity, morphological family sizes, role of word order, number of orthographic and phonological neighbors, word frequencies, overall vocabulary size, Ngram and Nphone distributions, clause and sentence lengths, which words are marked as profane, and many more. This is why it makes sense to call for meta-*mega*studies, rather than merely more multi-linguistic studies between languages. It is hard to construct cross-lingual experiments with matched stimuli. However, using megastudies it will be possible to compare general trends across very large word sets. For example, one will be able to answer questions about whether specific linguistic effects are more important (more heavily weighted with respect to explaining some performance measure) in one language than in another.

There is a potential practical outcome to this endeavor. When we can assign relative weights with respect to performance measures of the same aspects of many different languages, we might be able to develop a systematic measure of 'how difficult' a language is, perhaps at least with respect to another given language if not universally. We might be able to come up with a systematic, wide-ranging, and data-driven measure of linguistic competence. When we have characterized the ranges of relative weightings of different aspects of language for native speakers of a language, it should be possible to characterize how closely a second-language learner approaches that gold standard. This could be a much better method of characterizing language competence than methods currently used, which generally depend upon measuring performance with fairly narrow samples of spoken and written text.

As an example, native sensitivity to, say, orthographic neighborhood size presumably both ranges between some bounds, and serves as implicit measure of vocabulary knowledge (performance on a task cannot be affected by a word's neighbors if you do not know that words neighbors). Native-level second-language speakers should, at their best, fall within the very same range, or possibly within a different known range for perfect bilinguals of the two languages concerned. A person who can speak or write competently in some restricted setting might betray a vocabulary deficit from their insensitivity to such implicit measures of vocabulary similarity, just as they might betray a deficit in vocabulary access by failing to show an expected frequency effect of a known magnitude. Such tests need not be limited to single word measures of course: performance on connected text can equally be characterized. Since words and sentences have many characteristics, every word access that can be measured can participate in estimating multiple different measures. We don't need to have a test for sensitivity to word frequencies, and a separate test for sensitivity to morphological structure, and so on, since we can estimate sensitivity to multiple measures from a single word set.

When the proper parameters are known, one can imagine using eye-tracking technology to get fine-grained measures of sensitivity to many different aspects of single word and sentence reading simultaneously. A carefully defined text of sufficient length might contain enough information if closely measured to provide not just a very good estimate of general reading ability, but a quite detailed report of focal weaknesses in a second-language reader's mastery. Production is always more difficult to test, but with computer speech recognition of words and fine-grained measurement of production speeds, one might be able to equally characterize a range of the relevant aspects of productive ability from a reasonably short sample of freely-produced text.

*References*

Paulesu, E., Démonet, J. F., Fazio, F., McCrory, E., Chanoine, V., Brunswick, N., Cappa, S. F., Cossu, G., Habib, M., Frith, C. D. and Frith, U. (2001). Dyslexia: cultural diversity and biological unity. *Science*, 291(5511), 2165–2167. https://doi.org/10.1126/science.1057179

## The article

The "language-as-fixed-effect fallacy" occurs when a psycholinguist runs experiments on one or two languages, then draws inferences about human language processing that go beyond the specific languages being tested. At least this is what the term should mean. The solution to the problem is similar to that recommended by Clark (1973) for the more familiar sense (where the "language" being "fixed" is actually a set of test items): treat language as a random variable. In other words, just as a megastudy (Balota, Yap, Hutchison, & Cortese, 2012) can generalize about word processing in some specific language by testing a large sample of words from that language, so too should one be able to generalize about human word processing by testing a large sample of human languages. I call this a meta-megastudy.

The rest of this essay is devoted to unpacking this simple idea. Befitting the theme of this book, my goal is to raise questions, not to answer them. I first show how a more sophisticated approach to cross-linguistic research would benefit psycholinguistics by providing an increased ability to disentangle partially confounded variables, and then show how such an approach could be made practical by exploiting the Web to coordinate the running of independent experiments and the sharing of results.

### The need for meta-megastudies

The core logic is simple: if megastudies are good for the study of individual languages, then meta-megastudies are good for the study of cross-language differences. By now the benefits of megastudies to lexical research are well known (Balota et al., 2012; Keuleers & Balota, 2015). A lexical experiment can be thought of as an attempt to study how processing is affected by lexical variables, under various conditions. The problems are that lexical items already exist and so cannot be experimentally manipulated, that the variables describing them may not be categorical (e.g., grammatical vs. content words) but gradient (e.g., lexical frequency), and that they may also be partially confounded with each other (e.g., grammatical words tend to have higher frequencies than content words). Forcing a factorial design by cherry-picking subsets of matched items risks experimenter bias (Forster, 2000) without eliminating confounds with uncontrollable or unknown variables (Cutler, 1981). Megastudies attempt to ameliorate these problems by treating experiments as database-generation machines, permitting future researchers to analyze the results in terms of any variable they can think of, using statistical techniques, particularly regression, to tease apart partial confounds. Megastudy databases exist in an ever-growing number and variety of languages, including English (Adelman et al., 2014; Balota et al., 2007; Hutchison et al., 2013; Keuleers, Lacey, Rastle, & Brysbaert, 2012), Dutch (Ernestus & Cutler, 2015; Keuleers, Diependaele, & Brysbaert, 2010), French (Ferrand et al., 2010), Chinese (Sze, Liow, & Yap, 2014; Myers, 2015; Tse, Yap, Chan, Sze, Shaoul, & Lin, forthcoming), and Malay (Yap, Liow, Jalil, & Faizal, 2010).

Nevertheless, within any particular megastudy, the language itself remains a preexisting, nonmanipulable independent variable. Thus if we want to study how processing is affected by experience with a language as a whole, cross-linguistic comparisons seem like a relevant source of information. Psycholinguists have long recognized the importance of cross-linguistic studies (Bates, Devescovi, & Wulfeck, 2001; Costa, Alario, & Sebastián-Gallés, 2007; Cutler, 1985; Evans & Levinson, 2009; Jaeger & Norcliffe, 2009); ProQuest's Language and Linguistics Behavior Abstracts shows well over 9,500 hits for psycholinguistic* AND (cross-linguistic* OR typolog*). Yet many cross-linguistic studies are concerned less with typology than with within-speaker multilingualism (including some that use megastudy methods, e.g., Keuleers, Stevens, Mandera, & Brysbaert, 2015; Lemhöfer et al., 2008), and of the studies that do compare separate speech communities, virtually all involve just two languages, presumably for practical reasons. The challenges of cross-linguistic psycholinguistics are illustrated by the unusually ambitious study of Bates et al. (2003): collecting naming responses on 520 object pictures from speakers of seven different languages (Bulgarian, English, German, Hungarian, Italian, Mandarin, Spanish) required an astonishing 22 authors, affiliated with 10 different institutions.

Just like the lexical items within a language, languages differ from each other in many different variables, some categorical (e.g., whether the language is written with an alphabetic orthography) and some gradient (e.g., the size of the syllable inventory), and these variables may also be partially confounded (e.g., with some historical exceptions, syllable-based orthographies tend to be restricted to languages with relatively small syllable inventories: DeFrancis, 1989). The key justification for meta-megastudies is thus essentially arithmetical: just as for single-language megastudies, running a regression across multiple data points is a much more effective way to disentangle partial confounds than testing two data points at a time.

Suppose there are ten cross-linguistic variables at play in some psycholinguistic debate. According to a common rule of thumb (e.g., Vittinghoff & McCulloch, 2007), data from around 100 languages (ten times the number of variables) may suffice to tease them apart in a regression analysis, even if correlated (O'Brien, 2007). While collecting data from so many languages is clearly beyond the powers of a conventional research team, the burden could be split across 100 separate teams, each generating its own single-language database for its own purposes, and the researchers who use the entire cross-linguistic database to study the ten variables may not even be any of the original data collectors. This is in fact the standard procedure in linguistic typology, where researchers use compilations of existing grammars to test hypotheses that the grammar writers themselves may never have dreamed of (e.g., Haspelmath, Dryer, Gil, & Comrie, 2005). This procedure has proven so reliable that many of the observations first made tentatively on the basis of small cross-language samples (Greenberg, 1963) still hold up in more sophisticated analyses of much larger databases (e.g., Dryer, 1992, 2011).

By contrast, a conventional pairwise exploration of this same language sample would be both more burdensome and less satisfying. While roughly the same number of researchers would be required (at least one native speaker per language), they would have to be in intimate collaboration throughout, from design through write-up; otherwise we would merely have the current status quo of ad hoc two-language studies with incommensurate goals and methods. Using pairwise language comparison to test one cross-linguistic variable also requires keeping all other variables constant, which is impossible except, perhaps, for closely related dialects, or for particular subpopulations of a speech community, like literate versus illiterate participants in an orthographic study. Treating language as a categorical factor, as the traditional approach does, would also force artificial discretization on naturally gradient variables (e.g., dividing languages into those with small vs. large syllable inventories).

Of course, traditional cross-linguistic methods continue to provide rich insights into language differences and similarities, and the testing of universalist processing claims (i.e., about how the human mind naturally responds to particular language stimuli) would also benefit from greater attention to other approaches, including searching for defaults in first language acquisition or artificial grammar learning (Myers, 2012). The goal of this essay is merely to highlight the potential benefits of adding meta-megastudies to an ever-expanding toolkit.

This tool is not even entirely new. The aforementioned Bates et al. (2003), with its uniform picture-naming task applied to seven languages, already provides a database ready-made for meta-megastudies, as is the set of lexical decision and written word naming megastudies cited above (five languages and counting). Although the currently available cross-language sample is still small and typologically limited, the potential is already there.

## The challenges of meta-megastudies

Meta-megastudies aim to deal with one challenge, confounding in cross-linguistic variables, by taming another, cross-language sampling. In this section I address each in turn; challenges regarding infrastructure are saved for a later section.

### Confounded variables

One sign of the tendency to neglect confounding in cross-linguistic studies is that even though many such studies involve the same two languages, they ascribe their results to different variables, trusting the experimental design to make all other variables irrelevant. Studies on Chinese, for example, often compare it with English, but the observed processing differences are variously ascribed to differences in orthography (e.g., Feng, Miller, Shu, & Zhang, 2001), the phonemic status of tone (e.g., Klein, Zatorre, Milner, & Zhao, 2001), syllable inventory size (e.g., O'Seaghdha, Chen, & Chen, 2010), or the dominance of compounding in word formation (e.g., Myers, 2007). Of course it is not unreasonable to ascribe specific processing differences to specific language characteristics, but this reasoning itself depends on universalist assumptions. The idea that orthographic experience might influence tone perception may be implausible a priori, but language samples diverse enough to pull apart these variables (e.g., by adding Japanese, a non-tonal language with a Chinese-like orthography) would help settle the point empirically. For other variables, universalist processing models are not yet articulated enough to tell us what to expect; for example, to find out if compounding affects syllable processing or vice versa, our best bet may be to head into the (cross-linguistic) lab.

In some cases there may be several equally plausible explanatory variables. In the remainder of this subsection I drive home this point with a single case study: the apparent cross-linguistic differences in the decomposition of syllables into phonemes (extending the discussion of Myers, 2012). Namely, while there is good evidence that spoken word processing depends crucially on phonemes in English (and the other European languages that have been tested), both in perception (e.g., Benkí, 2003; Norris & Cutler, 1988) and in production (e.g., Fromkin, 1971; O'Seaghdha et al., 2010), the status of phonemes in Chinese (i.e., Sinitic languages like Mandarin and Cantonese) is much less clear than that of syllables.

Chinese linguists have classified syllables in terms of onset consonants and rimes for around 1,500 years (Malmqvist, 1994), Chinese poets have long used alliteration (Cai, 2008), and children in modern Chinese-speaking communities are taught syllable-decomposing orthography (e.g., Pinyin). The fact that Chinese speakers are capable of syllable decomposition does not show that they do so readily or routinely, however. Chinese characters themselves represent monosyllabic (and monomorphemic) units, not phoneme-sized units, and while a child's ability to detect word onsets helps when learning to read English, this is not true for children learning to read Chinese (McBride-Chang et al., 2008).

Adult Chinese speakers also show a surprising indifference towards the phoneme. Chen, Chen, & Dell (2002) and O'Seaghdha et al. (2010) found no reaction time effects in Mandarin due to onset consonant priming in a form preparation (implicit priming) task, as compared with the robust effects in English also reported by O'Seaghdha et al. (2010). Similarly, in a picture-word interference naming study on Cantonese, Wong and Chen (2008) failed to observe the facilitation due to shared onset consonants found in languages like Dutch (Levelt, Roelofs, & Meyer, 1999). Qu, Damian, & Kazanina (2012) also failed to find onset priming in a task naming pictures of colored objects, where the monosyllabic color name and the disyllabic object name did or did not share the initial consonant, and Yu, Mo, & Mo (2014) found none in a picture naming task where disyllabic object names did or did not have the same onset consonant in both syllables.

As noted earlier, phonemes are not entirely ignored in Chinese, and Yu and Shu (2003) buck the trend by reporting onset priming in a Mandarin picture-word interference task. Moreover, two of the studies mentioned above, Qu et al. (2012) and Yu et al. (2014), both found significant phoneme priming in event-related potentials (ERP). In an fMRI study dispensing with behavioral measures entirely, Yu, Mo, Li, & Mo (2015) had participants read aloud pairs of nonsense disyllables that shared syllable onsets, entire syllables (but not characters), or nothing, finding brain activation for the onset repetition condition, in different areas from those for the syllable repetition condition. The authors of these three neuroimaging studies ascribe the paucity of behavioral evidence for onset priming to inhibition at a late self-monitoring stage that is intended to reduce the risk of consonant errors. Nevertheless, since they see such inhibition as universal, they must also explain why overt segment priming occurs in English anyway, ascribing it to intrinsically stronger phoneme activation in this language. Thus despite disagreements in detail (cf. O'Seaghdha, Chen, & Chen, 2013; Qu, Damian, & Kazanina, 2013), the consensus remains that Chinese word production differs from that of English because the former is dominated by syllables, not phonemes.

In spoken Chinese word recognition, listeners clearly rely on subsyllabic information (e.g., Gandour, Xu, Wong, Dzemidzic, Lowe, Li, & Tong, 2003; Malins &

Joanisse, 2010), but the methods used so far are not designed to distinguish phonemic from acoustic processing. Relevant techniques are available; for example, Benkí (2003) quantified the perceptual independence of subsyllabic components in English by analyzing the patterns of identification errors in noise, but his method does not seem to have been used yet in Chinese. Currently all we have are intriguing hints from independent studies using different methods: while Tseng, Huang, and Jeng (1996) report that Mandarin listeners were *faster* to detect onset consonants in lexical than in nonce syllables, Fox (1984) found, in a categorical perception study on English onset consonants, that the influence of syllable lexical status increased with *slower* response times.

Finally, wordlikeness judgment experiments suggest that Chinese speakers are much less comfortable than English speakers with expanding their existing syllable inventory via novel combinations of phonemes. The spoken and written nonword wordlikeness judgments collected from English speakers by Bailey and Hahn (2001) imply an overall mean acceptance score around .42 (on a zero-to-one scale, transformed from a nine-point Likert scale) for 272 randomly generated nonce English monosyllables (differing from real syllables by one or two phonemes), as if the participants recognized that many English-like syllables can be generated by freely combining English segments. By contrast, in a wordlikeness megastudy on Mandarin speakers, Myers (2015) reports an overall acceptance rate of only .11 (in a binary judgment task) for 3,274 nonlexical syllables generated by combining Mandarin onsets, prevocalic glides, rimes, and tones (96% differing from real syllables by just one of these elements), presented in phonological transcription (Taiwan's equivalent of Pinyin) to avoid misperception. Methodological differences across these studies seem insufficient to explain such a large difference in results; Bailey and Hahn (2001) and Myers and Tsay (2005) found very little effect of modality (speech vs. writing) on wordlikeness judgments in English and Mandarin, respectively, and Bader and Häussler (2010) and Weskott and Fanselow (2011) show that conclusions drawn from linguistic acceptability judgments do not depend on response scale (whether binary, Likert, or continuous magnitude estimation).

Moreover, neighborhood density, which involves comparisons with entire words, and phonotactic probability, which depends on decomposition into phonemes and phoneme strings (Luce & Large, 2001), affect wordlikeness roughly equally in English; Bailey and Hahn (2001) report that phonotactic and orthotactic probabilities together accounted for about 17% of response variance for their auditory nonce syllables, with neighborhood density accounting for an additional 14% of the variance (as computed in a multiple regression taking both variables into account simultaneously; personal communication, T. Bailey, 14 September, 2016). By contrast, for spoken Cantonese wordlikeness judgments of 270 nonce syllables on a seven-point Likert scale, Kirby and Yu (2007) found that neighborhood density accounted for around 33% of the response variance, while phonotactic probability accounted for less than 2% of additional variance (again computed in a multiple regression taking both variables into account simultaneously; personal communication, J. Kirby, 14 September, 2016). A similar asymmetry can be seen in the Mandarin wordlikeness database used by Myers (2015) (available at http://lngproc.ccu.edu.tw/MWP): a mixed-effects logistic regression model with random slopes predicting binary response choice from log neighborhood density and log mean bigram conditional probability (both phoneme-based, ignoring tone) yields a larger standardized coefficient for neighborhood density ($\beta = 0.98$, $SE = 0.05$, $z = 17.86$, $p < .001$) than for phonotactic probability ($\beta = 0.16$, $SE = 0.02$, $z = 6.00$, $p < .001$). This difference in coefficients is significant by a likelihood ratio test comparing the above additive model with one assuming a single coefficient for the sum of the fixed variables, algebraically the same as an additive model with identical coefficients for both variables ($\chi^2(4) = 2086.2$, $p < .001$).

The reason for cross-language differences in syllable decomposition may seem obvious: English orthography spells out consonants and vowels, but Chinese orthography does not. While the evidence that experience with an orthographic system affects speech processing is mixed (Alario, Perre, Castel, & Ziegler, 2007; Rastle, McCormick, Bayliss, & Davis, 2011), it may be more relevant for certain tasks than for others; in Chinese in particular, Bi, Wei, Janssen, and Han (2009) found that characters affected spoken word form preparation only when the characters were actually read aloud. Nevertheless, focusing exclusively on orthography begs the question of how Chinese manages to get away with a non-phonemic orthography in the first place. A common answer is that an alphabetic orthography would be impractical because the Chinese lexicon is rife with homophones (e.g., Sampson, 2015). Thus in addition to orthography, we must also consider homophony as a possible explanation for cross-linguistic differences in syllable decomposition.

These two explanations, in turn, relate to a third and a fourth: morpheme size and syllable inventory size. Homophones arise so readily in Chinese because virtually all Chinese morphemes are monosyllabic, and there just are not that many Chinese syllables to go around: the syllable inventory in Mandarin (approximately 1,300, even taking tone into account; Myers, 2012) is around ten times smaller than that in English or Dutch (approximately 12,000 each; Levelt et al., 1999). This makes it ten times more feasible for Mandarin speakers simply to memorize their syllable inventory, even if they also weakly compose syllables out of, or decompose them into, phonemes, and even if Dutch speakers also memorize their top few hundred most common syllables, as Levelt et al. (1999) suggest.

Meanwhile, Kirby and Yu (2007) explain their Cantonese wordlikeness findings (where phonotactic probability played a much smaller role than in English) in terms of a fifth difference: the Cantonese lexicon uses much more of the syllable space defined by its subsyllabic components (36%) than does English (6%). The same applies to Mandarin: the 3,274 nonlexical syllables tested in Myers (2015) were chosen from all 4,516 logically possible combinations of onset, medial, rime, and tone, resulting in a 28% probability of hitting a real Mandarin syllable through random combinations of these subsyllabic components alone. If logically possible syllables are too likely to be lexical, decomposition into phoneme strings becomes an unhelpful strategy in making wordlikeness judgments (aside from the trivial benefit of detecting non-native phonemes, which could be done via low-level acoustic processing anyway).

Chinese also differs from English in banning consonant clusters and complex rimes; Chinese traditional phonological theory even treats the rime as a whole, an analysis supported by the highly restricted phonotactics of VX sequences (Light, 1977). More generally, as Chen, Dell, and Chen (2007) show, phoneme sequences are statistically more predictable in the Mandarin lexicon than in the English lexicon, making it more parsimonious to process them together, as they demonstrate in a connectionist model.

This makes six confounded variables (orthography, homophony, morpheme size, syllable inventory size, ratio of lexical to possible syllables, phoneme predictability), though for thoroughness we might also want to add the less obviously relevant variables alluded to earlier (lexical tone and compounding). It matters for psycholinguistic theory which of these variables prove to be truly explanatory for particular processing differences; even among the phonological variables, recall that Yu et al. (2015) found that onset repetition and syllable repetition activate distinct brain regions. These observations make it unlikely that cross-linguistic confounding can be avoided simply by recoding all of the variables as a single predictor (cf. Moscoso del Prado Martín, Kostić, & Baayen, 2004, in their replacement of lexical variables like type and token frequencies by a single measure of informational complexity).

Fortunately, the intrinsic distinctness of these six variables also implies that their tight confounding in English and Chinese should show cracks in a sufficiently large and varied cross-linguistic sample. Many languages are traditionally unwritten or have large illiterate populations, so it should not be too difficult to find speakers of languages with large syllable inventories who nevertheless have as little orthographic training in syllable decomposition as Mandarin speakers. As for the purely phonological variables, the WALS database (World Atlas of Language Structure; Haspelmath *et al.*, 2005) includes the parameters of consonant inventory (five levels), consonant-vowel ratio (five levels), and syllable structure (three levels). Crossing these in the online interface generates a table with 51 non-empty cells, a result that bodes well for a meta-megastudy dependent on there being sufficient variability in factors related to these.

I have no 51-language meta-megastudy to report here, but the shape that such a study may take is suggested by an analysis by Cohen-Goldberg (2012) of the seven-language picture naming database of Bates et al. (2003). Using the publicly available by-item means (Szekely et al., 2004: https://crl.ucsd.edu/experiments/ipnp/7lgpno.html), Cohen-Goldberg found that the presence of phonologically similar onset and coda consonants slowed down monosyllabic word naming in a data sample combining Bulgarian, English, German, and Hungarian responses. Clearly the phenomenon of onset-coda competition can shed light on cross-linguistic variation in syllable decomposition, but this was not the focus of Cohen-Goldberg' study. His selection criteria led him to exclude the other three languages in the database: Spanish and Italian (too few monosyllabic words) and unfortunately also Mandarin (its phonemically contrastive aspiration did not fit with his uniform segment coding scheme). The four languages that he did test were also treated as a fixed rather than random variable and no interactions with this variable were tested, so we cannot say whether the onset-coda competition effect varies systematically across these languages. Nevertheless, this sort of data set has potential. For example, if trial-level data were also made available, cross-trial onset priming could be quantified and tested for interactions with language; the literature review above would lead us to expect particularly weak priming in Mandarin. It may even be possible to detect correlations between the strength of onset priming and language-level variables like syllable inventory size, though of course seven languages are far too few to tease apart all six of the variables discussed above.

## Typological language sampling

Even after acknowledging the reality of cross-linguistic confounds, it may seem that the hard work has only just begun: choosing a sample of languages to compare, from the around 7,000 (Lewis, Simons, & Fennig, 2014) on the planet. After all, typological linguists expend considerable effort on sampling to ensure that inferences about the human language faculty are not skewed by historical descent or borrowing from geographical neighbors (Bickel, 2015; Cysouw, 2005).

Fortunately, meta-megastudies of processing are sufficiently different from typological studies of grammar to make language selection far simpler in the former case. First and foremost, unlike grammars, psycholinguistic processes are essentially automatic, not learned social conventions. Typologists only worry about sampling because the learning of grammatical features is enforced by speech communities, and thus can spread by descent or borrowing. By contrast, it is hard to see how psycholinguistic processes per se could be subject to learned social conventions (e.g., requiring members of a speech community to show a particular kind or degree of lexical frequency effect). Thus if some linguistic feature affects processing, we expect speakers of all languages with this feature to show roughly the same processing effect (aside from interactions with other features), whether or not the languages are related or geographical neighbors.

Moreover, even typologically rare language features (e.g., Chinese orthography) are revealing about general human cognition through their mere existence (see Newmeyer, 2005, for related notions). As long as the cross-language sample provides enough information about the independent variables of interest, typologically unbalanced distributions should not undermine the inferential power of regression analyses (Baayen, Davidson, & Bates, 2008).

Finally, the vast size of single-language megastudies already shows that psycholinguists consider diversity and comprehensiveness more important than representativeness. Due to history and borrowing, any extant lexicon provides as skewed a picture of a native speaker's powers of lexical productivity as extant human languages do of the human language faculty. This lexical skew bothers nobody; as Clark (1973) advised, items are routinely treated as a random variable.

Even with advances in technology and education, the scope of meta-megastudies will likely remain limited by cultural or economic factors, including access to laboratory-quality equipment and familiarity with arcane notions like quiz-taking (cf. Rice, Libben, & Derwing, 2002, and their still too futuristic suggestion that non-invasive neurolinguistic methods may help). Diversity and comprehensiveness also require taking sign languages into account. While many important lexical variables are just as definable in sign languages as in spoken languages, like lexical frequency (Fenlon, Schembri, Rentelis, Vinson, & Cormier, 2014) and neighborhood density (Caselli & Cohen-Goldberg, 2014), the sign/speech parameter is intrinsically confounded with many other lexical variables. Whether due to the visual-manual modality (Meier, 2002) or creolization (Singleton & Newport, 2004), even historically unrelated sign languages differ much less from each other than do spoken languages (Sandler & Lillo-Martin, 2006), with very similar constraints in both phonology (e.g., Sandler, 1999) and morphology (e.g., Aronoff, Meir, & Sandler, 2005).

## Infrastructure for meta-megastudies

While the advent of meta-megastudies may be as inevitable as the steady spread of megastudies across ever more languages, this spread would be more efficient, and the component studies more useful for meta-analysis, with improved infrastructure. To put it crudely, I think the key to speeding the growth of meta-megastudies is to exploit the Web to take crowdsourcing, that favorite buzzword of the megastudy literature, and add another: citizen science.

Crowdsourcing is when an elite group has the masses do menial tasks for them, as when psychologists (e.g., Graham et al., 2011), psycholinguists (e.g., Keuleers & Balota, 2015) or even theoretical linguists (e.g., Erlewine & Kotek, 2016) run experiments on the Web, collecting hundreds or thousands of responses from hundreds or thousands of participants. Technological advances are making megastudies ever easier to run, with Web apps like Science XL (http://www.sciencexl.org/home; Dufau et al., 2011), Tatool (http://www.tatool.ch; von Bastian, Locher, & Ruflin, 2013) and Pavlovia (https://pavlovia.org/; Peirce & MacAskill, 2018) supplementing private lab-written Web systems or proprietary services like Amazon Mechanical Turk (http://www.mturk.com) and Cambridge Brain Sciences (http://www.cambridgebrainsciences.com).

Citizen science, by contrast, is when the masses themselves do science, making small-scale observations (e.g., cataloging visitors to home bird feeders) that can be compiled and studied for large-scale patterns (Silvertown, 2009; Bishop, 2014). Of course like many buzzwords, crowdsourcing and citizen science blur together; the difference I want to highlight is the degree of responsibility given to the contributors. Meta-megastudies must distribute responsibility because they only become feasible when not only the raw response data, but the experiments themselves, are contributed by large numbers of people. The consequence is that meta-megastudies require a much less centralized working style than scientists are generally familiar with (Nielsen, 2012, suggests that this is the future of science more generally).

The Web already has a smattering of decentralized information compilation projects where contributors are treated like scholars rather than data points. The most famous is Wikipedia, which despite being written and edited by thousands of anonymous people, is often lauded for its accuracy (e.g., Clauson, Polen, Boulos, & Dzenowagis, 2008; Heylighen, 2007), albeit with caveats (e.g., Kupferberg & Protus, 2011). Within linguistics, WALS (created by a large but fixed group of experts) is being supplemented by projects like Terraling (http://www.terraling.com), which allows linguists to upload and share more detailed linguistic descriptions than is possible in WALS, with a particular focus on syntax. Even more interesting is the Endangered Languages Archive (ELAR: http://elar.soas.ac.uk/; Nathan, 2013), which explicitly encourages interactions between the creators of its language databases and the Web visitors who use them, including speakers of the archived languages themselves.

What I propose, then, is to merge the technology and philosophy underlying Web experimentation with those underlying Web databases. Meta-megastudy infrastructure would aim at achieving four key goals. First, it should make it easier for new language data to be added while maintaining methodological consistency. This could be achieved via a Web app with built-in functions for designing and running only a limited number of experiment types (e.g., lexical decision and picture naming). Many languages have never been studied experimentally at all; if ready-made tools embolden novice experimenters to contribute data on their native language, typological psycholinguistics would benefit enormously, whether or not the individual contributions are large enough to constitute megastudies themselves.

Second, the system should ensure data quality and ethical standards. This could be achieved using the tried-and-true Web approach of peer vetting. For example, the system could require experimenters to identify their credentials (e.g., via a link to a university homepage), visible to other experimenters and their own participants, and there could also be a feedback system whereby experimenters are publicly rated by their participants and fellow experimenters for ethical behavior and linguistic competence. If all else fails, transparently appointed moderators could lay down the law on particularly egregious offenders.

Third, the system should assist non-native-speaking typologists in their analyses. This could be achieved by encouraging experimenters to share not only their stimuli and results, but also key lexical variables, both item-level and language-level. For example, experimenters who upload pronunciation dictionaries could be rewarded with the automated calculation of phonological neighborhood densities and phonotactic probabilities. The words could also be parsed into syllables using universal algorithms (as in the Sylli tool of Iacoponi & Savy, 2011: http://sylli.sourceforge.net), which, besides providing the language's syllable inventory, would allow the generation of nonwords for tasks like lexical decision and wordlikeness via a syllable-based bigram-chain grammar (as in the Wuggy tool of Keuleers & Brysbaert, 2010: http://crr.ugent.be/programs-data/wuggy); generating written nonwords might be further simplified with phoneme-to-grapheme conversion (Rentzepopoulos & Kokkinakis, 1996). Automated morphological parsing may also be possible (e.g., Durrett & DeNero, 2013) for generating morphologically complex nonwords. Other algorithms could compute cross-linguistic variables like phoneme frequencies, starting with public databases (e.g., that built for the Automated Similarity Judgment Program by Brown, Holman, & Wichmann, 2013: http://asjp.clld.org) but gradually incorporating the lexicons contributed by other experimenters in the meta-megastudy system.

Finally, even though the system could be used solely as a private psycholinguistic cloud service, experimenters who choose to share their materials and results would provide an extra contribution visible to all, hopefully inspiring other experimenters to follow suit. Altruism might receive a further nudge through technical means, such as by making an experimenter's quota for running new experiments dependent on sharing previous ones.

To test the feasibility of such a meta-megastudy system, Tsung-Ying Chen and I have been developing a Web app aimed at implementing as many of the above goals as we can (Chen & Myers, 2021). We call it Worldlikeness, since it was originally designed for the collection and sharing of wordlikeness judgments across languages. Built in the Javascript-based Meteor platform (Coleman & Greif, 2013), Worldlikeness is hosted at https://worldlikeness.org/; it runs in any modern browser, including on mobile devices. As a Web experiment system, it helps experimenters design and run simple non-factorial experiments, and rewards participants with colorful graphic comparisons of personal statistics with group results. Experimenters need to register and provide contact information for their participants, who do not need to register at all, and both types of users control access to their data. As a Web database system, Worldlikeness facilitates the sharing of experimental materials and results, including previously collected data uploaded to the system. Depending on the privacy settings of the data providers, results may be available only to the original experimenter and the experimenter's collaborators, all other registered experimenters, or the public at large. Data sharing is encouraged via a quota system like that sketched above, among other devices. Researchers wanting to perform typological analyses across many languages can download all of the data available to them, or just search for a particular subset of interest; typological researchers thus need not be experimenters themselves. Among our key goals, automating tools for typological research has proven to be the most challenging; we are still working out the kinks in our algorithms for stimulus generation and language type marking.

Nevertheless, Worldlikeness has already proven itself in experiments on Mandarin and Taiwan Southern Min (Myers & Chen, 2016), Taiwan Sign Language (Lee, 2016), and Japanese and Vietnamese, with some of the results publicly available for download from the Worldlikeness database. As repeatedly emphasized in this essay, no single research team can contribute enough languages to create a meta-megastudy database on its own, so we are actively soliciting contributions. While wordlikeness is not a particularly popular task in psycholinguistics, the Worldlikeness app is already set up to be used for the lexical decision task as well. Not much modification would be necessary to adapt it for other common megastudy tasks, like written word naming or picture naming: Worldlikeness is not only free but open-source, and readers are more than welcome to repurpose its code.

## Conclusions

Are phonemes processing units in spoken word processing? What determines when readers access word meanings via phonological recoding or directly from the written form? How is morphological decomposition affected by semantic transparency? Quantitatively precise answers to theoretically significant psycholinguistic questions depend not just on the particular experimental design or particular test items, but also on the particular language: psycholinguistics is not psychophysics, and language experience matters. Yet two-language studies are not enough: languages differ from each other in too many ways to be sure that any cross-language result must be due to one factor and not another.

Meta-megastudies simply put a fancy name on a simple idea: generalizing the big data approach of single-language megastudies, where the variable confound problem has long been recognized, to cross-linguistic psycholinguistics. If the meta-megastudy approach could be made feasible, the benefits would clearly be enormous: claims about processing universals could be tested the same way typological linguists test claims about grammatical universals. Rather than observing a stimulus-response pattern in one language and assuming things work exactly the same way in all languages unless proven otherwise, stimuli and responses could be studied across a wide variety of languages and the general relationship induced via statistical modeling, with fewer sampling headaches than in typological studies of grammar.

Fortunately, technological advances are indeed making meta-megastudies feasible, the same advances that sparked the megastudy revolution itself. The computing and networking powers of the Web enable the development of systems that link experimenters not just with their participants, but with each other, using automation to ensure methodological consistency and to reward contributors with perks like stimulus creation, and Web-mediated social pressure to enforce ethics and quality. Psycholinguists know that cross-linguistic research is essential to understanding the full richness of the human mind, but the vast majority of the world's languages are still missing from this research. Give citizen scientists the right tools, and watch what happens.

## References

Adelman, J. S., Johnson, R. L., McCormick, S. F., McKague, M., Kinoshita, S., Bowers, J. S., Perry, J. R., Lupker, S. J., Forster, K. I., Cortese, M. J., Scaltritti, M., Aschenbrenner, A. J., Coane, J. H., White, L., Yap, M. J., Davis, C., Kim, J., & C. J. Davis. (2014). A behavioral database for masked form priming. *Behavior Research Methods*, 46(4), 1052–1067. https://doi.org/10.3758/s13428-013-0442-y

Alario, F. X., Perre, L., Castel, C., & Ziegler, J. C. (2007). The role of orthography in speech production revisited. *Cognition*, 102(3), 464–475. https://doi.org/10.1016/j.cognition.2006.02.002

Aronoff, M., Meir, I., & Sandler, W. (2005). The paradox of sign language morphology. *Language*, 81(2), 301–344. https://doi.org/10.1353/lan.2005.0043

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, 59(4), 390–412. https://doi.org/10.1016/j.jml.2007.12.005

Bader, M., & Häussler, J. (2010). Toward a model of grammaticality judgments. *Journal of Linguistics*, 46 (2), 273–330. https://doi.org/10.1017/S0022226709990260

Bailey, T. M., & Hahn, U. (2001). Determinants of wordlikeness: Phonotactics or lexical neighborhoods? *Journal of Memory & Language*, 44, 569–591. https://doi.org/10.1006/jmla.2000.2756

Balota, D. A., Yap, M. J., Cortese, M. J., Hutchison, K. A., Kessler, B., Loftis, B., Neely, J. H., Nelson, D. L., Simpson, G. B., & Treiman, R. (2007). The English Lexicon Project. *Behavior Research Methods*, 39, 445–459. https://doi.org/10.3758/BF03193014

Balota, D. A., Yap, M. J., Hutchison, K. A., & Cortese, M. J. (2012). Megastudies: What do millions (or so) of trials tell us about lexical processing? In J. S. Adelman (Ed). *Visual word recognition*, Vol. 1 (pp. 90–115). London: Psychology Press Psychology Press.

Bates, E., D'Amico, S., Jacobsen, T., Székely, A., Andonova, E., Devescovi, A., Herron, D., Lu, C-C., Pechmann, T., Pléh, C., Wicha, N., Federmeier, K., Gerdjikova, I., Gutierrez, G., Hung, D., Hsu, J., Iyer, G., Kohnert, K., Mehotcheva, T., Orozco-Figueroa, A., Tzeng, A., & Tzeng, O. (2003). Timed picture naming in seven languages. *Psychonomic Bulletin & Review*, 10 (2), 344–380. https://doi.org/10.3758/BF03196494

Bates, E., Devescovi, A., & Wulfeck, B. (2001). Psycholinguistics: A cross-language perspective. *Annual Review of Psychology*, 52, 369–96. https://doi.org/10.1146/annurev.psych.52.1.369

Benkí, J. R. (2003). Quantitative evaluation of lexical status, word frequency, and neighborhood density as context effects in spoken word recognition. *The Journal of the Acoustical Society of America*, 113(3), 1689–1705. https://doi.org/10.1121/1.1534102

Bi, Y., Wei, T., Janssen, N., & Han, Z. (2009). The contribution of orthography to spoken word production: Evidence from Mandarin Chinese. *Psychonomic Bulletin & Review*, 16(3), 555–560. https://doi.org/10.3758/PBR.16.3.555

Bickel, B. (2015). Distributional typology: Statistical inquiries into the dynamics of linguistic diversity. In B. Heine & H. Narrog (Eds). *The Oxford Handbook of Linguistic Analysis, 2nd edition* (pp. 901–923). Oxford: Oxford University Press.

Bishop, S. (2014). Science exposed. *Scientific American*, 311(4), 46. https://doi.org/10.1038/scientificamerican1014-46

Brown, C. H., Holman, E. W., & Wichmann, S. (2013). Sound correspondences in the world's languages. *Language*, 89 (1), 4–29. https://doi.org/10.1353/lan.2013.0009

Cai, Z.-Q. (Ed.) (2008). *How to read Chinese poetry: A guided anthology*. New York: Columbia University Press.

Caselli, N. K., & Cohen-Goldberg, A. M. (2014). Lexical access in sign language: A computational model. *Frontiers in Psychology*, 5, 428. https://doi.org/10.3389/fpsyg.2014.00428

Chen, J.-Y., Chen, T.-M., & Dell, G. S. (2002). Word-form encoding in Mandarin Chinese as assessed by the implicit priming task. *Journal of Memory and Language*, 46(4), 751–781. https://doi.org/10.1006/jmla.2001.2825

Chen, T.-M., Dell, G., & Chen, J.-Y. (2007). A cross-linguistic study of phonological units: Syllables emerge from the statistics of Mandarin Chinese, but not from the statistics of English. *Chinese Journal of Psychology*, 49(2), 137–144.

Chen, T.-Y., & Myers, J. (2021). Worldlikeness: A Web crowdsourcing platform for typological psycholinguistics. *Linguistic Vanguard*, 7(s1), 20190011. https://doi.org/10.1515/lingvan-2019-0011

Clark, H. (1973). The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior*, 12, 335–359. https://doi.org/10.1016/S0022-5371(73)80014-3

Clauson, K. A., Polen, H. H., Boulos, M. N. K., & Dzenowagis, J. H. (2008). Scope, completeness, and accuracy of drug information in Wikipedia. *Annals of Pharmacotherapy*, 42(12), 1814–1821. https://doi.org/10.1345/aph.1L474

Cohen-Goldberg, A. M. (2012). Phonological competition within the word: Evidence from the phoneme similarity effect in spoken production. *Journal of Memory and Language*, 67(1), 184–198. https://doi.org/10.1016/j.jml.2012.03.007

Coleman, T., & Greif, S. (2013). *Discover Meteor*. URL: http://www.discovermeteor.com

Costa, A., Alario, F. X., & Sebastián-Gallés, N. (2007). Cross-linguistic research on language production. In M. G. Gaskell (Ed.), *The Oxford handbook of psycholinguistics* (pp. 531–546). Oxford: Oxford University Press.

Cutler, A. (1981). Making up materials is a confounded nuisance, or: Will we be able to run any psycholinguistic experiments at all in 1990? *Cognition*, 10, 65–70. https://doi.org/10.1016/0010-0277(81)90026-3

Cutler, A. (1985). Cross-language psycholinguistics. *Linguistics*, 23, 659–667. https://doi.org/10.1515/ling.1985.23.5.659

Cysouw, M. (2005). Quantitative methods in typology. In R. Kohler, G. Altmann, & R. G. Piotrowski (Eds.) *Quantitative Linguistik: Ein internationales Handbuch* [Quantitative linguistics: An international handbook] (pp. 554–578). Berlin: Walter de Gruyter.

DeFrancis, J. (1989). *Visible speech: The diverse oneness of writing systems*. Honolulu: University of Hawaii Press.

Dryer, M. S. (1992). The Greenbergian word order correlations. *Language*, 68 (1), 81–138. https://doi.org/10.1353/lan.1992.0028

Dryer, M. S. (2011). The evidence for word order correlations. *Linguistic Typology*, 15(2), 335–380. https://doi.org/10.1515/lity.2011.024

Dufau, S., Duñabeitia, J. A., Moret-Tatay, C., McGonigal, A., Peeters, D., Alario, F. X., Balota, D. A., Brysbaert, M., Carreiras, M., Ferrand, L., Ktori, M., Perea, M., Rastle, K., Sasburg, O., Yap, M. J., Ziegler, J. C., & Grainger, J. (2011). Smart phone, smart science: How the use of smartphones can revolutionize research in cognitive science. *PloS One*, 6(9), e24974. https://doi.org/10.1371/journal.pone.0024974

Durrett, G., & DeNero, J. (2013). Supervised learning of complete morphological paradigms. Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL), pp. 1185–1195.

Erlewine, M. Y., & Kotek, H. (2016). A streamlined approach to online linguistic surveys. *Natural Language & Linguistic Theory*, 34(2), 481–495. https://doi.org/10.1007/s11049-015-9305-9

Ernestus, M., & Cutler, A. (2015). BALDEY: A database of auditory lexical decisions. *The Quarterly Journal of Experimental Psychology*, 68(8), 1469–1488. https://doi.org/10.1080/17470218.2014.984730

Evans, N., & Levinson, S. C. (2009). The myth of language universals: Language diversity and its importance for cognitive science. *Behavioral and Brain Sciences*, 32 (5), 429–492. https://doi.org/10.1017/S0140525X0999094X

Feng, G., Miller, K., Shu, H., & Zhang, H. (2001). Rowed to recovery: the use of phonological and orthographic information in reading Chinese and English. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 27(4), 1079–1100.

Fenlon, J., Schembri, A., Rentelis, R., Vinson, D., & Cormier, K. (2014). Using conversational data to determine lexical frequency in British Sign Language: The influence of text type. *Lingua*, 143, 187–202. https://doi.org/10.1016/j.lingua.2014.02.003

Ferrand, L., New, B., Brysbaert, M., Keuleers, E., Bonin, P., Méot, A., Augustinova, M., & Pallier, C. (2010). The French Lexicon Project: Lexical decision data for 38,840 French words and 38,840 pseudowords. *Behavior Research Methods*, 42(2), 488–496. https://doi.org/10.3758/BRM.42.2.488

Forster, K. I. (2000). The potential for experimenter bias effects in word recognition experiments. *Memory & Cognition*, 28(7), 1109–1115. https://doi.org/10.3758/BF03211812

Fox, R. A. (1984). Effect of lexical status on phonetic categorization. *Journal of Experimental Psychology: Human Perception and Performance*, 10(4), 526–540.

Fromkin, V. A. (1971). The non-anomalous nature of anomalous utterances. *Language*, 47(1), 27–52. https://doi.org/10.2307/412187

Gandour, J., Xu, Y., Wong, D., Dzemidzic, M., Lowe, M., Li, X., & Tong, Y. (2003). Neural correlates of segmental and tonal information in speech perception. *Human Brain Mapping*, 20(4), 185–200. https://doi.org/10.1002/hbm.10137

Graham, J., Nosek, B. A., Haidt, J., Iyer, R., Koleva, S., & Ditto, P. H. (2011). Mapping the moral domain. *Journal of Personality and Social Psychology*, 101(2), 366–385. https://doi.org/10.1037/a0021847

Greenberg, J. H. (1963). Some universals of grammar with particular reference to the order of meaningful elements. In J. H. Greenberg (Ed.) *Universals of language* (pp. 73–113). Cambridge, MA: MIT Press.

Haspelmath, M., Dryer, M. S., Gil, D., & Comrie, B. (Eds.) (2005). *The world atlas of language structure*. Oxford: Oxford University Press.

Heylighen, F. (2007). Why is open access development so successful? Stigmergic organization and the economics of information. In Lutterbeck, B., Bärwolff, M., & Gehring, R. A. (Eds.) *Open Source Jahrbuch 2007*. Berlin: Technical University of Berlin.

Hutchison, K. A., Balota, D. A., Neely, J. H., Cortese, M. J., Cohen-Shikora, E. R., Tse, C. S., Yap, M. J., Bengson, J. J., Niemeyer, D., & Buchanan, E. (2013). The semantic priming project. *Behavior Research Methods*, 45(4), 1099–1114. https://doi.org/10.3758/s13428-012-0304-z

Iacoponi, L., & Savy, R. (2011). Sylli: Automatic phonological syllabification for Italian. *INTERSPEECH* 2011, 641–644.

Jaeger, T. F., & Norcliffe, E. J. (2009). The cross-linguistic study of sentence production. *Language and Linguistics Compass*, 3/4, 866–887. https://doi.org/10.1111/j.1749-818X.2009.00147.x

Keuleers, E., & Balota, D. A. (2015). Megastudies, crowdsourcing, and large datasets in psycholinguistics: An overview of recent developments. *The Quarterly Journal of Experimental Psychology*, 68 (8), 1457–1468. https://doi.org/10.1080/17470218.2015.1051065

Keuleers, E., & Brysbaert, M. (2010). Wuggy: A multilingual pseudoword generator. *Behavior Research Methods*, 42(3), 627–633. https://doi.org/10.3758/BRM.42.3.627

Keuleers, E., Diependaele, K., & Brysbaert, M. (2010). Practice effects in large-scale visual word recognition studies: A lexical decision study on 14,000 Dutch mono-and disyllabic words and nonwords. *Frontiers in Psychology*, 1, 174. https://doi.org/10.3389/fpsyg.2010.00174

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(1), 287–304. https://doi.org/10.3758/s13428-011-0118-4

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. https://doi.org/10.1080/17470218.2015.1022560

Kirby, J. P., & Yu, A. C. L. (2007). Lexical and phonotactic effects on wordlikeness judgments in Cantonese. *Proceedings of the International Congress of Phonetic Sciences*, 16, 1389–1392.

Klein, D., Zatorre, R. J., Milner, B., & Zhao, V. (2001). A cross-linguistic PET study of tone perception in Mandarin Chinese and English speakers. *Neuroimage*, 13(4), 646–653. https://doi.org/10.1006/nimg.2000.0738

Kupferberg, N., & Protus, B. M. (2011). Accuracy and completeness of drug information in Wikipedia: An assessment. *Journal of the Medical Library Association*, 99(4), 310–313. https://doi.org/10.3163/1536-5050.99.4.010

Lee, H.-H. (2016). A comparative study of the phonology of Taiwan Sign Language and Signed Chinese. Unpublished National Chung Cheng University Ph.D. thesis.

Lemhöfer, K., Dijkstra, T., Schriefers, H., Baayen, R. H., Grainger, J., & Zwitserlood, P. (2008). Native language influences on word recognition in a second language: A megastudy. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 34 (1), 12–31.

Levelt, W. J., Roelofs, A., & Meyer, A. S. (1999). A theory of lexical access in speech production. *Behavioral and Brain Sciences*, 22(01), 1–38. https://doi.org/10.1017/S0140525X99001776

Lewis, M. P., Simons, G. F., Fennig, C. D. (Eds.). (2014). *Ethnologue: Languages of the world, Seventeenth edition*. Dallas, Texas: SIL International. Online version: http://www.ethnologue.com/17/

Light, T. (1977). The Cantonese final: An exercise in indigenous analysis. *Journal of Chinese Linguistics*, 5(1), 75–102.

Luce, P. A., & Large, N. R. (2001). Phonotactics, density, and entropy in spoken word recognition. *Language & Cognitive Processes*, 16(5/6), 565–581. https://doi.org/10.1080/01690960143000137

Malins, J. G., & Joanisse, M. F. (2010). The roles of tonal and segmental information in Mandarin spoken word recognition: An eyetracking study. *Journal of Memory and Language*, 62(4), 407–420. https://doi.org/10.1016/j.jml.2010.02.004

Malmqvist, G. (1994). Chinese linguistics. In G. Lepschy (Ed.), *History of linguistics: Volume I: The Eastern traditions of linguistics* (pp. 1–24). London: Longman.

McBride-Chang, C., Tong, X., Shu, H., Wong, A. M. Y., Leung, K. W., & Tardif, T. (2008). Syllable, phoneme, and tone: Psycholinguistic units in early Chinese and English word recognition. *Scientific Studies of Reading*, 12(2), 171–194. https://doi.org/10.1080/10888430801917290

Meier, R. P. (2002). Why different, why the same? Explaining effects and non-effects of modality upon linguistic structure in sign and speech. In R. P. Meier & K. Cormier (Eds.) *Modality and structure in signed and spoken languages* (pp. 1–25). Cambridge, UK: Cambridge University Press. https://doi.org/10.1017/CBO9780511486777.001

Moscoso del Prado Martín, F., Kostić, A., & Baayen, R. H. (2004). Putting the bits together: An information theoretical perspective on morphological processing. *Cognition*, 94(1), 1–18. https://doi.org/10.1016/j.cognition.2003.10.015

Myers, J. (2007). Generative morphology as psycholinguistics. In G. Jarema & G. Libben (Eds.), *The mental lexicon: Core perspectives* (pp. 105–128). Amsterdam: Elsevier. https://doi.org/10.1163/9780080548692_007

Myers, J. (2012). Chinese as a natural experiment. In G. Libben, G. Jarema, & C. Westbury (Eds.), *Methodological and analytic frontiers in lexical research* (pp. 155–169). Amsterdam: John Benjamins. https://doi.org/10.1075/bct.47.09mye

Myers, J. (2015). Markedness and lexical typicality in Mandarin acceptability judgments. *Language & Linguistics*, 16 (6). https://doi.org/10.1177/1606822X15602606

Myers, J., & Chen, T-Y. (2016). The time course of sociolinguistic influences on wordlikeness judgments. In A. Botinis (Ed.), *Proceedings of the 7th Tutorial and Research Workshop on Experimental Linguistics* (pp. 119–122). International Speech Communication Association.
Myers, J., & Tsay, J. (2005). The processing of phonological acceptability judgments. Proceedings of Symposium on 90–92 NSC Projects (pp. 26–45). Taipei, Taiwan, May.
Nathan, D. (2013). Access and accessibility at ELAR, a social networking archive for endangered languages documentation. In M. Turin, C. Wheeler & E. Wilkinson (Eds.) *Oral literature in the digital age: Archiving orality and connecting with communities*. Cambridge, UK: Open Book Publishers.
Newmeyer, F. J. (2005). *Possible and probable languages: A generative perspective on linguistic typology*. Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199274338.001.0001
Nielsen, M. (2012). *Reinventing discovery: The new era of networked science*. Princeton, NJ: Princeton University Press.
Norris, D., & Cutler, A. (1988). The relative accessibility of phonemes and syllables. *Perception & Psychophysics*, 43(6), 541–550. https://doi.org/10.3758/BF03207742
O'Brien, R. M. (2007). A caution regarding rules of thumb for Variance Inflation Factors. *Quality & Quantity*, 41, 673–690. https://doi.org/10.1007/s11135-006-9018-6
O'Seaghdha, P. G., Chen, J.-Y., & Chen, T.-M. (2010). Proximate units in word production: Phonological encoding begins with syllables in Mandarin Chinese but with segments in English. *Cognition*, 115, 282–302. https://doi.org/10.1016/j.cognition.2010.01.001
O'Seaghdha, P. G., Chen, J.-Y., & Chen, T.-M. (2013). Close but not proximate: The significance of phonological segments in speaking depends on their functional engagement. *Proceedings of the National Academy of Sciences*, 110(1), E3. https://doi.org/10.1073/pnas.1217032110
Peirce, J. W., & MacAskill, M. R. (2018). *Building experiments in PsychoPy*. London: Sage.
Qu, Q., Damian, M. F., & Kazanina, N. (2012). Sound-sized segments are significant for Mandarin speakers. *Proceedings of the National Academy of Sciences*, 109(35), 14265–14270. https://doi.org/10.1073/pnas.1200632109
Qu, Q., Damian, M. F., & Kazanina, N. (2013). Reply to O'Seaghdha et al.: Primary phonological planning units in Chinese are phonemically specified. *Proceedings of the National Academy of Sciences*, 110(1), E4. https://doi.org/10.1073/pnas.1217601110
Rastle, K., McCormick, S. F., Bayliss, L., & Davis, C. J. (2011). Orthography influences the perception and production of speech. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 37(6), 1588–1594.
Rentzepopoulos, P. A., & Kokkinakis, G. K. (1996). Efficient multilingual phoneme-to-grapheme conversion based on HMM. *Computational Linguistics*, 22(3), 351–376.
Rice, S., Libben, G., & Derwing, B. (2002). Morphological representation in an endangered, polysynthetic language. *Brain and Language*, 81(1), 473–486. https://doi.org/10.1006/brln.2001.2540
Sampson, G. (2015). A Chinese phonological enigma. *Journal of Chinese Linguistics*, 43, 679–691. https://doi.org/10.1353/jcl.2015.0014
Sandler, W. (1999). Cliticization and prosodic words in a sign language. In T. A. Hall and U. Kleinhenz (Eds.) *Studies on the phonological word* (pp. 223–255). Amsterdam: John Benjamins. https://doi.org/10.1075/cilt.174.09san
Sandler, W., & Lillo-Martin. (2006). *Sign language and linguistic universals*. Cambridge, UK: Cambridge University Press. https://doi.org/10.1017/CBO9781139163910

Silvertown, J. (2009). A new dawn for citizen science. *Trends in Ecology & Evolution*, 24(9), 467–471. https://doi.org/10.1016/j.tree.2009.03.017

Singleton, J. L., & Newport, E. L. (2004). When learners surpass their models: The acquisition of American Sign Language from inconsistent input. *Cognitive Psychology* 49, 370–407. https://doi.org/10.1016/j.cogpsych.2004.05.001

Sze, W. P., Liow, S. J. R., & Yap, M. J. (2014). The Chinese Lexicon Project: A repository of lexical decision behavioral responses for 2,500 Chinese characters. *Behavior Research Methods*, 46(1), 263–273. https://doi.org/10.3758/s13428-013-0355-9

Szekely, A., Jacobsen, T., D'Amico, S., Devescovi, A., Andonova, E., Herron, D., Lu, C.-C., Pechmann, T., Pléh, C., Wicha, N., Federmeier, K., Gerdjikova, I., Gutierrez, G., Hung, D., Hsu, J., Iyer, G., Kohnert, K., Mehotcheva, T., Orozco-Figueroa, A., Tzeng, A., Tzeng, O., Arévalo, A., Vargha, A., Butler, A. C., Buffngton, R., & Bates, E. (2004). A new on-line resource for psycholinguistic studies. *Journal of Memory and Language*, 51(2), 247–250. https://doi.org/10.1016/j.jml.2004.03.002

Tse, C.-S., Yap, M. J., Chan, Y.-L., Sze, W.-P., Shaoul, C., & Lin, D. (forthcoming). The Chinese Lexicon Project: A megastudy of lexical decision performance for 25,000+ traditional Chinese two-character compound words. *Behavior Research Methods*.

Tseng, C.-H., Huang, K.-Y., & Jeng, J.-Y. (1996). The role of the syllable in perceiving spoken Chinese. *Proceedings of the National Science Council, Part C: Humanities and Social Sciences*, 6 (1), 71–86.

Vittinghoff, E., & McCulloch, C. E. (2007). Relaxing the rule of ten events per variable in logistic and Cox regression. *American Journal of Epidemiology*, 165(6), 710–718. https://doi.org/10.1093/aje/kwk052

von Bastian, C. C., Locher, A., & Ruflin, M. (2013). Tatool: A Java-based open-source programming framework for psychological studies. *Behavior Research Methods*, 45(1), 108–115. https://doi.org/10.3758/s13428-012-0224-y

Weskott, T., & Fanselow, G. (2011). On the informativity of different measures of linguistic acceptability. *Language*, 87 (2), 249–273. https://doi.org/10.1353/lan.2011.0041

Wong, A. W.-K., & Chen, H.-C. (2008). Processing segmental and prosodic information in Cantonese word production. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 34 (5), 1172–1190.

Yap, M. J., Liow, S. J. R., Jalil, S. B., & Faizal, S. S. B. (2010). The Malay Lexicon Project: A database of lexical statistics for 9,592 words. *Behavior Research Methods*, 42(4), 992–1003. https://doi.org/10.3758/BRM.42.4.992

Yu, L., & Shu, H. (2003). Hanyu yanyu chansheng de yuyin jiagong jizhi [Phonological processing mechanism in Chinese speech production]. *Xinli Kexue*, 26 (5), 818–822.

Yu, M., Mo, C., & Mo, L. (2014). The role of phoneme in Mandarin Chinese production: Evidence from ERPs. *PloS one* 9 (9), e106486. https://doi.org/10.1371/journal.pone.0106486

Yu, M., Mo, C., Li, Y., & Mo, L. (2015). Distinct representations of syllables and phonemes in Chinese production: Evidence from fMRI adaptation. *Neuropsychologia*, 77, 253–259. https://doi.org/10.1016/j.neuropsychologia.2015.08.027

CHAPTER 3

# Pay no attention to that man behind the curtain

## Explaining semantics without semantics

Chris Westbury
University of Alberta

### Thesis

Hempel and Oppenheim (1948) drew a distinction in scientific explanation between the *explanandum*, statements describing the empirical phenomenon to be explained, and the *explanans*, statements describing the evidence that allow one to predict that phenomenon. To avoid tautology, these sets of statements must refer to distinct domains. The scientific explanation of semantics must be grounded in an explanans that appeals to entities from non-semantic domains, i.e. we have to explain semantics in terms of things that are not themselves semantic. I consider eight (non mutually-exclusive) candidate domains that could ground semantics, including affect, lexical or sub-word co-occurrence, mental simulation, and associative learning. Following Wittgenstein (1958), I propose that adjudicating between these different domains is difficult because of the reification of a word's 'meaning' as an atomistic unit. If we abandon the idea that the meaning of a word is an atomistic unit and instead think of word meaning as a set of dynamic and disparate embodied states unified by a shared label, many apparent problems associated with identifying a word meaning's 'true' explanans disappear. Lexical semantics can be considered as sets of weighted constraints that are individually sufficient for specifying and labeling a subjectively-recognizable location in the high dimensional state space defined by our neural activity.

### Commentaries on Westbury thesis

Ray Jackendoff commentary on Westbury thesis

Westbury says that "we have to explain semantics in terms of things that are not themselves semantic." Ultimately, of course, semantics has to be explained in terms of neural computation and the evolution of the human brain, tasks that for the

https://doi.org/10.1075/z.238.03wes

foreseeable future are probably unattainable. In the meantime, how should we proceed? A place to start is to ask what phenomena a theory of semantics has to account for (Jackendoff 2002, 2012). These include at least:

- Inference: How we judge that a conclusion follows from premises
- Reference: How we construe sentences as talking about the world
- Combinatoriality: How the meanings of words are constructed from smaller parts and how they in turn can be incorporated into larger semantic structures
- The possibility of (pretty good) translation: How words and sentences in two randomly chosen languages can be understood as "saying the same thing"
- The grounding of natural language meaning in nonlinguistic concepts, especially in concepts that are attributable to other primates

These goals, with the exception of reference and nonlinguistic grounding, are not addressed by any of the eight proxies referred to in Westbury's article.[1]

Westbury further urges us to "abandon the idea that the meaning of a word is an atomistic unit." I'm not sure who holds this position, aside perhaps from Jerry Fodor (1975). When a semanticist writes CAT, this can be considered an abbreviation for combinatorial mental representations that we don't yet understand very well. On the other hand, many aspects of the combinatorial structure of meaning *have* been elucidated, for instance:

- Argument structure: Many words (especially verbs, but also nouns like *friend* and prepositions like *across*) express functions whose arguments are instantiated by the meanings of their syntactic complements.
- Internal argument structure: Many words are systematically (though not necessarily exhaustively) related to other words, the meaning of one serving as an argument within the meaning of the other. Standard examples are the noun *butter* and the verb *butter (the bread)*; the verb *bake* and the noun *baker*; and the adjective *broken*, the intransitive verb *(the glass) broke*, and the transitive verb *break (the glass)*.
- Family resemblance structure: A word's meaning may involve a number of conditions, any of which is sufficient, but none of which is necessary for identifying a referent; stereotypical instances satisfy all or at least more of the conditions. Wittgenstein's (1953) discussion of *game* falls in this category.

1. By the way, Westbury's recommendation of grounding semantics in "embodied human experience" depends on being able to say how the body and your experiences are mentally encoded. You can't simply import your body or the outside world into your brain.

– Conditions in multiple domains: The word *book* denotes an entity that is simultaneously a physical object and a bearer of information (Pustejovsky 1995). Similarly, *shaking hands* is simultaneously a physical action and an expression of social solidarity. Such mixed configurations are widespread. It should be no surprise that meanings involving different nonlinguistic domains engage brain areas characteristic of those domains.

Westbury would no doubt say such analyses, of which there are many, are not *explanations* of semantics. But they nevertheless mark important progress in understanding how meanings work. A goal for the future is to further elucidate this structure, in the hope that we can eventually figure out what CAT is an abbreviation for, and in the hope that we can eventually figure out how the neurons learn, store, and use meanings in order to make inferences and describe the world.

*References*


Fodor, J. A. (1975). *The Language of Thought*. Cambridge, MA: Harvard University Press.

Jackendoff, Ray. (2002). *Foundations of Language*. Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198270126.001.0001

Jackendoff, Ray. (2012). *A User's Guide to Thought and Meaning*. Oxford: Oxford University Press.

Pustejovsky, James. (1995). *The Generative Lexicon*. Cambridge, MA: MIT Press.

Wittgenstein, Ludwig. (1953). *Philosophical Investigations*. Oxford: Blackwell.


## James Myers commentary on Westbury thesis

Meaning is treated with a degree of mysticism denied to other aspects of language; nobody searches for the pronunciation of life. Not even philosophers are immune. In Hilary Putnam's "Twin Earth" thought experiment, H2O on Earth and some other superficially identical but chemically distinct substance on Twin Earth are both called "water", and crucially also thought of in identical terms, by inhabitants of the respective planets. From this we are supposed to conclude that "water" really means different things on the two planets, despite their identical mental representations, because of differences in the mind-external world. If Putnam is right, then, none of the eight approaches to meaning advocated in this paper, alone or in combination, captures the essence of meaning, since they all operate mind-internally. The immediate response to Putnam is simple: like all philosophical thought experiments, his says a lot more about the nature of intuitions than about the nature of reality.

This doesn't mean that coming up with a non-mystical approach to meaning is trivial. The old behaviorist notion, promoted by Bloomfield and others of his era, that meanings are nothing more than links between externally observable stimuli and responses, is not merely counterintuitive (which to Bloomfield may have been a point in its favor), but also not very satisfying intellectually. Chomsky's critique of Skinner still seems to hold: theories of language founded on stimulus-response links are either correct but meaningless (of course utterances are triggered by the environment *somehow*), or meaningful but incorrect (principles derived from lever-pressing rats don't generalize very far). The solution isn't to return to mysticism (nor to follow Chomsky himself and ignore meaning entirely), but to do something like what is proposed in this paper: posit and test a wider variety of links.

One benefit of this kind of approach is that it provides an explanation for Putnam's intuition: the links between language and thought have to be supplemented with links between thought and reality. The big conceptual challenge, though, is that the division of labor between the first sort of link (psycholinguistics) and the second sort (cognitive psychology) is notoriously hard to draw, as debates over the modularity of language show. The result is, as the author implies at the end of his essay, that the word *cat* has too many links of too many different kinds to make it possible to sum up any more elegantly than with the word *cat* itself. This may not be mysticism, but for now it can do little more (as the saying goes) than elevate our confusion to a higher level.

## Dorit Ravid commentary on Westbury thesis

From the point of view of a developmental (psycho)linguist, the crucial question about lexical semantics is how these lexical concepts are learned by children in a speech community so as to enable communication within that community, relying on the seamlessly efficient co-interpretation of the same notion. Cognitive science regards the human brain as the most powerful learning device shaped by evolution (Griffiths et al., 2010), mapping out the external and internal world in terms of objects, people, places, states, properties, ideas, actions, events, and processes. These are encoded across an array of forms, from overt lexical units to periphrastic constructions (Goldberg, 1995). A useful way of conceptualizing lexical concepts is as knowledge structures specialized for symbolic representation (Evans, 2009). Evans regards lexical concepts as sense units inferred from the ambient language and stored as part of language knowledge, providing access to encyclopaedic knowledge structures, often encapsulating complex and informationally diffuse ideas. Lexical development thus lies in the crossroads of language and conceptual development at a time when the foundations of individual human knowledge are established (Tooby, Cosmides & Barrett, 2005), mediated by the typological structure of the ambient language. This means that in acquiring what may be called a "vocabulary", children learning a morphology-rich language will have to pay particular attention to the internal structure of words in order to learn to extract and express meaning.

Learning words crucially relies on their communicative environments (Medina, Snedeker, Trueswell, & Gleitman, 2011). Lexical meaning is derived directly from how language is actually used (Clark & Wong, 2002; Langacker, 2000), as already postulated by Wittgenstein and Morris. The syntactic, discursive, pragmatic and environmental contexts in which words are encountered are not merely helpful in lexical acquisition and processing, they are in fact inherent to determining the nature of a word's meaning, including context-specific shifts in senses (Pustejovsky, 1995). To gain the breadth and depth of word knowledge, words must be experienced frequently in their contexts (Ambridge et al., 2015; Sandoval & Gómez, 2013). That is, numerous, variegated encounters with words are critical for them to be interpreted, organized into categories and learned, including the creation of long-range relationships between similar environments to sustain new-word acquisition (Landauer & Dumais, 1997).

Saying that children "know" certain words or that their lexicon contains certain words relies on a given linguistic definition of what a word is and according to a list of words matching this definition. However, the systematic ways in which linguists use the notion of lexicon does not imply that children's knowledge relies on a similar systematic organization. These concepts do not reside in the child but rather constitute conventions according to which scholars describe and classify children's communicative abilities, taking their description of adult lexicon and morphology as model. These pre-defined linguistic concepts are used to construct stimuli for experimental tasks in which children's responses are elicited. The terminology we use also constrains what is considered relevant to observation and how observations are classified. For example, when we say that children have correctly acquired a word, the judgement of correctness is based on what is in the adult inventory instead of what constitutes communicative success.

## *References*

Ambridge, B., Kidd, E., Rowland, C. F., & Theakston, A. L. (2015). The ubiquity of frequency effects in first language acquisition. *Journal of Child Language*, 42, 239–273. https://doi.org/10.1017/S030500091400049X

Clark, E. V., & Wong, A. D-W. (2002). Pragmatic directions about language use: words and word meanings. *Language in Society*, 31, 181–212.

Evans, V. (2009). *How words mean: Lexical concepts, cognitive models, and meaning construction*. Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199234660.001.0001

Goldberg, A. (1995). *Constructions: A construction grammar approach to argument structure*. Chicago: Chicago University Press.

Griffiths, T. L., Chater, N., Kemp, C., Perfors, A., & Tenenbaum, J. B. (2010). Probabilistic models of cognition: Exploring representations and inductive biases. *Trends in Cognitive Sciences*, 14(8), 357–364. https://doi.org/10.1016/j.tics.2010.05.004

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 104(2), 211. https://doi.org/10.1037/0033-295X.104.2.211

Langacker, R. W. (2000). A Dynamic Usage-Based Model. In M. Barlow & S. Kemmer (Eds.), *Usage-Based Models of Language* (pp. 1–63). Stanford: CSLI.

Medina, T. N., Snedeker, J., Trueswell, J. C., & Gleitman, L. R. (2011). How words can and cannot be learned by observation. *Proceedings of the National Academy of Sciences*, 108(22), 9014–9019. https://doi.org/10.1073/pnas.1105040108

Pustejovsky, J. (1995). *The generative lexicon*. Cambridge, MA.: MIT Press.

Sandoval, M. & Gómez, R. L. (2013). The development of nonadjacent dependency learning in natural and artificial languages. *WIREs Cognitive Science*, 4, 511–522. https://doi.org/10.1002/wcs.1244

Tooby, J., Cosmides, L. & Barrett, C. (2005). Resolving the debate on innate ideas: Learnability constraints and the evolved interpenetration of motivational and conceptual functions. In Carruthers, P., Laurence, S. & Stich, S. (Eds.), *The innate mind: Structure and content*. NY: Oxford University Press, 305–337. https://doi.org/10.1093/acprof:oso/9780195179675.003.0018

## Russel Richie commentary on Westbury thesis

The author is right that a complete theory of (lexical-)semantics (henceforth, semantics) will have to appeal to domains outside of semantics. However, I think we should be careful not to throw the baby out with the bathwater: many useful explanations of semantics rely in some way on appeals to semantics.

Just as lay learners of a language find thesauri and dictionaries useful to understanding a lexicon, so, too, do cognitive scientists find these and other cyclical lexical data structures (e.g., word association norms) revealing of how we learn (Hills et al., 2009), represent (Steyvers & Tenenbaum, 2005), and process (Griffiths et al., 2007; Thompson & Kello, 2014) semantics in its various forms (for review of networks, including semantic networks, in cognitive science, see Baronchelli et al., 2013).

To put a finer point on it, meaning *is* often constructed out of other meanings (which the author emphasizes – the meaning of a word *isn't* atomistic).[2] We learn new meanings, concepts, words by hearing or reading about them, i.e., in terms of other meanings. We're able to state an infinite number of *new* concepts and sentences in terms of *old* concepts. To the extent that all this is true, explaining (some instances/aspects of) meaning will *require* appeal to (other instances/aspects of) meaning itself. In fact, this is largely the enterprise of compositional semantics in linguistics. But lexical semanticists ought to be interested in this, as well.

On a more philosophical note, what is the end-state of a science demanding explanations of X in non-X terms? If we require that everything have *some* explanation, then we can't ultimately have everything appeal to some Ur-cause or

2. *Do* many cognitive scientists still subscribe to atomism of concepts, as Westbury seems to believe? This is one area that would be interesting to probe in a survey of cognitive scientists a la Bourget and Chalmers's (2014) survey of professional philosophers.

Ur-explanation. That possibility aside, there seem to be two possibilities. We either have cycles of explanation, where A is explained in terms of B, B in C, and C in A, or we constantly create new explanatory domains, yielding explanatory turtles all the way down. The former seems like it doesn't really satisfy Hempel and Oppenheim's desideratum that the explanans be distinct from the explanandum. The latter – that we will never truly finish understanding the universe – is either depressing or exciting, depending on one's perspective.

*References*

Baronchelli, A., Ferrer-i-Cancho, R., Pastor-Satorras, R., Chater, N., & Christiansen, M. H. (2013). Networks in cognitive science. *Trends in Cognitive Sciences*, 17(7), 348–60. https://doi.org/10.1016/j.tics.2013.04.010

Griffiths, T. L., Steyvers, M., & Firl, A. (2007). Google and the mind: predicting fluency with PageRank. *Psychological Science*, 18(12), 1069–76. https://doi.org/10.1111/j.1467-9280.2007.02027.x

Hills, T., Maouene, M., Maouene, J., Sheya, A., & Smith, L. (2009). Longitudinal analysis of early semantic networks preferential attachment or preferential acquisition? *Psychological Science*, 20(6), 729–739. https://doi.org/10.1111/j.1467-9280.2009.02365.x

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. https://doi.org/10.1207/s15516709cog2901_3

Thompson, G. W., & Kello, C. T. (2014). Walking across Wikipedia: a scale-free network model of semantic memory retrieval. *Frontiers in Psychology*, 5(86). https://doi.org/10.3389/fpsyg.2014.00086

## Benjamin Tucker commentary on Westbury thesis

The article by Westbury, "Explaining semantics without semantics" provides a road forward, casting aside many of the previous assumptions about semantics. The author outlines eight possible domains for exploring and understanding semantics. Each of these candidates provides an interesting domain in which further experimental research could be performed, testing Westbury's many proposals. In this commentary I have selected one domain and will draw some parallels between these domains and casual speech.

One interesting possible candidate is *Domain 4: Action in the real world*. This domain as described in this paper, parallels arguments made by Tucker and Ernestus (this issue). In both cases the authors underscore the importance of investigating the real world consequences of language. How is language recognized by speakers in spontaneous conversations or what do co-occurrence patterns tell us about what a word means? These two questions could be joined by asking if these co-occurrence patterns are related in a way about reduction. One might predict that reduction would be more likely to occur when a word occurs in a common semantic environment. As an out of domain example, when attending a casual or family party, the way in which a person acts/dresses is going to be more informal and relaxed, whereas if you're attending a party with special dignitaries the interactions are going to be more formal and careful. Even the attendees personal appearance might change from jeans and a t-shirt to a suit. The acoustic realization of a word might vary greatly depending on its co-occurrence with other words or the context in which it is produced. In situations where the word is surrounded by its closest "friends" it might be more relaxed and as a result appears in a more relaxed/reduced form, whereas when the word finds itself in unfamiliar company the words' realization and appearance might differ.

These proposed domains have further connections to representation. Westbury indicates the need for multiple representations for a semantic label. This is further supported by work in casual speech where many researchers argue for the need of multiple representations to account for the inherent variability found in the speech signal (Ernestus, 2014). Further, the notion that access is not a single process, I am extending beyond just semantic access, finds support from research on casual speech – which indicates that many processes are involved from perception to access allowing the listener to achieve efficient access.

*References*

Ernestus, M. (2014). Acoustic reduction and the roles of abstractions and exemplars in speech processing. *Lingua*, 142, 27–41. https://doi.org/10.1016/j.lingua.2012.12.006

Tucker, B. V., & Ernestus, M. (2016). Why we need to investigate casual speech to truly understand language production, processing and the mental lexicon. *The Mental Lexicon*, 11(3), 375–400. https://doi.org/10.1075/ml.11.3.03tuc

## The article

*'A thing is identical with itself.' – There is no finer example of a useless proposition, which yet is connected with a certain play of the imagination. It is as if in imagination we put a thing into its own shape, and saw that it fitted.*
Ludwig Wittgenstein (1958)
*Philosophical Investigations*
p. 84

In the conclusion to Moliere's (1673) comedic play *Le Malade Imaginaire [The Imaginary Patient]*, a medical student is examined in pseudo-Latin by pompous doctors. When the student is asked why morphine puts a person to sleep, he is praised by his self-important examiners for answering that morphine puts a person to sleep because it contains a *virtus dormitiva*, a dormitive property. We laugh at this explanation, as Moliere intended us to, because it is obvious that the answer is devoid of content since it merely *re-labels* the phenomenon it pretends to *explain*. The idea that scientific explanation needs to explain a phenomenon *in terms other than itself* seems obvious. We would be unimpressed by a person who explained why the sky was blue by telling us that it is because the sky is blue. However, self-referential explanations such as this – what Bateson (1972) called 'dormitive principles' – are more common in the information sciences generally than in the physical sciences. The reason is that it is difficult to map explanations outside of the domain in information sciences. We want to explain information structures in terms of information structures. Dormitive principles are common in psychology in particular (where they often hide behind an implicit appeal to a homunculus, the magical 'last step' in which a special sensing part of the brain actually *senses* the representation that had theretofore been unsensed), and especially in theories of semantics. In this paper I will examine how self-referential explanations find a foothold in research into semantics, provide an overview of research directions that offer a way out of the circularity, and discuss what that research suggests about the meaning of a word.

In papers looking towards the future of a discipline, the focus is often on new techniques that are becoming or may soon become available. If I offer no suggestions in that direction here, it is in part because I hope to convince the reader that we are already living in a golden age for studying lexical semantics, closer to understanding what semantics is than we have ever been. My view of the future of research into semantics involves a modification that is mainly philosophical, or (if that term offends the empiricist in you) conceptual. The future of semantics lies in *adjusting our understanding of what it will mean to explain semantics*. That adjustment has direct implications for what kind of experiments we are going to want to design and what kind of theorizing those experiments will support. In that sense, it speaks to the future of semantic studies.

## What makes an explanation non-dormitive?

The issue of explanatory circularity is more subtle than the trivial examples above may make it seem.

One subtlety is that some dynamic systems (and most relevantly biological systems at various levels of description) may be properly described as self-organizing or autopoietic (see Maturana and Varela, 1980; Kugler & Turvey, 1987; Thelen & Smith, 1994; Mingers, 1995; Varela, Thompson, & Rosch, 1993). An explanation of such a system in terms of its own structure and dynamics would not be a dormitive property. Autopoeiesis may be relevant at both social and neural activation levels in understanding semantics (for an accessible introduction, see Elman, 2009). Many issues related to causality, explanation, and the emergence of non-random structure at different time scales are raised by dynamic systems theory. Here, rather than trying to untangle these issues here, we will simply note that they exist and sidestep them, referring the interested reader to the cited resources.

In order to understand another way in which the issue of explanatory circularity can be subtle, let us consider two more interesting answers to the question 'why is the sky blue?'.

We might be slightly impressed by a person who explained to us that the sky is blue because it transmits wave lengths about 470 nanometres long. This might be so even if we recognize that saying 'the sky transmits wave lengths about 470 nanometres long' is identical to saying 'the sky is blue'. The explanation is a still something of a dormitive principle, but at least it tells us a new fact (the wavelength) about blueness. That seems to make it a better explanation than the first. We have moved from pure tautology to synonymity, an identity mapping between two ways of describing the phenomenon that are, at least, grounded in different domains: *blueness* in *phenomenology*, and *wavelength* in *physics*.

We would surely be much more impressed by a person who explained that the sky is blue because:

a. Long wavelengths pass through air without interacting much with hydrogen and oxygen gas molecules in the air.
b. Short wavelengths are much more likely to absorbed by hydrogen and oxygen gas molecules in the air.
c. The short wavelengths that are absorbed by hydrogen and oxygen molecules are scattered in all directions, a phenomenon known as 'Rayleigh scattering'.
d. Those short scattered waves reach our retina because they are scattered.
e. Short wavelengths excite certain specialized receptors ('blue cones') in our retina.
f. Exciting blue cones leads to the subjective experience that English speakers have collectively agreed to call 'seeing blue'.

Following Hempel and Oppenheim (1948) we can call the set of statements describing the empirical phenomenon to be explained (in this case, the claim that 'the sky is blue') the *explanandum*, and we can call the statements describing the evidence that can explain (or, more precisely according to Hempel and Oppenheim, allow one to predict) that phenomenon (points a.–f. above) the *explanans*.

There are two points worth highlighting about the differences between the dormitive principle explanans ('[the sky is blue because] the sky is blue') and the best explanans ('[the sky is blue because] points a.–f. are true').

The first point is that the best explanans escapes tautology by appealing to causal forces that act in different explanatory domains than the domain implicit in the explandum. Instead of explaining the sky's colour in terms of the sky, its observed colour, or a re-description of that colour, the blueness is explained in terms of the physical properties of wavelengths, molecules, and the hydrogen- and oxygen-covered planet we inhabit: that is, explained in terms of empirically-accessible things that are not themselves blue.

The second point worth highlighting is that the final statement ('Exciting blue cones leads to the subjective experience that we have collectively agreed to call *seeing blue*') is simultaneously *psychological* and *problematic* (as are so many statements, alas!). It is *psychological* because it appeals to a private phenomenological experience. It is *problematic* because it is very close to a dormitive principle. It essentially says 'we call blue things *blue*', thereby attempting to explain the *meaning* of the phenomenon of blueness in terms of the *experienced meaning of the phenomenon of blueness*.

The problematic sentence pays lip service to breaking out of the domain by implying (perhaps in an attempt to follow Wittgenstein, 1958)[3] that the explanation for why blue things are *blue* is to be found not in the mind, but rather in the history of social interaction.[4] Our explainer might perhaps have made that even clearer by discussing the etymology of the word 'blue': e.g. by saying 'we call things that seem to be blue *blue* because the word is derived from the Old High German word *blao*', all the while hoping that the interrogator didn't ask him to explain why the Germans decided on just *that* word for that experience. ("Why did they call the sky *blao*? Because *the sky is blue* and they were speaking *Old High German*, dammit!")[5]

This amateur foray into middle school physics gives us clues about what we want from any explanation that is not just a dormitive principle. We want to ground that explanation in a set of causal mechanisms that unfold in domains:

a. That are distinct from the one we are trying to explain, and
b. That have causal regularities that are easier to track than any regularities in the domain we are trying to explain.

---

3. As I am doing here myself. I have included footnotes to make clear that much of what I have to say was said decades ago in Wittgenstein's (1958) *Philosophical Investigations*.

4. "…nothing is more wrong-headed than calling meaning a mental activity! Unless, that is one is setting out to produce confusion." (Wittgenstein, 1958, p. 172)

5. "How do I know that this colour is red? – It would be an answer to say: 'I have learnt English'." (Wittgenstein, 1958, p. 117)

It is helpful, if not easy, to break b.) down still further. What makes causal regularities in one domain *easier to track* than the correlated causal regularities in another domain? It is not easy to answer this question because there are many answers. The general idea is a broad one: If we feel *slightly less confused* in the second domain than we did in the first domain, we have made some scientific progress by mapping into the second domain. Since there are many ways to become slightly less confused, there are many ways that correlated causal regularities can be easier when tracked in one domain than in another. That said, perhaps the most important ways causal regularities can be easier to track are because:

a. The regularities apparent in the explanatory domain are themselves grounded in other well-understood domains, or
b. The regularities apparent in the explanatory domain help to cleave the phenomenon of interest into smaller pieces that at least *look to be* (and perhaps objectively *are*) more tractable than the phenomenon as a whole.

In this paper, I consider how we might map lexical semantics into explanatory domains that might help us understand semantics in much the same way that mapping blueness into physics helped humans to understand why the sky is blue. Although the same general principles of domain-mapping apply, explaining semantics will be much more difficult than explaining why the sky is blue, because semantics is subject to multiple more-or-less independent causal factors, or levels of constraint. My goal is not to outline these causal factors exhaustively (which is impossible to do because it is very unlikely that they have all been identified yet) but rather to consider some examples of the *kind* of causal factors that are relevant to lexical semantics, and, more generally, to consider how we might think about problems (nearly all problems in psychology) in which explanations require us to consider more than one cause, each of which accounts for a small part of the variance in the phenomenon we are trying to explain.

## Mapping lexical semantics into other domains

When it comes to explaining lexical semantics, we professional semanticists have too often and for too long allowed ourselves to fall back on dormitive principles. As Bergen and Chang (2013) put it: “Traditionally, linguists use quick-and-dirty approximations of meaning, often dodging the issue of what meaning is by *merely labeling word meanings* (for instance, the meaning of the word cat might be represented as CAT or as a collection of ungrounded semantic features, like [+FELINE, +DOMESTIC])” [Emphasis added] (p. 173). We use many analogous dormitive principles that are more subtle. When we treat human semantic judgments as an independent variable, we are correlating semantics with semantics, thereby explaining nothing. The same is true when we try to explain semantics in terms of rules or human-defined synonym/association sets,[6] both of which also 'help themselves' to the semantics they are trying to explain.

Everyone knows *why* we explain semantics in terms of semantics, or a pointer to it. We do it because the problem of semantics is so difficult. To cleave it into smaller, more tractable parts, we simplify the problem by ignoring the last step, which is *the actual linking of semantic structures to meaning*. We do this even though we know that the actual linking of semantic structures to meaning is *the* essential problem of semantics. We leave that part for future research because it is difficult and we assume that someone else will figure it out later. This is not an unusual thing to do. Assuming a yet-to-be-found 'unknown invisible' to get over the hard part of a scientific problem is a time-honored tradition in science, which has led to the discovery of such useful entities as phlogiston, the demons that take possession of the mentally ill, and the Cartesian animal spirits that cause motor movements by filling up our interfibrillar spaces.

What else might we do? Rather than going all the way up to semantics without actually doing any semantics, I suggest (following the discussion above) that semanticists need to be looking for domains into which we can map the hard part so it seems more tractable. I do not pretend this is a radical new view. In recent years, a lot of work has been done on mapping semantics into other domains. I want to argue that this work is all there is to explaining semantics. The future of semantic studies is to keep doing what we are doing, to recognize that we are doing exactly the right thing by doing many disparate things.

In this section I will briefly point out eight domains into which semantics has been mapped: linguistic unit co-occurrence, extensions to perceptual discriminative learning, word form, action in the real world, action in a simulated world, innate perceptual categories, metaphorical extensions of those categories, and affect. My goal is not to exhaustively review the work in mapping semantics into these eight domains, which would be impossible to do since all of them are underlain by extensive literatures. The goal of presenting them here is rather simply to motivate the idea that there are differences in each of these domains that do correlate with experienced semantic differences. In the next section I will discuss how this might help us think coherently about that difficult final step.

**6.** "The use of the word 'rule' and the use of the word 'same' are related to one another, they are cousins. If I teach anyone the use of the one word, he learns the use of the other with it." (Wittgenstein, 1958, p. 86).

*Domain 1: Co-occurrence of linguistic units*

Most late second-language learners know from experience that it is possible to understand the meaning of new words simply from seeing them used in context often enough. I clearly remember suddenly coming to understand the meaning of the French word *quotidien* (*daily*) in this way when I saw the word on a billboard at a metro stop in Paris. Co-occurrence models of semantic such as LSA (Landauer & Dumais, 1997), HAL (Lund & Burgess, 1996), HiDEx (Shaoul & Westbury, 2008, 2010, 2011), BEAGLE (Jones & Mewhort, 2007), WINDSORS (Durda & Buchanan, 2008) and others measure formalize this by bootstrapping semantics from patterns of co-occurrence of some linguistic unit (words, short phrases, or documents) in the text. Most recently, closely-related models have been proposed that do not simply count co-occurrence of words in a large corpus within a moving window, but rather try to predict each target words context in a large corpus using a neural network (Skip-gram; Mikolov, Chen, Corrado, & Dean, 2013; Mikolov, Sutskever, Chen, Corrado, & Dean, 2013). The models compile data about co-occurrence of words in a large corpus of text (where 'co-occurrence' is defined slightly differently in different models), and use the distance between the resulting co-occurrence vectors (second order co-occurrence) to compute measures of association. The vectors have many semantic-like properties. For example, it has been shown that distance between co-occurrence vectors in various models is sufficient for obtaining a passing grade on the TOEFL (which asks for synonyms of words; Landauer & Dumais, 1997), can equal student performance on grade school multiple choice biology tests (Lifchitz, Jhean-Larose, & Denhière, 2009), correlates with behavioral measures of lexical access (Buchanan, Westbury, & Burgess, 2001), and can be used to predict human judgments of, e.g., imageability (Westbury, Shaoul, Hollis, Smithson, Briesemeister, Hofmann, & Jacobs, 2013), subjective familiarity (Westbury, 2013), and affect measures such as valence and arousal (Westbury, Keith, Briesemeister, Hofmann, & Jacobs, 2015; Hollis, Westbury, and Lefsrud, 2017).

A recent analysis of the skip-gram matrix using principle component analysis (Hollis and Westbury, 2016) suggests that, despite it being constructed using only information about lexical co-occurrence, there is a great deal of information in the matrix other than what we would normally called 'semantic' information. This includes information about word frequency, word length, affect, and even form similarity of the target word to other words in the language (orthographic neighbourhood size). It had been suggested decades ago that frequency might carry semantic information. Harris (1970) wrote that "[i]f we consider words or morphemes A and B to be more different in meaning than A and C, then we will often find that the distributions of A and B are more different than the distributions of A and C. In other words, difference of meaning correlates with differences of distribution" (p. 13).

Recent work suggests that semantic information can be gleaned from patterns of co-occurrence using lower-level co-occurrence units than words. Baayen, Milin, Filipović Đurđević, Hendrix and Marelli (2011) and Baayen, Shaoul, Willits, and Ramscar (2016) both discuss and analyze the performance of a naïve discriminant model that models language learning using patterns of co-occurrence of three-character strings that may cross word boundaries (i.e. that may include a space among their three characters).

Co-occurrence models are a good example of a domain that demonstrates changes that correlate with semantic changes, but that are 'easier to track' than they are in semantics. In virtue of being defined in fully algorithmic terms with fully quantified theoretical entities, it is possible to have a good understanding of what is happening in co-occurrence models. Co-occurrence models also help us think about one of the great puzzles of lexical semantics, which is: How can we represent the meanings of abstract words that have no real-world referents? Many of the domains below deal with 'tacking' lexical semantic representations onto non-lexical representations (i.e. our language-independent notion of, say, a cat). Pure lexical co-occurrence can delimit the meanings of words independently of whether they have any non-lexical representations.

*Domain 2: Extensions to discriminative learning*

Along with their formal structure, co-occurrence models have an additional attraction, which is that they provide a direct bridge to another domain in which we can more easily track semantic changes than we can when semantics is confined to its own domain: the world of discriminative learning (Rescorla & Wagner, 1972). Discriminative learning has a long and fruitful history in psychology. It is mathematically well-specified in way that is coherently integrated with the formalisms of information theory (Rescorla, 1988) and it is supported by a huge body of empirical work. For many decades, discriminative learning was considered irrelevant to language, because of a dogmatic belief (nurtured by Chomsky and his acolytes) that 'language was special', unique in being impossible to explain using the normal learning principles that were being applied with so much success in other areas of psychology. This once deeply-entrenched dogma has recently begun to face serious challenge. As its name suggests, the naïve discriminant learning model discussed above, which can simulate many linguistic phenomena, was explicitly built using the standard discriminative learning principles first laid out in Rescorla and Wagner (1972). Word-word co-occurrence models are also entirely consistent with them. In one of the first papers exploring such a model, Landauer and Dumais (1997) noted that their model "accomplishes much the same thing as conditioning rules such as Rescorla and Wagner (1972) in that it makes the primary association better represent the informative relation between the entities rather than the mere fact that they occurred together" (p. 216).

*Domain 3: Word form*

Another intriguing thing about models that bootstrap language learning from *sub-word units* is that they can explain one of the mysteries of language: why are there patterns of non-etymologically-related, morphologically-irrelevant *word forms* that have shared meanings? This is the problem of sound symbolism, understanding why some strings seem to have inherent semantics from their form alone. Although sound symbolism has an ignoble history, there is little doubt that *some* word forms do correlate with some meanings. The most obvious examples are onomatopoeic words such as *sizzle, swoosh, ding-dong*, and *meow* that deliberately resemble the sound of their referents. Phonesthemes are families of words that share meanings and form without being etymologically related and without resembling their referent. A commonly-cited example in English is the category of 'gl' words – *glitter, glow, glimmer, gleam, glare, glisten, glint*, and *glance* – which are etymologically unrelated but all have to do with a short intense visual experience. Another example is that concrete words tend to be shorter than abstract words, in part because concrete words tend to be more common. Reilly, Westbury, Kean, and Peele (2012) showed that people were sensitive to this difference, and would reliably judge shorter nonwords as being more concrete than long nonwords. This work was recently replicated and extended using judgments about abstract and concrete foreign words instead of nonwords (Reilly, Hung, and Westbury, 2017).

Baayen et al. (2011) analyze phonesthemes as 'pseudo-affixes' (although as noted above, their model makes no distinction between lexical regularities and affixes). They emphasize that "a discriminative learning approach predicts that even small local consistencies in the fractionated chaos of local form-meaning correspondences will be reflected in the weights, and that they will codetermine lexical processing, however minute these contributions may be" (p. 56).

Certainly, most words do not have any sound symbolic value. It is equally certain that there are small local consistencies in the fractionated chaos of local form-meaning correspondence. Insofar as the brain makes use of this form similarity when it is available, this suggests that not all semantic representations are likely to be represented in the same way. The important general points, whose depths have probably not yet been plumbed, are that form-meaning correlations can sometimes provide reliable clues to word meaning and that much evidence suggests that humans are able to reliably detect these correlations when they are there.

### *Domain 4: Action in the real world*

One of the limitations of models of semantics based on lexical co-occurrence is that they are insulated from the real world. It is obvious that we do not need to wait for someone to utter the sentence *Cats are furry* in our presence in order to associate the word *cat* with the word *furry*. We just need to hang out with a cat.

It has been computationally and theoretically simplifying to develop co-occurrence vectors from text corpora alone. However, there is nothing in the formal structure of co-occurrence models that precludes them from incorporating co-occurrence of linguistic elements with non-linguistic elements of experience, and it is trivially obvious that real world co-occurrence can fix at least some lexical meanings: If you ask me what a cat is, I can just point to one. It is notable that many theorists of *language learning* have made much of the fact that humans are the only primates that have been observed to spontaneously and regularly exhibit pointing behavior (e.g. see Bloom (2000), p. 85).

Extending Morris's (1946) dispositional view of meaning, Osgood (1952) and Osgood, Suci, and Tannenbaum (1957) proposed that semantics was grounded in conditioned behavior towards real objects in the world. Their basic idea was that a word (or other sign) of a thing produces a (weakened) disposition to act in the same way as the disposition produced by the thing itself. Dispositions associated with signs in this way could modulate the dispositions associated with other signs or objects, creating more complex new dispositions that could in turn become associated with other signs.

Although empirical work in relating lexical semantics to real-world experience is difficult because the real world is complex and hard to control, there has recently been much success in studying the relation under controlled or simulated conditions (e.g. Roy & Pentland, 2002; Yu & Ballard, 2007; Smith & Yu, 2008; Andrews, Vigliocco, and Vinson, 2009; Johns & Jones, 2015; for reviews, see Vigliocco, Meteyard, Andrews, & Kousta, 2009; Smith, Suanda, and Yu, 2014).

### *Domain 5: Real world simulation*

I cited Bergen & Chang (2013) above for their observation that linguists have often represented meaning with arbitrary representations (CAT) or sets of ungrounded features. They continue on directly from that observation to write:

> But work on mental simulation suggests that what language users know is how to take a linguistic form (such as a word or grammatical structure) and *activate perceptual or motor representations corresponding to the denoted entities* and events, and vice versa. That is, *linguistic units drive mental simulations*: instead of fizzling out in static, formal, arbitrary representations (like CAT), meaningful words and constructions serve as interfaces that activate and constrain modality-rich mental simulations. (p. 173; Emphasis added)

There is now a great deal of evidence that lexical semantics is in part fixed by simulated meanings grounded in the domain of sensory-motor mechanisms (e.g. Fincher-Kiefer, 2001; Glenberg, 1997; Glenberg and Kaschak, 2002; Zwaan, 2004; for a review, see Barsalou, 2008). This was suggested long ago by a man who rarely gets the credit he deserves for pushing this idea hard: Sigmund Freud in his unjustly neglected (1891/1953) book *On Aphasia* (though Freud had forerunners for this idea; see Greenberg, 1997). In that book, Freud argued for a complex word form that was composed of several neural representations that were quasi-distinct (or, at least, theoretically separable following brain damage), each linked to a different sensory-motor modality:

> The word, then, is a complicated concept built up from various impressions, i.e. it corresponds to an intricate process of associations entered into by elements of visual, acoustic and kinaesthetic associations […] The idea, or concept, of the object is itself another complex of associations composed of the most varied visual, auditory, tactile, kinaesthetic, and other impressions.
>
> (Freud, 1891/1953, p. 77–78)

The idea that words and concepts are both represented in multi-dimensional ways in the brain raised for Freud many of the same points discussed in a wider context in this paper: that word meanings may be variably represented and therefore be more or less similar along several quasi-independent dimensions. He even goes so far as to suggest that this might result in there being distinctly different forms of word/concept representation in one person than in another: "one individual speaks, writes, and reads predominantly or exclusively with the help of kinaesthetic sensory impressions, while another may employ the visual element for the same purpose, etc." (p. 98).

Freud argued for his complex word form on the basis of evidence from aphasics, which provides much evidence to support the idea that there are many different and dissociable aspects to any semantic representation. Recent fMRI evidence from experiments designed specifically to test the hypothesis that word representations are in part also sensory-motor representations has provided clear experimental support for the idea. For example, Hauk, Johnsrude, and Pulvermüller (2004) used fMRI to show that when participants were asked to passively read action words related to specific body parts (e.g. *lick, pick*, or *kick*), there was increased activity in the relevant region of the somatotopic motor strip. We don't kick when we read the word *kick* but our brain gets a little hint of what an actual kick is (see also Aziz-Zadeh, Fiebach, Naranayan, Feldman, Dodge, & Ivry, 2008; Pulvermüller 1999, 2005; Speer, Reynolds, Swallow; & Zacks, 2009). Again, we see a mapping from lexical meaning into a domain that is more concrete and phylogenetically universal – and therefore more amenable to empirical study – that is not lexical meaning.

It is worth again pointing out the obvious: that some words can be represented this way very easily, and some cannot, again suggesting that not all semantic representations are likely to be represented in the same way neurologically, a point I will return to below.

*Domain 6: Embodied human experience*
A closely related domain onto which some semantics can be mapped comes from the founding assumption of cognitive semantics, that there are universal semantic categories derived from the universally shared commonalities of embodied human experience (Gibson, 1977; Johnson, 1987; Lakoff, 1987). Cognitive semanticists have argued that the nature of embodied human experience automatically instills us with certain semantic categories, such as *inside/outside* (and the related class of *container*), *part/whole, up/down, you/me*, and many more.[7] These semantic categories transcend any need for experiencing any particular mapping between the world and language, since they are directly experienced and understood in virtue of our very nature as living beings.

*Domain 7: Metaphorical extensions of embodied human experience*
The importance of these universal categories is enhanced by their natural and equally universal extension as metaphors, which some have claimed serve as the basis of many abstract semantic categories. Johnson (1991) noted that "metaphorical understanding is so pervasive and so deeply constitutive of our intentional interactions within our environment that we are virtually unaware either of its existence or of its metaphorical character" (p. 11). Rumelhart (1979) argued many years ago that metaphorical extensions were just 'business as usual' for the language system, writing that "the classification of an utterance as to whether it involves literal or metaphorical meanings is analogous to our judgment of whether a bit of language is formal or informal" (p. 72), in the sense that "it is a judgment that can be reliably made, but not one which signals fundamentally different comprehension processes" (p. 72). Rumelhart's theoretical speculation on this matter has recently received some empirical support from the finding that exposure to metaphorical extensions of sensory-motor-related words (e.g. 'The country lifted the veil on its nuclear problem') is correlated with activation in sensory-motor regions of the brain (Desai, Conant, Binder, Park, & Seidenberg, 2013).

Suggested metaphorical extensions of universal embodied categories are numerous, including, e.g., the characterization of both purpose and time as a path

7. "'This body has extension.' To this we might reply: 'Nonsense!' – but are inclined to reply 'Of course!' – Why is this?" (Wittgenstein, 1958, p. 90).

('I am going to win'), the conflation of height with quantity ('Fill her up'), and the characterization of anger in terms of dangerous animals ('Don't snap at me!'; Lakoff & Johnson, 1980; Lakoff, 1987).

*Domain 8: Affect*

Although my list is not intended (nor needs) to be exhaustive, the last domain I will consider is the domain of affect. Writing in 1877 (and serving as a direct inspiration for Freud's closely-related views on language), the German physician Adolf Kussmaul noted that "When an idea is expressed through speech, the emotions are the prime movers of the spoken word" (cited in and translated by Greenberg, 1997, p. 44). Experienced affect is something very close to lexical semantics: a directly experienced, unambivalently meaningful internal mental state that can be excited by a specific stimulus class.[8] Although affect might be considered as just another universal embodied category, evidence suggests that it may play a special role in semantic access. Affective dimensions have, for example, been shown to affect lexical access times (Kousta, Vinson, & Vigliocco, 2009; Westbury, Keith, Briesemeister, Hofmann, & Jacobs, 2015) and word recall probability (Adelman and Estes, 2013), and to play a role in organizing meaning judgments (Osgood, Suci, & Tannenbaum, 1957. This and other evidence has suggested to some that affective processing plays a direct role in lexical semantics (see discussion in Vigliocco, Meteyard, Andrews, & Kousta, 2009).

Affect is of particular interest in the present context, because an attempt has already been made to eliminate the homunculus in emotional access in much the way I am arguing we need to eliminate it in semantic access. The famous James-Lange theory of emotion, which is now widely accepted (Dagleish, 2004), argues against the idea that first an emotional representation is evoked, and then that representation is experienced. James (1890/1950) wrote that:

> Our natural way of thinking about […] emotions is that the mental perception of some fact excites the mental affection called the emotion, and that this latter state of mind gives rise to the bodily expression. My theory, on the contrary, is that *the bodily changes follow directly the perception of the exciting fact, and that our feeling of the same changes as they occur* IS *the emotion*. Common-sense says, we lose our fortune, are sorry and weep; we meet a bear, are frightened and run; we are insulted by a rival, are angry and strike. The hypothesis here to be defended says that this order of sequence is incorrect, that the one mental state is not immediately induced by the other, that the bodily manifestations must first be interposed between, […] Without the bodily states following on the perception, the latter would be purely

**8.** "Can I not say: a cry, a laugh, are full of meaning? And that means roughly: much can be gathered from them." (Wittgenstein, 1958, p. 147).

> cognitive in form, pale, colorless, destitute of emotional warmth. We might then see the bear, and judge it best to run, receive the insult and deem it right to strike, but we should not actually *feel* afraid or angry. (p. 449–450)

This view of affect is a close analogy to how I suggest we should think about semantics. There is no need to posit an additional canonical semantic representation that is separable from our embodied experience of the multiple and variable constraints that act on us when we experience a word. In the next section I will discuss this view further.

## Semantics as a mapping from multiple domains

I reference the eight domains above to motivate two closely-related ideas:

a. the idea that we do not have one single mental representation, but rather many mental representations per semantic label and
b. the idea that semantic access is not one single process at all.

### **a.** *Semantic representations are not unitary*

Our habit of stopping at a semantic representation's label ('the meaning of *cat* is CAT') has had an insidious side effect, which is that it has made it natural to assume (by directly implying) there is exactly one semantic representation for each word, or at least, each word meaning. There are two good, closely-related 'intuition pumps' (to borrow a term from philosopher Dan Dennett, 1991) we can contemplate to get us used to thinking otherwise: the concrete process of medical diagnosis, and the more abstract structure of regression models.

Medical diagnosis is just one of a number of complex taxonomic enterprises in which humans engage on a regular basis. Taxonomists since Aristotle have to face the problem that the world is messy and many categories have to be defined fairly loosely if they are to be useful. In facing this problem, they have sometimes been forced to create categories that assign exactly the same label to two cases that are completely disjunct according to objective criteria. There are many polythetic categories in the DSM-V (American Psychiatric Association, 2013) that allow for diagnosis of the same mental disorder with disjunct symptom sets.

Psychometric instruments and symptom lists are really just examples of a simple linear regression model over a set of observable characters, in which the beta weights have (most often) all been set to one. We can easily generalize the point about polythetic diagnostic categories to other regression models, since the same point holds for any of them: it is always possible to get the same fitted value using a disjunct set of predictors. We can add as many complications as anyone likes (say, add a thousand more predictors and make the regression model horribly non-linear) without altering the basic point that we can get to *the same dependent measure value* from *very different sets of independent characters* or *from very different weightings of the same set of characters*.

**b.** *Semantic access is not one single process*

The second claim I want to defend follows directly from the first. If access to any particular semantic structure is indeed some variably-weighted mapping from a multitude of domains, then it follows that semantic access is not a singular process. Many roads lead to ROME.[9]

Consider this question: What does it mean to say that two different scores on a psychometric instrument are *the same*, given that the same score can be obtained as a function of different (possibly entirely disjunct) responses? We call two scores on, for example, the Beck Depression Inventory (BDI-II) *the same* to indicate that two people are depressed to the same degree, even if they obtained their scores by attesting to two different (conceivably even entirely disjunct) sets of symptoms. However, if we saw the same disjunct sets of responses often enough, we might at some point change our mind about that judgment of identity. We might decide that the score obtained from one set and the exact same score obtained from the other set were in fact measuring different things, i.e. two distinct subtypes of depression in this case. At that point, we would not say that two scores on the two different subsets were 'the same' any more than we would say that *obtaining a* BDI-II *score of 29* is the same as *having 29 grains of rice*.[10] I want to argue here that the meaning of an ordinary word like *cat* is closely analogous to the meaning of an ordinary number like 29. We should not necessarily take *the identity of two labels* (the same number attached to different measurements) to indicate *the identity of those labels' meaning*. When we punt on semantics by assuming simplistically that the meaning of the word *cat* is some information structure CAT, we are implicitly taking for granted that there is in fact one singular and unitary underlying semantic structure CAT.

But what if CAT is *not* a singular or unitary representation? What if the word *cat* rather a label we apply to a complex, weighted set of *CAT clues* from disparate domains? Rumelhart (1979) explicitly suggested this many years ago, writing:

---

**9.** "Language is a labyrinth of paths. You approach from one side and know your way about; you approach the same place from another side and no longer know your way about." (Wittgenstein, 1958, p. 82).

**10.** "Now has '1' a different meaning when it stands for a measure and when it stands for a number? If the question is framed in *this* way, one will answer in the affirmative." (Wittgenstein, 1958, p. 148).

> My approach suggests that comprehension, like perception, should be likened to Hebb's (1949) paleontologist,[11] who uses his beliefs and knowledge about dinosaurs in conjunction with the clues provided by the bone fragments available to construct a full-fledged model of the original. In this case the words spoken and the actions taken by the speaker are likened to the clues of the paleontologist, and the dinosaur, to the meaning conveyed through these clues. (p. 78)

Rumelhart writes of '*the* meaning', though his paper argues against the view that words have just one meaning because meanings are so naturally extensible through metaphor. To the extent that we are deducing word meaning from a set of clues, our own individual experience of CAT at one time and in one context may be partially or completely disjunct from our individual experience of CAT at another time and in another context. If the predictors are numerous enough, the beta weights variable enough, and/or the process that combines everything complex enough, we may have a great many disjunct experiences that we recognize as a good exemplar of the meaning of the word *cat* even though they have no characteristics in common.

Is this likely? Since we have many quasi-independent domains into which we can map semantics which may have more or less weight in different contexts and/or for different words, I think it is. Those domains provide numerous quasi-independent predictors that may combine in a multitude of different ways to bring to mind what we call *the same meaning* or (what we usually take to be a synonym) *the same representation*, when we really mean is *the same label*.

A final intuition pump may help convince the skeptics. There are many examples of a multi-determined singular subjective states that might seem more concrete than lexical semantic access. A convenient example is thirst, since it is a universally experienced, unambiguously meaningful, easily identifiable, nameable, and well-understood state, thereby similar to a semantic representation in all but the last characteristic. We know that the phenomenological feeling of thirst is determined by multiple physiological factors, including decreased blood volume (*hypovolaemia*, detected by cells in the kidney), decreases in arterial pressure (detected in the arteries themselves), changes in the osmotic pressure needed to halt movement through cell walls (detected by circumventricular organs outside the blood-brain barrier, as well as by other detectors in the body), decreased salt in the blood (also detected by cells in the kidney and by changes in arterial pressure), moisture in the mouth, and cultural factors. Although of course many of these physiological causal mechanisms are strongly correlated and have causal relations to each other, they are quasi-independent. For example, in normal functioning the body loses water from both intracellular and extra-cellular spaces, but during episodes of emergency

11. Hebb (1949) had compared the act of perception to the paleontological re-construction of a complete dinosaur from the recovery of a few bones.

rapid fluid dumping (vomiting and diarrhea) a disproportionate amount water is lost from extracellular fluid (McKinley & Johnson, 2004), resulting in a thirst that must be abnormal inasmuch as it is driven disproportionately, compared to 'normal thirst', by detection of changes in the osmotic pressure. To us it just *feels like* thirst. We are not privy to the underlying differences in what is happening in our bodies, in the same way that we would not be privy to the differences that underlie two 'identical' scores on a psychometric instrument when we just read the number 29 on a clinical report.

If we depended on an analogy to 'classical' semantics, we would take all these signals of thirst as preludes to the 'activation' (whatever that means) of a final canonical 'representation' (whatever *that* means) THIRST, which would always lie just out of our grasp, accessible only to the experiencing part of our brains, Mr. Homunculus. This is a mistake. Just as in the James-Lange theory there is nothing more to emotions than *what they feel like*, there is nothing more to thirst than the subjective awareness that one or more of a number of specific physiological limits have been exceeded. A range of distinct brain/body states can correlate with that awareness. We can usefully label them all *thirst*, since they are functionally identical in demanding the same response, but it would be misleading to say they were all THIRST.

A slightly different way of saying this, more directly relevant to lexical semantics, is that the word *thirst* does not have a single referent. It maps on to a range of *different* brain/body states. That in doing so it *unifies* those distinct states is an important clue about the proper function of semantics.

## Explaining semantics without semantics

We can make a directly analogous argument about lexical semantic access, since we have identified several candidates for 'a quasi-independent domain onto which we can map semantics'. Thinking of the semantic experience of a word as a partially-activated network of quasi-independent cues makes it easy to imagine how we might be able to recognize similarities between word meanings, since in such networks similarity can be defined as a function of which cues are activated and to what degree. We can conceive of the representation of the semantics of any word as a vector whose length is the number of possible cues, which may contain many zeros or near-zero values signifying non-activated or little-activated cues, irrelevant to the meaning of the current word (as for example, motor activation might be for a very abstract word like 'libertarianism').

The co-occurrence models discussed above offer a concrete implemented example of how it possible to go from this kind of vector of activation values to a continuous measure of semantic similarity. This analogy is helpful for thinking about similarity in high dimensional spaces in a rigourous and structured way, but should not be taken to imply that similarity judgments are in fact represented as vectors. Co-occurrence models define similarity by the similarity of the vectors that represent each word. If we order the vectors (of length 300) of the word-2vec matrix (Mikolov, Chen, Corrado, & Dean, 2013; Mikolov, Sutskever, Chen, Corrado, & Dean, 2013) by their cosine similarity to the vector for the word *lettuce* we find that the four vectors that are closest are those for the words *spinach, tomatoes, cucumbers*, and *iceberg_lettuce*. The model 'knows' that these words are semantically related to the word *lettuce* only because it 'knows' or is able to compute that the vectors that represent the words have, on average, similar values across a long vector of values. Of course semantic similarity judgments in the brain would depend on other values (as discussed above) than simply lexical co-occurrence and represent values in ways that are likely to be dynamic and complex rather than as simple numbers. However, the general idea that a high-dimensional space naturally allows for graded similarity computations is independent of what features define those dimensions or how exactly similarity between those features is computed.

To the extent that experienced word meaning may draw upon differentially-activated cues from different domains in different contexts and at different times for different reasons, we should not necessarily expect to find one semantic representation per linguistic label but rather a class of varying mental states that map into the same linguistic label.[12] These states may be united not because they share some subset of features or present a common representation to that part of the organism that can actually experience meaning. They may be united rather simply because we have learned to map the same linguistic label onto them. In the terms we have been using, we might say they share one (presumably, heavily-weighted) feature, that one that we call 'a name'. In contrast to the commonly-held idea that semantics 'picks out' distinct individual entities in the world and/or each entity's unique, canonical psychological representation (CAT), semantics would then be a way of enabling us to use the single heavily-weighted feature 'name' to treat as *identical*, things that might otherwise be quite *distinct* (perhaps but not only: different mental representations of the same entity).

The idea that a name might be best conceived of as one heavily-weighted feature of a word's feature set gives us two (only apparently different) ways for different experiences of semantic activation to the experiences of 'the same semantic

**12.** "Understanding a sentence is much more akin to understanding a theme in music than one may think." (Wittgenstein, 1958, p. 145).

representation'. On the one hand, two feature sets might be the same because they are close on average over many their feature values, as discussed above. We might come to realize that the word *courgette* (which is commonly used in British English) means exactly the same thing as the word *zucchini* (used in American English), not because anyone has informed us about this identity, but because we have noticed that the (empirically-observed or linguistically-deduced) features of the former word are nearly identical to those of the latter. On the other hand, we might come to understand that two are the same, not because we have noticed their shared features (one might well hope to avoid ever encountering the features of the insipid zucchini) but rather because someone told us point-blank 'A zucchini *is* a courgette' (see discussion in Frege, 1948). Note that these 'two different ways to be similar' are in fact the same way, since in both cases the recognition of identify depends on computing average similarity over a vector of features. The only difference is that the similarity measure due to name identity is heavily skewed by one single heavily-weighted feature, which is the name itself.

The conception that representations can be unified just because of a shared label inverts the usual idea that there is a single specific semantic representation grounding a linguistic label. A single representation does not ground a word. A single word grounds many representations. We know this is true even if we stick with lexical semantics, ignoring the diverse features from multiple domains into which it can be mapped. Lakoff (1987) memorably titled a book to celebrate the fact that:

> the Australian aboriginal language Dyirbal [...] has a category, *balan*, that actually includes women, fire, and dangerous things. It also includes birds that are not dangerous, as well as exceptional animals, such as the platypus, bandicoot, echidna. This is not [...] a matter of classification by common properties. (p. 5)

It might be possible to argue that the meaning of *balan* is BALAN, some very complicated singular representation that ties together all these disparate things in a neat package. But it is also possible (surely more likely) that many distinct mental representations – of women, of fire, of bandicoots – are separately recognized as being *balan*, united only through having that common feature activated.

Plaster & Polinsky (2007) discuss this "well-known case of exotic noun classification" (p. 1) in detail. Their discussion is closely consistent the general idea being defended here, that semantics can be conceived of a set of differentially activated features that may be of disparate kinds. They propose an interesting (if somewhat speculative) historical case study of the ontogenesis of this single semantic category. Dixon (1972) had proposed a wholly semantic (dormitive principle) explanation of 'balan' (picked up by Lakoff, 1987) in which the assignment of the various entities to 'balan' follows a set of conceptual rules such as "If a noun has characteristic X (on the basis of which its class membership would be expected to be decided) but is, through belief or myth, associated with characteristic Y, then generally it will belong to the class corresponding to Y, not to X" (Plaster & Polinsky, 2007, p. 4). Plaster & Polinsky disagree with this explanation, and in doing so, attempt to map the explanation for the composition of the semantic category 'balan' into non-semantic domains. They argue that the development of the category 'balan' reflects the aggregate influence of multiple simple linguistic and perceptual regularities, such "the appearance, function or use of a particular referent" (p. 6), the frequency with which certain words are used, and the historical evolution of gender marking in Dyirbal. The exotic noun category, they argue, does not exist because it was ever intended by anyone to make a coherent semantic claim (their paper is titled "Woman are not dangerous things") but rather "can be accounted for on the basis of a small number of [morphologically marked] semantic features and the morphophonemic similarities of items and their type/token frequency, without recourse to complicated semantics." (p. 1). Following the discussion above of why the sky is blue, this explanation of 'balan' in terms of non-balanish things is a more satisfying explanation than the dormitive principle explanation championed by Dixon and Lakoff.

I began by arguing that explanations need to map from the explanadum to explanans statements that are grounded in different domains, and have pointed to some examples of how this might work for explanations of semantics. Now I have come full circle and return to the problem of the dormitive principle. When a phenomenon is composed of enough weighted characters that it may not even resemble *itself* from one observation to the next, we need a tool whose purpose is to render the many different manifestations of that phenomenon identically in order to unite them into a single category, much as a score of 29 on the BDI-II makes all who achieved that score 'the same' even if they have nothing in common with respect to their responses or as a need to drink makes all states of thirst the same. It is very useful to be able to mark radically different things as identical, because the world and our minds are both wild and complex, churning and changing, leaving us with hardly a thing to hold on to that doesn't manifest itself under multiple guises, and because very different things can be most adaptively dealt with using identical behavioral or conceptual responses.

But *a word* is, of course, exactly the right feature for uniting multiple distinct manifestations of a phenomenon into a single category. As noted above, a word is a high-weight feature that we can attach wherever it is convenient, and for whatever purpose. A word is what Cronbach and Meehl (1955) argued that all psychometric constructs are: "a principle for making inferences" (p. 297), a tool for forcing an inference of identity. This is the sense in which semantics really *is* a dormitive principle by its very nature: because the only meaning of *a word* (or, indeed any other sign) is *whatsoever is united under it*, however disparate those various entities may be and why ever they may be united.[13] The meaning of the word *cat* is not CAT. CAT is phlogiston. CAT is the tiny homunculus who squats inside your head so he can tell you what you are seeing. CAT is a thing that has been fitted perfectly into its own shape.

The meaning of *cat* is *cat*.

## Acknowledgements

Thanks to Geoff Hollis, Jeff Keith, and three anonymous reviewers for feedback on earlier versions of this manuscript.

## Funding

This work is part of a program of research funded by the Natural Sciences and Engineering Research Council of Canada.

## References

Adelman, J. S., and Estes, Z. (2013). Emotion and memory: A recognition advantage for positive and negative words independent of arousal. *Cognition*, 129(3), 530–5. https://doi.org/10.1016/j.cognition.2013.08.014

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders: DSM-5*. Washington, D.C: American Psychiatric Association. https://doi.org/10.1176/appi.books.9780890425596

Andrews, M., Vigliocco, G., & Vinson, D. (2009). Integrating experiential and distributional data to learn semantic representations. *Psychological Review*, 116, 463–98. https://doi.org/10.1037/a0016261

Aziz-Zadeh, L., Fiebach, C. J., Naranayan, S., Feldman, J., Dodge, E., & Ivry, R. B. (2008). Modulation of the FFA and PPA by language related to faces and places. *Social neuroscience*, 3(3–4), 229–238.

Barsalou, L. W. (2008). Grounded Cognition. *Annual Review of Psychology*, 59(1), 617–645. https://doi.org/10.1146/annurev.psych.59.103006.093639

Baayen, R. H., Milin, P., Filipović Đurđević, D., Hendrix, P., & Marelli, M. (2011). An Amorphous Model for Morphological Processing in Visual Comprehension based on Naive Discriminative Learning. *Psychological Review*, 118, 438–481. https://doi.org/10.1037/a0023851

13. "Every sign *by itself* seems dead. *What* gives it life? – In use it is *alive*." (Wittgenstein, 1958, p. 128).

Baayen, H., Shaoul, C., Willits, J., & Ramscar, M. (2016). Comprehension without segmentation: A proof of concept with naive discrimination learning. *Language, Cognition, and Neuroscience*, 31(1), 106–128. https://doi.org/10.1080/23273798.2015.1065336

Bateson, G. (1972). *Steps to an Ecology of Mind: Collected Essays in Anthropology, Psychiatry, Evolution, and Epistemology*. Chicago, Il.: University of Chicago Press.

Bergen, B., & Chang, N. (2013). Embodied Construction Grammar. In T. Hoffmann & G. Trousdale (Eds.), *Oxford Handbook of Construction Grammar* (pp. 168–190) Oxford, UK: Oxford University Press.

Bloom, P. (2000). *How children learn the meanings of words*. Cambridge, MA:MIT Press. https://doi.org/10.7551/mitpress/3577.001.0001

Bourget, D., & Chalmers, D. J. (2014). What do philosophers believe?. *Philosophical Studies*, 170(3), 465–500.

Buchanan, L., Westbury, C., & Burgess, C. (2001). Characterizing Semantic Space: Neighbourhood Effects in Word Recognition. *Psychonomic Bulletin And Review*, 8(3), 531–544. https://doi.org/10.3758/BF03196189

Dagleish, T. (2004). The emotional brain. *Nature Reviews Neuroscience*, 5, 582–589.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, 52(4), 281. https://doi.org/10.1037/h0040957

Dennett, Daniel. (1991). *Consciousness Explained*. Boston, MA: Little, Brown and Company.

Desai, R. H., Conant, L. L., Binder, J. R., Park, H., & Seidenberg, M. S. (2013). A piece of the action: Modulation of sensory-motor regions by action idioms and metaphors. *Neuroimage*, 83, 862–869. https://doi.org/10.1016/j.neuroimage.2013.07.044

Durda, K., & Buchanan, L. (2008). WINDSORS: Windsor improved norms of distance and similarity of representations of semantics. *Behavior Research Methods*, 40(3), 705–712. https://doi.org/10.3758/BRM.40.3.705

Elman, J. L. (2009). On the meaning of words and dinosaur bones: Lexical knowledge without a lexicon. *Cognitive science*, 33(4), 547–582. https://doi.org/10.1111/j.1551-6709.2009.01023.x

Frege, G. (1948). On sense and Reference. *The Philosophical Review*, 57:3, 209–230. https://doi.org/10.2307/2181485

Freud, S. (1891). *On Aphasia*. [E. Stengel, Trans.] New Southgate, NH: The Cheswick Press.

Fincher-Kiefer, R. (2001). Perceptual components of situation models. *Memory and Cognition* 29, 336–343. https://doi.org/10.3758/BF03194928

Gibson, J. J. (1977). *The Theory of Affordances. In: Perceiving, Acting, and Knowing: Toward an Ecological Psychology* (p. 127–143). [R. Shaw and J. Bransford, Eds.] Hoboken, NJ: John Wiley & Sons Inc.

Glenberg, A. M. (1997). What memory is for. *Behavioral and Brain Sciences*, 20, 1–55. https://doi.org/10.1017/S0140525X97000010

Glenberg, A. M. & Kaschak, M. P. (2002). Grounding language in action. *Psychonomic Bulletin & Review*, 9, 558–565. https://doi.org/10.3758/BF03196313

Greenberg, V. D. (1997). *Freud and his aphasia book: Language and the sources of psychoanalysis*. Ithaca, NY: Cornell University Press.

Harris, Z. (1970). *Papers on Syntax*. [H. Hiz, Ed.] Boston, MA: D. Reidel Publishing Company.

Hauk, O., Johnsrude, I., & Pulvermüller, F. (2004). Somatotopic representation of action words in human motor and premotor cortex. *Neuron*, 41(2), 301–307.

Hebb, D. (1949). *The Organization of Behavior*. New York, NY: Wiley.

Hempel, C. G. & Oppenheim, P. (1948). Studies in the Logic of Explanation. *Philosophy of Science* 15:2, 135–175. https://doi.org/10.1086/286983

Hollis, G., & Westbury, C. (2016). The principals of meaning: Extracting semantic dimensions from co-occurrence models of semantics. *Psychonomic Bulletin & Review*, 23(6), 1744–1756. https://doi.org/10.3758/s13423-016-1053-2

Hollis, G., Westbury, C., & Lefsrud, L. (2017). Extrapolating Human Judgments from Skip-gram Vector Representations of Word Meaning. *The Quarterly Journal of Experimental Psychology*, 70(8), 1603–1619. https://doi.org/10.1080/17470218.2016.1195417

James, W. (1890/1950). *The Principle of Psychology*, Volumes One and Two. New York, NY: Dover Publications Inc.

Johnson, M. (1987). *The Body in the Mind: The Bodily Basis of Meaning, Imagination, and Reason*. Chicago, IL: The University of Chicago Press. https://doi.org/10.7208/chicago/9780226177847.001.0001

Johnson, M. (1991). Knowing through the body. *Philosophical Psychology*, 4, 3–18. https://doi.org/10.1080/09515089108573009

Johns, B. T. & Jones, M. N. (2015). Generating structure from experience: A retrieval-based model of language processing. *Canadian Journal of Experimental Psychology*, 69(3), 233–251. https://doi.org/10.1037/cep0000053

Jones, M. N., & Mewhort, D. J. K. (2007). Representing word meaning and order information in a composite holographic lexicon. *Psychological Review*, 114, 1–37. https://doi.org/10.1037/0033-295X.114.1.1

Kousta, S.-T., Vinson, D. P., & Vigliocco, G. (2009). Emotion words, regardless of polarity, have a processing advantage over neutral words. *Cognition*, 112, 473–481. https://doi.org/10.1016/j.cognition.2009.06.007

Kugler, P. N. & Turvey, M. T. (1987). *Information, natural law, and the self-assembly of rhythmic movement*. Hillsdale, NJ: Erlbaum.

Kussmaul, A. (1877). *Die störungen der sprache: Versuch einer pathologie der sprache*. Leipsig, Germany: Vogel.

Lakoff, G. (1987). *Women, fire, and dangerous things: What categories reveal about the mind*. Chicago, Il: The University of Chicago Press. https://doi.org/10.7208/chicago/9780226471013.001.0001

Lakoff, G. & Johnson, M. (1980). *Metaphors we live by*. Chicago, Il: The University of Chicago Press.

Landauer, T. K. & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 104(2), 211–240. https://doi.org/10.1037/0033-295X.104.2.211

Lifchitz, A., Jhean-Larose, S., & Denhière, G. (2009). Effect of tuned parameters on an LSA multiple choice questions answering model. *Behavior Research Methods*, 41, 1201–1209. https://doi.org/10.3758/BRM.41.4.1201

Lund, K., & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, & Computers*, 28, 203–208. https://doi.org/10.3758/BF03204766

Maturana, H. & Varela, F. (1980). *Autopoiesis and Cognition: The Realization of The Living*. Drodrecht, Holland: D. Reidel Publishing. https://doi.org/10.1007/978-94-009-8947-4

McKinley, M. J. & Johnson, A. K. (2004). The physiological regulation of thirst and fluid intake. *News in Physiological Sciences*, 19, 1–6. https://doi.org/10.1152/nips.01470.2003

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). *Efficient estimation of word representations in vector space*. Retrieved from arXiv: 1301.3781.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In: *Advances in neural information processing systems* 26 (pp. 3111–3119). Cambridge, MA: MIT Press.

Mingers, J. (1995). *Self-producing Systems: Implications and Applications of Autopoiesis*, New York, USA: Plenum Press.

Moliere. (1673/2003). The Imaginary Invalid [La Malade Imaginaire]. Downloaded October 1, 2015 from: http://www.gutenberg.org/9/0/7/9070

Morris, C. W. (1946). *Signs, language, and behavior*. New York, NY: Prentice Hall. https://doi.org/10.1037/14607-000

Osgood, C. E. (1952). The nature and measurement of meaning. *Psychological Bulletin*, 49, 197–237. https://doi.org/10.1037/h0055737

Osgood, C. E., Suci, G. J., & Tannenbaum, P. H. (1957). *The Measurement of Meaning*. Urbana, Il: University of Illinois Press.

Plaster, K., & Polinsky, M. (2007). Women are not dangerous things: Gender and categorization. *Harvard Working Papers in Linguistics*, 12. Downloaded September 10, 2016 from: http://scholar.harvard.edu/files/mpolinsky/files/Dyirbal.pdf

Pulvermüller, F. (1999). Words in the brain's language. *Behavioral and Brain Sciences*, 22, 253–336. https://doi.org/10.1017/S0140525X9900182X

Pulvermüller, F. (2005). Brain mechanisms linking language and action. *Nature Reviews Neuroscience*, 6, 576–582. https://doi.org/10.1038/nrn1706

Reilly, J., Westbury, C., Kean, J., & Peele, J. (2012). Arbitrary Symbolism in Natural Language Revisited: When Word Forms Carry Meaning *PLoS ONE* 7(8), e42286. https://doi.org/10.1371/journal.pone.0042286

Reilly, J., Hung, J., & Westbury, C. (2017). Non-Arbitrariness in Mapping Word Form to Meaning: Cross-Linguistic Formal Markers of Word Concreteness. *Cognitive science*, 41(4), 1071–1089. https://doi.org/10.1111/cogs.12361

Rescorla, R. A., & Wagner, A. R. (1972). *A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement*. In A. H. Black & W. F. Prokasy (Eds.), *Classical conditioning II: Current theory and research* (pp. 64–99). New York, NY: Appleton-Century-Crofts.

Rescorla, R. A. (1988). Pavlovian conditioning: It's not what you think. *American Psychologist*, 43(3), 151–160. https://doi.org/10.1037/0003-066X.43.3.151

Roy, D. K. & Pentland, A. P. (2002). Learning words from sights and sounds: A computational model. *Cognitive Science*, 26, 113–146. https://doi.org/10.1207/s15516709cog2601_4

Rumelhart, D. E. (1979). Some problems with the notion of literal meanings. In: A. Ortony (Ed.) *Metaphor and Thought*. Cambridge, MA: Cambridge University Press, p. 71–82.

Shaoul, C., & Westbury, C. (2008). *HiDEx: The high dimensional explorer*. Edmonton, AB. Retrieved from http://www.psych.ualberta.ca/~westburylab/downloads.html

Shaoul, C., & Westbury, C. (2010). Exploring lexical co-occurrence space using HiDEx. *Behavior Research Methods*, 42(2), 393–413. https://doi.org/10.3758/BRM.42.2.393

Shaoul, C., & Westbury, C. (2011). HiDEx: The High Dimensional Explorer. In: *Applied Natural Language Processing and Content Analysis: Identification, Investigation, and Resolution*. Editors P. McCarthy and C. Boonthum. IGI Global, 230–246.

Smith, L. B., Suanda, S. H., & Yu, C. (2014). The unrealized promise of infant statistical word-referent learning. *Trends in Cognitive Science*, 18(5), 251–258. https://doi.org/10.1016/j.tics.2014.02.007

Smith, L. & You, C. (2008). Infants rapidly learn word-referent mappings via cross-situtaional statistics. *Cognition*, 106, 1558–1568. https://doi.org/10.1016/j.cognition.2007.06.010

Speer, N. K., Reynolds, J. R., Swallow, K. M., & Zacks, J. M. (2009). Reading stories activates neural representations of visual and motor experiences. *Psychological Science*, 20(8), 989–999.

Thelen, E. & Smith, L. B. (1994). *A Dynamic Systems Approach to the Development of Cognition and Action*. Cambridge, MA: MIT Press.

Varela, F., Thompson, E., and Rosch, E. (1993). *The Embodied Mind: Cognitive Science And Human Experience*. Boston, MA: MIT Press.

Vigliocco, G., Meteyard, L., Andrews, M., & Kousta, S. (2009). Toward a theory of semantic representation. *Language and Cognition*, 1(2), 219–247. https://doi.org/10.1515/LANGCOG.2009.011

Westbury, C. (2013). You can't drink a word: Lexical and individual emotionality affect subjective familiarity judgments. *Journal of Psycholinguistic Research*, 43(5), 631–49. https://doi.org/10.1007/s10936-013-9266-2

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. *Quarterly Journal of Experimental Psychology*, 68(8), 1599–1622. https://doi.org/10.1080/17470218.2014.970204

Westbury, C. F., Shaoul, C., Hollis, G., Smithson, L., Briesemeister, B. B., Hofmann, M. J., & Jacobs, A. M. (2013). Now you see it, now you don't: On emotion, context, & the algorithmic prediction of human imageability judgments. *Frontiers in Psychology*, 4:991. https://doi.org/10.3389/fpsyg.2013.00991

Wittgenstein, L. (1958). *Philosophical Investigations*. New York, NY: McMillan Publishing Co.

Yu, C. & Ballard, D. H. (2007). A unified model of early word learning: Integrating statistical and social cues. *Neurocomputing*, 70, 2149–2165. https://doi.org/10.1016/j.neucom.2006.01.034

Zwaan, R. A. (2004). The immersed experiencer: Toward an embodied theory of language comprehension. *Psychology of Learning and Motivation*, 44, 35–62.

CHAPTER 4

# Why we need to investigate casual speech to truly understand language production, processing and the mental lexicon

Benjamin V. Tucker and Mirjam Ernestus
University of Alberta / Radboud University

## Thesis

The majority of studies addressing psycholinguistic questions focus on speech produced and processed in a careful, laboratory speech style. This 'careful' speech is very different from the speech that listeners encounter in normal, casual conversations. Research on casual speech is necessary to show the validity of psycholinguistic conclusions based on careful speech. Moreover, research on casual speech has and will produce new insights and questions on the processes underlying communication and on the mental lexicon that cannot be revealed by research using careful speech. In order to understand research on casual speech a historic perspective is important. There are many examples of how casual speech differs from careful speech and we show that these differences may have important implications for psycholinguistic theories. There are many challenges that research on casual speech faces, which stem from the high variability of this speech style, its necessary casual context, and that casual speech is connected speech. Embracing these challenges and pursuing new opportunities, mostly in the form of new experimental methods that facilitate research on connected speech combined with advanced (still to be developed) statistical techniques, will allow real progress in psycholinguistic understanding.

## Commentaries on Tucker and Ernestus thesis

### Ray Jackendoff commentary on Tucker and Ernestus thesis

Tucker and Ernestus presciently call for more psycholinguistic research on natural connected speech. Their article cites numerous corpora of recorded casual speech that can serve as data for such research. In fact, collecting copious data is not that difficult. The real bottleneck is in coding the data: how to identify the phonemes and

https://doi.org/10.1075/z.238.04tuc

the morphemes; how to gloss the texts, how to translate them idiomatically into the language in which research is being conducted; what to measure, what to count; how to make the corpus properly searchable; how to train assistants to do all this tedious work. It is only with meticulous annotation by well-trained coders that a corpus can be useful for research purposes (see also my commentary on Myers above).

An important factor that Tucker and Ernestus stress is the variability inherent in the production of casual speech. This is however not a new challenge. The variationists, starting with William Labov (1966, 1969), have devoted decades of research to coping with inherent variation, including not just grammatical variables but sociolinguistic variables such as gender, age, ethnicity, and social setting (see for instance the journal *Language Variation and Change*). A mathematical treatment of variation was pioneered by Cedergren and Sankoff (1974). This was one of the earliest applications of multivariate statistics in linguistic research, providing a technique for assessing the respective contributions of multiple hypothesized factors to systematic variation across a community. However, what has not been addressed by the variationists is the psycholinguistic and/or neurolinguistic mechanisms that give rise to variation. Studying such mechanisms would be an interesting goal for future psycholinguistics and neuroscience. My hunch is that systematic variation will be found not just in language production, but in any kind of real-time motor activity and planning.

The issue of variation leads to another question posed by Tucker and Ernestus, a new perspective on the old question of storage vs. computation. When we find variability in casual speech, is it the result of storing alternate forms, or is it the result of alternative phonetic realizations of a single stored form? One would hope, for instance, that the variation between *walking* and *walkin'* is general across all verbs, and that one does not have to store two versions of every single verb. Perhaps, however, one does store two versions of the present participle affix, *-ing* and *-in'*, competing for application to verb stems. Or perhaps not: the variation may arise simply by virtue of on occasion simplifying articulation on the fly. For another case, it feels fairly safe to imagine that the variation between *what the hell* and *what the heck* (as well as more imaginative variants) is the result of storing both forms, in competition. But of course these are merely conjectures, awaiting psycholinguistic testing, as Tucker and Ernestus recommend.

## *References*

Cedergren, Henrietta, and David Sankoff. (1974). Variable rules: Performance as a statistical reflection of competence. *Language* 50, 333–355. https://doi.org/10.2307/412441

Labov, William. (1966). *The Social Stratification of English in New York City*. Washington, D.C.: Center for Applied Linguistics, 1966. 2006. Second edition: Cambridge/Cambridge U. Press.

Labov, William. (1969). Contraction, deletion, and inherent variability of the English copula. *Language* 45, 715–762. https://doi.org/10.2307/412333

## James Myers commentary on Tucker and Ernestus thesis

The fact that we, as humans, understand casual speech, does not imply that we, as scientists, will ever *truly* understand casual speech. Our brains have apparently solved this engineering problem, but engineering solutions need not have readily comprehensible scientific interpretations. For example, automatic speech recognition technology has made enormous strides in recognizing spontaneous speech, in part using so-called "deep learning" techniques that essentially throw enormous amounts of training data at gigantic artificial neural networks, which presumably mimics, to some very rough extent, what actual human brains do. The problem is that such models are merely designed to work, not to explain, so looking inside one of them in search of the things that scientists need to find, like causes, effects, representations, or principles, is about as baffling as looking inside actual brains. The authors seem to hint at just such an unsatisfying outcome to their research program when they say that modeling casual speech will require statistical techniques not yet invented. What if the statistical techniques they need are mathematically impossible?

They may be making the problem harder than it needs to be, however. Studying casual speech in a laboratory is intrinsically paradoxical. Chomsky notoriously criticized corpus linguists as being like physicists who base their theories solely on what they see passing by their window. True, some physicists actually do theorize about what passes by their window, among them the applied physicists known as climate scientists. Climate models are incredibly complicated, at least as complicated as successful casual speech models would have to be, but they are not used so much to test hypotheses about the general principles underlying climate, which are already reasonably well understood, as to apply these principles to make concrete predictions about the specific climate that we have now. In other words, like automatic speech recognition tools, their goals are more practical than explanatory. By contrast, more fundamental physicists, as Chomsky claimed, have made progress only by working with oversimplified laboratory models.

In the case of casual speech, what sort of fundamental principles might be suitable for laboratory oversimplification? At least one is completely uncontroversial: speech is a physical process. So why not build on the many existing models of the physics of speech (weights, springs, harmonics, and such)? Speech rate may vary, durations may vary, articulatory gestures may lag and overlap, but all of these things are subject to physiological constraints, not just psychological ones. Models that build in these constraints may make casual speech more amenable to a scientific interpretation.

## Dorit Ravid commentary on Tucker and Ernestus thesis

The tension between top-down linguistic theory and categorization, on the one hand, and evidence about actual language production (including, but not restricted to, casual speech phenomena) is overwhelming. The pursuit of knowledge about the role of language in cognition and culture requires that linguists segment data into well-motivated structure-meaning/function units, identify patterns and compute whatever is necessary to answer our questions. This works quite well when linguists inquire about their own minds or consult edited written language. Grammaticality well-formedness judgments – the intuitive decisions used by trained linguists as a privileged database for the development of linguistic theory – are not sufficiently representative of the language community at large, not being statistically validated, and made by sophisticated people who are not at all typical language users, attuned to subtle syntactic distinctions (Kitagawa & Fodor, 2006). Written language represents a form of language that is rich, variegated, syntactically complex, highly lexical, well-organized, coherent, and overtly marked for linguistic distinctions (Chafe, 1994). It leads us astray into assuming that language knowledge is highly abstract and formal. But spoken language production, even when constrained in narrative form, is nothing like that in any sense (Ravid & Chen-Djemal, 2015). Casual speech phenomena are part of this tension between edited and 'real' language.

According to Christiansen & Chater (2016) "Language is fleeting. As we hear a sentence unfold, our memory for preceding material is rapidly lost. Speakers, too, soon lose track of the details of what they have just said. Language processing is therefore `Now-or-Never`: If linguistic information is not processed rapidly, that information is lost for good". This means that much of language is compressed into quanta that can be easily accessed by language users and learners (Christiansen & Chater, 2017). For a developmental (psycho)linguist, there is no other option but to rely on data of this nature, despite its fuzzy and often contradictory nature, in order to construct reliable theories of language learning. Here is an example of what happens if we actually listen to casual speech production. When examining the distributions of grammatical subjects, we looked at the 1st person singular pronoun *ani* followed by a future tense verb in agreement with it. The written, edited form of such verbs starts with the letter *aleph*, standing for the initial glottal stop representing 1st person singular future tense inflection, e.g., *adaber* 'I will talk', *ekanes* 'I will enter', or *etlabesh* 'I will dress'. But in children's speech, and in casual adult speech, the 1st person singular actually start with *y*-, exactly like 3rd person future tense verbs. This is mostly due to the obligatory co-occurrence of the first person singular pronoun *ani* and the future tense verb in colloquial speech (Dattner, Kertes & Ravid, 2019). If we apply the categorical morpho-phonological criterion to trace the usage of grammatical subjects in acquisition, the actual distributions of person distinctions will be unreliably skewed.

### References

Chafe, W. (1994). *Discourse, consciousness, and time: The flow and displacement of conscious experience in speaking and writing*. Chicago: The University of Chicago Press.

Christiansen, M. H. & Chater, N. (2016). The now-or-never bottleneck: A fundamental constraint on language. *Behavioral and Brain Sciences*, 39, e62 [target article]. https://doi.org/10.1017/S0140525X1500031X

Christiansen, M. H. & Chater, N. (2017). Towards an integrated science of language. *Nature*, July, 1, 0163.

Dattner, E., Kertes, L., Zwilling, R., & Ravid, D. (2019). Usage patterns in the development of Hebrew grammatical subjects. *Glossa: A Journal of General Linguistics*, 4(1), 129, 1–28. https://doi.org/10.5334/gjgl.928

Kitagawa, Y., & Fodor, J. (2006). Prosodic influence on syntactic judgments. In G. Fanselow, C. Fery, R. Vogel, & M. Schlesewsky (Eds.), *Gradience in grammar: Generative perspectives* (pp. 336–358). Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199274796.003.0017

Ravid, D. & Chen-Djemal, Y. (2015). Spoken and written narration in Hebrew: A case study. *Written Language and Literacy*, 18, 56–81. https://doi.org/10.1075/wll.18.1.03rav

## Russel Richie commentary on Tucker and Ernestus thesis

How one prioritizes ecological validity relative to experimental control is likely closely related to whether one thinks cognition is component-dominant or interaction-dominant (Stephen & Mirman, 2010; Szary et al., 2015). Roughly, this debate concerns whether one thinks cognition is the summation of independent contributions from separate components, or the result of multiplicative interactions between components.

A believer of component-dominance might react to this article with the following: At present, there doesn't seem to be much consensus on how anything works in psycholinguistics, even in carefully controlled environments. We have many empirical phenomena – e.g., garden path effects and local coherence effects in sentence processing, rhyme effects and ganong effects in spoken word processing – that are elicited in controlled contexts, but there continues to be a plethora of models to explain those effects. There are a variety of parsers, from SOPARSE (Tabor & Hutchins, 2004) to Hale's (2011) rational parser, to account for sentence processing effects. There are a variety of word recognition models, from TRACE (McClelland & Elman, 1986) to TISK (Hannagan, Magnuson, & Grainger, 2013). We – the believer of component-dominance would argue – ought to first understand how speech works in carefully controlled environments before we make our jobs as scientists more complicated with more ecologically valid, and thus complex, settings. *We have to focus on one thing at a time.*

A believer of interaction-dominance, on the other hand, might respond as follows: Cognition isn't (mostly) the sum of its components or modules. Rather, cognition is (primarily) the result of *interactions* between parts. The effect of a component (say, a particular set of vocal articulators) on a cognitive system (say, speech production) depends enormously on the contributions of other components (another set of vocal articulators). In regression terms, the interaction of x0*x1 on y dominates the main, additive effects of x0 or x1 on y. Thus, an experiment holding x0 or x1 constant while varying the other variable will miss the most important factor explaining variance in y. For this reason, we must study wide ranges of variation in both x0 and x1 simultaneously – *we can't simply isolate a component of interest by controlling for all others, as we do in typical experiments. We must simultaneously study wide ranges of variation in combinations of contributing components, in the ranges in which they vary in the 'real world'*.

In any case, psycholinguistic theories *are* already constrained by considerations of ecological validity, to some extent. For example, a well-known challenge of TRACE is scaling up to human-sized lexicons, an issue that motivated TISK, which covers much of the same empirical phenomena with far fewer mental resources (Hannagan et al., 2013). Similarly, constraint-based theories of sentence processing hold that comprehenders utilize whatever information is at-hand in the rich contexts of real language use (Tanenhaus et al., 1995).

### *References*

Hale, J. T. (2011). What a rational parser would do. *Cognitive Science*, 35, 399–443. https://doi.org/10.1111/j.1551-6709.2010.01145.x

Hannagan, T., Magnuson, J. S., & Grainger, J. (2013). Spoken word recognition without a TRACE, *Frontiers in Psychology*, 4, 563. https://doi.org/10.3389/fpsyg.2013.00563

McClelland, J. L., & Elman, J. L. (1986). The trace model of speech perception. *Cognitive Psychology*, 18, 1–86. https://doi.org/10.1016/0010-0285(86)90015-0

Szary, J., Dale, R., Kello, C. T., & Rhodes, T. (2015). Patterns of interaction-dominant dynamics in individual versus collaborative memory foraging. *Cognitive Processing*, 16(4), 389–399. https://doi.org/10.1007/s10339-015-0731-8

Stephen, D. G., & Mirman, D. (2010). Interactions Dominate the Dynamics of Visual Cognition, 115(1), 154–165. https://doi.org/10.1016/j.cognition.2009.12.010

Tabor, W., & Hutchins, S. (2004). Evidence for self-organized sentence processing: digging-in effects. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, 30(2), 431–50. https://doi.org/10.1037/0278-7393.30.2.431

Tanenhaus, M., Spivey-Knowlton, M., Eberhard, K., & Sedivy, J. (1995). Integration of Visual and Linguistic Information in Spoken Language Comprehension. *Science*, 268(5217), 1632–1634. https://doi.org/10.1126/science.7777863

## Chris Westbury commentary on Tucker and Ernestus thesis

In a recent paper, Hollis and Westbury (2016) used principal component analysis (PCA) on a matrix derived from the co-occurrence-inspired word2vec model of semantics (Mikolov, Chen, Corrado, and Dean, 2013). In co-occurrence models, words are represented as vectors that characterize a word's usage history in a large corpus of text. Words with similar vectors have similar patterns of usage, and therefore, usually similar semantics. In the model used by Hollis & Westbury, the vectors were (arbitrarily, but following convention) of length 300.

In applying PCA, it is possible to extract as many components as there are values in each vector in the original matrix, i.e. up to 300 in this case. However, one of the purposes of PCA is to project the original matrix down to a smaller number of dimensions. It is obvious that we could individuate words with many fewer dimensions than 300, for a simple mathematical reason. If each dimension was able to cut down half the possibilities (the optimal case, as shown by Shannon, 1948), we could individuate all the words in a lexicon of, say, a million words (many times more than the average person uses) with just 20 dimensions, since $2^{20} = 1{,}048{,}576$. Such optimality might be too much to ask for in a complex and messy natural system like the history of language use. It may be impossible to mathematically extract ideal dimensions that each eliminate exactly half of the uncertainty in word identification. Moreover, the system is not trying to individuate *words* but *meanings*, and there are many more meanings than words. However, one might think it reasonable to expect only a few more dimensions, since the number of elements individuated grows exponentially with the number of available dimensions. With 25 dimensions, an optimal system would be able to individuate $2^{25}$ (33,554,432) meanings in a lexicon, and with as many as 30 dimensions, it would be able to individuate $2^{30} = 1{,}073{,}741{,}824$ meanings. It might seem almost beyond imagination that the semantic system would require as many as 100 dimensions, because $2^{100}$ seems like massive overkill for the purpose, optimally allowing for the individuation of 1.26e30 items, millions of millions more items than the number of words or meanings in the dictionary of any language.

Considerations such as these would lead one to expect that the PCA of a co-occurrence matrix might be able to easily load all of the variance in the first few PCs, certainly less than 100. However, this is not what we find. Many words and categorical semantic distinctions show large variance in very high rank PCs, even to the 300th PC. Why is this?

The answer is that the information derived from linguistic experience that is contained in the matrix is not merely about individuating word meanings. It also defines a complex of networks of relationships between word meanings. If you know only the intensional (dictionary) definition of the word *mammal* (e.g. 'a warm-blooded animal with hair that suckles its young'), you know almost nothing about mammals. For example, you might not be able to definitely say whether a cat (a hairless cat, say) or a human baby was a mammal or not. You also need an *extensional definition*, knowledge about the set of objects that a term describes. To understand it, the word *mammal* needs to be associated (as it is in the word2vec matrix) with a set of mammals (e.g. the closest neighbours of the word *mammal* include many names such as *elephant, sealion, otter*). Although this should be sufficient in theory for understanding the word, in practice there are many more relationships encoded in normal word use. For example, the close neighbors of the word *mammal* also include many examples of non-mammals (e.g. *reptile, birds*, and *invertebrate*) and many examples of supersets or subsets of mammals (e.g. *animals, primates, critter*, and *creatures*). It is not enough to understand merely these first-order relationships: to understand the word *mammal* in the colloquial sense of the word *understand*, you also need to know about the second-order relationships between the word's neighbours. Each neighbor is situated not just in a single network of the word *mammal*, but in a potentially very large set of inter-connected networks. It is strictly untrue to say that you don't what a mammal is if you don't know that a porpoise is more similar to a whale than it is to a panther or that a brachiosaurus and triceratops are more similar to each other than either is to mammal. However, you would surely be considered to have a very weak understanding of meaning of the word *mammal* if you did not also understand these relationships.

Considerations such as these suggest that the meaning of a word is not contained in the word itself, but rather in the wider networks of word relationships (Deacon, 1998), which contain much more information than merely what is needed to individuate intensional word meanings. However, there is still much more information about word meaning in the real world than what is contained in co-occurrence networks. In real speech, word meanings as normally understood can be modulated or even negated by many extra-lexical factors, including changes in pitch, volume, speech rate, and enunciation; uses of appropriate or inappropriate registers; focal word choices; management of turn-taking; adherence or violation of Gricean maxims; facial expressions; and many more. To fully understand linguistic semantics is necessarily to understand how these real-world parameters impinge upon language interpretations.

## *References*

Deacon, T. W. (1998). The Symbolic Species: The Co-evolution of Language and the Brain. WW Norton & Company.

Hollis, G., & Westbury, C. (2016). The principals of meaning: Extracting semantic dimensions from co-occurrence models of semantics. *Psychonomic Bulletin & Review*, 23(6), 1744–1756. https://doi.org/10.3758/s13423-016-1053-2

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Shannon, C. E. (1948). A mathematical theory of communication, Part I, Part II. *Bell Systems Technical Journal*, 27, 623–656. https://doi.org/10.1002/j.1538-7305.1948.tb00917.x

## The article

One of the central goals of psycholinguistic research is understanding the way in which language is recognized, produced and represented in the mental lexicon. As researchers have worked to achieve this goal, a substantial collection of hypotheses, theories, and computational models of lexical representation, comprehension and production have been generated and tested in (psycholinguistic) experiments and corpus analyses. Importantly, nearly all work focusing on speech has investigated read aloud speech or speech elicited under formal conditions (often in the laboratory, in front of a microphone, and directed towards strangers). Words are carefully articulated and occur in isolation, in simple sentences, or in sentences that reflect those sentences typically found in written language (books, newspapers, etc.). In this article, we use the umbrella term "careful speech" for all speech that is not spontaneously uttered or that has not been carefully recorded in order to reflect everyday speech. Warner (2011) surveyed the *Journal of Phonetics* (v.36, 2008) to verify the amount of research being performed on non-careful speech. In that survey of 36 articles, only 4 articles used non-careful speech; two of which investigated infant babbling and isolated word productions, and two studied North American indigenous languages, and used several speech genres. Wagner, Trouvain, & Zimmerer (2015) reported that in the 2007 and 2011 conference proceedings of the International Congress of Phonetic Sciences, only 9% and 19% of the papers respectively reported data which they classified as "unscripted" speech.

In this article, we argue that in order to truly understand the cognitive processes underlying communication, psycholinguistic research requires a concerted shift to more casual, spontaneous types of speech, which indicates, at least, whether our models, theories and hypotheses are also valid for the type of speech speakers and listeners most frequently use. Moreover, as we show in this paper, studies of spontaneous speech also raise new questions which have to be answered in order to establish a clear understanding of speech production, speech comprehension, and the mental lexicon.

Research on the processing of conversational speech requires experimental paradigms in which target words can be embedded in sentences, in all sentence positions. The focus on careful speech, however, has resulted in experimental paradigms suited for the study of words presented without context or with very little context. For instance, many studies have used the auditory lexical decision paradigm which is not well suited for investigating the processing of words in linguistic context. Other popular experimental paradigms such as phoneme monitoring, shadowing or word spotting use sentences rather than isolated words. However, many questions have been raised about these paradigms because they can induce strong task effects and participants rely heavily on the orthography of the words they hear to perform the task (e.g., Bates & Hiu, 1996; Connine & Titone, 1996; McQueen, 1996). Today, only a few experimental paradigms allow for the investigation of production and recognition of words in sentences: the map task, cross-modal priming, eye tracking, pupilometry, and recording brain activity. However, all these methods have serious shortcomings. Below, we discuss these shortcomings as well as introducing new opportunities for the research on casual speech.

So far, we have used the terms "spontaneous", "conversational", "casual speech", "non-careful speech" and "unscripted speech, as opposed to "careful speech". We believe that there is a continuum spanning the very careful, hyper-articulated styles of speech to the fast flowing, often hypo-articulated conversations with a close friend (e.g., Ernestus, Hanique, & Verboom, 2015; Tucker, 2007; Warner 2011; Wagner et al., 2015) and that all these different speech styles should be addressed in psycholinguistic research (see also Warner, 2012). For convenience, we focus on the extremes of the continuum: careful and casual speech. That is, we compare unscripted conversational speech uttered in informal conversations with careful speech, which we have defined as an umbrella term referring to carefully articulated speech often produced in formal situations.

Careful speech comprises read aloud speech, which has, among other things, the same semantic and syntactic properties of written text. Because our focus is not on the difference between written language versus spoken language, but on (unscripted) speech uttered in formal settings versus more spontaneously produced speech, we only consider read aloud speech if it provides information about careful speech in general (for instance with respect to pronunciation).

We first explain why casual speech has received so little attention in the past, and why the time is now ripe to start studying this form of speech. We then demonstrate the importance of investigating casual speech for psycholinguistics on the basis of different types of speech data and their implications. This section is not a thorough review of the literature, but a selection of the literature illustrating our points. We finally discuss some of the challenges facing research on casual speech, ways in which these challenges can be met, as well as some of the opportunities arising from these challenges. We conclude this paper with a short section on future directions for psycholinguistics and casual speech.

## Why casual speech has attracted so little attention within psycholinguistics

Research on first language acquisition has, for many years, shown that studying spontaneously uttered speech can provide valuable insight into language representation, comprehension and production. This research field was one of the first to build a large corpus of spontaneously uttered speech (MacWhinney, 2000). Similarly, the field of sociolinguistics has long studied spontaneously uttered dialogues (e.g., Labov, 1972) with interesting results. Nevertheless, research on adult speech processing has neglected casual speech. We identify several (related) explanations for the general research focus on careful speech.

First, research on how language users produce and process carefully articulated words in isolation and in simple sentences provides insights into the basic mechanisms underlying speech production and comprehension that are likely to play a part in everyday language communication. That is, in order to learn about speech production and processing in conversational settings, research on casual speech is not always strictly necessary. The advantage of investigating careful speech over casual speech is that with highly controlled speech stimuli, we can conduct better controlled experiments which allow for firmer conclusions (Xu, 2010). It is thus tempting to only investigate the processing of careful speech as it already provides so much information about the linguistic system.

A second possible explanation for the focus on careful speech may simply be the common belief that language outside the laboratory is very similar to careful speech: Everyday speech may have some special characteristics, but the differences between careful speech and casual speech are not substantial. More than half a century ago, studies were published showing that the opposite may be true, demonstrating differences in speech rate (Pollack & Pickett, 1963) and the extent of vowel reduction (e.g. Lindblom, 1963), and documenting extreme assimilation (Hocket, 1955; Stampe, 1973) and syllable deletion (Richter, 1930; from Warner, 2011) for casual speech that is not present in careful speech. In addition, in 1988, Mehta and Cutler indicated that the processing of casual speech may partly rely on mechanisms different from those underlying the processing of careful speech. They further argued for the need to study more casual speech styles in psycholinguistic research.

A third possible explanation for why modern psycholinguistic research has focused on careful speech is that historically this field has received inspiration from formal linguistic theories, many of which have traditionally focused on linguistic competence. How speakers produce words and sentences outside the laboratory and how listeners process words and sentences that do not represent careful speech were deemed uninteresting. This is because language users' performances in these situations may be heavily influenced by factors that are not part of their linguistic competence (Chomsky, 1965), for instance their working memory or their hearing acuity. A divide and debate around competence versus performance and their role in linguistics has been ongoing since the 1960s (e.g., Hymes, 1992). It is only very recently, especially with the rise in interest in individual differences and in language user groups other than highly educated young adults, that psycholinguists have formulated research programs that aim at developing complete theories of speech production and comprehension. These combine the processes and representations representing linguistic competence with factors that play a role in all types of behavior, such as working memory (e.g. Koch & Janse, 2016).

A final explanation for the neglect of casual speech may be that this speech style is hard to investigate. Traditional experimental paradigms are designed to test the production or processing of words in isolation, while the production or processing of casual speech requires researchers to test connected speech. Moreover, the laboratory environment in which most experiments are conducted does not favor the use of casual speech.

We believe that the time is now ripe to add casual speech as a research focus in speech production, speech comprehension, and the mental lexicon. There is increasing evidence indicating that casual speech is inherently different to careful speech, as we demonstrate in the next section. Moreover, we can make more progress if we extend our research focus to other registers than to careful speech alone. Finally, we also note that to facilitate research on casual speech, new experimental and statistical techniques are required.

## Relevance of differences between casual and careful speech for psycholinguistics

As briefly noted above, several studies have shown clear differences between careful and casual speech (e.g. Hocket, 1955; Pollack & Pickett, 1963; Lindblom, 1963; Richter, 1930; Warner, 2011), and that careful and casual speech may be processed differently (Mehta & Culter, 1988). The goal of this section is to demonstrate the value of studying casual speech to improve our understanding of language representation, production, and comprehension. We provide examples showing that casual and careful speech styles differ at many linguistic levels. Speech that is uttered in everyday situations may be very different in many respects from what the dictionary or grammar dictates should be said, from that what is often taught in an introductory linguistics course (which is typically based on careful speech), and from speech produced in an experimental or careful situation.

### *Differences in word choice*

One way in which casual speech is different from careful speech is the word types that are used. There are many content words that are generally assumed not to be sufficiently 'proper' to be used in formal situations. For example, it is unlikely that the American English *dude*, Dutch *vent* 'guy', Romanian *măi* 'dude' or French *pote* 'friend' occur in formal settings. Similarly, Torreira, Adda-Decker and Ernestus (2010) showed that swear words occur much more frequently in conversational French (e.g. *putain* 'whore' has a frequency of 0.79 occurrences per thousand words in the Nijmegen Corpus of Casual French) than in French broadcasted on the radio (as incorporated in the ESTER corpus, Galliano et al., 2005; where *putain* has a frequency of zero). This difference between careful and casual speech does not only hold for content words, but also for words that function more as discourse markers: in one conversation of Western Canadian English, the multifunctional *like* was used 181 times in a 25-minute conversation by one of two interlocutors (Podlubny, Geeraert, & Tucker, 2015).

More generally, nouns, articles, adjectives and prepositions are more frequent in careful than in casual speech situations (e.g., Heylighen and Dewaele, 2002; Biber, 1988; Biber et al. 1998). Careful speech, in contrast, is said to be characterized by relatively higher frequencies of interjections, verbs, adverbs and pronouns. Furthermore, previous research documented that in careful speech, word type/token ratios are higher, and longer words are used more often (e.g. Biber 1988; Biber et al. 1998). It therefore comes as no surprise that Bentum, Ernestus, ten Bosch and van den Bosch (submitted) show that the different subcorpora of the Spoken Dutch Corpus (Oostdijk 2000), with each subcorpus representing a different speech register, can be distinguished based on the likelihood of the words given the immediately preceding words.

These differences at the lexical level raise several (related) important questions about the mental lexicon, speech production and speech comprehension. A first general question is: what do these differences mean for lexical representation? Is the information about the frequencies of occurrence of the words lexically stored per speech register and, if so, is this frequency information stored in the form of a continuum, in line with our idea that speech registers form a continuum? Second, does the frequency of occurrence of a word in a given speech style just result from the lexical storage of frequency information specific to this speech register, or is there (in addition) active inhibition during speech production for specific speech styles (e.g. for a word like *dude* in careful speech)? A third general question relates to speech comprehension. We know that listeners make use of the predictabilities of words. For instance, lexical frequency is a good predictor for how easily listeners recognize words in isolation, while the conditional probability of a word given the preceding word is a better predictor if the word occurs in sentence context. Are listeners able to use the register-specific predictabilities of words, as they seem able to use modality specific probabilities of words (e.g., Gaygen & Luce, 1998)?

### *Differences at the syntactic level*

Another way in which casual speech differs from careful speech concerns syntactic structures. Unsurprisingly, casual speech is replete with false starts, filled pauses, hesitations, repetitions, and unfinished sentences. These are absent in careful speech and are typically avoided as much as possible in casual speech produced in formal situations (e.g., Torreira et al., 2010). Further, sentence structure tends to be more complex in careful speech than in casual speech (e.g., Biber 1988; Biber et al. 1998). Sentence length, for instance, helps to identify the different components of the Spoken Dutch Corpus (Wiggers & Rothkrantz, 2007). Other types of sentence structures may be more frequent in casual speech than in careful speech. To give an example, the pervasive use of dislocations, where one phrase, usually the subject, is set apart from the main clause, and within that clause, a pronoun is used (as in *Le chocolat, c'est bon* 'The chocolate, that is good'), is a key characteristic of spoken French (e.g., De Chat, 2007).

These differences at the syntactic level raise similar questions to the differences noted above at the lexical level. For instance, what are the implications for the mental lexicon, and what exactly are the mechanisms underlying these differences? Furthermore, listeners have been shown to expect certain word types (e.g. a past participle) given the preceding sentence structure (Viebahn, Ernestus & McQueen, 2015). Do listeners adapt these expectations as a function of speech style?

In addition, these differences have implications for research on casual speech. They show that if we would like to contrast the processing of careful and casual speech, we cannot simply present participants with exactly the same sentences pronounced either in a careful or in a casual way. Participants may show different processing when listening to syntactically careful sentences pronounced in a casual way, as opposed to listening to syntactically casual speech.

### *Differences in tones and intonation*

Casual speech may also differ from careful speech in how lexical tones and intonation are realized. In recent work, Brenner (2013) showed that the acoustic cues for individual tones in Mandarin casual speech are greatly reduced, and many of the standard cues for tone (like pitch) are often unreliable. Brenner (2015) investigated the perceptual implications of these results using two perception studies. The first study compared careful and casual speech with the pitch replaced by a synthetic whisper or hum and found that casual speech was more difficult to recognize. In a second experiment, participants were asked to transcribe either the full resynthesized utterance or a synthesized whispered form from casual speech. Brenner found that the utterances with pitch cues slightly reduced the transcription errors (e.g., from 16% to 12.5% for character recognition). The results from Brenner's studies indicate that recognizing tones in casual speech is more difficult than in careful speech. He also shows that tone perception in casual speech is a complex interaction of acoustic cues and not just based on pitch, which is in line with previous careful speech research (e.g., Taft & Chen, 1992; Fu and Zeng, 2000; Liu and Samuel, 2004; Chen & Tucker, 2013). Brenner's results suggest that in everyday conversations, listeners may rely on different cues for language processing than would be expected based on laboratory studies alone.

Differences between casual and careful speech have also been documented for intonation. De Ruiter (2015) showed that in German the information status of discourse referents is marked differently in unscripted narrations of picture stories than in read aloud speech. For instance, whereas in the scripted speech given referents were consistently de-accented, they were not in the unscripted speech. In addition, where scripted speech showed low boundary tones, unscripted speech showed low pitch accents in combination with high boundary tones to indicate continuity. The unscripted speech investigated by de Ruiter could also be classified as careful, because they are monologues elicited in the laboratory. This may indicate that the differences between read aloud careful speech and casual speech are even more substantial.

In three corpus studies of casual German and English, Schweitzer et al. (2015) investigated whether the exact intonation contour is affected by the identities of the words. They found that the frequency with which a given word occurs with a given accent type affects the amplitude of the accent. Moreover, this absolute frequency increases the variability in pitch accent shape, while the relative frequency with which a word occurs with a given accent has the opposite effect. These findings are unexpected in a production model which predicts that intonation contours are determined independently of the words. Schweitzer and colleagues therefore argued that intonational contours are stored as part of the lexical representations of words.

These studies have direct implications for our view of the mental lexicon. Most importantly, they raise questions about what is exactly stored in the mental lexicon. For instance, are lexical tones stored as lexical tones or are they specified in other ways that better correspond to their role in casual speech? Are words specified for their most likely intonation contour? The answers to these questions necessarily have direct implications for our models of speech production and comprehension.

*Phonological assimilation*

According to the phonological descriptions of many languages, mostly based on introspection and careful speech, categorical assimilation is a frequent phenomenon. That is, many languages contain productive processes that categorically change the feature of one segment to match a neighboring segment. As a consequence, a sentence like *a quick ru*[m] *picks you up* is completely ambiguous as to whether it refers to alcohol or to sports. This observation has initiated several theories and experimental work on how listeners cope with complete assimilation (e.g., Gaskell & Marslen-Wilson, 1998; Lahiri & Reetz, 2002). Recent work, however, casts doubt on the wide-spread assumption that categorical, fully productive assimilation is prevalent in languages (for an overview, see Ernestus, 2012). Many categorical assimilation processes are less frequent than previously assumed. Moreover, many processes are not always categorical: often the segment that is assumed to completely assimilate to its neighboring segment still contains some of its original acoustic properties (Ernestus, 2012). This shows that the strategies that listeners have been assumed to apply in order to cope with categorical assimilation may not play an important role in casual speech comprehension. Experiments testing how listeners process categorical assimilation show what the perception system does if put under pressure, but they may not reflect casual speech processing.

Regressive voice assimilation in Dutch is one of the processes that is more complex than previously thought, and that consequently raises questions about speech production. This process is generally assumed to be categorical: every obstruent followed by a voiced plosive is completely voiced. However, the study by Ernestus, Lahey, Verhees, & Baayen (2006), which is based on lively read aloud stories for the blind (and thus not even on casual speech), shows that this assumption is not correct; in no less than 25% of the productions where regressive voice assimilation should have occurred according to received wisdom (e.g. in *wetboek*, consisting of the parts *wet* /ʋɛt/ 'law' and *boek* /buk/ 'book'), the segment sequence showed progressive voice assimilation (/ʋɛtpuk/ rather than /ʋɛdbuk/). Furthermore, the study showed a correlation between the frequency of the word (often a compound) and the degree and direction of voicing. A finding like this has direct implications for models of speech production that are fully decompositional, such as Weaver++ (Levelt, Roelofs, & Meyer, 1999). In these models, effects of frequency are assigned at the morpheme level and they therefore have difficulty accounting for frequency effects across two lexeme boundaries of compounds.

*Single sound substitutions and the absence of sounds*

Not only assimilation, but also other processes that affect or delete single sounds may be more frequent in casual speech than in careful speech. Dilts' (2013) dissertation used the phonetic transcriptions (hand corrected force-aligned segmentation) provided in the Buckeye Corpus of Conversational Speech (Pitt, Dilley, Johnson, Kiesling, Raymond, Hume, & Fosler-Lussier, 2007) to examine reduction and found that 29,888 of the 137,319 content word tokens (i.e., 22%) were transcribed as lacking one or more sounds from the citation form. More generally, 38% of the content word tokens were transcribed in a form that did not match the citation forms provided by the corpus.

Another example concerns the realization of voiced and voiceless alveolar stops as flaps in North American English. In contrast to what may be concluded on the basis of introspection (which is likely to be influenced by orthography), the flap is the most common production in words like *atom* [æɾəm]: Warner and Tucker (2011) show that 99% of tokens of word-medial alveolar stops are flapped when the stop is followed by an unstressed vowel, even in a carefully articulated word list. In fact, a production containing an aspirated /t/ sounds awkward and foreign in North American English and the flap represents the careful realization. In line with this result, in a series of six experiments, McLennan, Luce, & Charles-Luce (2003) observed that listeners recognize words more easily if they are pronounced with a flap than with a careful stop.

One question raised by this variability is whether the different pronunciation variants (like /$t^h$/ versus flap in American English) result from some "phonological" rule operating during speech production and the inverse during comprehension, or whether the different pronunciation variants may be stored in lexical memory. Ranbom & Connine, (2007) addressed this question by investigating how quickly listeners recognize the /$t^h$/ and the flap variants of words, contrasting words that are mostly produced with /$t^h$/ with those that are mostly produced with a flap. They found that listeners recognize a word's flap variant more easily if the word occurs more often in the flap variant, and they take this as evidence that the different word pronunciation variants are lexically stored, together with their lexical frequencies. Similar frequency effects were observed for the processing of the variants of French words lacking the word-medial schwa (e.g., /fnɛtr/ for /fənɛtr/ *fenêtre* 'window' and /rnar/ for /rənar/ *renard* 'fox' (Brand & Ernestus, submitted)). Bürki, Ernestus & Frauenfelder (2010) showed that speakers need less time to start producing the reduced variant of a word (i.e. the variant without word-medial schwa), the more frequent the occurrence of this variant relative to the corresponding full variant. Brand & Ernestus (submitted) showed that these same frequency effects hold for perception: listeners recognize the reduced variant of a French schwa word more quickly if that word occurs more often without schwa. Following Ranbom & Connine (2007), both Bürki and colleagues and Brand and Ernestus interpret these frequency effects as supporting the hypothesis that a word may be lexically stored with several pronunciation variants.

This hypothesis has important implications for theories of the mental lexicon. It raises questions, for instance, about the number of pronunciation variants that are typically stored for a single word, about which types of pronunciation variants are likely to be stored (only those that differ from the full variants at the segmental level or also those that differ from the full variants in the exact realization of these segments?), and about how the pronunciation variants of a single word are lexically organized. These questions are directly relevant for theories on speech production and speech comprehension. For instance, if many pronunciation variants are stored, what role is left for computation (phonological rules)? Moreover, do the different pronunciation variants stored in the lexicon compete with each other during word recognition? This latter question is a good example of a question that would not have been raised if all research was restricted to careful speech.

### *Continua of pronunciation variation*

In casual speech, there is often a continuum between one realization of a segment and another, which again raises questions about lexical representation, speech production and speech comprehension. For instance, Warner and Tucker (2011) showed that, in addition to /nt/ and flap realizations, there are many pronunciation variants of the North American English flap that differ in strength. Importantly, the weakened flaps spanning this continuum between a real flap and an approximant-like consonant are very common in casual speech.

Whole words can also show continua of pronunciation variants. Greenberg (1999) reported an investigation of the Switchboard corpus of casual conversations (Godfrey, Holliman, & McDaniel, 1992) and found that the word *that* is produced in 117 different ways, with the form [ðæ] being the most frequent at 11%. Greenberg also found that the word *people* was produced 45 times in the corpus with 21 different realizations. The word *and* was produced 87 different ways; the top seven realizations for this word are [æn], [ɛn], [ɪn], [ən], [n̩], [n], with [ænd] in seventh position. Similar findings have been reported for other languages. For instance, Ernestus (2000: 141) reported 14 pronunciations for Dutch *natuurlijk* 'of course'. In a laboratory setting this range of variation found in casual speech is unlikely to occur.

This variation raises the question of how listeners arrive at lexical access. There is no known phonological model available that would allow for the listener to decode all these variants into their phonological forms as a form of allophonic variation. There are many possible alternative explanations. For instance, there could be a pattern matching system that simply finds the "best fit" to the acoustic input. It is unclear, however, whether such a simple mechanism can explain all findings (e.g. why is Dutch [ɛik] so easily recognized as *eigenlijk* 'actually'/ɛixələk/ even though there is also the word *eik* /ɛik/ 'oak tree'). Moreover, if we just assume a simple pattern matching mechanism for comprehension, we have to find another mechanism to explain the (consistent) variation in production. If we assume storage of all pronunciation variants, the same questions arise as those concerning the storage of pronunciation variants of words differing in single segments, including which pronunciation variants are stored and how pronunciation variants of the same word are treated during word recognition.

## Challenges and opportunities for research on the production of casual speech

In the previous section, we discussed some of the many differences between careful speech and casual speech. We also illustrated some of the questions arising from this research in relation to our understanding of psycholinguistics. Taken together, we show that it is imperative that more research be performed on casual speech, however we recognize that is not an easy process. In this section, we describe some of the challenges and opportunities for research on casual speech production.

*Experimental paradigms*

Speech production in psycholinguistics is commonly investigated with experimental paradigms like picture naming or sentence reading. These paradigms seldom elicit casual speech, as they are conducted in formal settings (typically a laboratory) and speakers have to perform a (formal) task. Because speakers tend to only produce casual speech if they feel free to say what they want, in the way they want, in informal settings, the elicitation of casual speech is extremely difficult in highly controlled experiments. This especially excludes all experimental paradigms that aim at measuring how much time speakers need to proceed from concept to the articulation of a word.

Many studies have attempted to draw information about the cognitive processes underlying speech production from the detailed acoustic characteristics of words (e.g. Bürki et al., 2011; Pluymaekers, Ernestus & Baayen, 2006). Variables that affect the acoustic properties of words must have a role in the speech production process. For instance, Gahl, Yao and Johnson (2012) argued on the basis of corpus research that words with many phonological neighbors tend to be phonetically reduced (shortened in duration and produced with more centralized vowels) in connected speech, in contrast to what has been found for more careful speech (Munson & Solomon, 2004). They argue that since words with many neighbors are harder to understand, reduction does not benefit the listener and is therefore likely to be speaker-driven. Further research is necessary to confirm this hypothesis.

A very different approach to research on speech production can be found in studies using ultrasound tongue imaging and electromagnetic articulography (e.g., Schönle, Gräbe, Wenig, Höhne, Schrader, & Conrad, 1987; Stone 1990). These techniques provide direct information on the position of the articulators during speaking. Although the speakers and their interlocutors can easily see and feel the equipment involved (in the case of articulography, the sensors, being small, only interfere minimally with articulation), they tend to accommodate their productions relatively quickly (e.g., Fowler and Turvey, 1981), especially when they are free to move their heads (Gick, 2002; Wrench & Scobbie 2011). As a consequence, the speech recorded with this type of equipment can also be said to be casual.

*Compiling and working with corpora of casual speech*

Corpora of casual speech that best approximate the speech obtained in controlled experiments are those in which the participants have to perform a task. One task that may make participants produce requested words in a (semi-) casual speech style is the map task, in which one participant has to describe a route indicated on a map to another participant, who has a slightly mismatching map without the route (e.g., Anderson et al., 1991; Engen, Baese-Berk, Baker, Choi, Kim, Bradlow, 2010). This task may elicit casual speech, for instance if the participants know each other very well and the setting is as informal as possible. If the map contains specific landmarks (e.g. a cathedral, a tree), these landmarks are likely to be mentioned and the researcher can thus elicit and compare specific words. The words mentioned in the route descriptions, however, are likely to differ in many respects due to their immediate contexts (e.g. in prosody, speech rate, the predictability of following word), which makes direct comparisons of the word tokens difficult.

Recently, many corpora have been compiled of truly casual conversations in many different languages (see, e.g. Ernestus & Baayen, 2011; Warner, 2012 for descriptions of some of these speech corpora). These corpora allow researchers to quickly identify bits of speech and investigate their syntactic constructions, acoustic characteristics, their words' frequencies of occurrence, and many other characteristics. Analyzing a recording of free casual conversation is even more difficult than analyzing recordings made with a map task because the speakers are free to say what they want and have no restrictions regarding their production. Nevertheless, the many studies based on these corpora show that those that are currently available are a valuable asset to psycholinguistics, but that more are necessary to truly understand the full range of variation possible across speech styles and languages (including signed languages) and the cognitive processes underlying them. For example, most of the data available are limited in terms of speech register; only a few data sets allow direct comparisons of more than two or three speech registers.

Many of these corpora have also begun to include video along with the audio and transcripts to enable the analysis and investigation of gestures during casual speech. Related to these gestural aspects of communication, it is also necessary to investigate the nature of casual signed languages, where differences between casual and careful language should also be explored, as in Tyrone and Mauk (2010). One possible advantage of investigating casual signed languages is that the articulatory aspects of signed languages are simpler to access.

Unfortunately, corpora come at a high cost in terms of their creation: it is relatively simple to record the data; the labor intensive aspect is manually tagging the speech (e.g., orthographic transcription, segmenting words and phonemes) for future analysis. Tagging speech data can, however, at least partly be done automatically. Increasingly, researchers use automatically generated transcriptions of speech corpora to speed up their research. This is a promising development, although the output of automatic transcribers has to be considered with care. Automatic speech recognizers available for research purposes still produce many errors in their orthographic transcriptions of casual conversations, among others, due to the overlap between speakers (e.g., Çetin & Shriberg, 2006) and the extensive variation between and within speakers (Benzeghiba et al., 2007). However recent advances indicate that significant improvement is on the horizon (Xiong et al., 2016).

The automatic generation of phonemic transcription consists of the machine choosing a given word in the orthographic transcription from the pronunciation variants listed in the machine's lexicon that best matches the acoustic signal (forced alignment). The resulting phonemic transcriptions are often as (un)reliable as those produced by human transcribers (for a discussion, see, e.g. Ernestus & Baayen, 2011). Note that this procedure implies that the automatically generated transcription can only contain word pronunciation variants that are listed in the lexicon and therefore cannot reveal new pronunciation variants.

The use of a corpus that is automatically phonemically transcribed makes it possible to quickly extract a large amount of data, which is extremely helpful when analyzing casual speech. Researchers can do this without having ever listened to or visually inspected the data. This can be a major disadvantage, however, because the researcher can miss important aspects of the pronunciation variation under investigation. We believe that research on casual speech requires that the researcher listens to the actual conversations. In addition, researchers should visually investigate the signal's acoustic characteristics because often the researcher, who is also a listener, reconstructs the full form of the speech when listening to casual, connected speech (Kemps, Ernestus, Schreuder & Baayen, 2004). This can make it very difficult to identify reductions such as [dʲætə] for 'do you have to…'.

*Statistical modeling of casual speech data*

The lack of control in production studies of casual speech determines the need for advanced statistical analysis. For instance, if we would like to know whether a given variable has an effect on the duration of vowels, the effects of all other possible variables on vowel duration (e.g. speech rate, the presence of accent, the word's frequency of occurrence) have to be partialed out. This may not be easy because the resulting statistical analyses, with many control variables, may result in data overfitting. Moreover, a variable that was not considered relevant at the time of the study may be found to be so later on, which could make the result of the study difficult to interpret. Note, however, that the latter also holds for highly controlled experiments where words are produced in isolation: if an important variable is not taken into account, the different conditions in the experiment may not be well matched and the results may consequently be difficult to interpret.

The last decade has seen a tremendous increase in the use of linear mixed effects models, also called multi-level models, in the analysis of language data (e.g., Baayen, 2008). These models make it possible to extract more patterns from noisy data than was possible in the past. The resultant models are complex, and new journal articles are regularly published indicating how these models should be applied in order to avoid overfitting and Type 1 errors (e.g. Bates, Kliegl, Vasishth & Baayen, submitted; Wurm & Fisicaro, 2014). These articles sometimes contradict each other and it is therefore still unclear what the best procedure for model fitting is. We recommend that researchers explicitly describe their fitting procedure and report only those statistical results that also emerge if the fitting procedure is (slightly) changed (e.g. with or without random slopes). This is also true for generalized additive mixed models, which some researchers have also started to use (e.g. Kryuchkova et al., 2012; Baayen, van Rij, de Cat, & Wood, to appear; Hastie & Tibshirani, 2002; Wood, 2006). In addition to the methods exemplified here, there are several other methods that can facilitate the analysis of casual speech, including Bayesian modeling (e.g., Kruschke, 2010, 2014) and classification techniques (e.g., Dilts, 2013; Tagliamonte & Baayen, 2012). While statistical approaches offer many opportunities, we do not believe that statistical solutions will overcome all the challenges presented in the analysis of casual speech.

## Challenges and opportunities for research on the comprehension of casual speech

Like research on speech production, those researching the comprehension of casual speech face many challenges, although they can benefit from many new opportunities. This section describes both the challenges and opportunities for different aspects of casual speech comprehension research.

### *The stimuli in comprehension experiments*

As is true for any study of speech comprehension, studies of the comprehension of casual speech must be based on sound knowledge of the characteristics of the type of speech presented. Whereas, for careful speech styles, this knowledge can be based on introspection or pronunciation dictionaries, this is not the case for casual speech, as argued above. Comprehension studies with laboratory-made stimuli must therefore follow production studies. An example is the work by Tucker (2011) on reduced word-medial stops. Tucker created stimuli in the laboratory that reflected the production findings of Warner & Tucker (2011). He then used a lexical decision task with words in isolation to compare the comprehension of reduced and unreduced stops.

We often do not know which properties of a casually produced sentence or word are important for speech comprehension. For instance, if we are interested in how listeners process left dislocation in casual speech, what do the stimulus sentences have to sound like for participants to accept them as representing real casual speech? One way to overcome this problem is to present listeners with stretches from spontaneously produced speech. Examples of these types of studies include: Brenner (2015), described above; Ernestus, Baayen & Schreuder (2002), who investigated how the amount of context influences the recognition of highly reduced word forms; Van de Ven, Ernestus & Schreuder (2012), who investigated what type of context listeners use most when processing reduced pronunciation variants of words; Podlubny, Nearey, & Tucker (2011) and Bernhard & Tucker (2015), who both investigated the contribution of specific acoustic cues such as duration, amplitude and pitch to the recognition of reduced words.

This solution, however, implies that researchers surrender some level of control when designing their experiments. It is very unlikely that they will find two stretches of speech in a corpus that only differ in the characteristic under investigation. Moreover, they may not find examples of, for instance, the words they are interested in because, while the variation in casual speech is much greater in some respects, it is smaller in others (as discussed above). As a result, a certain amount of noise is inherently introduced into these experiments. This noise can often be dealt with by incorporating additional control variables in the statistical modeling, as described above, including properties of the word, or information about the intonation contour. This, however, leads to the same statistical challenges discussed above. Moreover, this statistical solution requires large amounts of perception data to accurately test against the research questions.

*Experimental paradigms*

Another challenge for research on the comprehension of casual speech concerns the experimental paradigm. An experimental paradigm has to allow the researcher to investigate the comprehension of linguistic units (e.g., syntactic structures, words) in their natural context and it has to allow participants to listen to the stimuli as in a normal conversation. Traditional paradigms, however, present words in isolation or ask participants to perform unnatural metalinguistic tasks (e.g. lexical decision or phoneme monitoring).

One possibility that fulfills one of the two requirements is to play back a casual monologue and have participants perform a type of cross-modal identity priming during parts of the monologue. This would allow the researcher to test the listener's processing at predetermined points in the monologue where a phenomenon of interest occurs. So far, we have little experience with this experimental method and, as a consequence, it is unclear what the best interval is between the word in the acoustic signal and the word on the screen and how this interval should vary as a function of speech style. It is also unclear how sensitive this method may be. Moreover, we suspect that the metalinguistic task that participants have to perform may cause them to listen to the speech in a different way than they do in everyday informal settings.

Another experimental paradigm that can be used with spontaneous speech is the visual world paradigm, where an individual's eye-gazes are recorded while they are listening to speech and looking at words or objects presented on a computer screen. Brouwer, Mitterer and Huettig (2012), for instance, presented stretches of casual speech to participants, who had to indicate which of the four words presented on the screen appeared in the sentence. The researchers showed that if words are reduced, listeners penalize acoustic mismatches less strongly when listening to casual speech than when listening to fully articulated speech. One of the advantages of this experimental paradigm is that participants do not have to perform a metalinguistic task.

There are, however, also disadvantages. The visual world paradigm primes the listeners for the words presented on the computer screen (in the form of pictures or the printed orthographic transcriptions), which affects speech processing. Furthermore, the statistical modeling of eye movements may be an even greater challenge than the modeling of acoustic data (discussed above) because eye movement data do not consist of independent single points (e.g., vowel durations, single reaction times) but of time series of dependent data points. To overcome this problem, eye movements are traditionally reconverted to simple data points by averaging the percentage of looks to a given object over a given time window. This analysis method does not allow the researcher to spot differences in the exact shapes of the eye movement curves. Fortunately, new methods are being developed that can take the exact shapes of the curves into account (e.g., Baayen, et al., to appear; Mirman, Dixon, & Magnuson, 2008; Oleson, Cavanaugh, McMurray, & Brown, 2015).

Eye tracking equipment can also be used to measure the size of the pupil. Pupilometry (measuring pupil dilation or size) may provide information about the processing of casual speech as changes in pupil size have been claimed to reflect changes in cognitive load or the amount of effort put forth during a task. This technique has provided interesting information about many domains of cognition (e.g., Goldinger & Papesh, 2012), including language processing. A typical auditory task involves listening to a stimulus and responding to that stimulus, like repeating the word that was heard (e.g., Klingner, Tversky & Hanrahan, 2011; Zekveld, Kramer & Festen, 2010). One of the major challenges for experimental paradigms using this technique is that the pupils require time to return to a baseline or "resting" dilation diameter; this is often up to 3 seconds. Postponing presenting the next stimulus until after the reset time results in an unnatural experiment and makes it difficult to investigate the natural processing of casual speech. Several recent studies have, nevertheless, used pupilometry with connected sentences (Koch & Janse, 2016; van Rij et al., submitted). Van Rij et al. (submitted) present several statistical solutions for resolving many of the challenges of analyzing pupil dilation over the course of a sentence, which could be applied to stimuli of casual speech.

With some significant investment, it may also be possible to analyze the processing of casual speech with electroencephalography (EEG), functional near-infrared spectroscopy (fNIRS), functional magnetic resonance imaging (fMRI), or magnetoencephalography (MEG). Neurolinguistics has become a very popular area in psycholinguistics, but the number of studies using these techniques to investigate the processing of casual speech is extremely low. The probable reason is that these data are very difficult to analyze and, according to accepted wisdom, these studies should therefore contrast extremely well-controlled stimuli such they that are (nearly) identical in the different experimental conditions except for the single feature under investigation. Moreover, if target words are presented in sentences, they should occur in sentence-final position so that the processing of the target word does not coincide with the processing of the following word. These restrictions obviously exclude the study of casual speech. We are aware of only one project that, in contrast to this tradition, has investigated natural connected speech instead of highly controlled sentences. The study presented read aloud stories in a careful style of speech to participants in an fMRI experiment and investigated the effect of a word's predictability on processing (Willems, Frank, Nijhof, Hagoort and van den Bosch, 2016). We are convinced that neuro-imaging will provide useful data once reliable statistical methods are available to analyze this extremely noisy data, and we are excited by the attempts to solve these problems (e.g., Baayen et al., to appear; Mulder, ten Bosch & Boves, submitted).

## Where do we think psycholinguistic research should go?

This article has put forward the claim that real progress in theories of speech production, speech comprehension, and the mental lexicon can only be achieved if more studies investigate the processing of speech in its most natural context, that is, in casual conversations. These studies will provide information that cannot be revealed by studies on careful speech and will raise new and important questions.

Moreover, studies have to focus on casual speech in order to investigate whether hypotheses that are supported by studies on careful speech or on written language also hold for everyday language situations. We thus see the extension to casual speech as the last step in a research path that establishes the relevance of a given cognitive process, first in highly controlled conditions and then in conditions that are increasingly natural. This path can start, for instance, with research on the processing of printed words in isolation, subsequently broaden to printed words in sentences, printed words in stories, single word utterances, single word utterances reflecting the reduction found in casual speech, sentence-final words, sentence medial words, and end in research on words in casual conversations. We find that in nearly all of the psycholinguistic literature the last step is left out and it is this step that we explored and discussed in the present article, focusing specifically on what can be learned and gained from the investigation of casual speech in research on language representation, comprehension and production.

We provided several examples of experimental paradigms that have already been used to investigate casual speech. We also discussed several new possible experimental methods such as cross-modal identity priming, the visual-world paradigm, pupilometry, and neurolinguistic methods that may also provide useful avenues for researching casual speech. We believe that there are many other ways that production and comprehension could be explored in casual speech and encourage researchers to explore other alternative avenues.

This article focused on casual speech and discussed experimental methods in which speech is the only means of communication. In everyday conversations, however, speakers also convey messages with their facial expressions and their gestures. Moreover, speech may be produced differently and comprehended differently when the object the speech refers to is visibly present. The next step in the ecological validity path should therefore be casual speech in its multi-modal context (e.g. the work described by Drijvers & Özyürek, in press). Moreover, we expect that this type of speech will also produce new questions and insights not revealed by studies that only focus on casual speech.

Our exploration of how we can investigate everyday language processing showed that this is only possible with advanced statistical techniques. We therefore encourage researchers to learn and develop new statistical procedures, for instance, based on techniques developed in other disciplines, and apply them to language data. We believe that the field will profit greatly from strong collaborations between psycholinguists and mathematicians.

In conclusion, the study of speech production and comprehension and of the mental lexicon is awaiting rewarding challenges. Substantial extension to casual speech, that is, to language in an ecologically valid setting, will enrich our knowledge about daily language behavior. Doing so is likely to show that what is currently known based on careful speech is at best only half of the story.

## Funding

The writing of this article was partly funded by a Canadian Social Sciences and Humanities Research Council Insight grant to the first author and a vici grant from the Netherlands Organization for Scientific Research to the second author.

## References

Anderson, A. H., Bader, M., Bard, E. G., Boyle, E., Doherty, G., Garrod, S., & Sotillo, C. (1991). The HCRC map task corpus. *Language and Speech*, 34(4), 351–366. https://doi.org/10.1177/002383099103400404

Baayen, R. H. (2008). *Analyzing Linguistic Data. A Practical Introduction to Statistics Using R*. Cambridge University Press. https://doi.org/10.1017/CBO9780511801686

Baayen, R. H., van Rij, J., de Cat, C. and Wood, S. N. (to appear) Autocorrelated errors in experimental data in the language sciences: Some solutions offered by Generalized Additive Mixed Models. In Speelman, D., Heylen, K. and Geeraerts, D. (eds), *Mixed Effects Regression Models in Linguistics*. Berlin, Springer. Retrieved from http://arxiv.org/abs/1601.02043. https://doi.org/10.1007/978-3-319-69830-4_4

Bates, D., Kliegl, R., Vasishth, S. and Baayen, R. H. (submitted) Parsimonious mixed models.

Bates, E., & Liu, H. (1996). Cued shadowing. *Language and Cognitive Processes*, 11(6), 577–582. https://doi.org/10.1080/016909696386962

Bentum, M., Ernestus, M., ten Bosch, L. & van den Bosch, A. (submitted) How do speech registers differ in the predictability of words?

Benzeghiba, M., De Mori, R., Deroo, O., Dupont, S., Erbes, T., Jouvet, D., Fissore, L., Laface, P., Mertins, A., Ris, S., Rose, R., Tyagi, V., & Wellekens, C. (2007). Automatic speech recognition and speech variability: A review. *Speech Communication*, 49(10), 763–786. https://doi.org/10.1016/j.specom.2007.02.006

Bernhard, D., & Tucker, B. (2015). The effects of duration on human processing of reduced speech. *Canadian Acoustics*, 43(3).

Biber, D. (1988). *Variation across speech and writing*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511621024

Biber, D., Conrad, S., & Reppen, R. (1998). *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge University Press. https://doi.org/10.1017/CBO9780511804489

Brand, Sophie & Ernestus, Mirjam. (submitted) How do native listeners and learners of French comprehend French word pronunciation variants?

Brenner, D. (2013). The acoustics of Mandarin tones in careful and conversational speech. *The Journal of the Acoustical Society of America*, 134(5), 4246. https://doi.org/10.1121/1.4831619

Brenner, D. S. (2015). The phonetics of Mandarin tones in conversation. Retrieved from http://arizona.openrepository.com/arizona/handle/10150/578721

Brouwer, S., Mitterer, H., & Huettig, F. (2012). Speech reductions change the dynamics of competition during spoken word recognition. *Language and Cognitive Processes*, 27(4), 539–571. https://doi.org/10.1080/01690965.2011.555268

Bürki, A., Ernestus, M., Gendrot, C., Fougeron, C., & Frauenfelder, U. H. (2011). What affects the presence versus absence of schwa and its duration: A corpus analysis of French connected speech. *The journal of the Acoustical Society of America*, 130(6), 3980–3991. https://doi.org/10.1121/1.3658386

Bürki, A., Ernestus, M. & Frauenfelder, U. H. (2010). Is there only one "fenêtre" in the production lexicon? On-line evidence on the nature of phonological representations of pronunciation variants for French schwa words. *Journal of Memory and Language* 62, 421–437. https://doi.org/10.1016/j.jml.2010.01.002

Çetin, Ö., & Shriberg, E. (2006). Speaker overlaps and ASR errors in meetings: Effects before, during, and after the overlap. In *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings* (Vol. 1). https://doi.org/10.1109/ICASSP.2006.1660031

Chen, T.-Y., & Tucker, B. V. (2013). Sonorant Onset Pitch as a Perceptual Cue of Lexical Tones in Mandarin. *Phonetica*, 70(3), 207–239. https://doi.org/10.1159/000356194

Chomsky, N. (1965). *Aspects of the Theory of Syntax*. Cambridge, MA: MIT Press.

Connine, C. M., & Titone, D. (1996). Phoneme monitoring. *Language and Cognitive Processes*, 11(6), 635–646. https://doi.org/10.1080/016909696387042

De Chat, C. (2007). *French Dislocation. Interpretation, syntax, acquisition*. *Oxford Studies in Theoretical Linguistics*, 17. Oxford University Press, Oxford, (288pp).

Dilts, P. C. (2013). Modelling phonetic reduction in a corpus of spoken English using Random Forests and Mixed-Effects Regression (Thesis). Retrieved from https://era.library.ualberta.ca/downloads/5425k999s

Drijvers, L., & Özyürek, A. (in press) Visual context enhanced: The joint contribution of iconic gestures and visible speech to degraded speech comprehension. *Journal of Speech, Language, and Hearing Research*. https://doi.org/10.1044/2016_JSLHR-H-16-0101

Engen, K. J. V., Baese-Berk, M., Baker, R. E., Choi, A., Kim, M., & Bradlow, A. R. (2010). The Wildcat Corpus of Native-and Foreign-accented English: Communicative efficiency across conversational dyads with varying language alignment profiles. *Language and Speech*, 53(4), 510–540. https://doi.org/10.1177/0023830910372495

Ernestus, M. (2000). *Voice assimilation and segment reduction in casual Dutch: A corpus-based study of the phonology-phonetic interface*. Holland Institute of Generative Linguistics, Utrecht.

Ernestus, M. (2012). Message related variation: Segmental within speaker variation. In: A. C. Cohn, C. Fougeron, & M. Huffman (eds.), *The Oxford Handbook of Laboratory Phonology*, 92–102. OUP, Oxford.

Ernestus, M. & R. H. Baayen. (2011). Corpora and exemplars in phonology. In: J. Goldsmith, J. Riggle, & A. Yu (eds.), *The Handbook of Phonological Theory* (2nd ed.), pages 374–400. Wiley-Blackwell, Chichester, West Sussex. https://doi.org/10.1002/9781444343069.ch12

Ernestus, M., Baayen, R. H., & Schreuder, R. (2002). The recognition of reduced word forms. *Brain and Language*, 81, 162–173. https://doi.org/10.1006/brln.2001.2514

Ernestus, M., Hanique, I., & Verboom, E. (2015). The effect of speech situation on the occurrence of reduced word pronunciation variants. *Journal of Phonetics*, 48, 60–75. https://doi.org/10.1016/j.wocn.2014.08.001

Ernestus, M., Lahey, M., Verhees, F., & Baayen, R. H. (2006). Lexical frequency and voice assimilation. *Journal of the Acoustical Society of America*, 120, 1040–1051. https://doi.org/10.1121/1.2211548

Fowler, C. A., & Turvey, M. T. (1981). Immediate compensation in bite-block speech. *Phonetica*, 37(5–6), 306–326. https://doi.org/10.1159/000260000

Fu, Q.; Zeng, F. (2000). Identification of temporal envelop cues in Chinese tone recognition. *Asia Pacific Journal of Speech Language and Hearing*, 5: 45–57. https://doi.org/10.1179/136132800807547582

Gahl, S., Yao, Y., & Johnson, K. (2012). Why reduce? Phonological neighborhood density and phonetic reduction in spontaneous speech. *Journal of Memory and Language*, 66(4), 789–806. https://doi.org/10.1016/j.jml.2011.11.006

Galliano, S., Georois, E., Mostefa, D., Choukri, K., Bonastre, J.-F., and Gravier, J. (2005). ESTER phase II evaluation campaign for the rich transcription of French broadcast news. *Proc. Interspeech 2005*, 2453–2456.

Gaskell, Gareth and William Marslen-Wilson. (1998). Mechanisms of phonological inference in speech perception. *Journal of Experimental Psychology: Human Perception and Performance* 24: 380–396.
Gaygen, D. E., & Luce, P. A. (1998). Effects of modality on subjective frequency estimates and processing of spoken and printed words. *Perception & psychophysics*, 60(3), 465–483. https://doi.org/10.3758/BF03206867
Gick, B. (2002). The use of ultrasound for linguistic phonetic fieldwork. *Journal of the International Phonetic Association*, 32(02), 113–121. https://doi.org/10.1017/S0025100302001007
Godfrey, J. J., Holliman, E. C., & McDaniel, J. (1992). SWITCHBOARD: Telephone speech corpus for research and development. In *1992 IEEE International Conference on Acoustics, Speech, and Signal Processing, 1992. ICASSP-92* (Vol. 1, pp. 517–520). https://doi.org/10.1109/ICASSP.1992.225858
Greenberg, S. (1999). Speaking in shorthand – A syllable-centric perspective for understanding pronunciation variation. *Speech Communication*, 29, 159–176. https://doi.org/10.1016/S0167-6393(99)00050-3
Goldinger, S. D., & Papesh, M. H. (2012). Pupil dilation reflects the creation and retrieval of memories. *Current Directions in Psychological Science*, 21(2), 90–95. https://doi.org/10.1177/0963721412436811
Hastie, T. J. & Tibshirani, R. J. (2002). *Generalized additive models*. Vol. 43. CRC Press, 1990.
Heylighen, F., & Dewaele, J.-M. (2002). Variation in the contextuality of language: An empirical measure. *Foundations of Science*, 7(3), 293–340. https://doi.org/10.1023/A:1019661126744
Hockett, Charles F. (1955). *A manual of phonology*. Baltimore: Waverly Press.
Hymes, D. (1992). The concept of communicative competence revisited. Thirty years of linguistic evolution. *Studies in honour of René Dirven on the occasion of his sixtieth birthday*, 31–57.
Kemps, R., Ernestus, M., Schreuder, R., & Baayen, R. H. (2004). Processing reduced word forms: The suffix restoration effect. *Brain and Language*, 19, 117–127. https://doi.org/10.1016/S0093-934X(03)00425-5
Klingner, J., Tversky, B., & Hanrahan, P. (2011). Effects of visual and verbal presentation on cognitive load in vigilance, memory, and arithmetic tasks. *Psychophysiology*, 48(3), 323–332. https://doi.org/10.1111/j.1469-8986.2010.01069.x
Koch, X., & Janse, E. (2016). Speech rate effects on the processing of conversational speech across the adult life span. *The Journal of the Acoustical Society of America*, 139(4), 1618–1636. https://doi.org/10.1121/1.4944032
Kruschke, J. K. (2010). What to believe: Bayesian methods for data analysis. *Trends in cognitive sciences*, 14(7), 293–300. https://doi.org/10.1016/j.tics.2010.05.001
Kruschke, J. K. (2014). *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan*. Academic Press.
Kryuchkova, T., Tucker, B. V., Wurm, L. H., & Baayen, R. H. (2012). Danger and usefulness are detected early in auditory lexical processing: Evidence from electroencephalography. *Brain and Language*, 122(2), 81–91. https://doi.org/10.1016/j.bandl.2012.05.005
Labov, W. (1972). *Sociolinguistic Patterns*. University of Pennsylvania Press.
Lahiri, A., & Reetz, H. (2002). 'Underspecified recognition', in Carlos Gussenhoven, Natasha Warner, and Toni Rietveld (eds.), *Phonology & Phonetics: Laboratory Phonology VII*. Berlin, Mouton, pp. 637–676. https://doi.org/10.1515/9783110197105.637
Levelt, W. J. M., Roelofs, A., & Meyer, A. S. (1999). A theory of lexical access in speech production. *Behavioral and Brain Sciences*, 22, 1–38. https://doi.org/10.1017/S0140525X99001776

Lindblom, B. (1963). Spectrographic study of vowel reduction. *The Journal of the Acoustical Society of America*, 35(11), 1773–1781. https://doi.org/10.1121/1.1918816

Liu, S.; Samuel, A. G. (2004). Perception of Mandarin lexical tones when F0 information is neutralized. *Language & Speech*, 47: 109–138. https://doi.org/10.1177/00238309040470020101

MacWhinney, B. (2000). *The CHILDES Project: Tools for analyzing talk*. Third Edition. Mahwah, NJ: Lawrence Erlbaum Associates.

McLennan, C. T., Luce, P. A., & Charles-Luce, J. (2003). Representation of lexical form. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 29(4), 539–553.

McQueen, J. (1996). Word spotting. *Language and Cognitive Processes*, 11(6), 695–699. https://doi.org/10.1080/016909696387114

Mehta, G., & Cutler, A. (1988). Detection of target phonemes in spontaneous and read speech. *Language and Speech*, 31 (Pt 2), 135–156. https://doi.org/10.1177/002383098803100203

Mirman, D., Dixon, J. A., & Magnuson, J. S. (2008). Statistical and computational models of the visual world paradigm: Growth curves and individual differences. *Journal of Memory and Language*, 59(4), 475–494. https://doi.org/10.1016/j.jml.2007.11.006

Mulder, K., ten Bosch, L., & Boves, L. (submitted) Comparing different methods for analyzing ERP signals. https://doi.org/10.21437/Interspeech.2016-967

Munson, B., & Solomon, N. P. (2004). The Effect of Phonological Neighborhood Density on Vowel Articulation. *J Speech Lang Hear Res*, 47(5), 1048–1058. https://doi.org/10.1044/1092-4388(2004/078)

Oleson, J. J., Cavanaugh, J. E., McMurray, B., & Brown, G. (2015). Detecting time-specific differences between temporal nonlinear curves: Analyzing data from the visual world paradigm. *Statistical Methods in Medical Research*, 0962280215607411.

Oostdijk, N. (2000). The Spoken Dutch Corpus Project. *The ELRA Newsletter*, 5, 4–8.

Pitt, M. A., Dilley, L., Johnson, K., Kiesling, S., Raymond, W., Hume, E., & Fosler-Lussier, E. (2007). Buckeye Corpus of Conversational Speech (2nd release) [www.buckeyecorpus.osu.edu] Columbus, OH: Department of Psychology. *Ohio State University (Distributor)*.

Pluymaekers, M., Ernestus, M., & Baayen, R. (2006). Articulatory planning is continuous and sensitive to informational redundancy. *Phonetica*, 62(2–4), 146–159. https://doi.org/10.1159/000090095

Podlubny, R., Geeraert, K., Tucker, B. V. (2015). It's all about, *like*, acoustics. *Proceedings of the 18th International Congress of Phonetic Sciences*. Glasgow, UK: the University of Glasgow. Paper number 0477.

Podlubny, R., Tucker, B. V., & Nearey, T. (2011). 'Sorry, what was that?': The roles of pitch, duration, and amplitude in the perception of reduced speech. Poster presented at the Nijmegen Spontaneous Speech Workshop (Nijmegen, NL).

Pollack, I., & Pickett, J. M. (1963). Intelligibility of excerpts from conversational speech. *Language and Speech*, 6, 165–171. https://doi.org/10.1177/002383096300600305

Ranbom, L. J., & Connine, C. M. (2007). Lexical representation of phonological variation in spoken word recognition. *Journal of Memory and Language*, 57(2), 273–298. https://doi.org/10.1016/j.jml.2007.04.001

Richter, E. (1930). Beobachtungen über Anglitt und Abglitt an Sprachkurven und umgekehrt laufenden Phonogrammplatten. In Paul Menzerath (ed.) *Berichte über die I. Tagung der Internationalen Gesellschaft für experimentelle Phonetik*, 87–90. Bonn: Scheur.

Ruiter de, L. E. (2015). Information status marking in spontaneous vs. read speech in story-telling tasks – Evidence from intonation analysis using GToBI. *Journal of Phonetics* 48, 29–44. https://doi.org/10.1016/j.wocn.2014.10.008

Schönle, P. W., Gräbe, K., Wenig, P., Höhne, J., Schrader, J., & Conrad, B. (1987). Electromagnetic articulography: Use of alternating magnetic fields for tracking movements of multiple points inside and outside the vocal tract. *Brain and Language*, 31(1), 26–35. https://doi.org/10.1016/0093-934X(87)90058-7

Schweitzer, K., Walsh, M., Calhoun, S., Schütze, H., Möbius, B., Schweitzer, A., & Dogil, G. (2015). Exploring the relationship between intonation and the lexicon: Evidence for lexicalised storage of intonation. *Speech Communication*, 66, 65–81. https://doi.org/10.1016/j.specom.2014.09.006

Stampe, D. (1973). A Dissertation on Natural Phonology. PhD Diss. University of Chicago.

Stone, M. (1990). A three-dimensional model of tongue movement based on ultrasound and X-ray microbeam data. *The Journal of the Acoustical Society of America*, 87(5), 2207–2217. https://doi.org/10.1121/1.399188

Taft, M., & Chen, H. C. (1992). Judging homophony in Chinese: The influence of tones. *Advances in Psychology*, 90, 151–172. https://doi.org/10.1016/S0166-4115(08)61891-9

Tagliamonte, S. A., & Baayen, R. H. (2012). Models, forests, and trees of York English: Was/were variation as a case study for statistical practice. *Language Variation and Change*, 24(2), 135–178. https://doi.org/10.1017/S0954394512000129

Torreira, F., Adda-Decker, M., and Ernestus, M. (2010). The Nijmegen Corpus of Casual French. *Speech Communication*, 52:201–221. https://doi.org/10.1016/j.specom.2009.10.004

Tucker, B. V. (2007). *Spoken word recognition of the reduced American English Flap*. The University of Arizona. Retrieved from http://hdl.handle.net/10150/194987

Tucker, B. V. (2011). The effect of reduction on the processing of flaps and /g/ in isolated words. *Journal of Phonetics*, 39(3), 312–318. https://doi.org/10.1016/j.wocn.2010.12.001

Tyrone, M. E., & Mauk, C. E. (2010). Sign lowering and phonetic reduction in American Sign Language. *Journal of Phonetics*, 38(2), 317–328. https://doi.org/10.1016/j.wocn.2010.02.003

van Rij, J., Natalya, P., van Rijn, H., Wood, S. N., & Baayen, R. H. (submitted). Pupil dilation to study cognitive processing: challenges and solutions for time course analyses.

Ven, M. van de, Ernestus, M. & Schreuder, R. (2012). Predicting acoustically reduced words in spontaneous speech: The role of semantic/syntactic and acoustic cues in context. *Laboratory Phonology* 3, 455–481.

Viebahn, M., M. Ernestus, & J. McQueen. (2015). Syntactic predictability in the recognition of carefully and casually produced speech. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 41 (6), 1684–1702.

Wagner, P., Trouvain, J., & Zimmerer, F. (2015). In defense of stylistic diversity in speech research. *Journal of Phonetics*, 48, 1–12. https://doi.org/10.1016/j.wocn.2014.11.001

Warner, N. (2011). Reduction. In M. van Oostendorp, C. Ewen, E. Hume, & K. Rice (Eds.), *The Blackwell Companion to Phonology: General issues and segmental phonology* (Vol. 1, 1866–1891). John Wiley & Sons. https://doi.org/10.1002/9781444335262.wbctp0079

Warner, N. (2012). Methods for studying spontaneous speech. In A. Cohn, C. Fougeron, & M. Huffman (eds.), *The Oxford Handbook of Laboratory Phonology*. Oxford: Oxford University Press. 621–633.

Warner, N., & Tucker, B. V. (2011). Phonetic variability of stops and flaps in spontaneous and careful speech. *The Journal of the Acoustical Society of America*, 130(3), 1606–1617. https://doi.org/10.1121/1.3621306

Wiggers, P., & Rothkrantz, L. J. M. (2007). Exploratory analysis of word use and sentence length in the Spoken Dutch Corpus. In V. Matoušek & P. Mautner (Eds.), *Text, Speech and Dialogue* (366–373). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-74628-7_48

Willems, R. M., Frank, S. L., Nijhof, A. D., Hagoort, P., & Bosch, A. van den. (2016). Prediction during natural language comprehension. *Cerebral Cortex*, 26(6), 2506–2516. https://doi.org/10.1093/cercor/bhv075

Wood, S. N. (2006). *Generalized Additive Models*. New York: Chapman & Hall/CRC. https://doi.org/10.1201/9781420010404

Wrench, A. A., & Scobbie, J. M. (2011). Very high frame rate ultrasound tongue imaging. In Proceedings of the 9th international seminar on speech production (ISSP) (pp. 155–162).

Wurm, L. H., & Fisicaro, S. A. (2014). What residualizing predictors in regression analyses does (and what it does not do). *Journal of Memory and Language*, 72, 37–48. https://doi.org/10.1016/j.jml.2013.12.003

Xiong, W., Droppo, J., Huang, X., Seide, F., Seltzer, M., Stolcke, A., … Zweig, G. (2016). The Microsoft 2016 Conversational Speech Recognition System. *arXiv:1609.03528 [Cs]*. Retrieved from http://arxiv.org/abs/1609.03528

Xu, Y. (2010). In defense of lab speech. *Journal of Phonetics*, 38(3), 329–336. https://doi.org/10.1016/j.wocn.2010.04.003

Zekveld, A. A., Kramer, S. E., & Festen, J. M. (2010). Pupil response as an indication of effortful listening: The influence of sentence intelligibility. *Ear and Hearing*, 31, 480–490. https://doi.org/10.1097/AUD.0b013e3181d4f251

CHAPTER 5

# Hebrew adjective lexicons in developmental perspective

## Subjective register and morphology

Dorit Ravid, Amalia Bar-On, Ronit Levie and Odelia Douani
Tel Aviv University

### Thesis

The "holy grail" (or one of the major ones) of developmental psycholinguistics is to understand how words – with their structure, complex relationships, breadth and depth of meaning – are learned and processed by native speakers. Developmental scientists are especially interested in ways of assessing AoA of words or word classes, such as the class of Adjectives. One possibility is to use the morphological structure of words as a criterion. However, it is not entirely clear what constitutes morphological complexity and if this could be a valid criterion for lexical knowledge. Obviously, frequency plays a major role in this process, however frequency has many faces and offers challenges such as the interface of corpora with language users, usage of spoken and written corpora, insufficiently sensitive methods of judging frequency across different word classes, and language-specific issues, such as homography in Semitic orthographies. It is especially challenging to objectively assess frequency in relation to children's language acquisition and development. One subjective way to assess AoA, which circumvents the problems raised by objective frequency counts, is to use the ranking of words by their linguistic register (defined as 'levels of linguistic usage') by expert judges. A classification of Adjectives by register can reveal the underlying relationship between word structure, lexical semantics and levels of usage.

### Commentaries on Ravid et al. thesis

Ray Jackendoff commentary on Ravid et al. thesis

The authors propose that relative age of acquisition for adjectives can be correlated with frequency, which in turn can be correlated with gradation of register.

https://doi.org/10.1075/z.238.05rav

Furthermore, they conjecture that gradation of register can be determined by polling speech-language clinicians for subjective judgments of register.

It is not clear to me how to contextualize this undertaking. Do we have independent evidence for age of acquisition, against which these hypotheses can be evaluated? Where does age of acquisition for adjectives fit in with age of acquisition for nouns and verbs? How does age of acquisition vary across individuals, and is *order* of acquisition nevertheless relatively invariant?

Perhaps most important, even assuming a valid correlation between clinicians' judgments and age of acquisition, what is responsible for the spread in age of acquisition among adjectives? Could it simply be a matter of difference in exposure? After all, it is unlikely that children hear many "rare adjectives with high lexical specificity, serving highly educated Hebrew users with professional expertise". Alternatively, is the spread partly a function of cognitive development? That is, are some adjective meanings simply beyond the grasp of 5-year-olds? And is this something special about adjectives, or are similar issues of specialization and cognitive limitation found with nouns and verbs as well?

The result in this study of greatest interest to me is the distribution of morphological types: the most extensive variation in morphological patterns is found in the "core adjective lexicon," those adjectives judged most frequent and hypothesized to be learned earliest. The variety of forms diminishes as one moves up the scale of registers, until the morphological form of "super-literate" adjectives is nearly uniform across the entire class. This distribution is reminiscent of the well-known observation that irregular forms are likely to be high-frequency. However, the present result is more ramified, in that it is not simply a matter of regular vs. irregular. Rather, there is a competition in the high-frequency range among numerous alternatives, dwindling in the low-frequency range to a single highly dominant pattern plus a very few stragglers. One wonders if this situation is possible because of the rich derivational morphology of Hebrew, and whether counterparts might be found in other languages.

## James Myers commentary on Ravid et al. thesis

This study reveals interesting structural and behavioral patterns associated with subjective estimates of register level. Nevertheless, one would expect that a measure called "age of acquisition" would measure the age of acquisition. With very rare exceptions, however, when studies use this term, they actually mean subjective estimates by adults of the age of acquisition. One class of exceptions are the very few studies (e.g., Morrison, *et al.*, 1997, in *The Quarterly Journal of Experimental Psychology*) that look at when actual children first begin to use or show comprehension of certain words. The present study represents another kind of exception:

subjective adult estimates of something else (here, register) are presumed to correlate with the age of acquisition, or more directly, with subjective adult estimates of the age of acquisition. Of course, everyone knows that working with adults is easier than working with children (though the authors here depended on expert adults, who are presumably even harder to corral than children, let alone the regular adults who they rejected for providing insufficiently reliable data). And of course, researchers who offer the adult-based proxies have also argued that they correlate well with other adult-based proxies and with actual child data.

The problem is that almost every lexical variable correlates with every other lexical variable, while still not being identical. When offering subjective adult measures for child-oriented variables, researchers highlight the correlations and not the lack of identity; when arguing that lexical frequency is distinct from age of acquisition (or familiarity or word length or complexity, subject to Menzerath's Law), they highlight the lack of identity and not the correlations. Surely a study of the actual age of acquisition in actual Hebrew-acquiring children (and college students and post-graduate professionals) would yield values for this study's adjectives that are correlated with, yet not reducible to, the expert register judgments. Such values would then allow the researchers to show that such-and-such a behavioral phenomenon in adult lexical processing is more likely to derive from the brain changes associated with learning such-and-such a word type at such-and-such an age, and not from the word's register level (or vice versa, since sociolinguistic influences on lexical processing are important to study in their own right).

Scientific inference is hard, particularly in fields that really can only be studied via non-experimental methods (participants come into the lab already knowing their language, so lexical experiments are not true experiments, since the key variables are not manipulable). If you truly care about the age of acquisition, go test kids.

## Russell Richie commentary on Ravid et al. thesis

Usage of register, familiarity, formality and other subjective proxies of frequency is an interesting, likely fruitful avenue for future research in language research. A few thoughts concerning usage of register and similar notions occur to me:

First, 'register', as with any sociolinguistic concept, is surely complex and difficult to reduce to a 5 point scale. It seems likely that register has several dimensions, corresponding to relevant axes of power in a particular society: gender, religion, race, age, class, and so forth. Even if we restrict register to something like 'formality' in the sense of language use that has carefully prescribed rules of use in a particular context, it seems likely that we will discover language that, prima facie, seems informal yet fits that definition. For example, internet memes often have a highly formulaic character where one takes a frame and replaces only a variable or two, yet would we consider them 'formal' or 'high register'?

Second, it would seem difficult to use register to investigate individual and even group-level differences in language input: even if a broad selection of users of a language could agree which words are low register and which are high register, different children receive input in different distributions across register levels – the child of highly educated parents likely receives higher register input than the child of less educated parents. Would the investigator ask parents to estimate how often they hear words from different register levels by giving examples from each level? That seems unlikely to be more accurate than current approaches of estimating frequency directly from child-directed speech in CHILDES or other corpora.

Finally, as the authors note, frequency surely does play a role in language development. Indeed, most theoretical models of language development rely on frequency as a predictor of learning in some way. If we want to generate detailed predictions from such models, we must supply them with empirically-motivated estimates of frequency of different linguistic items in the child's input. Thus, even if we think register is a good proxy of frequency, we will eventually need to figure out how to precisely model a word's frequency based on its register. And given that a child will hear a particular linguistic item anywhere from a handful of times to thousands or millions of times, it would seem that we would need incredibly fine-grained estimates of register (again, finer-grained than 5 levels) to capture the variation in frequency.

Thus, while register may prove to be a useful construct to developmental linguists, it seems it will come with challenges of its own.

## Benjamin Tucker commentary on Ravid et al. thesis

The article by Ravid, Bar-On, Levie and Douani presents a method in which subjective register, assigned by expert raters, can be used to investigate the acquisition of register by language learners. Further, this study provides a useful baseline for researchers interested in investigating languages that do not currently have large corpora to generate data and hypotheses about register and they argue that this may constitute a potential proxy for standard lexical predictors like frequency.

I think that in order for these results to carry forward into the next decade and inform our understanding of representation and comprehension, we need to begin exploring these domains, much like the meta-megastudies paper, in other languages and exploring how to apply these techniques to other languages of the world. One of the outstanding questions for me in reading this paper is that I wonder how it might be applied to a wider range of languages without the historical depth available for a language like Hebrew, with a relatively new or no history of literacy and without a high degree of literacy in the language. I believe this is an important issue and touches on some of the issues discussed in the article by Myers. Extending these types of methods to endangered or under-documented languages of the world is immensely important to the understanding of how language works. Generally, expanding to a larger selection of languages has great potential in allowing researchers to investigate lexical representation in language. Many of the world's languages and language families have been ignored in the psycholinguistic literature likely due to the limitation involved in running those experiments. A method such as the one proposed by Ravid et al. has the potential to be extremely useful if it can be applied in a more general way.

The discussion of endangered or under-documented languages brings up some additional questions. The authors discuss register at the level of literate language, would they predict differences in their study if they were to focus on spoken registers? Would these be different registers? If so, would we as speakers then have a split lexicon, one for reading and writing and another for spoken language? Modifying this type of measure to something based on spoken language might also provide a path along which other researchers can make use of these types of measures for a much larger variety of languages.

## Chris Westbury commentary on Ravid et al. thesis

Psycholinguistics relies heavily on ratings of lexical measures such as frequency, concreteness, age of acquisition, valence, arousal, and many more. There is a tendency to assume in the field me that such measures are 'gold standards', very good estimates of true underlying values. However, this assumption does not stand up to scrutiny. Reliability correlations between independent estimates of human judgments almost never exceed 0.8 (suggesting that human judgments account for at most 64% of the variance in the very same judgments made by other humans) and are regularly as low as 0.6, suggesting that independent judgments account for much less than half of the variance in each other. The variance in individual word ratings often have 95% confidence intervals that cover a large portion of the rating scale.

None of this will be surprising to anyone familiar with the psychometric literature. Cronbach's alpha values (a measure of reliability of psychometric instruments that is a generalization of split half correlation) for most psychometric instruments usually hover in the same range, between 0.6 and 0.8, rarely higher. This is in no way a reflection on the competence of psychologists, but rather a mathematical inevitability arising from the causal structure of the biological world, in which many independent causes each account for a small amount of variance in the phenomena of interest. This causal structure necessarily confines the biological sciences to relying on statistical estimation rather than precise measurement.

The idea of a single 'true' value of a judged lexical measure is an illusion. The judged value will depend on the validity of the measure; the source of those measures; on the age, education, location, and linguistic experience of the population to which they are being applied; on the historical time period; and on many other factors. One of those factors is the method that is used to generate the estimates. This paper discusses the utility of using expert judges of register as a proxy for AofA, using the clever idea that register is easier for linguistically-savvy judges to agree on than AofA. Methods such as this, that take advantage of correlational structure to produce estimates, raise questions of ontological reality, i.e. questions about which factors are 'real' and which are in fact just proxies of something else. Such questions are especially salient in psycholinguistics, because many proposed lexical factors are strongly correlated with each other, with the result that any one factor might easily be well predicted by a linear or non-linear combination of other factors. For example, recent work has questioned the ontological status of such presumably basic measures as lexical frequency (Baayen, 2010) and concreteness (Westbury, Shaoul, Hollis, Smithson, Briesemeister, Hofmann and Jacobs, 2013; Westbury, Cribben, and Cummine, 2016). In both cases, doubts were raised on the basis of the fact that the measure's alleged effects could be eliminated with the introduction of other correlated measures that the authors preferred.

Does psycholinguistics need to grapple more directly with issues of ontological primacy by agreeing on explicit criteria that allow us to adjudicate in a systematic way between competing incompatible claims to such primacy? One possible answer is simply: *Yes, some constructs are real and some are not, and we need to definitively decide which are which*. A second possible answer is simply the opposite: *No, we do not need to worry about this issue, since all of the constructs are simply human conveniences*. Instead of fighting over which constructs are real and which are not, we should recognize that the reality is that we have a strongly correlated set of predictors, and how we demarcate that correlational structure is up to individual investigators. The middle ground between these two answers is that answer which is most clearly consistent with standard psychometric theory. In their foundational paper on this issue of construct validity, Cronbach & Meehl (1955) lay out the various kinds of construct validity and then try to map out "the logic of construct validation" (p. 290). They argue that constructs can only be defined in terms of a network of associations (their "nomological network") that links constructs to each other and to empirically-grounded observations, and that allows for prediction of at least some of those observed variables. Under Cronbach and Meehl's view, "the best construct is the one around which we can build the greatest number of inferences, in the most direct fashion" (p. 288). Cronbach and Meehl's pragmatic approach to construct validity should be highlighted more explicitly in the psycholinguistic literature, so that the field as a whole can come to explicit agreement about why we prefer some predictors to a linear or non-linear combination of their correlates.

*References*

Baayen, R. H. (2010). Demythologizing the word frequency effect: A discriminative learning perspective. *The Mental Lexicon*, 5(3), 436–461. https://doi.org/10.1075/ml.5.3.10baa

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, 52(4), 281. https://doi.org/10.1037/h0040957

Westbury, C. F., Shaoul, C., Hollis, G., Smithson, L., Briesemeister, B. B., Hofmann, M. J., & Jacobs, A. M. (2013). Now you see it, now you don't: On emotion, context, and the algorithmic prediction of human imageability judgments. *Frontiers in Psychology*, 4. https://doi.org/10.3389/fpsyg.2013.00991

Westbury, C. F., Cribben, I., & Cummine, J. (2016). Imaging imageability: Behavioral effects and neural correlates of its interaction with affect and context. *Frontiers in Human Neuroscience*, 10. https://doi.org/10.3389/fnhum.2016.00346

## The article

Most psycholinguists would agree that gaining information about lexical knowledge in learners is critical for the construction of theories of language acquisition (Douglas, 2003; Kennedy, 2014). *Frequency* has been a cornerstone of psycholinguistic research on the lexicon, as well as other domains of language learning, usage and change (Diessel, 2007; Gries & Divjak, 2012; Jurafsky, 2003). However, in seeking reliable information about lexical distributions of words across individuals' development, corpora frequency has been found wanting (Balota, Pilotti & Cortese, 2001; Brysbaert & Ghyselinck, 2006). We propose an alternative to objective frequency counts in corpora – subjective ranking by experts of lexical items' *register* in the sense of 'levels of linguistic usage', which has been independently linked to AoA (Ravid & Berman, 2009; Ravid & Vered, 2017).

### Adjectives

Adjectives are of particular interest in vocabulary development as the smallest (often absent) and most diverse lexical category in many languages (Dixon & Aikhenvald, 2014; Schachter & Shopen, 2007). As relational terms, adjectives show up later on in child speech than nouns and verbs (Casseli et al., 1995; Salerni et al., 2007), and they constitute a low-frequency class compared to other content words in children's early lexicons (Tribushinina et al., 2014; Sandhofer, Smith, &

Luo, 2000). A full array of adjectival categories is far from present even in 6-year-olds (Blackwell, 2005; Blodgett & Cooper, 1988), suggesting it coincides with the consolidation of a literate lexicon and its cognitive correlates (Dockrell & Messer, 2004; Ravid & Levie, 2010). The size and makeup of the adjective category can thus be taken as a yardstick for language development and proficiency. A first step towards this goal would be to amass a corpus of a language's adjectives and classify it according to a well-motivated criterion that would reflect developmental goals. Given the morphological richness of Hebrew, adjective morphology would be a first choice for such a developmentally-oriented classification.

*The morphology of Hebrew adjectives*

Modern Hebrew adjectives are particularly diverse morphologically. They make use of all three morphological structures in the language: Non-linear (root-and-pattern) forms, linear (stem-and-suffix) forms, and reduplicative forms.

*Non-linear adjective morphology*

Many adjective forms are based on the typically Semitic discontinuous root and pattern structure – both verbal or nominal (Ravid, 1990, 2003; Ravid & Schiff, 2006), e.g., *menumas* 'polite' (root *n-m-s*, pattern[1] *meCuCaC*). *Verbal patterns*. The participial (so-called *beynoni* 'middle') patterns of the seven *binyan* verb conjugations constitute one of the central word-formation mechanisms in Hebrew (Berman, 1978; Ravid, 2012). In addition to their inflectional role as present-tense temporal stems (Ravid et al., 2016), *beynoni* patterns productively derive adjectives or nouns (Berman, 1993). For example, *madhim* 'amazing' (root *d-h-m*, pattern *maCCiC*) serves not only as the present-tense stem in the temporal paradigm of the *hif'il* verb 'amaze', but also as an adjective in its own right.

Verb-related present-tense patterns fall into two distinct groups. One consists of the three resultative passive participle patterns *CaCuC* (*saduk* 'cracked', the resultative counterpart of the *qal* and *nif'al* conjugations), *meCuCaC* (*mekulkal* 'spoiled', the present-tense form of the *pu'al* conjugation), and *muCCaC* (*musmax* 'qualified', the present-tense form of the *huf'al* conjugation) (Berman, 1994; Ravid, 2004). A second group consists of the *beynoni* patterns of the five non-passive *binyan* conjugations – *CoCeC*[2] (*qal*, e.g., *bolet* 'prominent'), *niCCaC*[3] (*nif'al*, e.g., *ne'eman*

**1.** Patterns are represented as templates with C's denoting root radicals (Ravid, 1995). These can be surrounded by prefixes or suffixes.

**2.** As the semantically basic, most prevalent and most structurally diverse *binyan* conjugation, *Qal* has several present-tense allomorphs in addition to the default *CoCeC*, the most prominent of which is *CaCeC*, e.g., *ayef* 'tired'.

**3.** Although *nif'al* is considered a passive *binyan* (e.g., *nixtav* 'written'), much evidence points to its having a middle, non-agentive character such as *hitpa'el* (Berman, 1993; Ravid & Vered, 2017).

'loyal'), ma*CCiC* (*hif'il*, e.g., *maksim* 'charming'), *meCaCeC* (*pi'el*, e.g., *medake* 'depressing'), and *mitCaCeC* (*hitpa'el*, e.g., *mictayen* 'outstanding').

*Nominal patterns*

Hebrew Adjectives also productively share patterns with nouns. These are mainly *CaCCan* (e.g., *akshan* 'stubborn', cf. noun *safran* 'librarian'), and *CaCiC* (e.g., *ta'im* 'tasty', cf. *pakid* 'clerk'), as well as less productive nominal patterns such as *CaCaC* (adjective *yashar* 'honest', noun *naxash* 'snake'), the traditional category of handicaps *CiCeC* (e.g., *iver* 'blind'), and *CaCoC* (*gadol* 'big'), including color adjectives (e.g., *yarok* 'green').

*Linear adjective morphology*

Denominal adjective derivation (e.g., *kav-i* 'line-ar'), estimated to include over 1/3 of the adjectives in Hebrew (Ravid & Shlesinger, 1987), is part of Hebrew nominal word-level morphology expanding word stems by linear suffixation (Ravid, 2006). Modern denominal adjectives originate in a small Classical Hebrew class of *i*-suffixed nouns denoting ethnic origin, e.g., *yevusi* 'of Yevus'. These evolved in Medieval Hebrew and even more so in Modern Hebrew into the large, full-fledged class of adjectives derived by the suffixation of *-i* to a noun base (Berman, 1978).

*Reduplication*

Reduplication involves the repetition of part of the stem or its entirety, e.g., Agta *ulu / ululu* 'head/s' (Spencer, 1991). In Hebrew it works from left to right, e.g., adjective *shaxor / shxarxar* 'black / darkish'. Compared with the two widespread non-linear and linear systems that serve Hebrew inflection and derivation (Ravid, 2003, 2006), reduplication constitutes a third and rather minor structural device, generally restricted to diminutive expression in nouns and adjectives[4] (Hora et al, 2006).

*Adjectives in complex structures*

Adjectives moreover participate in two complex structures – taking quantifying prefixes of Aramaic origin, as in *tlat-xodshi* 'thrice-monthly',[5] and adjective-headed compounds, as in *kcar ruáx* 'short (of) spirit = impatient' (Ravid & Levie, 2010).

**4.** Note however that this is not the full story, as reduplication processes can apply to derived roots (e.g., *k-d-r-r* 'dribble' from *kadur* 'ball', or?-*v-r-r* 'bring air in' from *avir* 'air').

**5.** We excluded prefixed adjectives whose heads can be used as independent adjectives without the quantifying prefix, e.g., *cdadi* 'side'.

## Morphology as a developmental criterion

Studies have identified a connection between morphological adjective categories and the course of their acquisition in Hebrew. Accordingly, Ravid and Levie (2010) constructed a morpho-semantic adjective scale that took into account category semantics, structural complexity coupled with adjective distributions derived from naturalistic data and experimental elicitation. The lowest rank on this 1–4 scale consisted of a core list of mainly short, morphologically simplex Classical adjectives bearing primary adjectival meaning (*tov* 'good', *xam* 'hot'), prevalent in the speech of toddlers and their caregivers (Berman, 1985; Ravid & Nir, 2000; Tribushinina et al., 2014). Rank 2 on the scale constituted resultative passive adjectives such as *metukan* 'fixed', the earliest productive adjective category in preschool Hebrew (Berman, 1993; Ravid, 1995). Rank 3 consisted of adjectives with non-linear form, including non-passive *beynoni* participles (e.g., *macxik* 'funny'), nominal patterns (e.g., *shakran* 'liar'), and reduplicative adjectives (e.g., *CCaCCaC* (e.g., *shmanman* 'fatso'), occurring in the usage of schoolgoing children (Ravid, Levie & Avivi Ben Zvi, 2003). Rank 4 was dedicated to linear *i*-suffixed denominal adjectives such as *tipusi* 'typical', a later-developing, school-oriented category (Levin, Ravid & Rapaport, 2001; Ravid & Zilberbuch, 2003), requiring familiarity with a large, multifaceted nominal lexicon typical of literary prose, informative and expository texts (Ravid, 2004).

Ravid and Levie's (2010) analysis of adjectives in a corpus of spoken and written Hebrew texts generally confirmed the 1–4 scalar order. Across the school years there was a steady age-related increase in adjective ranking on the scale, in number of adjective rank levels, adjective category size and semantic abstractness and sophistication of specific adjectives, especially between adolescence and adulthood, and mostly in written expositions. Denominal adjectives (rank 4) were clearly found to associate with older age, literacy and advanced text production skills. However, this classification also proved to be too coarse to detect reduplicative adjectives, also missing complex adjectives, which were included in the syntactic rather than morphological analyses. Most important, ranks 1–3 contained adjectives that could be associated with different ages and language skills, e.g., everyday *shavur* 'broken' and lexically specific *natuy* 'slanted' in rank 2. Moreover, distinct groups with different semantic values shared the same patterns, e.g., *CaCiC* with a basic early adjective such as *ta'im* 'tasty', but also a productive class of potential adjectives such as *axil* 'edible', atypical of young children's language. Thus, Hebrew morphological categories cannot serve as the sole markers of adjectival AoA.

To rank Hebrew adjectives in a way that would reliably reflect language richness or proficiency, it was necessary to create an independent measure that would take into account individual adjectives and cluster them into meaningful groups, while relating this measure to adjective morphological categories so as to determine the viability of the previous classifications.

## Familiarity as index

Word frequency, associated with an individual's developing lexicon, would probably be the first choice of developmental researchers in classifying adjectives by AoA, as words that occur more frequently in the language also tend to be learned earlier on (Brysbaert & Cortese, 2011; Juhasz, 2005; Kuperman & van Dyke, 2013). Moreover, earlier lexical learning implies more exposure to the word in diverse communicative settings, and hence more familiarity with its specific shades of meaning, its morphological modulations, and its syntactic and pragmatic co-occurrence neighbors. One prominent way of gaining information about word frequency in the language is through *objective frequency* of occurrence in databases. However, objective frequency might not reflect the true distributions in the language because of a bias towards written over spoken sources (Balota et al., 2001) or towards a discourse genre such as sports, the news or the sciences (Biber & Conrad, 2009; Carroll, 1970). Important to the current analysis, objective frequency is most often obtained from adult corpora, and therefore cannot reflect the linguistic experience of children and adolescents (Brysbaert & Ghyselinck, 2006). While the CHILDES databases provide excellent information on frequencies in child-directed and child speech, most of them concern young children in conversation with caregivers (MacWhinney, 2000). Thus, obtaining information from objective frequencies about the long developmental course over later childhood and adolescence of a category of words such as adjectives is not really an option.

### *Hebrew adjectives in texts*

Attempts to construct a reliable adjective corpus in Hebrew have met with difficulties similar to those pointed above. For example, the Wintner Corpus (Itai & Wintner, 2008) is based on the 1990–1991 issues of the Israeli daily newspaper *Ha-Aretz*, and thus biased towards words frequent in the dated Hebrew lexicons of highly literate and sophisticated adult readers. Unsurprisingly, it does not contain adjectives such as *viral* or *autistic*, while *sovyeti* 'Soviet' occurs with high frequency (Benelli & Grinboim, 2015).

Obtaining an adjective list from Hebrew textual sources poses two additional, language-specific challenges. First, typological and historical factors have rendered the Hebrew adjective category extremely fuzzy. Classical Hebrew, lacking a dedicated morphological class of adjectives, used present tense verb participles and ethnic nominals to denote states, attributes and properties of nouns (Gai, 1995). As a result, current Hebrew adjectives not only share structural properties with both nouns and verbs, but in fact many adjectives are ambiguous across two or three content-word classes, defined only by syntactic factors in discourse, e.g., *metapes*, which might be interpreted as noun 'climber', present-tense verb 'climb/s', or adjective 'climbing'. Thus, a form such as *mag'il* might be the present-tense form in the verb paradigm where past and future tense would respectively mean 'caused / will cause disgust', or an adjective meaning 'disgusting'. Distinguishing adjectives from nouns and verbs makes the reliable identification of adjectives in corpora problematic. In fact, the Wintner Corpus contains about 1,000 words labeled as adjectives that do not appear in the Avneyon (2007) dictionary used in the current endeavor (Benelli & Grinboim, 2015). Such discrepancies would render a corpus-derived Hebrew adjective list rather unreliable.

A second Hebrew-specific problem is homography, which runs rampant in the default vowel-poor Hebrew orthography, so that a string such as לבנוני might be read as either *levanóni* 'Lebanese' or *lavnuni* 'whitish' (Bar-On & Ravid, 2011). The fact that several function words such as the definite article *ha* or the subordinating marker *she* are attached to the following written word increases homography tenfold (Ravid, 2012). Disambiguating homographic strings requires detailed manual analyses taking into account the semantic and morpho-syntactic environment of a large proportion of the words under consideration (Bar-On, Dattner & Ravid, submitted). Homography too contributes to the unreliability of objective frequency counts in written Hebrew corpora.

*Subjective judgments* of accrued experience in hearing, reading or using certain words constitute an alternative way of obtaining information about word frequency. True, subjective familiarity and frequency rating too may be contaminated by measures like the word's abstractness and orthographic similarity to other words (Thompson & Desrochers, 2009), inflectional and derivational entropies (Baayen, Feldman & Schreuder, 2006), as well as contextual factors in the elicititation, such as clarity of meaning (Balota et al., 2001). Nonetheless, taking into account participants' age, educational backgrounds, literacy skills, clinical conditions and other variables, several studies have shown a strong relation between subjective frequency ratings and objective frequency estimates from corpora (Auer, Burnstein & Tucker, 2000; Balota et al, 2004). In fact, as pointed out in Kuperman & Van Dyke (2013), subjective assessments of word frequency have been found to be superior over objective frequency gained from corpora.

In addition, unlike objective frequency, subjective frequency ratings can be elicited with language and literacy development in mind. The last years have seen the increasing use of language experts' subjective assessment of word familiarity as a viable alternative to objective frequency rating (Thompson & Desrochers, 2009).

Importantly, the ability to make valid assessments is related to the linguistic and reading experience of raters (Auer, Burnstein & Tucker, 2000; Kuperman & Van Dyke, 2013). Thus, more proficient language users make better judgment regarding word frequency, as they are familiar with more, and rarer words (Alderson, 2007). In the current study, we adopted linguistic *register* as a criterion for the subjective ranking of adjectives by language experts.

## Register

Ferguson (1994: 16) defines register as "the linguistic differences that correlate with different occasions of use". This definition was extended to 'levels of linguistic usage' (Ravid & Berman, 2009) as developmentally, gaining command of linguistic register involves learning the range of expressive options available in the target language, and being able to map relevant linguistic forms in accordance with communicative context. Several studies have found a relationship between register sensitivity and language development (Andersen, 1992; Clark, 2003). Ravid & Berman (2009) demonstrated that register can provide a yardstick for evaluating language development across the school years in Modern Israeli Hebrew, a language that shows marked distinctions in levels of usage between the everyday colloquial style and more formal types of expression due to its multilayered history (Shlesinger, 2000).[6] They showed that high register, semantically sophisticated, and often morphologically complex lexical items occurred with greater frequency in texts of older speaker/writers, especially in written expositions. Recently, Ravid & Vered (2017) showed in a structured elicitation task across adolescence that correct performance on neutral-register passive Hebrew verbs was significantly earlier than high-register passive verbs.

These findings have led us to adopt register as our independent measure as a proxy for frequency and familiarity assessment, hypothesizing that lower and neutral register adjectives should be more prevalent and thus learned earlier, while high register adjectives should be more typical of the lexicons of older, highly literate speaker/writers. We also assumed that we would find a relationship between the morphological structures of adjectives and their register ranking.

---

**6.** For example, 'tree' is denoted by both neutral register *ec*, deriving from Biblical Hebrew, and high-register *ilan*, from Mishnaic Hebrew.

## Method

The methodology of the current study consisted of two major steps: (i) compiling the adjective list, and (ii) eliciting the experts' register ranking judgements.

### *Step 1: Compiling the adjective list*

All words marked as adjectives in Avneyon (2007) were identified, with a few newer adjectives (e.g., *viráli* 'viral') added from the media. Literary ad-hoc adjectives considered unfamiliar even to highly literate Hebrew users (e.g., *tafush* '(probably) stupid'), highly specific vocational loan adjectives (e.g., *osmofíli* 'osmophilic'), and highly ambiguous class items were excluded from the list. The master list of adjectives, our study corpus, contained 3,747 Hebrew adjectives. While we cannot claim that it contains all of the current adjectives in Hebrew, all items included in it are unambiguously Hebrew adjectives. The analyses below thus treat adjective lemma types (rather than wordform tokens occurring in discourse corpora), and for the current purposes they disregard inflectional adjective morphology, a topic that has gained its own set of developmental studies (Ravid & Schiff, 2012).

### *Morphological structures*

Table 1 presents the morphological make-up of the adjectives in the master list, with examples. The lists consists of 19 different structures serving adjectives in nine major categories: (1) verb-based resultative passives; (2) verb-based non-passives; (3) nominal-based patterns; (4) nominal non-productive patterns; (5) denominal *-i* suffixed adjectives; (6) diminutive reduplicative adjectives; (7) complex and compound adjectives; (8) singletons; and (9) miscellaneous slang terms.

### *Step 2: Eliciting register judgments from language experts*

Two pilot studies preceded the final elicitation of register judgment, designed to determine the expert judge population, the minimal number of expert judges necessary for inter-judge agreement, and questionnaire length. The pilots showed that native-speaking, well-educated Hebrew adults of different academic backgrounds could not reach adequate agreement on adjective register ranking. In contrast, native Hebrew speaking speech-language clinicians with linguistic and psycholinguistic training were familiar with the notion of *language register*, and their responses showed much less variability, being more concentrated around a specific range of register ranks for most adjectives. The population selected as expert judges of language register rank was thus that of speech-language clinicians. The pilot studies also determined that the minimal number of judges per adjective was seven, and that 100 adjectives per questionnaire was an appropriate number for a single person to judge.

**Table 1.** Morphological make-up of the adjective list

| Category | Sub-category | Morphological structure | Example |
|---|---|---|---|
| Denominal | | *i*-suffixed | *ma'asi* 'practical' [*ma'ase* 'action' + *i*] |
| Verbal patterns Present-tense *beynoni* | Resultative passives | *CaCuC* | *sagur* 'closed' |
| | | *meCuCaC* | *mesudar* 'neat' |
| | | *muCCaC* | *muklat* 'recorded' |
| | Non-passive participles | *CoCeC* | *bolet* 'prominent' |
| | | *CaCeC* | *ayef* 'tired' |
| | | *niCCaC* | *nifrad* 'separate' |
| | | *maCCiC* | *mafxid* 'frightening' |
| | | *meCaCeC* | *meza'azéa* 'shocking' |
| | | *mitCaCeC* | *mictaber* 'compiled' |
| Nominal patterns | | *CaCCan* | *aclan* 'lazy' |
| | | *CaCiC* | *ta'im* 'tasty' |
| Non-productive | Handicapped | *CiCeC* | *iver* 'blind' |
| | | *CaCaC* | *xaxam* 'wise' |
| | Incl. colors | *CaCoC* | *raxok* 'distant' |
| Diminutive | Reduplicative | *CCaCCaC* | *kxalxal* 'blueish' |
| Complex | | | *du-parcufi* 'two-faced' |
| Singleton | | | *meheyman* 'credible' |
| Miscellaneous slang | | | *ahbal* 'moron' |

*Participants and procedure*

A set of 38 different questionnaires (37 containing 100 adjectives each, and one containing 50 adjectives) was compiled out of the randomized master list of 3,747 adjectives. They were disseminated electronically to 329 expert judges by email and via Facebook for volitional register judgment. Judges were instructed to intuitively classify the list of 100 adjectives into 5 register ranks, following examples from Rank 1 (basic register, e.g., *adom* 'red'), Rank 2 *barur* 'clear', Rank 3 *mevix* 'embarrassing', Rank 4 *meheyman* 'credible' and Rank 5 (very high register, *virtuózi* 'virtuose').

*Data analysis*

The 3,747 adjectives in 38 questionnaires were ranked by register rank on a scale of 1 to 5 by 329 language experts. Of these, 19 questionnaires were each ranked by seven expert judges, and 19 by more than seven experts, altogether yielding 32,329 rankings. Given the variance among experts' ranking of each adjective, among the rankings of each expert judge, and among all rankings, a cross-classification multi-level model was constructed to analyze the variance deriving from measurements in the different groups and levels (Level 1 – a single ranking of an adjective by an expert judge, Level 2 – expert judge, Level 2 – adjective). The mean of each adjective's ranking was calculated, as well as the mean for each expert judge's ranking. At Level 2, the mean distance of a single expert judge's ranking from his/her peers was calculated as a measure of expert judge agreement with peers. And at Level 2, the mean and variance values were calculated for each adjective across expert judges.

**Table 2.** Intra-class correlations (ICC) for adjective categories, register ranking and variance

| **Morphological structure** | **Level 2 Adjectives** | **Level 2 Expert judges** | **# Adjectives in category** | **# Expert judges** | **# Rankings** | **Mean register ranking** | **Variance in ranking** |
|---|---|---|---|---|---|---|---|
| 1. *i*-suffixed | .49 | .16 | 1,436 | 329 | 12,401 | 3.57 | 0.62 |
| 2. *CaCuC* | .54 | .13 | 416 | 329 | 3,649 | 3.06 | 0.61 |
| 3. *meCuCaC* | .48 | .16 | 603 | 329 | 5,176 | 3.09 | 0.57 |
| 4. *muCCaC* | .40 | .23 | 178 | 321 | 1,582 | 3.02 | 0.61 |
| 5. *CoCeC* | .48 | .17 | 94 | 298 | 805 | 2.65 | 0.62 |
| 6. *CaCeC* | .64 | .14 | 56 | 243 | 446 | 2.48 | 0.60 |
| 7. *niCCaC* | .56 | .09 | 68 | 278 | 572 | 3.17 | 0.51 |
| 8. *maCCiC* | .48 | .18 | 93 | 304 | 786 | 2.70 | 0.59 |
| 9. *meCaCeC* | .45 | .16 | 81 | 300 | 696 | 2.72 | 0.57 |
| 10. *mitCaCeC* | .32 | .20 | 23 | 139 | 189 | 2.66 | 0.74 |
| 11. *CaCCan* | .41 | .20 | 102 | 288 | 890 | 2.70 | 0.64 |
| 12. *CaCiC* | .56 | .14 | 121 | 322 | 1,103 | 2.64 | 0.54 |
| 13. *CiCeC* | .64 | .14 | 12 | 73 | 94 | 2.86 | 0.62 |
| 14. *CaCaC* | .69 | .07 | 24 | 156 | 239 | 1.97 | 0.45 |
| 15. *CaCoC* | .68 | .08 | 43 | 202 | 356 | 2.21 | 0.51 |
| 16. *CCaCCaC* | .46 | .04 | 28 | 168 | 233 | 2.76 | 0.65 |
| 17. Complex | .30 | .28 | 278 | 329 | 2,447 | 3.59 | 0.68 |
| 18. Singleton | .72 | .08 | 55 | 253 | 491 | 2.41 | 0.50 |
| 19. Miscel. | .38 | .33 | 36 | 219 | 325 | 2.36 | 1.05 |
| **Total** | **.42** | **.16** | **3,747** | **329** | **32,529** | **3.20** | **0.61** |

### *ICC analyses*

Table 2 presents intra-class correlations (ICC), central tendency and dispersion values for experts and for adjective categories, so as to determine the multi-level structure of the rankings (over 0.05). The rightmost two columns present the two values central to this analysis for each of the structural categories: the mean register rank and the amount of variance among expert judges, so that the higher the variance, the less agreement among expert judges regarding register ranking. Following these analyses, 10 expert judges with consistently outlier rankings (two-sided confidence interval at a 95% level of confidence) were removed from the following analyses so as to achieve a classification of the adjectives into meaningful groups with low variance (= high agreement) among expert judges.

**Table 3.** Comparing clustering in an ascending number of groups

| | 4 groups | 5 groups | 6 groups | 7 groups | 8 groups | 9 groups | 10 groups |
|---|---|---|---|---|---|---|---|
| AIC | 12601 | 12342 | 12168 | 12070 | 11982 | 11899 | 11829 |
| BIC | 12682 | 12442 | 12286 | 12207 | 12138 | 12073 | 12022 |
| Loglikelihood | −6287 | −6155 | −6065 | −6013 | −5966 | −5921 | −5883 |
| Entropy | .73 | .69 | .73 | .74 | .75 | .75 | .78 |
| # of adjectives | | | | | | | |
| 1 | 1639 | 1779 | 1174 | 1046 | 1244 | 325 | 258 |
| 2 | 62 | 630 | 505 | 406 | 134 | 215 | 919 |
| 3 | 798 | 832 | 223 | 121 | 225 | 581 | 507 |
| 4 | 1248 | 452 | 1385 | 1246 | 340 | 924 | 256 |
| 5 | | 54 | 404 | 214 | 396 | 389 | 215 |
| 6 | | | 56 | 683 | 31 | 805 | 687 |
| 7 | | | | 31 | 320 | 31 | 341 |
| 8 | | | | | 1057 | 358 | 502 |
| 9 | | | | | | 119 | 47 |
| 10 | | | | | | | 15 |
| Vuong-lo-Mendell | < .001 | .003 | < .001 | .03 | < .001 | .01 | .23 |
| Lo-Mendell-Rubin | < .001 | .003 | < .001 | .03 | < .001 | .01 | .24 |
| Test for n-1 versus n clusters | < .001 | < .001 | < .001 | < .001 | < .001 | < .001 | < .001 |

**Table 4.** The mean probability of belonging to a group (in percentages) from 4 to 9 groups

| | 4 groups | 5 groups | 6 groups | 7 groups | 8 groups | 9 groups |
|---|---|---|---|---|---|---|
| 1 | .87 | .74 | .80 | .81 | .79 | .77 |
| 2 | .94 | .80 | .86 | .81 | .83 | .83 |
| 3 | .82 | .80 | .84 | .85 | .88 | .79 |
| 4 | .83 | .87 | .77 | .78 | .73 | .74 |
| 5 | | .95 | .83 | .81 | .85 | .80 |
| 6 | | | .95 | .81 | .97 | .77 |
| 7 | | | | .96 | .71 | .95 |
| 8 | | | | | .82 | .77 |
| 9 | | | | | | .85 |

### *Latent class analyses*

The second step of analysis was a Model Based Latent Class Analysis (LCA) that enables the identification of unobservable subgroups that are similar, based on observed characteristics – in the current case, mean register ranks and variance in ranking for each of the adjectives. The purpose of this analysis was to find the optimal number of register / variance groups in the data, unrelated to their morphological classification. The methodology developed to determine the optimal number of groups was based on a multiple-test criterion for which the objective is to optimize classification goodness of fit subject to highest matching probability. This was done to ensure higher group-division of data having insignificant contribution to the goodness-of-fit (Lo-Mendell-Rubin test; Lo, Mendell and Rubin 2001), whereas items have higher probability to belong in the assigned group rather than in another (Nylund, Asparouchov & Muthen, 2007). Table 3 shows that according to the goodness-of-fit criterion, the 10th group does not contribute additional explanation of the variance beyond nine groups ($p$ = .24). This is also indicated by the AIC and BIC values (lower values indicate a better quality of the model) and the entropy value (the closer to 1, the better quality of group classification). Table 4 shows that the probability of an adjective belonging to a group is not lower than 80% up to 7 groups, decreasing beyond that. Taken together, a structure of seven latent groups was selected for our analysis.

## Results

### *Quantitative analyses*

Table 5 and Figure 1 show the grouping of all 3,747 adjectives on the master list into the seven groups emerging from the analysis described above. They depict the sizes of the adjective groups and their differing register ranks, ranging between a low 1.44 to a high 4.51. Most important, they indicate that five of these groups – 1, 2, 4, 5, and 6 – had low expert judge variance values ranging between 0.21 to

**Table 5.** Number of adjectives, percentages out of the total number of adjectives ($N$ = 3,747), mean register levels and variance in seven latent groups

| | Group 1 | Group 2 | Group 3 | Group 4 | Group 5 | Group 6 | Group 7 |
|---|---|---|---|---|---|---|---|
| Mean register rank | 2.49 | 4.51 | 3.18 | 3.71 | 1.44 | 3.12 | 3.08 |
| Variance | 0.49 | 0.22 | 1.58 | 0.52 | 0.21 | 1.03 | 2.46 |
| N adjectives | 1046 | 406 | 121 | 1246 | 214 | 683 | 31 |
| Percentage | 28% | 11% | 3% | 33% | 6% | 18% | 1% |

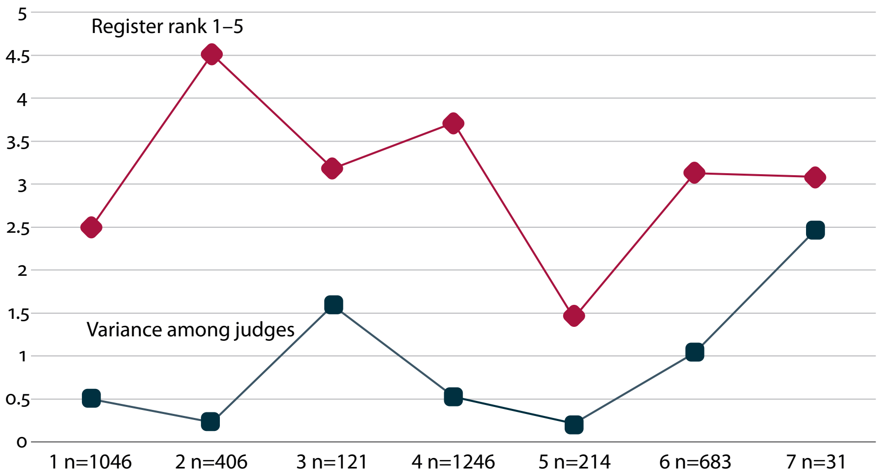



**Figure 1.** Seven latent group of register rank and expert judges' agreement

1.03, where high variance across judges means lower agreement, that is, judges' evaluations is sparse around the mean value. These five groups showed middle to high agreement among expert judges. As adjectives come in groups, we decided to exclude small group which also showed low agreement as expressed in the sparsity of the judges' evaluations

A final master list of 3,595 adjectives divided into five groups with high expert judge agreement and different register rankings constituted the focus of our further analyses. Given their register ranking and the lexical semantics of their adjective members, these five groups were taken to designate five developmentally consecutive "adjective lexicons". Table 6 and Figure 2 show the number and proportions of each lexicon out of the total of adjectives, with examples.

**Table 6.** The five adjective lexicons of Hebrew

| Lexicon | percentage | Register | Examples |
|---|---|---|---|
| Core | 6% | Basic | *xam* 'hot', *xadash* 'new', *rek* 'empty' |
| Basic | 29% | Neutral | *ragzan* 'cranky', *zahir* 'careful', *ani* 'poor' |
| Mature | 19% | Elevated | *megusham* 'awkward', *biológi* 'biological', *atum* 'opaque' |
| Educated | 35% | High | *xamakmak* 'elusive', *dagul* 'eminent', topográfi 'topographic' |
| Super-literate | 11% | Very high | *kamai* 'primeval', *menuce* 'feathered', *taktíli* 'tactile' |

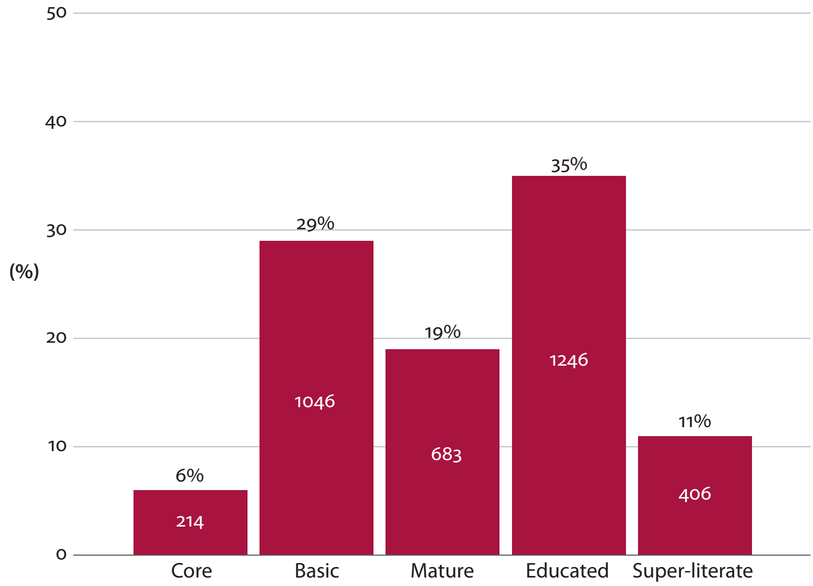



**Figure 2.** The five adjective lexicons of Hebrew, by linguistic register (*N* = 3,595)

## The Five adjective lexicons: Semantic and morphological analyses

Below we present semantic and morphological analyses of the adjectives in each of the five lexicons. We based our semantic analyses on Blackwell's (2005) and Tribushinina et al's (2014) semantic coding of adjectives in CDS and CS corpora. However, as we show below, this classification mostly applied to the lower register lexicons, and did not cover the most prevalent abstract adjective classes. This is where morphological categorization made a contribution towards distinguishing the lexicons. Figures 3–7 depict the morphological makeup of the five adjective lexicons.

### *The Core adjective lexicon: Lowest register (*M *= 1.44)*

This was the smallest lexicon, with 214 adjectives constituting 6% of the total of adjectives in the final master list, and the lowest mean register rank of 1.44. It contained what are universally regarded as the most basic adjectives in toddlers' and young children's perception of objects and people (Pitchford & Mullen 2001), prevalent in their speech and in caregivers' child addressed speech (Blackwell, 2005; Tribushinina et al., 2014). Semantically, this list included everyday, mostly externally perceived, often concrete semantic categories of adjectives, each containing

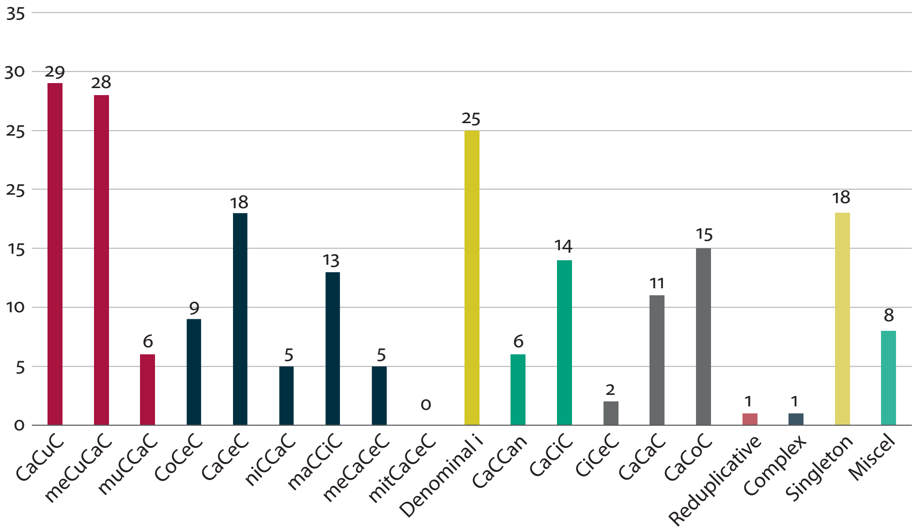



**Figure 3.** The Core adjective lexicon, by morphological category (*N* = 214)

a small number of the most prominent adjectives in the category. Following Tribushinina et al.'s (2014) classification, these categories designated age (*xadash* 'new'), behavioral properties (*macxik* 'funny'), color terms (*adom* 'red'), conformity (*ragil* 'usual'), consistency (*rax* 'soft'), modals (*carix* 'necessary'), ordinal numbers (*rishon* 'first'(, physical properties (*shaket* 'quiet'), physical states (*met* 'dead'), quantitative values (*rek* 'empty'), taste (*matok* 'sweet'), shape (*agol* 'round'), spatial values (*gavóha* 'tall'), and temperature (*xam* 'hot'). In addition, there were adjectives denoting basic internal states, sensations and emotions (*acuv* 'sad') and basic evaluative adjectives (*xamud* 'cute').

Morphologically, Figure 3 shows that the Core Adjective Lexicon contained the widest array of different structures. Most prominent (about ¼ of all adjectives in this lexicon) were the resultative *CaCuC* (*xamuc* 'bitter') and *meCuCaC* (*meluxlax* 'dirty), followed by about 10% *-i* suffixed adjectives, mostly simplex (*naki* 'clean'). Almost all non-passive *beynoni* participles were also represented in this lexicon, most prominently the *qal* allomorph *CaCeC* (*shamen* 'fat') and the *hif'il beynoni maCCiC* (*mafxid* 'scary'), in addition to *CaCoC* colors (*shaxor* 'black'), basic *CaCiC* adjectives (*ca'ir* 'young'), and less productive nominal patterns, e.g., *CaCaC* (*kacar* 'short'). While there were relatively numerous singletons (*kal* 'light'), this lexicon did not contain complex adjectives.

*The Basic adjective lexicon: Neutral register (*M = *2.49)*
This lexicon, scoring a mean of 2.49 out of 5 on register, was the second largest on the master list, containing 1046 (29%) of the adjectives. This was the adjective lexicon necessary to denote main and inherent properties of nouns, taken to represent the broad adjective lexicon shared by all Hebrew speakers with elementary education. There were far fewer concrete and externally perceived adjectives in this lexicon, relating to less prominent physical facets of objects (*zakuf* 'upright'), of the environment (*me'unan* 'cloudy', *mehavhev* 'blinking'), and of people (*mezukan* 'bearded'). Many of these external adjectives were lexically specific (*mezoham* 'filthy' rather than dirty, *zehe* 'identical' vs. *shone* 'different'). The majority of the Basic adjectives were abstract, expressing internal human attributes and states (*yediduti* 'friendly', *bayshan* 'shy'), advanced modals (*mesugal* 'able') and evaluations (*doxe* 'repulsive', *shaguy* 'mistaken', *yahir* 'arrogant'). Many adjectives expressed facets of objects and events that can only be captured by thought (*pit'omi* 'sudden', *meyaceg* 'representative', *mezuyaf* 'fake'), or require understanding of scientific principles (*nozli* 'liquid', *dalik* 'flammable', *sidrati* 'serial')

This lexical growth was accompanied by the concurrent changes in the morphological makeup of the Basic versus the Core Adjective lexicons (Figure 4). Whereas the Core Lexicon had 54% non-passive participles, nominal patterns and singletons, the Basic Lexicon had half that many: Non-passive *beynoni* participles (*ashem* 'guilty', *me'axzev* 'disappointing') took up about 16% of this lexicon, and nominal patterns denoting human qualities (*xashdan* 'suspicious', *zahir* 'careful') contributed another 10%, with about 1% singletons.

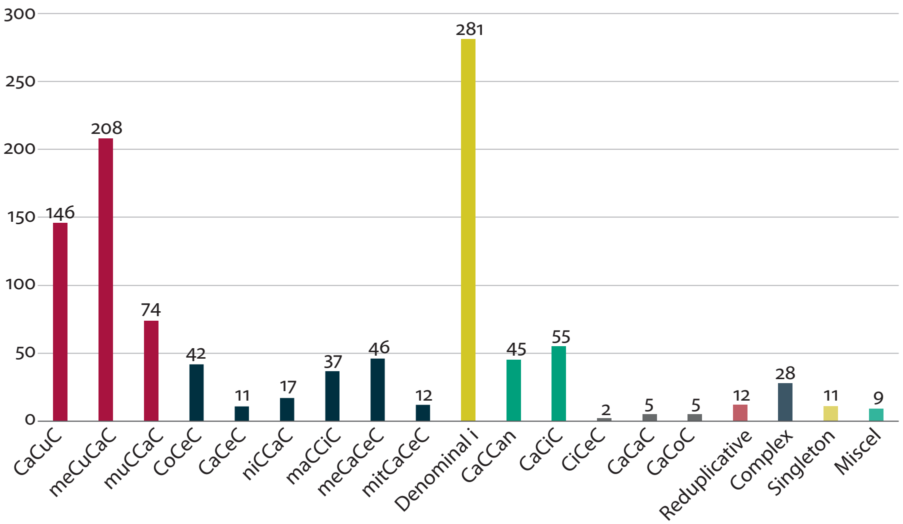



**Figure 4.** The Basic adjective lexicon, by morphological category ($N$ = 1,046)

The Basic Lexicon was characterized by the two prominent morphological structures of Modern Hebrew. There was a big chunk (40%) of adjectives in verb-derived resultative passive *CaCuC, meCuCaC* and *MuCCaC*, expressing the *states* of objects, people, events and concepts (*katuv* 'written', *memushax* 'lasting', *mufla* 'magical'); and 25% denominal adjectives denoting the *attributes* of objects, people, events and concepts (*arci* 'earthy', *hitpatxuti* 'developmental', *mikco'i* 'vocational').

*The Mature adjective lexicon: Elevated register (*M = *3.12)*

The Mature Adjective Lexicon, with a mean register rank of 3.12 out of 5, containing 683 (19%) adjectives, was a vocabulary detailed enough to express qualities and properties of humans, objects, places and concepts in the modern world, taken to serve literate Hebrew speakers with a highschool education.

Semantically, this lexicon had hardly any of the physical and external attributes found in the two lower register lexicons. It had a small number of adjectives that required familiarity with special properties of materials and objects, such as *mushxal* 'threaded' and *kalúa* 'woven'. Otherwise, this lexicon was all about abstract human states and attributes (*gluy lev* 'frank', *meyusar* 'tormented'), including qualities relevant to the adult world (*shtuf zima* 'salacious', *pasivi* 'passive'). Numerous adjectives testified to a systematic, school-based understanding of objects, systems, concepts and events (*xadish* 'modern', *ezraxi* 'civilian', *xofef* 'concurrent'). Again, this semantic shift was reflected in the morphological makeup of the Mature Lexicon (Figure 5). Close to 40% of the Mature adjectives were *i*-suffixed denominals (*xatuli* 'feline', *vatrani* 'lenient', *mishpati* 'legal') – among them, for the first time, foreign-based

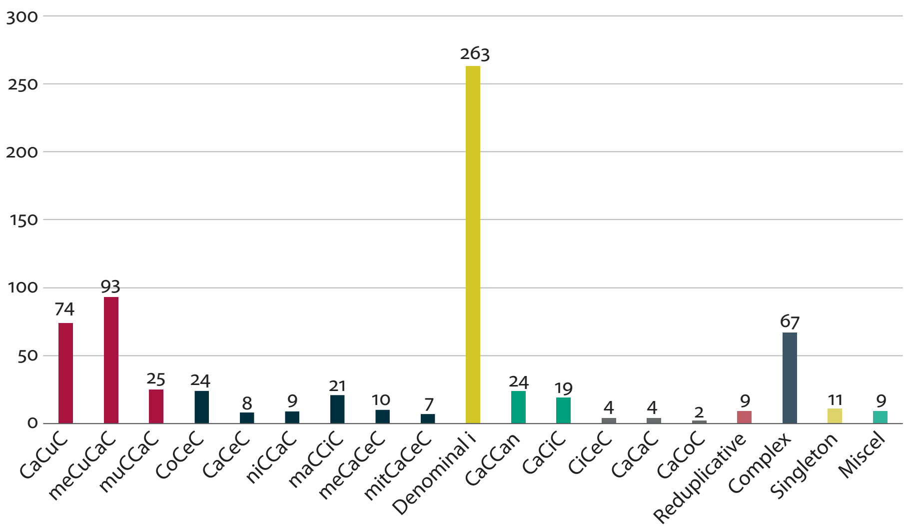



**Figure 5.** The Mature adjective lexicon, by morphological category (*N* = 683)

adjectives (*sensacióni* 'sensational', *globáli* 'global'); and another big chunk (28%) were passive resultatives. About 10% of this lexicon consisted of complex and compound adjectives (*du-mini 'bi-sexual'*, *ktan koma* 'small (of) stature'). Under ¼ of the adjectives here belonged to non-passive participles and nominal patterns.

*The Educated adjective lexicon: High register (*M *= 3.71)*
This was the largest adjective lexicon, with a mean register rank of 3.71 out of 5, consisting of 1246 adjectives (35% of the final master list), a lexicon of lexically specific, morphologically complex adjectives serving educated Hebrew users. To gain command of the Educated Adjective Lexicon, one must have a rich nominal lexicon and well-developed derivational morphology abilities. Morphologically, its makeup resembled that of the Mature Lexicon, underscoring the composition of literate adjectives in Hebrew. The largest component was 42% denominal adjectives, many of them foreign-based (*vulkáni* 'volcanic', *ambiciózi* 'ambitious'), and native Hebrew (*trashi* 'lapidarian'), requiring familiarity with both the sciences and Classical Hebrew. A third of the Educated adjectives were passive resultatives (*mesuvag* 'classified', *gadúa* 'severed'), and about 11% were complex adjectives (*nesu panim* 'distinguished', *kcar ro'i* 'short sighted').

*The Super-Literate adjective lexicon: Very high register (*M *= 4.51)*
This second-smallest lexicon (406 adjectives, 11% of the total) with the highest mean register rank (4.51 out of 5) was taken to be the experts' adjective vocabulary. The Super-Literate Adjective Lexicon contained rare adjectives with high

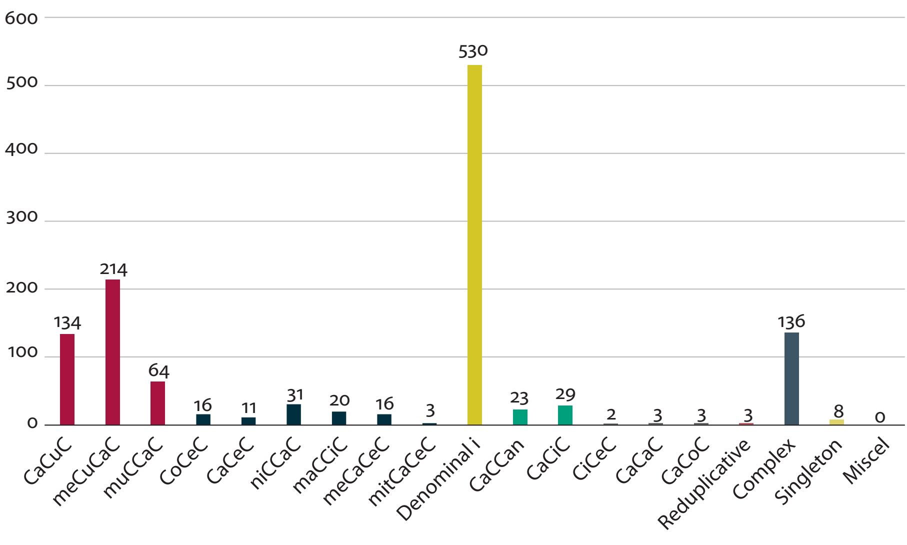



**Figure 6.** The Educated adjective lexicon, by morphological category (*N* = 1,246)

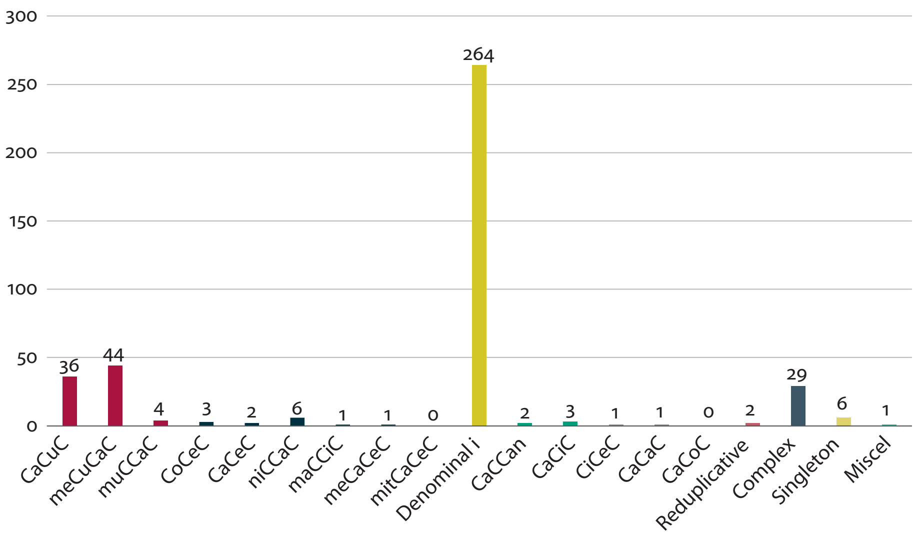



**Figure 7.** The Super-literate adjective lexicon, by morphological category ($N = 406$)

lexical specificity, serving highly educated Hebrew users with professional expertise. It required broad and deep conceptual knowledge, acquaintance with words from Classical Hebrew, and terminology from foreign languages. Paradoxically, this Lexicon did not require much knowledge of morphology, as over 2/3 of it was composed of denominal adjectives, overwhelmingly foreign-derived (*kompulsívi* 'compulsive', *puritáni* 'puritan', and Hebrew *nici* 'hawkish'). About 20% were rare resultative passive adjectives such as *merubad* 'stratified' or *kahuy* 'torpid').

## Discussion

The current study set out to propose subjective ranking by register as an independent measure of language knowledge across development. Carried out by language experts, with mostly high agreement among raters, a linguistic register scale from 1–5 by successfully distinguished five groups of over 3,500 Hebrew adjectives, which were respectively labeled the Core, Basic, Mature, Educated and Super-Literate Hebrew adjective lexicons.

The first major finding was to what extent language experts' judgments were in line with known attributes of language in acquisition. The judges were not asked about the estimated frequency of the adjectives on the list, nor about the assessed familiarity of Hebrew speakers of different age levels with them. Rather, the judges expressed their own perception about the contexts of usage of the ranked adjectives, implying their developmental characteristics (Ravid & Berman, 2009). Thus, most adjectives in the two lower register lexicons (*xazak* 'strong', *mo'il* 'helpful') were basic, inherent, and concrete, pertaining to a broad array of communicative contexts and intelocutors.[7] Increasing register rank in the next two lexicons indicated a growth in adjective abstractness and specificity (*karuy* 'mined', *xaluci* 'pioneering'), more restricted, literate, communicative settings, and knowledge of academic arenas, scientific topics, literature, history, and the arts. The lexicon with the highest register rank was mostly loan-based, requiring a high degree of linguistic and academic sophistication (*míti* 'mythical', *epidémi* 'epidemical'). Thus, while this study did not directly test the development of adjectives in Hebrew, it was developmentally-oriented, motivated and anchored in previous research, and yielded adjective lexicons that can be associated with the kind of Hebrew typically used at different age and cognitive levels. The measure used in achieving this classification fulfilled the requirements noted in the introduction by being independent of frequency and familiarity ratings.

*Semantics*

A second interesting finding related to the semantic changes characterizing the adjectives in the five consecutive lexicons, which revealed the need to go beyond early childhood in conducting AoA lexical studies. In the current study, the adjective categorization in Tribushinina et al (2014) and Blackwell (2005) meaningfully classified the adjectives in the two lower register lexicons, but was not sufficient for capturing the qualitative distinctions among adjectives in the higher-rank lexicons. These advanced adjectives were either more lexically specific alternatives within the same categories (comparable, perhaps, to English *catlike* and *feline*), or else they fitted in the Tribushinina et al category labeled "Other" (exemplified by *electric*). Thus, a more sophisticated classification, taking into account the cognitive, literacy and linguistic gains of adolescence is necessary to account for the changes in adjective semantics.

The lexical semantics of adjectives in the three higher register lexicons reflects a profound cognitive shift during the school years. The higher-register adjectives express scientific knowledge (*amórfi* 'amorphic'), sophisticated understanding of human relationships (*mefuyas* 'placated'), and the explosion of abstract nominals, encoding processes, activities, events, properties, states, concepts and ideas in literate expression, accompanied by adjectives denoting multiple facets of these abstract nominals (*dikduki* 'grammatical') (Nippold & Duthie, 2003; Ravid & Levie, 2010). These findings too underscore the need to further investigate language development beyond the early years (Berman, 2004; Nippold, 2016).

7. A small-scale pilot at Tel Aviv University (Vichner, 2016) already indicates that the adjectives found in 30 storybooks targeted at children aged 1–8 years belong in the two lower register lexicons (Core and Basic).

### *Morphology*

A third, typologically important finding was the concurrent change in the distribution of Hebrew morphological categories side by side with the changes from general and concrete to specific and abstract semantics. The lower register lexicons, and in particular the Core lexicon, contained the fullest array of structural categories, including *beynoni* present participles in all *binyan* conjugations and all nominal patterns, including the least productive. The higher the register rank, the fewer the number of morphological categories in the lexicons – the two highest register lexicons contained mainly two, and then one structural device. We believe these differences reflect the interface of lexicon and morphology in the acquisition of the Hebrew adjective class, with the U-shape well known to researchers of language development. The small size of the Core lexicon and the small number of adjectives in each morphological category indicate that they are mostly learned as non-analyzed amalgams denoting primary and salient noun attributes in early childhood. The proliferation of adjectives in the Basic lexicon (close to 1/3 of all adjectives on the master list) indicates the emergence of adjectival morphology in preschool and the early school years (Berman, 1985). In particular, this change points to the intensive learning of forms and meanings of resultative passive adjectives together with verb morphology (Berman, 1993, 1994; Ravid et al., 2016), and the inception of the class of denominal adjectives (Ravid & Nir, 2000). The decline in less productive adjectival forms and the increase in the highly productive resultatives and denominals in the Mature and Educated lexicons reflect the consolidation and automatization of Hebrew adjectival morphology as an accessible and flexible device during the school years (Ravid, 2004; Ravid & Zilberbuch, 2003). At the second end of the developmental U-shape, where mostly foreign-based denominal adjectives are learned as genre-related concepts and terms, morphology again plays a lesser role.

The analysis of the morphological makeup of the five adjective lexicons provides new information about the full developmental paths of adjective morphology. Modern Hebrew has come into its own in using two morphological devices for the expression of adjectival notions – *i*-suffixed denominals in the expression of nominal *attributes*, and verb-derived passive resultatives in the expression of nominal *states*.

The category of *i*-suffixed denominals presents clear linear development, increasing exponentially from 10% in the Core lexicon to two-thirds of the highest register lexicon. The findings of the current study thus support the centrality of denominal adjectives, the largest class in the three literate lexicons, as the main Hebrew mechanism conveying noun attributes, with specific syntactic affinity to abstract nominals (Ravid & Berman, 2010; Ravid & Levie, 2010; Ravid & Zilberbuch, 2003). The category undergoes its own evolution across the five lexicons from non-derived denominals in the Core lexicon (*acbani* 'nervous'), to Hebrew-based stems in the literate lexicons (*hatxalati* 'initial'), and finally to foreign stems in the Super Literate lexicon (*spiráli* 'spiral').

Resultative passive adjectives occupy at least 20% of all five lexicons, with a non-linear, protracted acquisition path. Previous studies pointed to the early childhood learning of resultative morphology (Berman, 1993, 1994), but the current study shows this category is in fact learned in two main waves. Resultative adjectives occupy ¼ of the core lexicon, increasing to 40% in the Basic lexicon as its largest category. This reflects the first wave of learning the present-tense forms of three different passive *binyan* conjugations to express the result of actions (e.g., *harus* 'destroyed' entailed by verb *haras* 'destroy', *mekupal* 'folded' entailed by *kipel* 'fold'). This proportion drops to 28% in the Mature lexicon, rising again to 33% in the Educated lexicon, reflecting a second wave of acquisition, with rarer and more passive-like resultatives coming in (*mushtal* 'implanted', *mekulas* 'eulogized') (Ravid & Vered, in press). The smallest proportion of resultatives (20%) is found in the Super-literate lexicon. The internal evolution of this category in the two highest-register lexicons indicates the acquisition of flexible command of root and pattern morphology, with highest-register resultatives derived directly from nouns (and other bases) rather than verbs, e.g., *memutag* 'branded' from *mutag* 'brand', *menuce* 'feathered' from *noca* 'feather', *memusad* 'committed' from *mosad* 'institution'. Overall, the largest resultative class is *meCuCaC*, related to *pi'el*, the most productive *binyan* conjugation (Berman, 1993), followed by *CaCuC*, related to *qal*, the *binyan* with the largest number of verb types (Ravid et al., 2016), and finally *MuCCaC*, the most passive-like resultative, related to *hif'il* (Ravid & Vered, in press).

Two other structures are of interest here. The first is the nominal pattern *CaCiC* (*yaxid* 'single') that actually hosts two different morphological entities and accordingly occupies different developmental time points. The first wave of *CaCiC* adjectives shares no specific adjectival meaning, unlike most productive noun and adjective patterns in Hebrew (Ravid, 1990). These mostly show up in the Core lexicon (*xarif* 'hot (taste)', *na'im* 'pleasant'). The second wave involves the morphologically complex potential adjectives (comparable to English *-able* adjectives), such as *amid* 'durable' or *pagía* 'vulnerable'. These occur with higher frequency in the Mature lexicon, and especially so in the morphologically complex Educated lexicon, associated with the high school and college years. This is the first empirical piece of evidence regarding the late development of the potential adjective category.

A second late occurring adjectival device labeled 'complex' adjectives, including adjective-headed compounds (*axuz dibuk* 'held (by the) devil = possessed') and prefixed adjectives (*rav-gili* 'multi-age'), showed up mostly Mature and Educated lexicons, providing support for previous findings (Ravid & Levie, 2010). Complex adjective numbers are probably greater than found here, as quantifier prefixes can create more than one complex adjective (e.g., *xad-sitri* 'one way', *du-sitri* 'two way'), and they also attach to free-standing adjectives (*du-cdadi* 'bi-directional'). This, however, would not change their high-register rating as part of the complex adjective lexicons of adolescent and adult Hebrew users.

*Register as proxy for frequency*

The subjective linguistic register rating of adjectives by language experts was successful as an independent proxy measure of lexical frequency, echoed in the concurrent evolution of adjective morphological devices. We now have in our possession five adjective lexicons arranged by developmental order, from which the next stage of the project will draw the stimuli for AoA and clinical tasks. Although the efficacy of subjective register ranking was demonstrated here for Hebrew, a language with great historical depth and discrepancies between spoken and written styles (Ravid, 1995, 2012), we believe it can be adapted to other languages, subject to their typology and specific contexts of language variation. Three examples come to mind. For Arabic, a widely-spoken language with acute diglossia, written Standard Arabic and the spoken dialects would clearly frame the extreme poles of register, with criteria spanning the lexicon, morphology and syntax (Saiegh-Haddad & Spolsky, 2014). Although French diglossia is less well-recognized, spoken and written French morpho-syntax would be a good candidate for register distinctions (Gadet, 2009). And for English, the sweeping historical re-lexification responsible for its current double Latinate-Germanic lexicon has already been linked to register distinctions across development (Bar-Ilan & Berman, 2007).

## References


Alderson, J. C. (2007). Judging the frequency of English words. *Applied Linguistics*, 28(3), 383–409. https://doi.org/10.1093/applin/amm024

Andersen, E. S. (1992). *Speaking with style: The sociolinguistic skills of children*. London: Routledge.

Auer, E. T., Bernstein, L. E., & Tucker, P. E. (2000). Is subjective word familiarity a meter of ambient language? A natural experiment on effects of perceptual experience. *Memory & cognition*, 28, 789–797. https://doi.org/10.3758/BF03198414

Avneyon, E. (2007). *The Sapir Hebrew-Hebrew dictionary*. Or Yehuda: Hed Artzi. [in Hebrew]

Baayen, R. H., Feldman, L. B., & Schreuder, R. (2006). Morphological influences on the recognition of monosyllabic monomorphemic words. *Journal of Memory and Language*, 55, 290–313. https://doi.org/10.1016/j.jml.2006.03.008

Balota, D. A., Pilotti, M., & Cortese, M. J. (2001). Subjective frequency estimates for 2,938 monosyllabic words. *Memory & Cognition*, 29, 639–647. https://doi.org/10.3758/BF03200465

Balota, D., Cortese, M., Sergent-Marshall, S., Spieler, D., & Yap, M. (2004). Visual word recognition for single-syllable words. *Journal of Experimental Psychology: General*, 133(2), 283–316. https://doi.org/10.1037/0096-3445.133.2.283

Bar-Ilan, L. & Berman, R. A. (2007). Developing register differentiation: The Latinate-Germanic divide in English. *Linguistics*, 45, 1–36. https://doi.org/10.1515/LING.2007.001
Bar-On, A., & Ravid, D. (2011). Morphological decoding in Hebrew pseudowords: A developmental study. *Applied Psycholinguistics*, 32, 553–581. https://doi.org/10.1017/S014271641100021X
Bar-On, A., Dattner, E., & Ravid, D. (submitted) Resolving homography in reading Hebrew garden-path sentences: A developmental perspective.
Benelli, N., & Grinboim, Y. (2015). *Between familiarity and frequency: A psycholinguistic study of Hebrew adjectives*. Tel Aviv University ms. [in Hebrew]
Berman, R. A. (1978). *Modern Hebrew structure*. Tel Aviv: University Publishing Projects.
Berman, R. A. (1985). *Acquisition of Hebrew*. Hillsdale, NJ: Lawrence Erlbaum.
Berman, R. A. (1993). Marking of verb transitivity by Hebrew-speaking children. *Journal of Child Language*, 20, 641–669. https://doi.org/10.1017/S0305000900008527
Berman, R. A. (1994). Formal, lexical, and semantic factors in acquisition of Hebrew resultative participles. *Berkeley Linguistic Society*, 20, 82–92.
Berman, R. A. (2004). Between emergence and mastery: The long developmental route of language acquisition. In R. A. Berman (Ed.), *Language development across childhood and adolescence: Psycholinguistic and crosslinguistic perspectives* (pp. 9–34). Amsterdam: John Benjamins. https://doi.org/10.1075/tilar.3.05ber
Berman, R. A., & Ravid, D. (2009). Becoming a literate language user: Oral and written text construction across adolescence. In D. R. Olson & N. Torrance (Eds.), *Cambridge handbook of literacy* (pp. 92–111). Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511609664.007
Biber, D., & Conrad, S. (2009). *Register, genre, and style*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511814358
Blackwell, A. A. (2005). Acquiring the English adjective lexicon: Relationships with input properties and semantic typology. *Journal of Child Language*, 32, 535–562. https://doi.org/10.1017/S0305000905006938
Blodgett, E. G., & Cooper, E. B. (1988). Talking about it and doing it: Metalinguistic capacity and prosodic control in three to seven year olds. *Journal of Fluency Disorders*, 13, 283–290. https://doi.org/10.1016/0094-730X(88)90019-8
Brysbaert, M., & Cortese, M. (2011). Do the effects of subjective frequency and age of acquisition survive better word frequency norms? *The Quarterly Journal of Experimental Psychology*, 64(3), 545–559. https://doi.org/10.1080/17470218.2010.503374
Brysbaert, M., & Ghyselinck, M. (2006). The effect of age of acquisition: Partly frequency related, partly frequency independent. *Visual Cognition*, 13, 992–1011. https://doi.org/10.1080/13506280544000165
Carroll, J. B. (1970). An alternative to Juilland's Usage Coefficient for lexical frequencies. *ETS Research Bulletin Series*, 1970(2), i–15.
Caselli, M. C., Bates, E., Casadio, P., Fenson, J., Fenson, L., Sanderl, L., & Weir, J. (1995). A cross-linguistic study of early lexical development. *Cognitive Development*, 10, 159–199. https://doi.org/10.1016/0885-2014(95)90008-X
Clark, E. V. (2003). *First language acquisition*. Cambridge: Cambridge University Press.
Diessel, H. (2007). Frequency effects in language acquisition, language use, and diachronic change. *New Ideas in Psychology*, 25, 108–127. https://doi.org/10.1016/j.newideapsych.2007.02.002
Dixon, R. M. W., & Aikhenvald, A. Y. (2014). *Adjective classes: A cross-linguistic typology*. Oxford: Oxford University Press.

Dockrell, J. E., & Messer, D. (2004). Later vocabulary acquisition. In R. A. Berman (Ed.), *Language development across childhood and adolescence: Psycholinguistic and crosslinguistic perspectives*, (pp. 35–52). Amsterdam: John Benjamins. https://doi.org/10.1075/tilar.3.06doc

Douglas, F. M. (2003). The Scottish corpus of texts and speech: Problems of corpus design. *Literary and linguistic computing*, 18(1), 23–37. https://doi.org/10.1093/llc/18.1.23

Ferguson, C. A. (1994). Dialect, register, and genre: Working assumptions about conventionalization. In D. Biber & E. Finegan (Eds), *Sociolinguistic perspectives on register* (pp. 15–30). New York: Oxford University Press.

Gai, A. (1995). The category 'adjective' in Semitic languages. *Journal of Semitic Studies*, 1, 1–9. https://doi.org/10.1093/jss/XL.1.1

Gadet, F. (2009). Stylistic and syntactic variation: introduction. In K. Beeching, N. Armstrong & F. Gadet (Eds.) *Sociolinguistic variation in contemporary French* (pp. 115–120). Amsterdam: Benjamins. https://doi.org/10.1075/impact.26.09gad

Gries, S. T. & Divjak, D. (Eds.) (2012). *Frequency effects in language learning and processing*. Berlin: De Gruyter Mouton. https://doi.org/10.1515/9783110274059

Hora, A., Avivi-Ben Zvi, G., Levie, R., & Ravid, D. (2006). Acquiring diminutive structures and meanings in Hebrew: an experimental study. In I. Savickiene & W. U. Dressler (Eds). *The acquisition of diminutives* (pp. 295–317). Amsterdam: Benjamins. https://doi.org/10.1075/lald.43.13hor

Itai, A., & Wintner, S. (2008). Language resources for Hebrew. *Language Resources and Evaluation*, 42, 75–98. https://doi.org/10.1007/s10579-007-9050-8

Juhasz, B. J. (2005). Age-of-acquisition effects in word and picture identification. *Psychological Bulletin*, 131, 684–712. https://doi.org/10.1037/0033-2909.131.5.684

Jurafsky, D. (2003). Probabilistic modeling in psycholinguistics: Linguistic comprehension and production. In R. Bod, J. Hay and S. Jannedy (Eds.), *Probability theory in linguistics* (pp. 39–96). Cambridge, MA: The MIT Press.

Kennedy, G. (2014). *An introduction to corpus linguistics*. London & New York: Routledge. https://doi.org/10.4324/9781315843674

Kuperman, V., & Van Dyke, J. A. (2013). Reassessing word frequency as a determinant of word recognition for skilled and unskilled readers. *Journal of Experimental Psychology: Human Performance and Perception*, 39, 802–823.

Levin, I., Ravid, D., & Rappaport, S. (2001). Morphology and spelling among Hebrew-speaking children: From kindergarten to first grade. *Journal of Child Language*, 28, 741–769. https://doi.org/10.1017/S0305000901004834

Lo, Y., Mendell, N., & Rubin, D. (2001). Testing the number of components in a normal mixture. *Biometrika*, 88, 767–778. https://doi.org/10.1093/biomet/88.3.767

MacWhinney, B. (2000). *The CHILDES Project: Tools for analyzing talk. Vol. 2: The Database*. 3rd ed. Mahwah, NJ: Lawrence Erlbaum.

Morrison, C. M., Chappell, T. D., & Ellis, A. W. (1997). Age of Acquisition norms for a large set of object names and their relation to adult estimates and other variables. *The Quarterly Journal of Experimental Psychology Section A*, 50, 3, 528–559.

Nippold, M. A. (2016). *Later language development: School-age children, adolescents, and young adults* (4th Ed.). Austin, TX: Pro-Ed.

Nippold, M. A., & Duthie, J. K. (2003). Mental Imagery and Idiom Comprehension: A Comparison of School-Age Children and Adults. *Journal of Speech, Language, and Hearing Research*, 46, 788–799. https://doi.org/10.1044/1092-4388(2003/062)

Nylund, K. L., Asparouhov, T. & Muthen, B. (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling: A Monte Carlo simulation. *Structural Equation Modeling*, 14, 535–569. https://doi.org/10.1080/10705510701575396

Pitchford, N. J. & Mullen, K. T. (2001). Conceptualization of perceptual attributes: A special case for color? *Journal of Experimental Child Psychology*, 80, 289–314. https://doi.org/10.1006/jecp.2001.2634

Ravid, D. (1990). Internal structure constraints on new-word formation devices in Modern Hebrew. *Folia Linguistica*, 24, 289–346. https://doi.org/10.1515/flin.1990.24.3-4.289

Ravid, D. (1995). *Language change in child and adult Hebrew: A psycholinguistic perspective*. New York: Oxford University Press.

Ravid, D. (2003). A developmental perspective on root perception in Hebrew and Palestinian Arabic. In Y. Shimron (Ed.), *Language processing and acquisition in languages of Semitic, root-based morphology* (pp. 293–319). Amsterdam: Benjamins. https://doi.org/10.1075/lald.28.14rav

Ravid, D. (2004). Later lexical development in Hebrew: Derivational morphology revisited. In R. A. Berman (Ed.), *Language development across childhood and adolescence: Psycholinguistic and crosslinguistic perspectives* (pp. 53–82). Amsterdam: Benjamins. https://doi.org/10.1075/tilar.3.07rav

Ravid, D. (2006). Word-level morphology: A psycholinguistic perspective on linear formation in Hebrew nominals. *Morphology*, 16, 127–148. https://doi.org/10.1007/s11525-006-0006-2

Ravid, D. (2012). *Spelling morphology: The psycholinguistics of Hebrew spelling*. New York: Springer. https://doi.org/10.1007/978-1-4419-0588-8

Ravid, D., & Berman, R. A. (2009). Developing linguistic register across text types: The case of Modern Hebrew. *Pragmatics & Cognition*, 17, 108–145. https://doi.org/10.1075/pc.17.1.04rav

Ravid, D. & R. Berman. (2010). Developing noun phrase complexity at school-age: A text-embedded cross-linguistic analysis. *First Language*, 30, 3–26. https://doi.org/10.1177/0142723709350531

Ravid, D., & Levie, R. (2010). Hebrew adjectives in later language text production. *First Language*, 30, 27–55. https://doi.org/10.1177/0142723709350529

Ravid, D., & Nir, M. (2000). On the development of the category of adjective in Hebrew. In M. Beers, B. van den Bogaerde, G. Bol, J. de Jong, & C. Rooijmans (Eds.), *From sound to sentence: Studies on first language acquisition* (pp. 113–124). Groningen: Center for Language and Cognition.

Ravid, D., & Schiff, R. (2006). Roots and patterns in Hebrew language development: Evidence from written morphological analogies. *Reading and Writing*, 19, 789–818. https://doi.org/10.1007/s11145-006-9004-3

Ravid, D. & R. Schiff. (2012). From dichotomy to divergence: Number/gender marking on Hebrew nouns and adjectives across schoolage. *Language Learning*, 62, 133–169. https://doi.org/10.1111/j.1467-9922.2011.00675.x

Ravid, D., & Shlesinger, Y. (1987). On the classification and structure of *-i* suffixed adjectives. *Hebrew Linguistics*, 25, 59–70. [in Hebrew]

Ravid, D., & Vered, L. (2017). Hebrew verbal passives in Later Language Development: The interface of register and verb morphology. *Journal of Child Language*, 44(6), 1309–1336. https://doi.org/10.1017/S0305000916000544

Ravid, D., & Zilberbuch, S. (2003). The development of complex nominals in expert and non-expert writing: A comparative study. *Pragmatics and Cognition*, 11, 267–297. https://doi.org/10.1075/pc.11.2.05rav

Ravid, D., Levie, R., & Avivi-Ben Zvi, G. (2003). Morphological disorders. In L. Verhoeven & H. van Balkom (Eds.), *Classification of developmental language disorders: Theoretical issues and clinical implications* (pp. 235–260). Mahwah, NJ: Erlbaum.
Ravid, D., Ashkenazi, O., Levie, R., Ben Zadok, G., Grunwald, T., Bratslavsky, R., & Gillis, S. (2016). Foundations of the root category: Analyses of linguistic input to Hebrew-speaking children. In R. A. Berman (Ed.), *Acquisition and Development of Hebrew: From Infancy to Adolescence* (pp. 95–134). Amsterdam: John Benjamins. https://doi.org/10.1075/tilar.19.04rav
Saiegh-Haddad, E. & Spolsky, B. (2014). Acquiring literacy in a diglossic context: Problems and prospects. In Saiegh-Haddad, E. & Joshi, M. (Eds.) *Handbook of Arabic literacy: Insights and perspectives* (pp. 225–240). Dordrecht: Springer. https://doi.org/10.1007/978-94-017-8545-7_10
Salerni, N., Assanelli, A., D'Odorico, L., & Rossi, G. (2007). Qualitative aspects of productive vocabulary at the 200- and 500-word stages: A comparison between spontaneous speech and parental report data. *First Language*, 27, 75–87. https://doi.org/10.1177/0142723707067545
Sandhofer, C., Smith, L. B., & Luo, J. (2000). Counting nouns and verbs in the input: Differential frequencies, different kinds of learning? *Journal of Child Language*, 27, 561–585. https://doi.org/10.1017/S0305000900004256
Schachter, P., & Shopen, T. (2007). Parts-of-speech systems. In T. Shopen (Ed.), *Language typology and syntactic description: Clause structure, Second Edition – Volume I* (pp. 1–60). Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511619427.001
Shlesinger, Y. (2000). *Journalistic Hebrew: Stylistic aspects of Israeli newspaper sections*. Beer-Sheva: Ben-Gurion University of the Negev Press. [in Hebrew]
Spencer, A. (1991). *Morphological theory*. Oxford, UK: Blackwell.
Thompson, G. L., & Desrochers, A. (2009). Corroborating biased indicators: Global and local agreement among objective and subjective estimates of printed word frequency. *Behavior Research Methods*, 41, 452–471. https://doi.org/10.3758/BRM.41.2.452
Tribushinina, E., Van den Bergh, H., Ravid, D., Aksu-Koç, A., Kilani-Schoch, M., Korecky-Kröll, K., Leibovitch-Cohen, I., Laaha, S., Nir, B., Dressler, W. U., & Gillis, S. (2014). The first year of adjectives: A growth curve analysis of child speech and parental input. *Language, Interaction and Acquisition*, 5, 185–226. https://doi.org/10.1075/lia.5.2.02tri
Vichner, A. (2016). *Adjectives in young children's storybooks*. Tel Aviv University ms. [in Hebrew]

CHAPTER 6

# Functionalism in the lexicon

## Where is it, and how did it get there?

Russell Richie
University of Connecticut

### Thesis

Natural language lexicons, in both the forms and meanings of their words, are structured in such a way that is beneficial for human cognitive and behavioral needs, like categorization and communication. However, to say that natural language lexicons possess the structure they do *because* such structure is suited to human needs requires identifying how natural language lexicons culturally evolve under pressure from such needs. This has yet to be demonstrated for most putative adaptations in the lexicon.

### Commentaries on Richie thesis

Ray Jackendoff commentary on Richie thesis

The starting point for this investigation is the assertion that "[n]atural language lexicons … are structured in such a way that is beneficial for human cognitive and behavioral needs." This leaves open two questions: what does "structure" mean, and how do we (non-question-beggingly) assess human needs?

In the case of phonology, the answers are not too difficult: the pressure to communicate a message most reliably but with least effort apparently results in the Zipfian distribution of word length, in which the most frequently used words are the shortest. Regardless of the exact factors leading to this distribution, this can be taken to be a form of structure in the lexicon.

The story is not so clear when it comes to semantics. What counts as "need," i.e. what we need differentiated vocabulary *for*, is much more context-dependent. Psycholinguists (to use a term that in the larger scheme of things is needed by hardly anybody) need a different repertoire of categories than economists, carpenters,

https://doi.org/10.1075/z.238.06ric

farmers, chefs, and theologians. More generally, one might expect that every specialized domain of discourse will give rise to a specialized vocabulary that optimizes communication.

Richie, however, focuses on the distribution of vocabulary in more everyday domains such as colors, kinship terms, and object names (though think of the vast variety of specialized tools and utensils). One might add other such domains such as verbs of locomotion (*run, jog, wander*), verbs of manner of speaking (*mutter, grumble, scream*), and more generally all the verb classes enumerated by Levin (1993). As for another domain, Richie also proposes that "the proper conceptualization of spatial relations requires a semantic space." There is no question that it does, as revealed in a profusion of publications such as Bloom et al. (1996); Levinson (2003); Landau (2017); Li and Gleitman (2002), and many others. Each of these domains may be said to structure the lexicon, dividing up semantic space in task-appropriate fashion.

Another way in which the lexicon is structured, touched on only tangentially by Richie, involves explicit relations among lexical items themselves, as outlined in Jackendoff and Audring (this volume). The strongest relations are among items that share elements of meaning and elements of phonology. For instance *complexity* is related to *complex* on both counts; moreover, the suffix *-ity* is related semantically and phonologically to the suffix in *adversity, profundity, specificity*, and so on. Other items are related only phonologically: the *conductor* of a train does not *conduct* it; a *honeymoon* has little to do with honey or moons. Still others are related semantically but not phonologically, such as *lion* and *tiger*. Given enough instances of such relationships, the lexicon also contains *schemas* that codify the features that related items have in common, as well as how they differ. These relationships arguably play a role in identifying and processing words, and especially in learning new ones – hence filling important cognitive "needs" in Richie's sense.

## *References*

Bloom, Paul, Mary A. Peterson, Lynn Nadel, and Merrill F. Garrett (eds.). (1996). *Language and Space*. Cambridge, MA: MIT Press.

Jackendoff, Ray, and Jenny Audring. (this volume) Morphological schemas: Theoretical and psycholinguistic issues.

Landau, Barbara. (2017). Update on "What" and "Where" in spatial language: A new division of labor for spatial terms. *Cognitive Science* 41 (S2), 321–350. https://doi.org/10.1111/cogs.12410

Levin, Beth. (1993). *English Verb Classes and Alternations: A Preliminary Investigation*. Chicago: University of Chicago Press.

Levinson, Stephen C. (2003). *Space in language and cognition: Explorations in cognitive diversity*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511613609

Li, Peggy, and Lila Gleitman. (2002). Turning the tables: Language and spatial reasoning. *Cognition* 83, 265–294.

## James Myers commentary on Richie thesis

One of the first people to link evolutionary notions to language was Darwin himself, who cited the genealogical approach to linguistic classification in *Origin of Species* in order to make his own proposal about biological classification seem more intuitive. The central challenges noted in the paper will also be familiar to biologists: When making functionalist claims, how do we go beyond just-so stories? How is the abstract principle of natural selection (which, as Richard Dawkins and many others have argued, applies to any system involving replication, variation, and feedback from the latter to the former) actually implemented in temporal, causal mechanisms?

From a biologist's perspective, however, the linguists have it much easier. The author mentions, but does not explore, two rich sources of data about potential functionalist mechanisms, both almost totally unavailable in biology. The first involves the emergence of novel linguistic systems. Biologists almost never encounter a species truly new to the world, and certainly none that spring up *de novo* (unless created in the lab). Emerging languages may be "rare and diverse", as the author notes, but rare is better than nothing. Rarity may even be the wrong way to describe the number of new sign languages that have already been discovered (not to mention the thousands of home sign systems, only a handful of which have been studied, by Goldin-Meadow and others). As for diversity, it is itself a valuable source of information about how function shapes form, as typologists (both linguistic and biological) know well. Many types of linguistic system are emerging around us all the time, not just sign, but also creoles, slang, writing systems, even artificial languages as mutated by their learners in the lab and at language creation conventions.

A second valuable source of evidence is history. Where biologists rely on the chance preservation of fossils, languages have long been written down precisely in order to preserve them, and with the gradual expansion of literacy across the world, more and more languages are leaving tracks. Writing also has a finer temporal grain than biological fossilization, providing evidence about changes within a single lifetime, and even about the earliest stages of now-mature systems. For example, the well-documented history of Chinese writing shows how a recursively generated lexical system can emerge from an initial set of holistic icons.

All in all, linguists are better positioned to test the mechanisms implementing natural selection than biologists. To misquote another bit of historical text, methinks the author doth protest too much.

## Dorit Ravid commentary on Richie thesis

This commentary views lexicons as part of the complex, dynamic, adaptive system that is language in any of its manifestations – usage (communication), change, acquisition, based on interrelated patterns of experience, social interaction, and cognitive processes (Beckner et al., 2009). Therefore, the way to determine the emergence and consolidation of a mental lexicon in a child is the careful investigation of how words occur in their environment, i.e., in corpora consisting of Child-Directed Speech or children's own output, Child Speech (Ashkenazi, Ravid & Gillis, 2016; Kidd, Lieven & Tomasello, 2010).

Elman (2003) formulated the idea that development in naturally noisy environments is the driving force in language learning, with young children aided rather than hindered by limited cognitive resources ("starting small"). The assumption is thus that lexical development takes place under constant pressure from the changing nature of the language input and the ability to predict more relationships and regularities. For example, Mariscal (2009) showed that Spanish-speaking children construct abstract agreement categories based on a dynamically changing confluence of sources in the input, such as noun phonology and the shape of determiners, pronouns, and adjectives.

### *References*

Ashkenazi, O., D. Ravid & S. Gillis. (2016). Breaking into the Hebrew verb system: a learning problem. *First Language*, 36, 505–524. https://doi.org/10.1177/0142723716648865

Beckner, C., Blythe, R., Bybee, J., Christiansen, M. H., Croft, W., Ellis, Nick C., Holland, J., Ke, J., Larsen-Freeman, D., & Schoenemann, T. (2009). Language is a complex adaptive system. Position paper, *Language Learning*, 59, Supplement 1, 1–27. https://doi.org/10.1111/j.1467-9922.2009.00533.x

Elman, J. L. (1993). Learning and development in neural networks: The importance of starting small. *Cognition*, 48, 71–99. https://doi.org/10.1016/0010-0277(93)90058-4

Kidd, E., Lieven, E. V. M. & Tomasello, M. (2010). Lexical frequency and exemplar-based learning effects in language acquisition: evidence from sentential complements. *Language Sciences*, 32, 32–142. https://doi.org/10.1016/j.langsci.2009.05.002

Mariscal, S. (2009). Early acquisition of gender agreement in the Spanish noun phrase: starting small. *Journal of Child Language*, 36, 143–171. https://doi.org/10.1017/S0305000908008908

## Benjamin Tucker commentary on Richie thesis

The article by Richie attempts to summarize some research and discuss future directions of a functionalist model of the lexicon. At the outset this is a large task in an article, the author focuses on many topics related to functionalism and discusses the potential functional nature of the lexicon. This paper is not a full review of the literature on functional applications to the lexicon and focuses largely on semantics with some mention of phonology along the way. When attempting to explain the lexicon in this manner it is very difficult to leave out other domains of functionality as they are part of a whole that functions in a manner specific toward the purpose – production or comprehension. As a result, a functional account of the lexicon, of any sort will always be incomplete, unless it considers the full functionality of the lexicon, whether it is functional or not. However, this provides a good starting point for future consideration of the other aspects of a functional lexicon.

It is also important to consider how both aspects of production and comprehension inform our knowledge of the lexicon. It is common in much of the research on the lexicon to simplify the discussion and assume that there is one lexicon and therefore results in both production and comprehension experiments must tell us something about the nature of lexical representation. However, researchers often work in either production or comprehension and fail to consider both domains. It is only when we consider both domains together that researchers of the lexicon may achieve a clearer picture of the structure of the lexicon.

One example of the differences between domains could be found in the oft used predictor "neighborhood density". In this example there are differences in the findings depending on the modality and the domain used to explore neighborhood density. First consider comprehension, researchers investigating visual comprehension find that higher neighborhood density is facilitatory (e.g., Ziegler et al., 2003) but in the auditory domain higher neighborhood density is inhibitory (e.g., Luce & Pisoni, 1998). Second if we vary the domain, we would observe that in production the effect of neighborhood density is variable such that some studies indicate that higher neighborhood density leads to longer vowel productions (Wright, 2004; Munson and Solomon, 2004) and vice versa (Gahl et al., 2012). If this is indeed true, what is the functional nature of neighborhood density in the lexicon? Are there separate lexicons for visual and auditory modalities or separate lexicons for production and comprehension? Maybe there is one lexicon but if so, what is its nature that would result in these differing results for the same predictor?

Additionally from a phonetic perspective this functionalist approach raises several questions. Would this functional perspective then apply in such a way that non-genetically related languages should arrive at some of the same categories or some of the same words because the words maximize informativity? Another way to think about this might be that if maximizing informativity is the goal should all languages have similar phonotactic distributions as a result of the drive for informativity?

*References*

Gahl, S., Yao, Y., and Johnson, K. (2012). Why reduce? Phonological neighborhood density and phonetic reduction in spontaneous speech. *Journal of Memory and Language*, 66 (4): 789–806. https://doi.org/10.1016/j.jml.2011.11.006

Luce, P. A., & Pisoni, D. B. (1998). Recognizing Spoken Words: The Neighborhood Activation Model. *Ear and Hearing*, 19(1), 1–36. https://doi.org/10.1097/00003446-199802000-00001

Munson, B. and Solomon, N. P. (2004). The Effect of Phonological Neighborhood Density on Vowel Articulation. *Journal of Speech, Language, and Hearing Research*, 47(5):1048–1058. https://doi.org/10.1044/1092-4388(2004/078)

Wright, R. (2004). Factors of lexical competition in vowel articulation. In Local, J., Ogden, R., and R. T., editors, *Papers in Laboratory Phonology* 6, pages 75–87. Cambridge University Press, Cambridge.

Ziegler, J. C., Muneaux, M., Grainger, J. (May 2003). Neighborhood effects in auditory word recognition: Phonological competition and orthographic facilitation. *Journal of Memory and Language*, 48 (4), 779–793. https://doi.org/10.1016/S0749-596X(03)00006-8

## Chris Westbury commentary on Richie thesis

One of the difficulties with evolutionary explanations is that they co-exist uneasily with normal human causal understanding. One reason for this is that humans prefer to identify one or a few directed causes, whereas by the very nature of their encoding in DNA, evolutionary adaptation can reflect the parallel effects of multiple, often separately-focused 'micro-causes', i.e. the summed effects of many small increments in fitness. A standard classroom example for illustrating how this can play out is that genes that cause the fitness-lowering disease sickle cell anemia offer a higher fitness benefit of protection from malaria, with the result that protection from malaria is an indirect cause of increased incidence of sickle cell malaria in malaria-prone environments (Gelpi and King, 1976; Flint, Hill, Bowden, et al., 1986).

It is natural to assume that the sole purpose of the lexicon is for communicating meaning using speech or the later-developing writing. Consistent with this, this paper identifies several natural conceptual categories of essential and basic terms in organizing human life that are also seen among the earliest words learned by modern humans. For example, the 1000 words judged as earliest learned in the 30,000 Kuperman, Stadthagen-Gonzalez, and Brysbaert (2012) age of acquisition (AofA) norms include many color terms, kinship terms, and number terms, along with other word clusters that seem basic to navigating human life from an early age: food names, animal names, toileting terms, and names of body parts. Although most of these conceptual distinctions can be made by non-human primates (at least, numbers, animals, colors, and social roles), non-human primates seem to be fit to their environments without being able to name them (with the exception of some predator calls).

An interesting question in the context of considering the functional structure of the human lexicon is: What was that neural machinery we now use for accessing *linguistic meaning* doing before it was being used to support communication? It is hard to say exactly how long humans have had speech, but estimates range between a few hundred thousand years and about a dozen hundred thousand years – a short enough time period to rule out major re-organization of our brains through natural selection. Many people have identified cognitive and motor homologue precursors of speech production, pointing out that speech builds on top of functions still found in non-humans for perception, memory, categorization, and decision-making, among other functions (e.g. Hickock, 2009). Fewer have speculated (perhaps because it is difficult to move much beyond mere speculation) about *conceptual homologues* to lexical semantics, which may shed further light on functionalism in the lexicon

Vygotsky (1962) famously suggested that children could use self-speech as a means of self-control. Taking this idea seriously suggests that one might expect to see several other categories of words that do in fact appear in the earliest 1000 words in the Kuperman et al. AofA norms: words that can serve as quantifiers and negators (*little, more, some, yes, no, big, huge, wrong*), basic emotion-related terms (*mad, sad, happy, unhappy, angry, scared, scary, okay, hurt, good, yucky, icky, cry*), pronouns (*you, me, we, he, them, him, myself, yourself, everyone*), and verbs related to self-presentation (*do, pretend, hush, forget, want, need, ask, try, wish, cannot, watch, give, must, choose, sorry, smile*). Notwithstanding Wittgenstein's (1958) arguments about the impossibility of a private language, such words seem (one may, at least, speculate) to suggest a very sparse protolanguage that might allow an individual to adaptively modulate their private emotional experience and interactions with others by predicating the personal world: *you happy, him angry, yes good, everyone scared (don't cry)*, and so on.

Perhaps the first quasi-linguistic predicates were exactly these kind of self-directed predicates, adaptive precisely because they did enable the modulation of self-presentation by managing internal states. Lupyan and Bergen (2016) outline the cognitive benefits that might accrue simply from generalizing this ability to predicate, most notably that "[g]rowing evidence suggests that labels (category names) play a causal role in category learning in children" (p. 4) because "language plays a role in learning the same/different relation" (p. 5). They conclude by suggesting that the "evolution of language […is the] evolution of a control system for programming our own and other people's minds".

## *References*

Gelpi, A. P. & King, M. C. (1976). Association of Duffy blood groups with the sickle cell trait. *Human Genetics*, 32: 65–68. https://doi.org/10.1007/BF00569977

Flint, J., Hill, A. V. S., Bowden, D. K., Oppenheimer, S. J., Sill, P. R., Serjeantson, S. W., Bana-Koiri, J., Bhatia, K., Alpers, M. P., Boyce, A. J. and Weatherall, D. J. (1986). High frequencies of α-thalassaemia are the result of natural selection by malaria. *Nature*, 321(6072), pp. 744–750. https://doi.org/10.1038/321744a0
Hickok, G. (2009). The functional neuroanatomy of language. *Physics of life reviews*, 6(3), 121–143. https://doi.org/10.1016/j.plrev.2009.06.001
Kuperman, V., Stadthagen-Gonzalez, H., & Brysbaert, M. (2012). Age-of-acquisition ratings for 30,000 English words. *Behavior Research Methods*, 44(4), 978–990. https://doi.org/10.3758/s13428-012-0210-4
Lupyan, G., & Bergen, B. (2016). How language programs the mind. *Topics in cognitive science*, 8(2), 408–424. https://doi.org/10.1111/tops.12155
Wittgenstein, L. (1958). *Philosophical Investigations*. New York, NY: McMillan Publishing Co.

## The article

One of the most important questions in all of cognitive science is: why is human cognition the way it is? As with all questions of "why", there is a range of suitable, non-exclusive answers, some more proximal and some more distal, even ultimate. For example, in answering, "Why is the human eye structured the way it is?", an answer giving a more distal or ultimate cause would be "Because the eye is adapted to its function of enabling vision". For example, the eye has a lens because it enables the eye to direct incoming light onto a smaller retinal surface area, without reducing the intensity of the stimulus. In contrast, an answer giving a more proximal cause would be "Because the evolution of the eye was subject to selective pressures for effective vision". In other words, the proximate cause identifies a mechanism by which the structure in question arose. Returning to the lens example, at some point in our evolutionary history, there must have been some organisms with lens, and some without. The adaptive advantage of the lens led to individuals possessing lens reproducing at greater rates, leading to a predominance of lenses among today's extant organisms (human and otherwise). As with the structure of the human body, explaining the whys of human behavior can be done with proximal and distal causes (Scott-Phillips, Dickins, & West, 2011). This is true in language as in other areas of cognition and behavior, and answering why language has the structure it has is a primary goal of functionalist linguistics (Haspelmath, 1999; Bybee, 1999, 2006). Functionalists seek both distal/ultimate causes like "Because language structure is suited to the needs of language users" and proximal causes like "Because it adapted/developed in a certain way over ontogenetic, historical, or phylogenetic timescales." Functionalists have mostly carried out this enterprise in the linguistic domains of syntax (Croft, 1995), morphology (Cutler, Hawkins, & Gilligan, 1986; Bybee, Pagliuca, and Perkins (1990)), and phonology (Lindblom 1986). However, in the past decade, and even more so in the past five years, research into functionalism in the *lexicon* has flourished. That is, researchers have found various ways that the meanings and forms of *actual lexicons, whose words, in the aggregate*, efficiently support communication, learning, memory storage, retrieval, or other cognitive functions, thus providing a distal cause to answer "Why do we have the words we have?" An integrative review of this burgeoning literature, however, has yet to appear. This paper will fill that gap, and suggest future directions of research, including additional ways functionalism in the lexicon might be assessed, and, crucial for the hypothesis, how functional, user-adapted structure in the lexicon might have arisen via language use and acquisition, and possibly thus provide an answer of proximal cause to the question of why we have the words we have.

## Goals of the current review

Before beginning the review in earnest, it is perhaps important to further demarcate the goals of this review. As stated above, others have reviewed or studied at length functionalism in many domains of language, including phonology (Lindblom, 1986, but see also work by, i.a., John Ohala, Ian Maddieson, Matthew Aylett, and Alice Turk). To the extent that the lexicon and phonology have a close relationship – for example, words in the lexicon follow phonological patterns, and language learners extract these patterns largely on the basis of real, individual words – the functionalism in phonetics and phonology that Lindblom and others have found (e.g., that vowel systems are sensitive to perceptual distinctiveness among their vowels) entails a certain functionalism in the word-forms in actual lexicon.[1] But there is a further way in which functionalism in the forms of words could surface, and which is our focus in this review. That is, the *actual* words in a lexicon, *in the aggregate*, might reflect functionalist organization. Notice that this is somewhat different from claims that, for example, phoneme inventories, or phonotactic restrictions, are functionally motivated – these latter claims make no reference to individual or sets of words in a lexicon. This distinction should become clearer as we discuss specific claims of functionalism in the *aggregate* lexicon.

There is another neighboring, recently growing literature that is even more relevant: work demonstrating wide-spread (within and across languages) iconicity (resemblances between word form and word meaning, onomatopoeia like 'pow' being the clearest example of such) and what these authors term 'systematicity'

1. Readers may notice that the upcoming section FUNCTIONALISM IN THE PHONOLOGICAL LEXICON is half as long as the following section FUNCTIONALISM IN LEXICAL-SEMANTICS. This is not to imply any imbalance in importance between phonology/form and meaning. Instead, it might be that, as just stated, some functionalism in the forms of actual words in a lexicon is already provided by functionalism in phonology.

(statistical regularities in word form that predict word function, e.g., phonetic/ phonological cues to word classes like noun or verb). Both iconicity and this 'systematicity' can be seen as examples of functionalism in the lexicon: they can support language users' needs for easy and effective word learning and use. However, we do not review this work here, as this body of literature has been reviewed elsewhere (e.g., Dingemanse, Blasi, Lupyan, Christiansen, & Monaghan, 2015) while the work reviewed below has not.

With this fairly circumscribed domain in mind, it may also be worthwhile to indicate who might be interested in this review, and why. Indeed, cognitive scientists of all stripes might find something relevant to their interests. It is likely fairly clear that psychologists, philosophers, and linguists, whose work forms the bulk of the review, may be interested. But so too might anthropologists – we will see that functionalism accounts for both the universality and diversity in lexicons around the globe – and computer scientists – efficient design and use of data structures, lexicons being such a structure, is a chief concern to them. Perhaps even neuroscientists might be interested; though the paper will not make much reference to the brain, it may turn out that certain functional needs, satisfied by the particular lexicons we have, are best understood at the level of the brain, rather than the mind.

Having clarified the scope of the paper, as well as the likely audience, we now turn to discussing cases of functionalist structure in the forms and meanings of natural language lexicons.

## Functionalism in the phonological lexicon

Any discussion of functionalism in the lexicon must begin with one of its earliest proponents, Harvard linguist George Kingsley Zipf. Zipf (1949) is often credited with popularizing[2] Zipf's Law, one form of which states that an inverse relationship obtains between word frequency and word length, Zipf explained this relationship with his principle of least effort: the average sentence length can be shortened, saving effort, if the most frequent concepts to be communicated are assigned short codes (as in Huffman Coding, a method of data compression). Only a few years later, Martinet (1952) claimed that sound contrasts will tend to disappear, rendering the forms in the lexicon simpler and more homogeneous, when the 'functional load', or importance of that contrast to distinguishing words, is low, meaning communication efficacy will not be sacrificed (as much) when lexical forms are

2. Though he didn't claim to have discovered it. That honor appears to belong to French stenographer Jean-Baptiste Estoup (Lelu, 2014).

collapsed onto each other.[3] However, Martinet's 'functional load hypothesis' was long doubted for lack of evidence (King, 1967), and Zipf's observed relationship between word length and frequency was seen as irrelevant to language and notions of effort or economy, as simple models like random typing, where one randomly presses alphabetic keys and the space bar, also produced such a relationship (Miller, 1957). Recent work, though, has breathed new life into these old ideas.

Piantadosi, Tily, and Gibson (2011), building off Zipf (1949), pointed out that assigning short words to frequent messages *would* be optimal if words were independently sampled from a stationary distribution. But clearly that is not the case: "eggs" is much likelier than, say, "mud" at the end of the sentence "I went to the store to buy milk and ___". Thus, perhaps words that are predictable *from context* are shorter. To investigate this, Piantadosi et al. used corpora for 10 languages to estimate each word's average surprisal, a measure of predictability, conditioned on preceding n-gram context ($n = 2$: "and ___"; $n = 3$: "milk and ___"). Indeed, average conditional surprisal predicted word length better than did frequency, and was still a significant predictor when frequency was partialed out, consistent with the general spirit of Zipf's original hypothesis that word forms are adapted for efficient, i.e., effort-minimizing, communication. A related benefit of a lexicon structured in this way is that the rate of information per unit time is kept relatively constant (a la the Uniform Information Density hypothesis, Jaeger and Tily, 2011). A constant information rate can make optimal use of the speech channel by maximizing the amount of information conveyed, without over- or under-taxing the channel capacity of speech or our cognitive systems.

Just as the classic theory by Zipf (1949) was refined and bolstered by Piantadosi et al. (2011), Martinet (1952)'s hypothesis has received even more recent support, as increased computational power and data have made more comprehensive analyses possible. Wedel, Kaplan, and Jackson (2013), using a database of a large number of phoneme mergers from a diverse set of languages, found that the number of minimal pairs significantly predicted probability of phoneme merger (such that if a phonemic contrast distinguished more minimal pairs, the contrast was less likely to merge), consistent with Martinet's hypothesis. They also found that higher phoneme probability was positively correlated with merger, and that this effect was stronger for phonemes that distinguished no minimal pairs. This is also suggestive of a functional lexicon, as more probable items are more predictable and hence less informative, and thus can be more easily eliminated without hurting communication. Later that same year, the Wedel group carried out even subtler work also

**3.** At first blush, this may seem like a hypothesis purely about phonology rather than the lexicon. However, this is a hypothesis about how phonology and word forms in the lexicon are affected by characteristics of the lexicon as a whole (i.e., minimal pair count).

consistent with the functional load hypothesis. Wedel, Jackson, and Kaplan (2013) found that (1) a contrast's count of minimal pairs from the same syntactic class, and (2) the count of minimal pairs of similar frequency, were even better predictors of merger probability. These findings are expected given effects like Piantadosi et al (2011)'s. That is, words from the same syntactic class occur in similar contexts, and thus must be distinguished better than words from different syntactic classes. For example, it is important to distinguish 'bat' and 'mat' as both are nouns and hence can appear at the end of the sentence "I went to the store and bought a ___", and whereas distinguishing 'bat' and 'bad' is less important, as one is a noun and the other an adjective, and hence they occur in different sentential contexts. Likewise, when two words are of similar frequency, it is more difficult to predict which word will occur than if one word is much more likely than the other. Hence, cues to word identity must be preserved.

Broadening beyond these classic theories of Zipf and Martinet, novel functionalist proposals concerning the forms of words in the lexicon have also emerged in recent years. Piantadosi, Tily, and Gibson (2009), for example, showed that stressed syllables are more informative of word identity than non-stressed syllables: stressed syllables contain a variety of different phoneme strings, whereas non-stressed syllables tend to contain only a few (often simply the schwa vowel /ə/, pronounced 'uh'). Given that stressed syllables are louder and longer and hence more perceptible, having more cues to word identity in this part of the word could be seen as efficient use of the communication channel. Similarly, Graff (2012) showed that the number of minimal pairs for a given contrast is a function of the confusability of the members of the contrast (confusability drawn from the data of Miller and Nicely, 1955). For example, there are many minimal pairs distinguished by the non-confusable contrast /b/ vs /t/ and relatively few minimal pairs distinguished by the very confusable contrast /f/ vs /θ/.[4] Piantadosi, Tily, and Gibson (2012) show that easy-to-process linguistic units like short and phonotactically probable forms (the combination of phonemes in 'pan' is more probable than the combination of phonemes in 'sprang') tend to be reused: short and phonotactically probable word-forms (like 'ran') have more meanings, and short and phonotactically probable syllables are used in more unique words. Far from creating ambiguity, re-using 'easy' forms in this way is actually efficient, given that context always provide disambiguating information as to the intended identity or meaning of a word.

One final functionalist proposal that somewhat straddles the phonological and semantic divide comes from Ferrer i Cancho and Sole (2003). Besides finding that

4. This difference in minimal pair count between these two contrasts would in turn make /f/ vs /θ/ but not /b/ vs /t/ a likely candidate for merger in English, according to the Wedel et al. (2013) findings reviewed above.

word frequency and word length are inversely related, Zipf (1949) also found that word frequencies follow a power law, such that the product of a word's frequency and its frequency rank is a constant. Zipf also accounted for this phenomenon with a principle of least effort: speakers want a single word that can mean anything, while listeners want a different word for everything.[5] Compromising these needs leads to a few words that dominate in frequency, and a large number of words that occur a few times. Despite the appeal of this idea, Zipf did not provide a rigorous test of it, which Ferrer i Cancho and Sole (2003) attempt to provide. In their work, they imagine that language users possess a form-meaning mapping matrix $A = \{A_{ij}\}$, where $A_{ij}$ is 1 if the $i$-th word-form can refer to the $j$-th meaning. The matrix, and the consequent word probabilities, evolve to minimize a cost function, wherein speakers pay a cost proportional to the entropy of signals they must convey (a formalization of the speaker's cost of retrieving a word), and listeners pay a cost proportional to the (expected) entropy over referents given a word (a formalization of the listener's cost of retrieving a meaning given a word). There is a single parameter $0 < \lambda < 1$ which trades off the cost between speakers and listeners, and the authors show that for a very particular setting of this parameter which roughly balances speakers' and listeners' costs, $\lambda = 0.41$, they obtain a Zipfian distribution of word frequencies, thus lending some support to Zipf (1949)'s idea that Zipf's Law arises from balancing speakers and listeners' needs.

## Functionalism in lexical-semantics

While research into functionalism in the *forms* of words has flourished, so too has research into functionalism in the *meanings* of words. As with the phonological literature, this work draws on ideas that date back decades but have recently received new support. Eleanor Rosch, pioneer of research into categorization, carried out some of the earliest and most prominent work on utility of category systems, work that continues to inspire current work on functionalism of lexicons. While Rosch's focus was on category systems generally, and not necessarily how *lexicons* carve up meaning, her ideas are nonetheless applicable to the current question, insofar as the lexicon maps onto such (efficient or utilitarian) category systems. Rosch's (1978) central claim, the kernel of which is shared with most functionalist accounts of the lexicon, was that "the task of category systems is to provide maximum information with the least cognitive effort" (p. 190). According to her, maximizing information

**5.** While both Zipf (1949) and Ferrer i Cancho and Sole (2003) view the competition as between speakers and listeners, one needn't have that view. In fact, viewing the competition as between the needs of effort and the needs of clarity/informativeness may be more desirable, as getting a message across, i.e., being informative, is in the interest of both speaker and listener.

and minimizing effort are at odds. For example, a maximally informative category system would have many extremely fine categories which allow precise prediction of many properties of an object based on its category membership. Such a system might have one category for "water poured into a tall container by a young man", and another for "water poured into a tall container by a young woman". However, such a system would likely be overkill for many purposes of categorization – both of the foregoing categories of water are drinkable, able to put out fires, etc. – and would be hard to learn and remember, i.e., effortful. Rosch formalized her notion of informativity with the measure of cue validity (e.g., Beach, 1964):[6] the conditional probability of membership in a particular category (e.g., 'dog') given a particular feature value (e.g., 'barks'). Rosch argued that basic-level categories like 'chair' (as opposed to subordinate categories like 'kitchen chair' or superordinate categories like 'furniture') are 'basic' (learned earlier, used preferentially)[7] because they have the greatest cue validity. Rosch illustrated this in a number of ways. For one, participants are better able to identify overlapping images of, for example, chairs than images of furniture, and not significantly better with overlapping images of kitchen chairs. Similarly, participants list few features in common to furniture, many to chairs, and not significantly more to kitchen chairs. Both pieces of evidence suggest that chairs and other basic level categories have more within-category similarity and hence cue validity than superordinate categories, and not much less than subordinate categories. At the same time, subordinate categories have lower total cue validity than do basic categories, because subordinate categories share most features with contrasting subordinate categories – 'kitchen chairs' are rather similar to 'office chairs'. Thus, Rosch gave preliminary evidence that the categories we tend to use – basic level categories – are those that are most informative. Crucially, though, she did not include any measure of effort in her formalization, a hole that, we will see, was only plugged some years later.

Since the forward-thinking work of Rosch (1978), improvements to her and her colleagues' formalization of category utility have been made, but the fundamental ideas of category utility generally, and informativity and effort specifically, have remained remarkably constant. One improvement to these ideas came from Corter

6. As with Zipf, it may seem a little odd to credit Rosch with so much in the development of functionalism when some of the critical ideas, like cue validity, predated her work. This choice was made for two reasons. First, one earlier reference that Rosch herself cites, Beach (1964), did not claim that category systems are structured to be utilitarian, like Rosch (1978) does. Second, while Beach (1964) has been less than 100 times according to Google Scholar, the reprinted Rosch (1978) has been cited more than 1000 times. Thus, due to the arguably greater impact that Rosch has had on the field, we are behooved to focus on her influence.

7. When people see a picture of a particular chair, they tend to say 'chair', and not 'furniture' or 'kitchen chair'.

and Gluck (1992). As they point out, cue validity does not take into account the conditional probability of feature values given category membership. That is, it does not account for the ability to predict an item's features based on the category it is in, which is a key function of categories. Recognizing this, Corter and Gluck developed a metric of the utility of a particular category that incorporates the probability of a category's occurrence in the world, the conditional probability of a particular feature value given membership in the given category, and the probability of that feature value in the world. Corter and Gluck show how, given feature structures for musical instruments, fruit, and furniture, their metric correctly predicts the basic level, independently measured as the level at which people are fastest to name visually presented objects. While using a different metric than Rosch (1978), the conclusion is very similar: we have the categories we have (and by extension, words mean what they mean) because they maximize informativeness.

Further research on functionalism in how the lexicon carves up semantic space has taken off in the last several years, beginning with Regier, Kay, and Khetarpal (2007). They sought to resolve a long-standing debate concerning the supposed universality of color lexicons. Some had claimed that color lexicons are organized around universal prototypical colors (Berlin & Kay, 1969). The boundaries of color categories are projected from these universal prototypes and thus tend to lie in similar positions in color space across different languages. Others (e.g. Roberson, Davidoff, Davies, & Shapiro, 2005), however, deny that such foci are the basis for color naming and instead maintain that color categories are defined at their boundaries by local linguistic convention, which is free to vary considerably across languages. Regier et al., (2007) sought a different account of cross-linguistic patterns of color lexicons: that color lexicons tend to maximize within-category similarity (similarity between every pair of elements in a category) and minimize between-category similarity (similarity between every pair of elements from different categories), a previously proposed normative principle of categorization (Garner, 1974).[8] They hypothesized that this principle combined with the fact that similarity space for color percepts is uneven could predict cross-linguistic patterns of naming.

Two pieces of evidence supported this prediction. First, artificially generated color lexicons that fell at global well-formedness maxima closely resembled attested natural language color lexicons. In other words, real lexicons resembled artificially constructed lexicons that maximized the sum of average within-category similarity and between-category dissimilarity. Second, systematic distortions (rotations of the color lexicons in color perceptual space) resulted in lower well-formedness

8. It is relatively intuitive why this principle is normative: if the purpose (or at least one purpose) of categorization is to treat members of a class as equivalent, then this is most easily achieved, or perhaps most rational, when the members are most like each other, but most unlike members of other categories.

patterns than attested color lexicons. Their results thus accommodated both universality – all humans are subject to the same irregularly shaped perceptual color space and principle of categorization – and language-specificity – there are multiple (near-)optimal[9] ways (differing, for example, on the number of basic color terms) to carve the perceptual color space while meeting the needs of efficient categorization.

Khetarpal, Majid, and Regier (2009) then extended this work to spatial lexicons (words for 'in', 'next to', 'through', etc.). To construct a similarity space for spatial concepts, Khetarpal et al. took 71 line-drawn scenes of the Topological Relations Picture Series designed by Melissa Bowerman (e.g., Bowerman & Pederson, 1992), and instructed English-speaking and Dutch-speaking participants to group the scenes into piles such that the scenes in a pile depict similar spatial relations. Similarity was then just the percentage of participants that put two scenes in the same pile. These pairwise similarities allowed the authors to use multidimensional scaling to reconstruct a similarity space, where each scene is a point and the distance between points/scenes is the similarity between them. As in Regier et al (2007), Khetarpal et al. (2009) then (a) computed the well-formedness of the lexicons of 9 typologically diverse languages, and (b) compared the well-formedness of these attested lexicons to well-formedness of lexicons systematically distorted from the actual lexicons. Again, the attested spatial lexicons overwhelmingly had greater well-formedness than corresponding, distorted lexicons, suggesting that, as in color lexicons, spatial lexicons of natural languages follow functional principles of categorization.

Continuing this work, Khetarpal, Neveu, Majid, Michael, and Regier (2013) expanded Khetarpal et al. (2009)'s dataset with two more languages (English and an Amazonian language, Maijɨki), and tested a functionalist account similar to Khetarpal et al. (2009)'s against a competing account of categories, according to which natural human categories simply carve out connected[10] regions of semantic space, (e.g. Croft, 2003; Haspelmath, 2003). Whereas Khetarpal et al. (2009)'s account formalized the functionalist principle as maximizing within-category similarity while minimizing between-category similarity, Khetarpal et al. (2013)'s formalization was slightly different. They formalized the informativeness of a

**9.** I thank Whitney Tabor and other readers of earlier drafts for bringing to attention an issue regarding how to quantify optimality. I discuss this further in ISSUES WITH FUNCTIONALIST ACCOUNTS OF THE LEXICON. For now, where work reviewed herein specifically claims optimality, I repeat the use of the terminology "(near)-optimal". Otherwise, I revert to less constraining language like "efficient" or "adaptive". The distinction is not critical for the viability of the Functional Lexicon Hypothesis.

**10.** A region is connected if one can travel from any part of the region to any other part without leaving the region. To use a geographical metaphor, the continental USA is connected, but the entire USA is not.

lexicon as its ability to support the listener's reconstruction of the speaker's intended meaning given an uttered term. This quantity of informativeness is high when, for every lexical item, the intended concept (here, spatial relation) is highly similar to the other concepts encodable by the same lexical item.[11] Khetarpal et al. tested whether natural language spatial lexicons have higher informativeness than artificially constructed lexicons that still satisfy the connectedness constraint. Indeed, artificially constructed spatial lexicons had far lower informativeness, measured as described above, than their corresponding naturally occurring spatial lexicons, reinforcing the claim that spatial lexicons are shaped by functional – here, communicative – constraints.

Regier and colleagues have since extended their enterprise to even more abstract semantic domains, including kinship (Kemp & Regier, 2012) and number (Xu & Regier, 2014). Moreover, these more recent works have recognized and incorporated into their analyses the observation that Rosch (1978) initially made, that it is not just informativeness of a lexicon that is important, but also its simplicity, and that these two constructs often trade off – a system with one category for everything would be compact but uninformative, but a system with one category for every thing or every two or three things would be informative but cumbersome. Here we review just the kinship work, but the number work is similar. Kemp and Regier (2012) define complexity as the minimum number of rules (using primitives like PARENT and FEMALE and rules including conjunction, disjunction, and transitive closure)[12] needed to define all the kinship terms in a lexicon. The communicative cost[13] of referring to a specific referent (say, *younger brother*) is defined as the expected number of extra bits needed to fully specify the speaker's intended referent (again, *younger brother*) when one utters a basic kinship term ('brother'), which is determined by the relations covered by a particular term ('brother' covers *younger brother* and *older brother* in English) and the probability of referring to those particular relations (which Kemp & Regier estimated from text corpora).[14]

**11.** Notice the persistence of the importance of high within-category similarity.

**12.** To illustrate transitive closure, consider the following example: the transitive closure of the relation PARENT is the set containing my PARENTS, my PARENTS' PARENTS, and so forth.

**13.** Communicative cost is the inverse of informativity. An informative expression is *not communicatively costly* as it runs *less* risk of miscommunication and requires *less* inference from the listener.

**14.** Because they estimated need probability from bigrams like "my grandmother" and "my daughter", they did not get a direct estimate of the need probabilities for unique, unambiguous relations like paternal grandmother vs maternal grandmother. Hence, for the current work, they made the additional simplifying assumption that the corpus-estimated probability for a term like 'my grandmother' was evenly distributed among the specific referents it covers (see Figure 3C of Kemp and Regier, 2012).

The total communicative cost of a kinship system is the sum of the products of the corpus-estimated need probabilities of particular relations (just the relative frequencies of relations in a corpus) and the costs associated with communicating that specific relation. Complexity and informativity were thus computed for real languages, and compared to the complexity and informativity of hypothetical, permutated languages. The former vastly outperformed the latter, suggesting that kinship term systems achieve a near-optimal trade-off between complexity and informativity, consistent with the functional lexicon hypothesis.

More recently, the most recent work from the Regier group (Regier, Kemp, & Kay, 2015) has seen a single metric of lexicon complexity and a single metric of informativity applied to multiple semantic domains. Of course, this is a much more stringent test of these functionalist ideas. As in Khetarpal et al. (2013) and Kemp and Regier (2012), Regier et al. (2015) conceive informativity as the ability of the listener to reconstruct the speaker's intended message. However, whereas these previous papers treated the speaker's and listener's internal representation of messages as single familial or spatial relations, Regier et al treat these representations as probability distributions to allow for uncertainty in the speaker and listener's representations.[15] Hence, they formalize reconstruction accuracy for a given message as the Kullback-Leibler (KL) divergence between the speaker's distribution and the listener's distribution.[16] As in Kemp and Regier (2012), the informativity of an entire lexicon is then just the sum of the product of need probabilities of particular objects (how often it must be talked about) and the KL divergences for those objects. To measure lexicon complexity, Regier et al. use number of terms for their analyses of color and object[17] lexicons, and the same representation language as in Kemp & Regier (2012) for kinship lexicons. To briefly summarize their results, they find that, using these metrics, human color, kinship, and object lexicons (or at least a small slice of the object lexicon) are (near-)optimal for their respective levels of complexity (again, here just quantified as the number of terms in the lexicon).

---

**15.** It is interesting that they characterize probability distributions as formalization of uncertainty about a referent. While this is sensible, one could also imagine that a speaker/listener is very certain about a meaning, and a probability distribution represents not uncertainty, but rather the specificity of the meaning. That is, a probability distribution with low variance would represent a highly specific concept: highly specific concepts (e.g., golden retriever) do not vary in their instantiation as much as do highly general concepts (e.g., animal).

**16.** KL divergence can be intuitively understood as the loss of information when one distribution (here, the listener's) is used to approximate another (here, the speaker's).

**17.** Objects represented using binary feature vectors, as in "platypus: 1-warm blood, 0-live birth, 1-beak, 0-flies".

Finally, Xu, Regier, and Malt (2015) applied this measure of informativity (based on need probabilities and KL divergence) and simplicity (number of terms) to analyze the tension, in container names (e.g., 'bottle', 'cup'), between communicative efficiency and semantic chaining, the historical process whereby "[a] name for one referent is extended to a conceptually related referent, and from there on to other referents, producing a chain of exemplars that all bear the same name" (pg. 1). The tension just mentioned is that chaining can produce categories where the opposite ends of the chain bear little similarity to each other, contra the principle of communicative efficiency laid out above that requires high within-category similarity. Xu et al. first show, with pile sort data [like the spatial relations pile sort data of Bowerman and Pederson (1992), but with container objects], historical corpora, and three computational models of categorization (one each for chaining, clustering, and majority vote categorization processes),[18] that container names show evidence of chaining in English, Spanish, and Mandarin. Specifically, the chaining model better predicted the diachronic development of extension of container names than did the other two models (although only 3–10% better than the clustering model for the three languages). At the same time, they show that the container lexicons from these languages exhibit higher informativity than the vast majority of hypothetical, non-attested, but comparable (in size and number of categories) container lexicons. Thus, despite the existence of chaining in this domain, and the potential of chaining for reducing communicative efficiency, container lexicons nonetheless find communicatively efficient ways to refer to container objects. Left open, however, is the question of what kind of categorization model captures the empirical synchronic and diachronic facts about *both* chaining *and* efficient categorization. It may be that some chaining-clustering hybrid model is needed, as the chaining model alone would seem to lead to inefficient lexicons, while the clustering model alone does not capture the diachronic facts of chaining. We expand on the broader importance of more work integrating functionalist pressures with other dynamical processes (of which chaining is an example), in the final section *THE EMERGENCE OF FUNCTIONALISM IN THE LEXICON: CURRENT AND FUTURE DIRECTIONS*.

**18.** The chaining model adds a new exemplar to whatever category contains the single exemplar to which the new exemplar is most similar. The clustering model is much like a typical 'exemplar' model of categorization – it assigns a new exemplar to the category whose exemplars are most similar to the new exemplar overall. The majority vote model assigns a target item to the category that has the most exemplars, without reference to any intrinsic relations among exemplar.

## Issues with functionalist accounts of the lexicon

Many of the above claims of functionalism in the lexicon have not gone without criticism. For example, Graff (2012) argued that Piantadosi et al. (2009)'s finding that stressed syllables are more informative for word identity is entirely expected given the phonetic properties of stress. That is, stress itself leads to the preservation of contrasts that would otherwise be neutralized in stressless syllables (Giavazzi, 2010). Notice this is a question of how the lexicon got the way it is, i.e., giving a proximal answer to "why are words the way they are?". Piantadosi et al. (2009)'s account of their finding would seem to be that some syllables were already stressed and others not stressed, and then some optimization mechanism (even if non-teleological) found that putting more phonemic contrasts in stressed syllables allowed more efficient communication. Graff (2012) argues instead that putting more phonemic contrasts in stressed syllables is simply a direct consequence of how stress or lack thereof affects pronunciation, i.e., lack of stress causes underarticulation and hence merger of contrasts. As we will see, this type of criticism – not of the empirical findings but of the presumed, optimizing mechanism that gave rise to them – will be common. As was mentioned at the beginning of the paper, providing a plausible optimization mechanism is a key part of any functional account. See the upcoming section on emergence of functional lexicons for more discussion of this point.

Ferrer i Cancho and Moscoso del Prado Martin (2011) similarly show that Piantadosi et al. (2011)'s finding of a positive relationship between word length and word information content (conditional surprisal) can be generated by the Miller (1957) random typing models (also called monkey models). In other words, they argue, "a linear correlation between information content and word length may simply arise internally, from the units making a word (e.g., letters) and not necessarily from the interplay between words and their context as suggested in [Piantadosi et al, 2011])". Moscoso del Prado Martin (2013) similarly argues that Piantadosi, Tily, and Gibson (2012)'s finding of preferential reuse of 'easy' linguistic units is expected in a purely random system. In fact, Moscoso del Prado Martin shows that such effects found in a corpus are weaker than those found in a purely random system, calling into question an optimization account of re-use of 'easy' linguistic units.

Piantadosi, Tily, and Gibson (2013) have since made several counterarguments to Ferrer i Cancho and Moscoso del Prado Martin (2011), though not yet to Moscoso del Prado Martin (2013). Three are particularly worth mentioning. First, they point out that monkey models are not a realistic model of language production: it is clearly not the case that we utter a sentence by randomly selecting a single character/phoneme at a time. However, there is an alternative interpretation of Ferrer i Cancho and Moscoso del Prado Martin (2011)'s point, one that seems to be implicit in much of the monkey model work downplaying the importance of Zipf's Law and related findings. That is, monkey models may just be a demonstration that an exceedingly simple, baseline process like them – even if monkey models are ultimately not the 'right' or accurate process – can give rise to Zipf's Law and related phenomena, and hence such phenomena are probably not reflective of any interesting optimization process. Regardless of whether this alternative interpretation was intended, additional rebuttals by Piantadosi et al. (2013) are not so easy to dismiss. For one, a large body of behavioral work suggests that people *do* reduce predictable content (e.g., Mahowald et al., 2013; see Jaeger and Tily 2011 for review). Additionally, and perhaps more importantly, recall that Piantadosi et al. (2011) found that information content predicted word length *better* than did frequency, a result that monkey models do not predict; in Ferrer i Cancho and Moscoso del Prado Martin (2011)'s monkey model, frequency and information content are mathematically identical. For these reasons, Piantadosi et al. (2011)'s original hypothesis that some communication optimization process shapes word lengths based on information content still seems plausible.

Piantadosi (2014) has also leveled criticisms against Ferrer i Cancho and Sole (2003)'s account of Zipf's law of word frequencies. He points out that they assume that all meanings are equally frequent, which seems likely to be untrue (we encounter and hence talk about aardvarks in the world far less often than we do cats). Perhaps more importantly, it assumes, without empirical or logical motivation, that the speakers' difficulty is proportional to the entropy in the set of signals. A more reasonable *a priori* choice might have been the conditional entropy in the set of signals given a particular meaning. This has a more intuitive appeal, as it seems to fit a psychological model of word production more: one conceives of a meaning, and then picks a signal/form that is linked to that meaning – one doesn't select a signal from the entire set of signals. As it happens, Ferrer i Cancho and Sole (2003) show that substituting the conditional entropy in signals given a meaning into their cost function removes the key result of Zipf's Law of word frequencies due to optimization of the cost function. Thus, their and Zipf's functional explanation of Zipf's (power) Law of word frequencies may be suspect.

Functional lexicon accounts may also be subject to criticisms leveled at rational, probabilistic accounts of cognition more generally (since functional lexicon accounts typically involve some rational and probabilistic model of language use). For example, Marcus and Davis (2013) show that such accounts are often plagued by two related problems: task selection and model selection. The problem of task selection can generally be understood as the problem of generalizability of an account/model to different tasks in a domain. For example, Battaglia, Hamrick and Tenenbaum (2013) showed that a model of probabilistic inference closely predicted human predictions of stability of a computer-generated block tower, leading Battaglia et al. to conclude that "[i]ntuitive physical judgments can be viewed as a form of probabilistic inference over the principles of Newtonian mechanics" (p. 5). However, when Marcus and David (2013) re-applied that same model to human behavior on the balance beam problem, which involves the exact same physical principles, the model predicted correct inferences, whereas humans perform woefully on that task. An analogous issue in functional lexicon accounts might be how well specific formalizations of optimality tested on, for example, kinship terms (Kemp and Regier, 2012), but *intended to be domain-general*, extend to different semantic domain like emotions or smells. As discussed, Regier et al. (2015) have already noticed this problem to an extent, but clearly work on functionalism in the lexicon could do with broader application of each model to various tasks/domains (see upcoming section on other ways to assess functionalism for more discussion of this point).

The model selection problem, in contrast, concerns how one specifies the components of one's model, including the choice of probabilities of events, which can come from three sources: (1) real world frequencies, (2) experimental participants' judgments, and (3) the assumptions that feed into mathematical models, e.g., that data are Gaussian vs Poisson distributed. Seemingly innocuous but erroneous estimation of probabilities by one of these methods can lead to surprisingly flawed model predictions. For example, Griffiths and Tenenbaum (2006) asked subjects to predict a movie's total gross given that it grossed X dollars after an unknown period of time. Griffiths and Tenenbaum's model made the erroneous assumption that a movie's gross is uniformly distributed across time, when in fact a movie's gross is heavily front-loaded. Updating their model with this corrected assumption would have led to rather different model predictions and hence rather different conclusions, namely, that participants overestimated future movie earnings, and that their reasoning was not optimal.

Again, similar problems may face functional lexicon accounts. Several of the arguments rely on estimation of real-world probabilities: for example, Kemp and Regier (2012)'s argument of optimality of kinship terms relies on corpus-based estimates of the probability of needing to refer to a particular familial relation. Might the need probabilities, and hence the estimates of optimality, have varied if a different corpus were used (Kemp and Regier admit this point themselves, and say it might even be a point of cross-cultural/linguistic variation)? Worse, might the method of estimating need probability based on corpus frequencies be inherently problematic? For example, given that kinship terms are ambiguous – English 'brother' can mean *older brother* or *younger brother* – corpus frequencies can't possibly assign different need probabilities for the relations covered by a single term (and as mentioned earlier, Kemp and Regier do in fact assign the same need probability to younger and older brother, maternal and paternal grandmother, etc.; see Figure 3C in their paper).

Another related statistical issue was raised in footnote 9, regarding how to quantify and assess lexicons for optimality. The basic result recurring throughout, for example, the Regier line of work is that naturally attested lexicons are more optimal than some surrogate data control cases, but that does not imply they are (near-)optimal. To address near-optimality, one usually establishes an upper bound formally. Certain data figures (e.g., Figure 10 in Regier et al., 2015) make an informal case that attested lexicons exist on this optimality frontier, but the Regier group do not do any rigorous statistical test to show that attested systems are not significantly different from the theoretical optimum. And, if they were to do such an analysis and find that attested systems *are* significantly different from the optimum, it's not clear what would count as *near* optimal.

Other fundamental issues concerning the nature of word meanings and category systems have not quite been addressed or incorporated in functionalist lexicon work. First, much of the preceding work treats concepts and word meanings as being rigid in boundaries, and unchanging across contexts. However, these assumptions have been challenged since at least Wittgenstein (1953). First, Wittgenstein pointed out that, even for a mundane concept like 'game', it is nigh-impossible to list the necessary and sufficient features that cover all examples of 'game' while excluding nongame sports (like hunting) or something like rotely kicking a ball against a wall. Instead, Wittgenstein suggested that different examples of a concept bear a "family resemblance" to each other, "a complicated network of similarities overlapping and criss-crossing" (Wittgeinstein, 1953, p. 66; the similarities here being things like *being fun* or *having rules*). In this sense, there is no rigid boundary for the concept 'game' (or possibly any other, even 'technical' concepts; see Murphy, 2002), where one side of the boundary contains 'games' and the other contains 'not-games': different things can be more or less 'game'-like, and in different ways. Since Wittgenstein (1953), cognitive scientists have either endorsed this nonclassical view of concepts (Rosch, 1978; Murphy, 2002) or admitted the above and related deep problems of the classical, essentialist view of concepts (Margolis & Laurence, 2014).

A second, related issue that has almost entirely remained unaddressed is the context-dependent nature of word meanings. That is, even accepting fuzziness of category boundaries per Wittgenstein (1953), what is considered a 'game', indeed the meaning of 'game', could change from context to context (Casasanto & Lupyan, 2015; Yee & Thompson-Schill, 2016). For example, six feet may be a good example of 'tall' for a person, but not for an elephant. This point, too, was suggested by Wittgenstein in his idea of 'language games'. The idea of language games is that all language is used for some particular purpose. Wittgenstein gives the example of some workers who only use four words: *slab, block, pillar*, and *beam*. When one worker says *beam* and points at a beam, another worker knows he means: "bring me that beam", or "stand that beam upright", or whatever else is appropriate in that situation. However, it is not that "bring me that beam" is the true meaning of "beam." In the context, just saying "beam" is enough, and the long, propositional form is unnecessary and possibly odd in that actual situation. In other words, the context or purpose determines the rules of the language-game, and the meaning of the words uttered.[19] Interestingly, Rosch (who wrote her undergraduate honors thesis on Wittgenstein, 1953),[20] one of the first to discuss functionalism in the lexicon/category systems, also raised the issue of context-sensitivity of meaning, although she did not suggest that it presented a problem for functionalist analyses (Rosch, 1978). It is still not clear whether current functionalist work, which has almost always treated word meaning as having carefully delineated, static boundaries (Regier et al., 2015 being an exception, where word meanings are probability distributions), simply provides an acceptable approximation to the true nature of word meanings, or whether ignoring the fuzziness and context-sensitivity of word meanings leads to deeply flawed conclusions about functionalism in the lexicon.[21]

Finally, for functionalist accounts of lexical semantic fields, one's account is only as good as one's characterization of the semantic domain. In the case of color, a relatively low-level "perceptual" domain whose structure is well-understood (saturation, hue, lightness), the characterization that Regier et al. (2007) have developed seems suitable enough. For more complex and abstract domains, however, it is less clear. For example, as discussed, Khetarpal et al. (2009, 2013) used pairwise similarity data to reconstruct a semantic space for spatial terms. What exactly is the nature of this space? Do its dimensions have a meaningful interpretation? Is a continuous space even the right way to conceptualize the meanings of spatial relations? A space was not Regier and company's preferred mode of representation for kinship terms; instead they opted for a representation language with primitives and rules for combining them. Perhaps the same might be suitable for spatial relations? But perhaps conceptual/semantic spaces are at least a part of the story. Gärdenfors (2000), in his book *Conceptual Spaces*, discusses a geometric semantics of spatial

**19.** One could possibly trace these ideas even earlier in Wittgenstein's writings, to his *Tractatus Logico-Philosophicus* (1921/1922), "Only the proposition has sense; only in the context of a proposition has a name meaning" (*TLP*, 3.3).

**20.** I thank an anonymous reviewer for alerting me to this history.

**21.** Interestingly, the fuzziness and context-sensitivity of our concepts might actually be thought of *as functionalist*. That is, one might argue that the world we live in, the one that shapes our concepts and word meanings, is itself not neatly carved, but rather fuzzy, and that the relevant nature or identity of a thing changes from context to context. See Murphy (2002, pg. 21, box "The Necessity of Category Fuzziness") for a tentative development of this idea.

prepositions developed by Zwarts (1995), according to which spatial prepositions map a reference object to a *region* relative to an anchor object. However, Gärdenfors (2000) points out that the meaning of spatial prepositions can not simply reduce to a region relative to an anchor object; for example, for the prepositions "throughout", "about", and "over" (as in covering), besides the reference object being in a particular region relative to the anchor object, the anchor object must be distributed over or extended all over the reference object. So perhaps the proper conceptualization of spatial relations requires a semantic space – spatial relations are *about* physical space, after all – and augmentation from, for example, logical/set-theoretic rules as in Kemp and Regier (2012). Clearly, the proper conceptualization of the meaning of spatial terms is not a simple matter, but more sophisticated functionalist accounts of spatial lexicons will eventually have to begin with such an improved conceptualization.

## Future directions – additional ways to assess functionalism in the lexicon

The first two sections of the paper marshaled a wide variety of evidence that the semantic and phonological organization in natural human lexicons is well-suited for human needs; that is, that the lexicon possesses structure that could aid communication, learning, etc. However, it is clear that more work remains to be done, particularly in light of Marcus and Davis's (2013) finding that certain models of rationality or optimality in cognition crumble under even the most minor of extensions to similar tasks or domains. In this section, then, we discuss how future work investigating functionalism in the lexicon might proceed.

On the semantic side of the lexicon, perhaps the clearest direction to follow, indeed one noticed upon by nearly everyone in the field, is to explore functionalism in the organization of new semantic domains. Cross-linguistic data exists on the organization for many domains, including odor (Majid & Burenhult, 2014; Wnuk & Majid, 2014 even has the pairwise similarity data needed for, e.g., Regier et al. (2007)'s analysis), emotions (Boster, 2005), artifacts (Malt, Sloman, Genarri, Shi & Wang, 1999), body parts (Majid, Enfield, & van Staden, 2006), plants and animals (Berlin, 1992; Atran, 1995). While investigation spreads to additional domains, a second important question can begin to be answered. That is, while the principles of efficient semantic organization reviewed here are largely domain-general, there also may be domain-specific principles, e.g., the unevenness of color perceptual space (Regier et al., 2007), impacting what exactly makes for efficient semantic organization. As more domains are explored, investigation of such domain-specific principles can begin in earnest.

The investigation of novel domains may also be aided by recruiting additional techniques for characterizing the structure of those domains. A few methods stand out. First, there are many corpus-based methods for reconstructing semantic spaces, perhaps the best known among these being Latent Semantic Analysis (LSA; Landauer, Foltz, & Laham, 1998) and probabilistic alternatives like Latent Dirichlet Allocation (Griffiths, Steyvers, & Tenenbaum, 2007; Blei, 2012). In essence, this approach counts word-word or word-document co-occurrences in a corpus to yield a high-dimensional, vectorial representation of a word's meaning, such that intuitively similar words (say, 'cat' and 'dog') occupy nearby points in space. Just as Wedel et al. (2013) found that a contrast's count of minimal pairs in the same syntactic class improved prediction of the likelihood of that contrast's merger, the count of minimal pairs within a certain threshold of distance in LSA space may likewise improve prediction of merger. However, LSA may be somewhat limited in its applications, as it treats word meanings as points in space, whereas much of the literature reviewed above relies on a conception of word meanings as covering regions in a space, or at least having some kind of referential range (e.g., recall the Regier group's work where informativity relied on the overlap between the speaker's distribution over a range of possible references and the listener's distribution over the same range). A recent extension of LSA by Richie, Kauffman, and Tabor (2014) resolves this. In their approach, a word type covers a region in space by virtue of the distribution of its *tokens* in that space. Basically, word-word or word-document vectors (we refer to them as 'type vectors') are first calculated as in traditional LSA. As usual, these type vectors can be understood as reflecting the general or typical meaning of a word. Then, in a second pass through a corpus, each *token* of a word is associated with its own 'context vector', simply the sum of the type vectors of the words in its context (in our work, a moving 15-word window around the token). The collection of context vectors for a particular word type will have a particular distribution and particular relations – most relevant here, overlap with other collections of context vectors. These computed overlap relations enabled by Richie et al.'s approach thus may yield reconstructions of semantic space that are amenable to the investigations that Regier and others have undertaken (which, again, often rely on notions of overlap between the meaning the speaker intended and the meaning the listener understood).

There are a few benefits of such corpus-based approaches. First, this approach could be used for any language for which there exist large enough corpora – it does not require careful, labor-intensive experimentation to yield, for example, pairwise similarity data for terms in the lexicon. Because of this, it could be used to examine functionalism in a greater number of semantic domains. Second, it could be used to study the change in lexical-semantic fields over time by using corpora from different time periods. Indeed, Sagi, Kaufmann, and Clark (2011) used a similar method to study the narrowing of terms like 'deer' (which at an earlier point meant 'animal') and the broadening of terms like 'dog' (which used to mean 'a (specific) powerful breed of dog'). Such data could be critical for testing different accounts of the emergence of functionalism in lexical-semantic fields.

The second set of tools that may be useful for characterizing the structure of lexicons comes from graph theory and network science, which have been used to characterize the structure of both lexical-semantic networks (i.e., networks where semantically-related words like 'cat' and 'dog' are connected; Steyvers & Tenenbaum, 2005) and phonological networks (i.e., networks where phonologically-related words like 'cat' and 'bat' are connected; Vitevitch & Goldstein, 2014), as well as the dynamics involving such networks, including growth (Hills, Maouene, Riordan, Sheya, & Smith, 2010; Hills, Maouene, Maouene, Sheya, & Smith, 2009) and search (Griffiths, Steyvers, & Firl, 2007; Thompson & Kello, 2014). The semantic networks reconstructed by, for example, Steyvers and Tenenbaum (2005) and Hills et al. (2009, 2010), may be amenable to analyses of functionalism like those presented. For example, one principle of functional organization that recurred was that a category system should have high within-category similarity, and low between-category similarity. High within-category similarity may have a network equivalent in the property of clustering, or the extent to which neighboring nodes A and B share their other neighbors. A category (say, 'animal') whose members ('dog', 'cat', 'fish') cluster together would be the same one whose members are all similar to each other.

The network approach may also afford completely novel analyses of functionalism. For example, network science offers a way to understand/measure robustness of networks when the networks are damaged (i.e., when nodes are removed). One way in which network structure is quantified is in its *mixing*, or how node properties affect how nodes are connected. In a network exhibiting *assortative mixing by degree*, for example, high degree nodes (those with a lot of connections) tend to connect to other high degree nodes. Newman (2002) found that removing nodes with a high-degree in networks with *dis*assortative mixing by degree (where high degree nodes connect to low degree nodes) greatly disrupted the ability to traverse a path from one node to another node in the system. In contrast, network traversal was not disrupted as much when high-degree nodes were removed from a network with assortative mixing by degree. In other words, networks with assortative mixing by degree are able to remain relatively connected in the face of targeted attacks to the system. How might this be relevant to the lexicon? For one, it is well-known that patients with brain damage sometimes exhibit semantically-selective deficits, having more difficulty, for example, with naming pictures of animals compared to pictures of fruit/vegetables or artifacts (Caramazza & Mahon, 2003). One interpretation of such results is that certain nodes or clusters of nodes in these patients' semantic networks are impaired somehow. Network science could thus enable us to determine whether (lexical-)semantic networks are structured such that they can withstand damage due to brain insult better than possible alternative semantic networks. Similarly, Kleinberg (2000) found that only certain small-world networks (networks with short paths from any two nodes, but also have high clustering of nodes) enabled efficient navigability – finding a short path from node A to node B – based on a simple search algorithm. It may turn out that human (lexical-)semantic networks similarly enable efficient navigation from concept to concept. In sum, then, a network approach may reveal that the structure of lexical networks facilitates certain psycholinguistic processes.

Related to the above point of seeking novel characterizations of the semantic and phonological lexicons, another way in which investigations of functionalism might be extended is by expanding or refining notions of functionalism. For example, current operationalizations of effort focus solely on, for example, how many rules are needed to represent the meanings of all kinship terms (Kemp & Regier, 2012) or numeral terms (both exact, like 'three', and approximate, like 'handful'; Xu & Regier, 2014) in a language. The assumption these works make is that the more rules a system requires, the harder it is to learn and store the lexical system. But perhaps not all rules are created equal – it seems reasonable that, at least in the case of kinship relations, conjunction is a more complex rule than transitive closure – or maybe vice-versa, the point being that rule complexity may be as important as number of rules. And perhaps there is some cost to using a diversity of rules rather than re-using a more limited number of rules repeatedly. That is, it seems plausible that learning a new term that makes use of often used rules will be easier than learning a new term that makes use of a novel rule. Moreover, it seems plausible that to the extent that human memory utilizes compression (Brashears, 2013), a system that distributes use of rules across kinship terms more uniformly will not compress as much as, and hence require more storage capacity than, a system that distributes rules less uniformly.

Other seemingly important aspects of language use have largely gone unanalyzed in the functional lexicon literature. Much of the above work focuses on informativity and effort in learning and storage, but clearly other functions are important, functions like memory retrieval during language production. Ferrer i Cancho and Sole (2003) are the only authors to seriously consider the effort involved in retrieval, and, as discussed in the last section, their operationalization of this difficulty is at best unmotivated, and at worst unrealistic. Future work would do well to consider the costs of retrieval, and operationalize such costs with a realistic and empirically motivated function like conditional entropy in signals given a meaning (as in production), as argued by Piantadosi (2014) and discussed in ISSUES WITH FUNCTIONALIST ACCOUNTS OF THE LEXICON.

## The emergence of functionalism in the lexicon: Current and future directions

Recall that functionalist explanations for language structure, as they have been described so far, are really distal causes or answers to the question "Why do human languages have the words they have?" Thus, they are not a full account of why the lexicon is the way it is. Instead, as with most "why" questions, there are additional explanations that are more proximal to the end result. In the final section of this paper, we consider such proximal causes or answers. The general form of these answers in language and cognition will be the same as they are in evolutionary biology: "Everything is the way it is because it got that way" (Thompson, 1917). Thus, we will consider here processes of language use, learning, and evolution that could give rise to functional lexicons.

Before considering possible processes like this, it is worth emphasizing again the importance of this enterprise. As Haspelmath (1999) points out, much skepticism of functionalist accounts stems from what non-functionalists see as a 'teleology fallacy', that identification of a functionalist explanation for a language structure is sufficient explanation of its origins. As just discussed, this is not the case. Further, as seen in the section *ISSUES WITH FUNCTIONALISM*, much of the criticism against functionalist accounts (e.g., Ferrer i Cancho & Sole, 2012 against Piantadosi et al., 2011; Graff (2012) against Piantadosi et al., 2009) is not against claims about how the lexicon is organized synchronically (i.e., at a given point in time), but rather with how it emerged diachronically (over time). That is, did the lexicon emerge via some adaptation mechanism sensitive to a particular function, consistent with a functionalist account, or as a "happy accident", inconsistent with a functionalist account. To illustrate this last point further, we can consider a study by Kurumada, Meylan, and Frank (2013). They found that language learners have an easier time segmenting words from continuous speech when those words follow a Zipfian distribution of word frequency, as compared to a uniform distribution (typically used in such artificial language learning experiments). The idea is that the handful of highly frequent words in the Zipfian distribution give the learner a way to 'break in' to the speech stream (see also Bortfeld, Morgan, Golinkoff, & Rathbun, 2005). Should we conclude that words follow a Zipfian distribution *because* such a distribution aids word segmentation (an ultimate cause)? Probably not: word frequencies are likely much more sensitive to the semantic and syntactic fit of a word to the message and sentence one wants to express (Piantadosi, 2014). Hence, it is difficult to imagine a psychologically plausible mechanism of word segmentation, learning, and use whereby the needs of word segmentation could exert an influence on word frequencies. In other words, the benefit for word segmentation that language learners get from Zipfian distributions of word frequency is likely what I have called a "happy accident".

The foregoing discussion should make clear why identifying mechanisms of functional lexicon emergence is of the utmost importance. To summarize, arguing that a particular property of the lexicon is the way it is because it is beneficial for some human need (e.g., uninformative/frequent words are short because this minimizes average message length, i.e., speaker effort) requires identifying a mechanism of language emergence or change that is sensitive to pressures for optimizing that particular structure (e.g., maybe over repeated uses of a word, language users keep reducing the word's length until communication fails or becomes difficult).[22] We now turn to considering in earnest how functional lexicons may have emerged.

As of the writing of this paper, this is a question that has usually only been briefly mentioned at the end of papers. Only a few authors – mostly those investigating functionalism in the phonological forms in the lexicon – have seriously considered how functional lexicons may have emerged. Piantadosi et al. (2011) suggest that the positive relationship between information content and word length may be due to lexicalization of routine reduction of predictable content. That is, it is already well-documented that language producers shorten predictable content (Fowler and Housum 1987; Jaeger and Tily 2011 for review). For example, Mahowald et al. (2013) showed that speakers use shortened synonyms ('info' instead of 'information', 'math' instead of 'mathematics') in contexts where the word is predictable. If such shortening is done enough, the form of the word may become shortened in the lexicon, rather than in the moment of production. Despite the intuitive appeal of such an account, it unfortunately is currently somewhat vague, being a purely verbal model, and to date it has escaped more explicit treatment by, for example, computational modeling.

Modeling has been somewhat successful with other functionalist phenomena. For example, Graff (2012) described a spreading activation model of word production, based on Martin (2007)'s, that could account for phenomena of lexical contrast maintenance (of which his finding that perceptible contrasts have more minimal pairs is an example, as is Wedel et al. (2013)'s that sound contrasts with many minimal pairs resist merger). In this model, there is a layer for phonemes (e.g., a node for /c/, a node for /a/, a node for /t/), a layer for word-forms (e.g., a node for 'cat'), and a layer for concepts (e.g., a node for 'furry four-legged feline'). To speak, the model activates a concept, and activation spreads to the corresponding words at the word layer, and then to phonemes corresponding to these words. Additionally, in Graff's implementation, activation can flow back from the phoneme layer to the word layer, and the words in the word layer laterally inhibit each

22. This is not meant to be the single, 'right' theory of reduction – it is merely intended to give a flavor of a functional characteristic and how it could emerge.

other, such that highly active words strongly inhibit all other words. To simulate lexicon evolution, every generation, synonyms for existing concepts are added and linked up to the network at the beginning of the generation, a concept is activated at random, and activation consequently flows throughout the network, and, finally, the synonyms with the weakest activation after activation of their respective concept are discarded at the end of each generation. Simulations of 1000 generations like this generated lexicons with low average similarity between words. Simulations without lateral inhibition between words, however, generated high average similarity between words. Why does lateral inhibition have this effect? When a target word that is similar to many other words is activated, the target word activates its phonemes, which activate these many other, similar words in the lexicon, which then laterally inhibit the target word, reducing its activation, and making it more likely to be discarded at the end of a generation.

Graff (2012)'s model is useful for understanding the general architecture of a possible lexical contrast mechanism, although it does not explain his specific finding that perceptible contrasts have more minimal pairs, as all the phonemes in his model are essentially the same – they are just numbered nodes. Nor does the model explain Wedel and company's results – the model does not even allow for two different phonemes to merge into one. One way to simulate Graff's finding, then, might be to add a distinctive feature layer beyond the phoneme layer, and allow feedback from the feature layer to the phoneme layer. Confusable sounds tend to have similar articulations: the highly confusable /f/ and /θ/ are both [−voice] and [+continuant]; /m/ and /n/ are both [+voice] and [+nasal]. If a language had a confusable minimal pair like 'ram' and 'ran', then activation of a word like 'ram' would feed all the way down to the feature layer, and back up to minimal pairs like 'ran' that share features, inhibiting 'ram' and possibly leading to its death in the language. But if a language had a non-confusable minimal pair like 'ram' and 'rack', then activation of 'ram' should not activate 'rack' as much, leading to less inhibition of 'ram' from 'rack', and the continued coexistence of that minimal pair.

Notice that the mechanism of lexicon evolution in Graff's model is *purely speaker-driven*. That is, lexical contrast maintenance in his account is just driven by the word production process. This should come as somewhat surprising given Graff's claim that his minimal pair finding is evidence of *communicative efficiency*, as the model really has no communication in it, no interaction between a speaker and a listener. If the minimal pair finding were really evidence of *optimization* of communicative efficiency per se, then we should expect the mechanism of emergence to involve something like language users being less likely to continue using the word forms that get misunderstood, which would be those words with phonological neighbors differing by more confusable sound contrasts (e.g., if 'ran' already existed in the lexicon, then 'ram' would be more likely to be misheard and discarded from the lexicon than would 'rack'). Graff thus may have unintentionally undermined his own account of communicative efficiency.

While Graff (2012)'s model only emulated the behavior of a single individual, naturalistic language use and learning occurs in the interactions between two or more individuals. Emulating this interaction is the domain of *agent-based modeling* (ABM). Use of agent-based modeling could thus be useful for understanding the processes of language use and learning that give rise to functional lexicons. Indeed, a few authors have used ABM's to tackle this issue (Hurford, 2003; Zuidema & Westermann, 2003). However, this work has been of somewhat limited utility so far. Hurford (2003) attempted to use ABM's to explain why synonymy is rare, but homonymy (two words with the same pronunciation, but different meaning) is common, but he provides only a handful of anecdotes rather than a substantial empirical case. To the contrary, the work reviewed in the first section of this paper shows that language users *do* avoid homonymy somewhat, resisting merging sound contrasts that would create too much homonymy (Wedel, Jackson, & Kaplan, 2013; Wedel, Kaplan, & Jackson, 2013). And recent work by Xu and Kemp (2015), evaluating the Google Million corpus from the 1890's to the 1990's, shows that synonym pairs from the 1890's tended to change in *parallel* ways, staying similar in meaning. In other words, contrary to Hurford (2003)'s claims, empirical evidence for synonymy avoidance is lacking. (At least, this is true on diachronic timescales. On ontogenetic timescales, however, the well-known mutual exclusivity constraint exerts its influence, leading children to prefer not to map new words to meanings for which they already have a word (Markman & Wachtel, 1988). This is essentially a synonymy avoidance mechanism.)

Zuidema & Westermann (2003) attempted something similar to Hurford (2003). They first formalized a utility of communicative success based on (1) form-meaning mapping matrices for production and comprehension, (2) form-form confusion matrices, and (3) meaning-meaning value matrices (how valuable is it that meaning $m$ was understood when meaning $m^*$ was intended). A hill-climbing algorithm, which randomly changes entries in a random matrix of a random agent, and then accepts the change if it yields higher communicative success, led to lexicons shared throughout the population with one-to-one mappings of form and meaning, with forms that were maximally distinctive from each other (allowing easy discrimination). Such lexicons do indeed have *some* basis in real human lexicons: as just discussed, (limited) homonymy avoidance is well-documented, although the existence and nature of synonymy avoidance is less certain. They then attempted to explain how actual individual events of language use could lead to such lexicons. They show that such lexicons do emerge if agents repeatedly pair up and make one step through the hill-climbing algorithm. Of course, real humans don't have access to the matrices in other humans' heads, so they can't employ the hill-climbing algorithm in this way. Zuidema and Westermann (2003) claim that they have obtained similar results with more realistic constraints on communication, where agents update their lexicons only on the basis of knowing whether communication succeeded or failed, but they give extremely few details on these simulations. Thus, we are left with little additional insight into the real processes of emergence of functional lexicons.

Other agent-based modeling work may have some relevance for explaining the emergence of functionalism in lexical semantics. Much of this work has focused on studying the emergence of color lexicons (Steels & Belpaeme, 2005; Baronchelli, Gong, Puglisi, & Loreto, 2010; Loreto, Mukherjee, & Tria, 2012). In Baronchelli et al. (2010) and Loreto et al. (2012), agents must carve up the perceptual color space into perceptual categories, label one or more perceptual categories into a single linguistic category, and then communicate to another agent which of two or more colors in a scene they are talking about. In addition, agents were endowed with the human Just Noticeable Difference (JND), an empirically-derived function of how close in wavelength two colors can be and still be discriminated by a human perceiver, based on the wavelength of the two colors. Under these relatively simple conditions, Baronchelli et al. (2010) found remarkable agreement between real human color lexicons and simulated lexicons. Specifically, the color categories that emerged in simulations with the human JND were significantly less dispersed than those that emerged in simulations with a uniform, non-human JND, and this difference agreed quantitatively with that observed in the World Color Survey, between real human lexicons and randomly rotated lexicons (rotations like those in Regier et al., 2007). Loreto et al. (2012) found that the time a population of agents required to conventionalize a color name in a particular region of the color space corresponded with the well-known implicational hierarchy of human color lexicons: if a population has a name for red, it also has a name for black and white (but not vice versa), if it has a name for green, it also has a name for red (but not vice versa), and so on. In other words, the model conventionalized terms fastest for color terms at the top of the cross-linguistically attested implicational hierarchy. In sum, the Baronchelli et al. (2010) and Loreto et al. (2012) models reproduce human color lexicons remarkably accurately.

Might the color lexicons produced in these models show (near-)optimal carving of color space as Regier et al. (2007) showed human color lexicons do? Further, recall that human lexicons can differ in the number of color terms, yet still be well-formed category systems. Do the models show such variability in the (near)-optimal carvings they attain? Still further, can the models show transitions from one efficient carving to another (perhaps due to contact with other languages)? As yet, modeling has not taken these directions, making these questions ripe for future work. However, before this is possible (at least) one issue must be resolved. That is, the Baronchelli et al. (2010) and Loreto et al. (2012) model uses a uni-dimensional color spectrum based just on wavelength, whereas the lexicon data in Regier et al. are two-dimensional, based on hue and lightness. The model may thus have to be expanded to handle two-dimensional color representation.

More recently, experimental work has begun to investigate how communication may lead to efficient lexicons. Or, rather, how it might *not*. Silvey, Kirby, and Smith (2013) and Thompson, Silvey, Kirby, and Smith (2014) asked one group of participants to label (using novel words) items drawn from a novel, arbitrary continuous semantic space individually, and asked another group of participants to communicate back and forth about the same set of items using novel words. They found that, even though the participants in the individual condition had few constraints on how they categorized the items (other than their own memory and the similarity space of the items, which affected the communication condition as well), their categorization systems had higher within-category similarity than the systems of the communication condition. Put another way, even though the communication condition – in contrast to the individual condition – would have *benefited* from imposing a labeling that maximized within-category similarity, the categorization systems developed in the communication condition nonetheless did not maximize this quantity.

Work similar to the foregoing work on ABM's and experiments of communication has investigated how the process of iterated cultural transmission and learning transforms language, often leading to 'the appearance of design without a designer' (e.g., Kirby, Dowman, Griffiths, 2007; Kirby, Cornish, & Smith, 2008; Xu, Dowman, & Griffiths, 2013). For example, Kirby et al. (2008) trained an initial generation of adults on an uncompositional artificial language, and these adults then transmitted the language to a second generation, who transmitted it to a third, and so on. Critically, each generation did not interact directly with the next – each generation simply named objects on a computer screen, and their recorded responses were later shown to the next generation – and nor did participants know that their responses would be used to train subsequent participants. Hence, communication was largely excised from the experiment. Through this experimental 'whisper down the alley', learners unknowingly endowed the language with compositionality, allowing it to be both learnable (cf. a language with different, holistic forms for every possible combination of meanings) and expressive (cf. a language with a single form that can mean anything). Of course, this unintentional, *non-teleological* mechanism of cultural language evolution sounds exactly like a likely mechanism of emergence of functional lexicons. It may be, then, that functional lexicons emerge through the process of repeated cultural transmission and learning, *without any need for communication per se*. Thus, any structure in the lexicon that was found to be efficient for communication, but arose through pure iterated learning (without 'communication' between participants), could not be said to exist *because* it is efficient for *communication*. Notice that this is still a functional explanation, albeit one appealing to the needs of learning and transmission rather than communication. To test this explanation, experiments similar to the above could be conducted, where instead of training participants on uncompositional artificial languages, participants could be trained on lexicons with suboptimal lexical semantics, or lexicons with suboptimal phonological forms. The Regier group is now working along these lines (Carstensen et al., 2015). Indeed, Carstensen et al. have found that iterated learning transforms randomly partitioned color and spatial relation lexicons into lexicons of greater informativeness (according to the measure of informativeness laid out in Regier et al., 2015) and similarity to real color and spatial relations lexica.

One final possible hypothesis of the origins of functional lexicons is worth discussion. An influential theory of categorization, *prototype theory*, holds that items are categorized on the basis of their similarity to a category's prototype (e.g., Posner & Keele, 1968, 1970; Reed, 1972; Rosch, 1973; Rosch, Simpson, & Miller, 1976; Gärdenfors, 2000). Gärdenfors (2000) pointed out that, if categorization follows this principle, such that any item belongs to the same category as the *closest* prototype, it can be shown that this rule will generate a partitioning of the semantic space into a so-called *Voronoi tessellation*. If the space is metric (like the familiar Euclidean space), then the partitions will be convex, that is, all points on the line segment connecting any two points in a partition will also be in the same partition. Notice that such convex regions are not merely connected, but also relatively compact, i.e., the similarity between items within a category/convex region is high.[23] Recall that this is an often-invoked property of functional category systems and lexicons (Rosch, 1978; Regier et al., 2007; Khetarpal et al. 2013). Thus, it may be that this particular functional aspect of lexicons follows not from pressures for utilitarian communication or categorization, but merely from the basic operation of categorization itself, making this aspect of lexicons a 'happy accident'. The results of the Silvey et al. (2013) and the Thompson et al. (2014) mentioned above may actually support this hypothesis, as they found that mere repeated categorization by an individual generated more efficient category systems than pairs of communicating individuals.

We've discussed so far three different hypotheses of the emergence of functional lexicons: emergence through (1) communication (which the ABM literature has typically focused on), (2) cultural transmission and learning (the hypothesis investigated by the iterated learning literature), and (3) the basic operation of

23. Admittedly, it is still possible to have long and thin but convex categories, which are not particularly compact, but convexity still entails more compactness than mere connectedness, which was the alternative hypothetical principle of category systems that Khetarpal et al. (2013) investigated, and which was insufficient for explaining the structure of spatial relations lexicons.

categorization (according to prototype theory). How could we discriminate between these hypotheses? First, it is important to bear in mind that they may not all be mutually exclusive: active pressures from all three processes may converge on the same kinds of lexicons. However, they may make some different predictions. For one, if the basic categorization principle is the only responsible factor, then we would expect that functional category systems and lexicons should emerge in the absence of communication or iterated learning (as in the Silvey et al., 2013 and the Thompson et al., 2014), and we should observe no hastened emergence of functionalism when communication or iterated learning is added in.

It is important to keep in mind the limitations of the approaches primarily used by the ABM and iterated learning literatures: computational models, artificial language learning experiments, and experiments where participants must invent novel communication systems to solve a joint action task (for reviews of the modeling literature, see Steels, 2011; for reviews of the experimental literatures, see Scott-Phillips & Kirby, 2010 or Galantucci, Garrod, & Roberts, 2012). As it stands, these approaches, individually or as a group, are somewhat limited in various ways in their ability to give solid evidence for one theory or another. The experimental literature necessarily works with impoverished languages, and usually with adults; showing that functionalism emerges in such languages is a far cry from showing how it actually emerged in full-blown human languages over generations of adult-to-child transmission chains. A burgeoning literature on the emergence of natural human languages may assist with this issue (Brentari & Coppola, 2012), but such naturally emerging languages are rare and diverse, making it difficult to draw conclusions about the roles of various causal factors. In contrast, computational models can only tell what the consequences are of a certain set of assumptions: they can not tell us how humans *actually* behave. In other words, *if* the model assumptions were true of humans, *then* the model would tells us something about how humans behave. Unfortunately, it is too common to see various aspects of a model stipulated without any empirical justification. Even in just the Baronchelli et al. (2010) and Loreto et al. (2012) models there are several questionable assumptions: perceptual categories do not evolve independently of language use, synonyms are immediately erased when another word is successful, and, of course, category boundaries are rigid (a color is either in one category or another – it can not be more or less representative of a category). One pessimistic conclusion from this discussion is that we may never be able to be particularly confident about how functional lexicons emerged. An alternative, more optimistic conclusion is that this situation is really a challenge, a call for greater integration between existing and forthcoming methods and data (perhaps including some suggested in a previous section, particularly historical corpora), so that different approaches can cover each other's weaknesses.

## Conclusion

In the introduction to this review, we discussed two complementary types of answers to questions of *why*: ultimate and proximate cause. In the body of this review, we suggested that the question "Why are human lexicons the way they are?" may have an ultimate causal answer of "Because such lexicons and/or the category systems they map onto are (near-)optimal or efficient for learning, communication, or other cognitive functions" and a proximate causal answer of "Because lexicons and/or category systems developed through repeated use and learning under pressures for efficiency (in learning, communication, etc.)". Most of the evidence we reviewed suggested that the lexicon possesses a structure that *is* functional in various respects. And in fact, what criticisms exist typically do not question that the lexicon possesses such functionality. Instead, what critics doubt, and what proponents have in many cases yet to demonstrate, is that functionality in the lexicon emerged due to pressures for such functionality (e.g., that communicatively efficient lexicons emerged due to pressures for efficient communication). For functionalist accounts to really be answers to the question "Why is the lexicon the way it is?", there must be demonstrations that functionalism in the lexicon emerged, or at least *could* have emerged, due to pressures for said functionalism. This is the direction future research will need to take.

## References

Atran, S. (1995). Classifying nature across cultures. In E. E. Smith & D. N. Osherson (Eds.), *Thinking: An invitation to cognitive science, vol. 3* (pp. 131–174). Cambridge, MA: MIT Press.
Baronchelli, A., Gong, T., Puglisi, A., & Loreto, V. (2010). Modeling the emergence of universality in color naming patterns. *Proceedings of the National Academy of Sciences of the United States of America*, 107(6), 2403–7. https://doi.org/10.1073/pnas.0908533107
Battaglia, P. W., Hamrick, J. B., & Tenenbaum, J. B. (2013). Simulation as an engine of physical scene understanding. *Proceedings of the National Academy of Sciences of the United States of America*, 110(45), 18327–32. https://doi.org/10.1073/pnas.1306572110
Beach, L. R. (1964). Cue probabilism and inference behavior. *Psychological Monographs: General and Applied*, 78(5–6), 1. https://doi.org/10.1037/h0093853
Berlin, B., Kay, P. (1969). *Basic Color Terms: Their Universality and Evolution* (University of California Press, Berkeley, CA).
Berlin, B. (1992). *Ethnobiological classification: Principles of categorization of plants and animals in traditional societies*. Princeton: Princeton University Press. https://doi.org/10.1515/9781400862597
Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, 55(4), 77. https://doi.org/10.1145/2133806.2133826

Bortfeld, H., & Morgan, J. (2005). Mommy and me: Familiar names help launch babies into speech-stream segmentation. *Psychological Science*, 16(4), 298–304. https://doi.org/10.1111/j.0956-7976.2005.01531.x

Boster, J. (2005). Emotion categories across languages. In H. Cohen & C. LeFebvre (Eds.), *Handbook of categorization in cognitive science* (pp. 187–222). Amsterdam: Elsevier. https://doi.org/10.1016/B978-008044612-7/50063-9

Bowerman, M. & Pederson, E. (1992). Cross-linguistic studies of spatial semantic organization. In *Annual Report of the Max Planck Institute for Psycholinguistics*, pp. 53–56.

Brashears, M. E. (2013). Humans use compression heuristics to improve the recall of social networks. *Scientific Reports*, 3, 1513. https://doi.org/10.1038/srep01513

Brentari, D., & Coppola, M. (2013). What sign language creation teaches us about language. *Wiley Interdisciplinary Reviews: Cognitive Science*, 4(2), 201–211. https://doi.org/10.1002/wcs.1212

Bybee, J. L., Pagliuca, W., & Perkins, R. D. (1990). On the asymmetries in the affixation of grammatical material. In W. Croft, K. Denning, & S. Kemmer (Eds.), Studies in Typology and Diachrony: Papers Presented to Joseph H. Greenberg on his 75th Birthday, 1–42. Amsterdam: John Benjamins. https://doi.org/10.1075/tsl.20.04byb

Bybee, J. (1999). Usage-based phonology. In M. Darnell, E. Moravcsik, F. Newmeyer, M. Noonan, and K. Wheatley (Eds.), Functionalism and formalism in linguistics, Volume I: General papers. Amsterdam: John Benjamins. 211–242. https://doi.org/10.1075/slcs.41.12byb

Bybee, J. (2006). From usage to grammar: The mind's response to repetition. *Language*, 82(4), 711–733. https://doi.org/10.1353/lan.2006.0186

Caramazza, A., & Mahon, B. Z. (2003). The organization of conceptual knowledge: the evidence from category-specific semantic deficits. *Trends in cognitive sciences*, 7(8), 354–361. https://doi.org/10.1016/S1364-6613(03)00159-1

Carstensen, A., Xu, J., Smith, C. T., & Regier, T. (2015). Language evolution in the lab tends toward informative communication. In Noelle, D. C., Dale, R., Warlaumont, A. S., Yoshimi, J., Matlock, T., Jennings, C. D., & Maglio, P. P. (Eds.) *Proceedings of the 37th Annual Meeting of the Cognitive Science Society*. Austin, TX: Cognitive Science Society.

Casasanto, D. & Lupyan, G. (2015). All concepts are ad hoc concepts. In E. Margolis & S. Laurence (Eds.) *Concepts: New Directions*, pp. 543–566. Cambridge: MIT Press.

Corter, J. E., & Gluck, M. A. (1992). Explaining basic categories: Feature predictability and information. *Psychological Bulletin*, 111(2), 291–303. https://doi.org/10.1037//0033-2909.111.2.291

Croft, W. (1995). Autonomy and Functionalist Linguistics. *Language*, 71(3), 490–532. https://doi.org/10.2307/416218

Croft, W. (2003). *Typology and universals: Second edition*. Cambridge, UK: Cambridge University Press.

Cutler, A. and Hawkins, J. A., & Gilligan, G. (1986). The suffixing preference: a processing explanation. *Linguistics*, 23. 723–758.

Dingemanse, M., Blasi, D. E., Lupyan, G., Christiansen, M. H., & Monaghan, P. (2015). Arbitrariness, iconicity, and systematicity in language. *Trends in cognitive sciences*, 19(10), 603–615. https://doi.org/10.1016/j.tics.2015.07.013

Ferrer i Cancho, R., & Sole, R. V. (2003). Least effort and the origins of scaling in human language. *Proceedings of the National Academy of Sciences of the United States of America*, 100(3), 788–91. https://doi.org/10.1073/pnas.0335980100

Ferrer-i-Cancho, R., & Moscoso del Prado Martín, F. (2011). Information content versus word length in random typing. *Journal of Statistical Mechanics: Theory and Experiment*, 2011(12), L12002. https://doi.org/10.1088/1742-5468/2011/12/L12002

Fowler, C. A., & Housum, J. (1987). Talkers' signaling of "new" and "old" words in speech and listeners' perception and use of the distinction. *Journal of Memory and Language*, 26(5), 489–504. https://doi.org/10.1016/0749-596X(87)90136-7

Galantucci, B., Garrod, S., & Roberts, G. (2012). Experimental Semiotics. *Language and Linguistics Compass*, 6(8), 477–493. https://doi.org/10.1002/lnc3.351

Gärdenfors, P. (2000). *Conceptual spaces: The geometry of thought*. MIT press.

Garner, W. R. (1974). *The Processing of Information and Structure* (Erlbaum, Potomac, MD).

Giavazzi, M. (2010). The phonetics of metrical prominence and its consequences for segmental phonology. PhD thesis, Massachusetts Institute of Technology.

Graff, P. (2012). Communicative efficiency in the lexicon. PhD Thesis, Massachusetts Institute of Technology.

Griffiths, T. L., Steyvers, M., & Firl, A. (2007). Google and the mind: predicting fluency with PageRank. *Psychological Science*, 18(12), 1069–76. https://doi.org/10.1111/j.1467-9280.2007.02027.x

Griffiths, T. L., Steyvers, M., & Tenenbaum, J. B. T. (2007). Topics in Semantic Representation. *Psychological Review*, 114(2), 211–244. https://doi.org/10.1037/0033-295X.114.2.211

Griffiths, T. L., & Tenenbaum, J. B. (2006). Optimal predictions in everyday cognition. *Psychological Science*, 17, 767–773. https://doi.org/10.1111/j.1467-9280.2006.01780.x

Haspelmath, M. (1999). Optimality and diachronic adaptation. *Zeitschrift Für Sprachwissenschaft*, 18.2, 180–205. https://doi.org/10.1515/zfsw.1999.18.2.180

Haspelmath, M. (2003). The geometry of grammatical meaning: Semantic maps and cross-linguistic comparison. In M. Tomasello (Ed.), *The new psychology of language, vol. 2* (pp. 211–242). Mahwah, NJ: Erlbaum.

Hills, T., Maouene, M., Maouene, J., Sheya, A., & Smith, L. (2009). Categorical structure among shared features in networks of early-learned nouns. *Cognition*, 112, 381–396. https://doi.org/10.1016/j.cognition.2009.06.002

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.

Hurford, J. R. (2003). Why synonymy is rare: Fitness is in the speaker. In Banzhaf, W., Christaller, T., Dittrich, P., Kim, J. T., & Ziegler, J. (Eds.), Advances in artificial life- Proceedings of the 7th European Conference on Artificial Life (ECAL), lecture notes in artificial intelligence vol. 2801, 442, Berlin: Springer Verlag.

Jaeger, T. F., & Tily, H. (2011). On language "utility": processing complexity and communicative efficiency. *Wiley Interdisciplinary Reviews: Cognitive Science*, 2(3), 323–335. https://doi.org/10.1002/wcs.126

Kemp, C., & Regier, T. (2012). Kinship categories across languages reflect general communicative principles. *Science*, 336(6084), 1049–54. https://doi.org/10.1126/science.1218811

Khetarpal, N., Majid, A., & Regier, T. (2009). Spatial terms reflect near-optimal spatial categories. In N. Taatgen et al. (Eds.), Proceedings of the 31st Annual Meeting of the Cognitive Science Society.

Khetarpal, N., Neveu, G., Majid, A., Michael, L., & Regier, T. (2013). Spatial terms across languages support near-optimal communication: Evidence from Peruvian Amazonia, and computational analyses. In M. Knauff, M. Pauen, N. Sebanz, and I. Wachsmuth (Eds.), Proceedings of the 35th Annual Meeting of the Cognitive Science Society.

King, R. (1967). Functional load and sound change. *Language* 43, 831–852. https://doi.org/10.2307/411969

Kirby, S., Cornish, H., & Smith, K. (2008). Cumulative cultural evolution in the laboratory: an experimental approach to the origins of structure in human language. *Proceedings of the National Academy of Sciences of the United States of America*, 105(31), 10681–6. https://doi.org/10.1073/pnas.0707835105
Kirby, S., Dowman, M., & Griffiths, T. L. (2007). Innateness and culture in the evolution of language. *Proceedings of the National Academy of Sciences of the United States of America*, 104(12), 5241–5. https://doi.org/10.1073/pnas.0608222104
Kleinberg, J. M. (2000). Navigation in a small world. *Nature*, 406(6798), 845. https://doi.org/10.1038/35022643
Kurumada, C., Meylan, S. C., & Frank, M. C. (2013). Zipfian frequency distributions facilitate word segmentation in context. *Cognition*, 127, 439–453. https://doi.org/10.1016/j.cognition.2013.02.002
Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes*, 25(2–3), 259–284. https://doi.org/10.1080/01638539809545028
Lelu, A. (2014). Jean-Baptiste Estoup and the origins of Zipf s law: a stenographer with a scientific mind (1868–1950). *BEIO, Boletín de Estadística e Investigación Operativa*, 30(1), 66–77.
Lindblom, B. (1986). Phonetic universals in vowel systems. In *Experimental Phonology*, Ed. by John Ohala & Jeri Jaeger, 13–44. Orlando: Academic Press.
Loreto, V., Mukherjee, A., & Tria, F. (2012). On the origin of the hierarchy of color names. *Proceedings of the National Academy of Sciences of the United States of America*, 109(18), 6819–24. https://doi.org/10.1073/pnas.1113347109
Mahowald, K., Fedorenko, E., Piantadosi, S. T., & Gibson, E. (2013). Info/information theory: speakers choose shorter words in predictive contexts. *Cognition*, 126(2), 313–8. https://doi.org/10.1016/j.cognition.2012.09.010
Majid, A., & Burenhult, N. (2014). Odors are expressible in language, as long as you speak the right language. *Cognition*, 130(2), 266–70. https://doi.org/10.1016/j.cognition.2013.11.004
Majid, A., Enfield, N. J., & van Staden, M. (2006). Parts of the body: Cross-linguistic categorisation (Special Issue). *Language Sciences*, 28(2–3).
Malt, B. C., Sloman, S. A., Gennari, S., Shi, M., & Wang, Y. (1999). Knowing versus naming: Similarity and the linguistic categorization of artifacts. *Journal of Memory and Language*, 40, 230–262. https://doi.org/10.1006/jmla.1998.2593
Marcus, G. F., & Davis, E. (2013). How robust are probabilistic models of higher-level cognition? *Psychological Science*, 24(12), 2351–60. https://doi.org/10.1177/0956797613495418
Margolis, E., & Laurence, S. (2014). Concepts. In E. N. Zalta (Ed.) *The Stanford Encyclopedia of Philosophy* (Spring 2014 Edition). URL = <http://plato.stanford.edu/archives/spr2014/entries/concepts/>.
Markman, E. M., & Wachtel, G. F. (1988). Children's use of mutual exclusivity to constrain the meanings of words. *Cognitive psychology*, 20(2), 121–157. https://doi.org/10.1016/0010-0285(88)90017-5
Martin, A. 2007. The Evolving Lexicon. PhD Thesis, University of California, Los Angeles.
Martinet, A. (1952). Function, structure, and sound change. *Word*, 8, 1–32. https://doi.org/10.1080/00437956.1952.11659416
Miller, G. A. (1957). Some effects of intermittent silence. *The American Journal of Psychology*, 70(2), 311–314. https://doi.org/10.2307/1419346
Moscoso del Prado Martin, F. (2013). The missing baselines in arguments for the optimal efficiency of languages. *Csjarchive.cogsci.rpi.edu*, 1032–1037. Retrieved from http://csjarchive.cogsci.rpi.edu/Proceedings/2013/papers/0203/paper0203.pdf

Murphy, G. (2002). *The big book of concepts*. Cambridge, MA: MIT press. https://doi.org/10.7551/mitpress/1602.001.0001
Piantadosi, S. T. (2014). Zipf's word frequency law in natural language: A critical review and future directions. *Psychonomic Bulletin & Review*, 1–35. https://doi.org/10.3758/s13423-014-0585-6
Piantadosi, S. T., Tily, H., & Gibson, E. (2011). Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences of the United States of America*, 108(9), 3526–9. https://doi.org/10.1073/pnas.1012551108
Piantadosi, S. T., Tily, H., & Gibson, E. (2012). The communicative function of ambiguity in language. *Cognition*, 122(3), 280–91. https://doi.org/10.1016/j.cognition.2011.10.004
Piantadosi, S. T., Tily, H. J., & Gibson, E. (2009). The communicative lexicon hypothesis. In Proceedings of the 31st Annual Conference of the Cognitive Science Society, pp. 2582–2587.
Piantadosi, S., Tily, H., & Gibson, E. (2013). Information content versus word length in natural language: A reply to Ferrer-i-Cancho and Moscoso del Prado Martin (2011). *arXiv Preprint arXiv:1307.6726*, (2011), 1–8. Retrieved from http://arxiv.org/abs/1307.6726
Posner, M. I., & Keele, S. W. (1968). On the genesis of abstract ideas. *Journal of Experimental Psychology*, 77, 353–363. https://doi.org/10.1037/h0025953
Posner, M. I., & Keele, S. W. (1970). Retention of abstract ideas. *Journal of Experimental Psychology*, 83, 304–308. https://doi.org/10.1037/h0028558
Reed, K. (1972). Pattern Recognition. *Cognitive Psychology*, 407, 382–407. https://doi.org/10.1016/0010-0285(72)90014-X
Regier, T., Kay, P., & Khetarpal, N. (2007). Color naming reflects optimal partitions of color space. *Proceedings of the National Academy of Sciences of the United States of America*, 104(4), 1436–41. https://doi.org/10.1073/pnas.0610341104
Regier, T., Kemp, C., & Kay, P. (2015). Word meanings across languages support efficient communication. In B. MacWhinney & W. O. Grady (Eds.), *The handbook of language emergence* (pp. 237–263). Hoboken, NJ: Wiley. https://doi.org/10.1002/9781118346136.ch11
Richie, R., Kaufmann, S., Tabor, W. (2014). An LSA-based method for estimating word meaning specificity: An application to an account of Zipf's Law. Poster presented at the 9th Annual Mental Lexicon Conference, Niagara-on-the-Lake, Ontario, Canada.
Roberson, D., Davidoff, J., Davies, I. R. L., & Shapiro, L. R. (2005). Color categories: Evidence for the cultural relativity hypothesis. *Cognitive Psychology*, 50, 378–411. https://doi.org/10.1016/j.cogpsych.2004.10.001
Rosch, E. (1973). On the internal structure of perceptual and semantic categories. In T. E. Moore (Ed.), Cognitive development and the acquisition of language. New York: Academic Press.
Rosch, E., Simpson, C., & Miller, R. S. (1976). Structural bases of typicality effects. *Journal of Experimental Psychology: Human Perception and Performance*, 2, 491–502.
Rosch, E. (1978). Principles of categorization. In E. Rosch & B. B. Lloyd (Eds.), Cognition and categorization. Hillsdale, NJ: Erlbaum. Reprinted in: Margolis, E. and Laurence, S. (Eds.) (1999). *Concepts: Core readings*. Cambridge, MA: MIT Press.
Sagi, E., Kaufmann, S., & Clark, B. (2011). Tracing semantic change with Latent Semantic Analysis. In K. Allan & Robinson, J. A. (Eds.), *Current Methods in Historical Semantics* (pp. 161–183). Berlin, Germany: Mouton de Gruyter. https://doi.org/10.1515/9783110252903.161
Scott-Phillips, T. C., & Kirby, S. (2010). Language evolution in the laboratory. *Trends in Cognitive Sciences*, 14(9), 411–7. https://doi.org/10.1016/j.tics.2010.06.006
Scott-Phillips, T. C., Dickins, T. E., & West, S. A. (2011). Evolutionary theory and the ultimate–proximate distinction in the human behavioral sciences. *Perspectives on Psychological Science*, 6(1), 38–47. https://doi.org/10.1177/1745691610393528

Silvey, C., Kirby, S., & Smith, K. (2013). Communication leads to the emergence of sub-optimal category structures. In M. Knauff, M. Pauen, N. Sebanz, & I. Wachsmuth (Eds.), *Proceedings of the 35th Annual Conference of the Cognitive Science Society*, 1312–1317. Austin, TX: Cognitive Science Society.

Steels, L., & Belpaeme, T. (2005). Coordinating perceptually grounded categories through language: a case study for colour. *The Behavioral and Brain Sciences*, 28(4), 469–89; discussion 489–529. https://doi.org/10.1017/S0140525X05000087

Steels, L. (2011). Modeling the cultural evolution of language. *Physics of Life Reviews*, 8(4), 339–56. https://doi.org/10.1016/j.plrev.2011.10.014

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. https://doi.org/10.1207/s15516709cog2901_3

Thompson, G. W., & Kello, C. T. (2014). Walking across Wikipedia: a scale-free network model of semantic memory retrieval. *Frontiers in Psychology*, 5, 86. https://doi.org/10.3389/fpsyg.2014.00086

Thompson, D. (1917). *On Growth and Form*. Cambridge University Press: Cambridge, UK.

Thompson, B., Silvey, C., Kirby, S., & Smith, K. (2014). The effect of communication on category structure. In *Evolution of Language: Proceedings of the 10th International Conference (EVOLANG10)* (pp. 537–538).

Vitevitch, M. S., Chan, K. Y., & Goldstein, R. (2014). Insights into failed lexical retrieval from network science. *Cognitive Psychology*, 68, 1–32. https://doi.org/10.1016/j.cogpsych.2013.10.002

Wedel, A., Jackson, S., & Kaplan, A. (2013). Functional load and the lexicon: Evidence that syntactic category and frequency relationships in minimal lemma pairs predict the loss of phoneme contrasts in language change. *Language and Speech*, 56(3), 395–417. https://doi.org/10.1177/0023830913489096

Wedel, A., Kaplan, A., & Jackson, S. (2013). High functional load inhibits phonological contrast loss: a corpus study. *Cognition*, 128(2), 179–86. https://doi.org/10.1016/j.cognition.2013.03.002

Wittgenstein, Ludwig. (1921/1922). *Tractatus Logico-Philosophicus*. London: Routledge and Kegan Paul Ltd.

Wittgenstein, L. (1953/2010). *Philosophical Investigations*. Blackwell Publishing.

Wnuk, E., & Majid, A. (2014). Revisiting the limits of language: the odor lexicon of Maniq. *Cognition*, 131(1), 125–38. https://doi.org/10.1016/j.cognition.2013.12.008

Xu, J., Dowman, M., & Griffiths, T. (2013). Cultural transmission results in convergence towards colour term universals. *Proceedings of the Royal Society B: Biological Sciences*, 280, 1758. https://doi.org/10.1098/rspb.2012.3073

Xu, Y., & Regier, T. (2014). Numeral systems across languages support efficient communication: From approximate numerosity to recursion. In P. Bello et al. (Eds.), Proceedings of the 36th Annual Meeting of the Cognitive Science Society. Austin, TX: Cognitive Science Society.

Xu, Y. and Kemp, C. (2015). A computational evaluation of two laws of semantic change. In Noelle, D. C., Dale, R., Warlaumont, A. S., Yoshimi, J., Matlock, T., Jennings, C. D., & Maglio, P. P. (Eds.) *37th Annual Conference of the Cognitive Science Society*. Austin, TX: Cognitive Science Society.

Xu, Y., Regier, T. and Malt, B. C. (2015). Historical semantic chaining and efficient communication: The case of container names. *Cognitive Science*. https://doi.org/10.1111/cogs.12312

Yee, E., & Thompson-Schill, S. L. (2016). Putting concepts into context. *Psychonomic Bulletin & Review*, 23(4), 1015–1027. https://doi.org/10.3758/s13423-015-0948-7
Zipf, G. (1949). *Human Behavior and the Principle of Least Effort*. Addison-Wesley, New York.
Zuidema, W., & Westermann, G. (2003). Evolution of an optimal lexicon under constraints from embodiment. *Artificial Life*, 1–14. https://doi.org/10.1162/106454603322694834
Zwarts, J. (1995). The Semantics of Relative Position. *Proceedings of Semantics and Linguistic Theory*, 5, 405–422. https://doi.org/10.3765/salt.v5i0.2710

CHAPTER 7

# Morphological schemas

## Theoretical and psycholinguistic issues

Ray Jackendoff [1,2] and Jenny Audring[3]
[1]Tufts University / [2]MIT / [3]Leiden University

### Thesis

1. A full account of language that includes morphology has to take seriously not only the productive patterns, as in much of current linguistic theory, but *all* the patterns, from completely productive to altogether marginal.
2. Such an account has to focus on what linguistic units are stored in the lexicon. It is necessary to give up the common assumption that the lexicon is nothing but an unstructured list of irregularities. Rather, the lexicon is full of patterns among stored items. Encoding such patterns *as patterns* requires formal mechanisms that express how lexical items are alike and how they differ. The result is a rich network of lexical relations.
3. Among these formal mechanisms are *schemas* – abstractions over stored items – plus explicit links to their instances. Schemas have the formal structure of lexical items, except that they contain variables at points where their instances differ. Schemas thereby relate lexical items to one another; they perform the function sometimes attributed to "lexical redundancy rules."
4. This leads to a major reframing: **Rules of grammar are not a separate component from the lexicon. Rather, they are part of the lexicon**, differing only in that they contain variables. Furthermore, the lexicon contains links that specify precisely the relations among lexical items.
5. A subset of schemas can be used generatively, to create new linguistic objects. This generative use takes over the role of productive rules in classical generative grammar.
6. The upshot is another major reframing: **Linguistic theory comes now to be primarily about relations among stored items rather than derivations of novel items**. While derivations don't go away, they are identified with the generative use of productive schemas, and they "ride on top" of the network of lexical relations, which in turn constitute the bulk of linguistic knowledge.

https://doi.org/10.1075/z.238.07jac

7. These reframings have consequences for theories of language processing. Basically, unlike standard approaches to grammar, this theory can be incorporated directly into a theory of the processor. In particular, when the processor is constructing complex structures in working memory, it accesses schemas in the same way as it accesses words. Like words, schemas have a resting activation proportional to frequency, and they are subject to priming. Just as multiple words compete in working memory for the status of "what is heard," so do syntactic and morphological schemas compete during parsing. In addition, spreading activation follows the links among items, including the links of words to schemas.
8. A morphological schema provides a potential decomposition of a perceived input. When a word can be accessed both by the whole-word and the compositional route, both can take place at once, contingent on relative frequency of the word, its base, and the schema.
9. The challenges for the future are (a) to apply the theory of schemas and links to the full panoply of morphological (and syntactic) phenomena that have been central to the literature for decades, as well as phenomena that have been neglected; (b) to explore further the implications of the theory for psycholinguistics, including such effects as masked priming and neighborhood density; (c) to use the theory to help design new experiments and new computational models of language processing.

*Reference*

Jackendoff, R., and Audring, J. (2020). *The Texture of the Lexicon: Relational Morphology and the Parallel Architecture*. Oxford: Oxford University Press.

## Commentaries on Jackendoff and Audring thesis

### James Myers commentary on Jackendoff and Audring thesis

A central problem for linguistic theory is the finding that grammar is fuzzier than one would expect from a discrete combinatorial system, yet still a lot more discrete and combinatorial than any purely bottom-up, lexically driven approach has been able to capture thus far. The particular solution advocated here involves schemas that are "fuzzy" in being memorized wholes of potentially any degree of size or complexity, while still being discrete and combinatorial in having parts that can be linked across them in the formation of productive analogies, or even via symbolic variables that stand for entire classes. Nevertheless, the fuzzy/discrete divide is a hard one to cross, and the authors are right to use the term "challenge" to describe their ambitious goal to "apply the theory of schemas and links to the full panoply of morphological (and syntactic) phenomena that have been central to the literature for decades". Lexical generalizations aren't as fuzzy as their model would seem to permit, or else morphological theory would likely never have been invented in the first place; syntax, if they are really ambitious enough to look there too, is even less fuzzy and involves phenomena that would require mechanisms for recursion, island constraints, and whatnot. There is also the practical challenge that the human brain, at least when situated within scientists, seems to prefer working with discrete top-down theories than with fuzzy bottom-up ones.

So it may help to step back and ask *why* language is both fuzzy and discrete. A plausible old idea is that the ultimate mechanics of language are indeed fuzzy, as activation states in a complex neural net, but discrete combinatorial systems have emerged out of it, via natural selection operating on the fine structure of this neural net. As has been pointed out by William Abler and other authors, quasi-atomic units obeying quasi-rigid rules have emerged in a variety of natural systems, including the genetic code and atoms themselves (though such emergence is hardly inevitable, or else all molecules could reproduce themselves and all animals could talk). Such systems betray their origins in a deeper layer of fuzziness (epigenetics and other biochemical complexities, the transmutation of elements), and recognizing this, as the authors emphasize here, is indeed an important step in making theoretical progress. Nevertheless, focusing on the fuzziness per se seems to miss the forest for the trees. Whether grammars are implemented via schemas or not, they have been honed by natural selection to "try" to rise above them.

## Dorit Ravid commentary on Jackendoff and Audring thesis

A central statement in the paper is that the principle of economy, on which so much of modern linguistics has been predicated, is "the wrong criterion for a theory of lexical storage in the brain", and that "the brain embraces redundancy" (p. 476), advocating a full-entry theory. How is such a lexicon learned, given the importance of morphological structure and semantics in its organization? Ackerman & Malouf (2013) have a solution, based on their word-and-paradigm perspective, where a word form is regarded as a configuration of recurrent elements distributed in complex words as members of a paradigm (p. 431). Under this view, the research questions at hand are how such complex forms are organized in paradigms, and how this organization is learned. Ackerman & Malouf propose the measure of integrative complexity in morphology, limiting the enumerative complexity of paradigms so as to allow them to be learned by native speakers – although the top-down view by linguists may consider too complex for children's learning (p. 435). This metric requires that morphological systems enable speakers to accurately infer from their experience with words about unknown forms of these words.

This approach goes hand in hand with two further views. First, that the division between inflection and derivation, and even between lexicon and syntax, is graded rather than polar, especially in morphology-rich languages. From a cognitive-semantic perspective, Evans (2009) enumerates a broad array of forms that encode lexical concepts, from bound morphemes, through words of all kinds, to phrases and syntactic constructions. From a typological perspective, Aikhenvald (2007) demonstrates exceptions for each morphological property across languages, culminating in the observation that what constitutes inflection or derivation must be established in relation to a particular language. Second, there is a close affinity between lexical frequency and the formation of grammatical generalizations (Ambridge et al., 2008; Borovsky, Elman & Fernald, 2012), with vocabulary size being the single most powerful predictor of children's grammatical development (Caselli et al., 1995; Devescovi et al., 2005). The evidence suggests that both derivational morphology, which creates morphological families and enables the formation of new words, and inflectional morphology, which expresses syntactic and pragmatic relations, sustain and are fed by the emergence and consolidation of syntax (Borovsky et al., 2016). In other words, there is a continuum between lexical, derivational and inflectional elements (Haspelmath & Sims, 2010). This makes sense, as content words fill designated syntactic positions in clauses and phrases, constitute the heads of syntactic phrases, and provide the stems for morphological inflection.

## *References*

Ackerman, F., & Malouf, R. (2013). Morphological organization: The low conditional entropy conjecture. *Language*, 89, 429–464. https://doi.org/10.1353/lan.2013.0054

Aikhenvald, A. Y. (2007). *Typological dimensions in word-formation*. In: Shopen, Timothy, (ed.) *Language typology and syntactic description. Grammatical Categories and the Lexicon*, Vol. 3. Cambridge: Cambridge University Press, 1–65.

Ambridge, B., Pine, J. M., Rowland, C. F., Young, C. R. (2008) The effect of verb semantic class and verb frequency (entrenchment) on children's and adults' graded judgements of argument-structure overgeneralisation errors. *Cognition*, 106, 87–129. https://doi.org/10.1016/j.cognition.2006.12.015

Borovsky, A., Elman, J. L., & Fernald, A. (2012). Knowing a lot for one's age: Vocabulary skill and not age is associated with anticipatory incremental sentence interpretation in children and adults. *Journal of Experimental Child Psychology*, 112, 417–436. https://doi.org/10.1016/j.jecp.2012.01.005

Borovsky, A., Ellis, E. M., Evans, J. L., & Elman, J. L. (2016). Semantic structure in vocabulary knowledge interacts with lexical and sentence processing in infancy. *Child Development*, 87, 1893–1908. https://doi.org/10.1111/cdev.12554

Caselli, M. C., Bates, E., Casadio, P., Fenson, J., Fenson, L., Sanderl, L., & Weir, J. (1995). A cross-linguistic study of early lexical development. *Cognitive Development*, 10(2), 159–199. https://doi.org/10.1016/0885-2014(95)90008-X

Devescovi, A., Caselli, M. C., Marchione, D., Pasqualetti, P., Reilly, J., & Bates, E. (2005). A cross linguistic study of the relationship between grammar and lexical development. *Journal of Child Language*, 32, 759–786. https://doi.org/10.1017/S0305000905007105

Evans, V. (2009). *How words mean: Lexical concepts, cognitive models, and meaning construction*. Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199234660.001.0001
Haspelmath, M. & Sims, A. D. (2010). *Understanding morphology*. 2nd edition. London: Hodder Education.

## Russell Richie commentary on Jackendoff and Audring thesis

Network analysis has been a popular tool in characterizing lexical knowledge, acquisition, and processing (for reviews see Baronchelli et al., 2013 and Vitevitch et al. 2014). For example, the scale-free and small-world character of lexical-semantic networks has motivated a preferential attachment-inspired model of lexical acquisition (Steyvers & Tenenbaum, 2005). A word's PageRank in a lexical-semantic network is a better predictor of its ease of lexical access than its raw frequency or its in-degree in the same network, a finding that motivated particular models of memory search (Griffiths et al., 2007). If Jackendoff and Audring's theory that grammar and lexicon are one and the same in terms of connected schemas, then the door is open to using network analysis to characterize *all* of linguistic knowledge and use in similar ways to the above studies. What old morphological and syntactic puzzles might find new explanations in graph theoretic terms? What new puzzles would present themselves? In lexical-semantic networks, clustering or community structure in the network reflect categories of different kinds – what would community structure in a network of schema reflect?

A challenge for this approach, however, would be to construct large databases of schema, in the same way that large databases of lexical networks have been constructed from thesauruses, word association networks, WordNet, and the like. Whereas the latter datasets are relatively easy to construct by asking participants to, e.g., give synonyms or words that come to mind for a particular stimulus word, it is harder to imagine how a large dataset of schema could be easily crowdsourced. It may be that construction of schema databases for various languages may be more analogous to construction of phrase structure grammars, i.e., requiring many hours of work from skilled linguist annotators. At the same time, if Jackendoff and Audring's schemas better fit natural language data than traditional grammars, then perhaps they ought to be easier construct in than PSG's. In addition, it may be possible to write algorithms that convert at least some of existing treebank and phrase structure grammar resources into the format of schemas.

### *References*

Baronchelli, A., Ferrer-i-Cancho, R., Pastor-Satorras, R., Chater, N., & Christiansen, M. H. (2013). Networks in cognitive science. *Trends in Cognitive Sciences*, 17(7), 348–60. https://doi.org/10.1016/j.tics.2013.04.010

Griffiths, T. L., Steyvers, M., & Firl, A. (2007). Google and the mind: predicting fluency with PageRank. *Psychological Science*, 18(12), 1069–76. https://doi.org/10.1111/j.1467-9280.2007.02027.x
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. https://doi.org/10.1207/s15516709cog2901_3
Vitevitch, M. S., Goldstein, R., Siew, C. S. Q. and Castro, N. (2014). Using complex networks to understand the mental lexicon. *Yearbook of the Poznań Linguistic Meeting*, 1, 119–138 https://doi.org/10.1515/yplm-2015-0007

## Benjamin Tucker commentary on Jackendoff and Audring thesis

The paper by Jackendoff and Audring takes an initial step toward creating an integrated theory of lexical representation. In the present paper, the authors focus on lexical representation and morphology, which they hope will highlight features of linguistic organization. A secondary goal of this article is to raise questions to researchers and modelers so that they can help fill out the picture of this approach to lexical representation. As a result, this commentary will underscore some of the questions that their Relational Morphology approach raises from my perspective.

First, I would be curious to see examples of how these schemas would work in morphophonology, particularly for allomorphy. For example, is there a separate schema or some form of a single schema for each of the three allomorphs of the plural morpheme in English -/s/, -/z/ and -/ez/? The authors general approach seems to suggest that they would favor separate schemas for each allomorph but would that actually be necessary? Further, how would schemas for reduplication be represented?

Second, in thinking about casual speech (briefly noted by the authors in the conclusion and discussed in Tucker and Ernestus (this issue)): how would schemas work in recognition when there are reductions of the word and a whole segment or even a syllable has been left out? This is also a more general question for models of word comprehension. However, I believe this is potentially relevant to Relational Morphology because in real spoken language the nature of the morpheme is likely different, resulting in a type of allomorph which only occurs as a reduction.

Third, the main examples presented here are in English, likely for obvious expository reasons. I would like to see these applied to both agglutinative and polysynthetic languages. In other words, how does this approach extend to something other than English? Further, the authors refer to syntactic application as well – would the syntactic application be the solution of isolating languages?

One area of interest that this approach raises, and is noted by the authors, is computationally implementing such an approach. I believe the implementation of such a schema in a computational model is still open but conceivably schemas could be included as lexical items in something like the end-to-end spoken word comprehension models DIANA (ten Bosch et al., 2014) or NDL (Arnold et al., 2016). Both models could allow for schemas to be built into the competition networks (following fairly different assumptions) and the influence of these schemas on the model prediction could then be tested.

The authors briefly outline one potential experiment that could be run to test aspects of their approach and have left open many doors that other researchers could explore, perhaps over the next decade. These doors include experiments that might fill in the picture of how this approach might work. The door has also been left open to finding evidence indicating that this may or may not be the correct approach – allowing researchers to create a fully integrated model of lexical representation.

### *References*

Arnold, D., Tomaschek, F., Sering, K., Lopez, F., Baayen, R. H. (2017). Words from spontaneous conversational speech can be recognized with human-like accuracy by an error-driven learning algorithm that discriminates between meanings straight from smart acoustic features, bypassing the phoneme as recognition unit. *PLOS ONE* 12(4): e0174623. https://doi.org/10.1371/journal.pone.0174623

Ten Bosch, L., Boves, L., & Ernestus, M. (2015). DIANA, an end-to-end computational model of human word comprehension. In *18th International Congress of Phonetic Sciences (ICPhS 2015)*. University of Glasgow.

## Chris Westbury commentary on Jackendoff and Audring thesis

The idea of a 'schema' stands at the center of a fundamental divide in the psychological sciences.

On one side of that divide, the development of schemas is regarded as a fundamental and necessary step in scientific explanation. When the structure of a problem has been laid out in schematic terms, some believe that the problem is solved, or nearly solved, especially if they assign causal efficacy to the schema. If schemas and rules can actually *do* things, then recognizing the schematic structure of a system is akin to explaining that system.

On the other side of the divide, schemas and rules are regarded as problematic, especially when they are assigned causal efficacy. From this point of view, schemas and rules are nothing more than post-hoc descriptions of recognized partial-regularities in a complex system. As such, they cannot serve as *explanations*. Nevertheless, they may serve as cogent *descriptions* of what it is that needs to be explained.

It is easy to understand why high-level schematic summaries are attractive and can be so useful (as has been very often discussed in the context of outlining Marr and Poggio's (1976) and Dennett's (1978) three levels of explanation): because the schematic level of explanation is often the most succinct and straightforward way of describing the dynamics of a system. Consider the original 8-bit Super Mario Brother's game. It certainly makes sense to explain the nature of the game to someone in terms of the things that Super Mario can do (the rules of the Super Mario game): run back and forth, jump, bonk his head on blocks, throw himself at coins, save Princess Peach, and so on. However, such an explanation is not, at least in my view, a scientific explanation. The scientific explanation of the Super Mario Brothers game would have to explain the game in terms that did not mention Mario or princesses or coins or blocks, but instead talked about algorithms, electronic circuits, and pixels on a screen, the elements of Super Mario that actually do have causal efficacy.

The main reason that schematic summaries cannot serve as scientific explanation is because they are, by design, unfalsifiable when properly formulated (i.e. when they do in fact describe true regularities). This is in part because they are also generally unquantifiable: it is not always easy – or even possible – to decide which of two competing schemas is a better description of a system. Since schemas must, by their very nature as high-level simplifications, focus on some elements of a system and ignore others, there is a danger that investigators will over-fit their schemas by focusing on confirmatory evidence and not on disconfirming evidence. Discussion then devolves into a matter of opinions about what is important or interesting and what is not. Statistical analyses were invented largely to take such discussions out of the realm of opinion. While statistical models (like all other models) rely on simplification too, the model that fits the data better, by some well-defined objective quantitative criterion, is the better model.

In the study of morphology now there is a cleave between statistical and schematic explanations. For example, Baayen, Milin, Đurđević, Hendrix, & Marelli (2011) begin their statistical analysis of the complex morphological system of the Serbian language by stating baldly that "In current theoretical morphology, […] the morpheme does not play an important role" (p. 2), in part on the basis of the fact that morphemes are often weak discriminative cues in language.

Although proponents of schemas-as-explanation may wish to simply disregard the very criteria that statistical morphologists used to assess their models, in doing so they risk isolating themselves within a self-referential game detached from scientific explanation. History is full of examples of schemas whose utility or truth-value was simply disconfirmed when scientific tools were brought to bear on them. The problem with schemas is that what you are able to see is not always what you get.

*References*

Baayen, R. H., Milin, P., Đurđević, D. F., Hendrix, P., & Marelli, M. (2011). An amorphous model for morphological processing in visual comprehension based on naive discriminative learning. *Psychological Review*, 118(3), 438. https://doi.org/10.1037/a0023851
Dennett. (1978). *Brainstorms*. Cambridge: MIT Press.
Marr, D. & Poggio, T. (1976). *From Understanding Computation to Understanding Neural Circuitry*. Artificial Intelligence Laboratory. A.I. Memo. Massachusetts Institute of Technology. AIM-357.

## The Article

We propose a theory of the lexicon in which rules of grammar, encoded as declarative schemas, are lexical items containing variables. We develop a notation to encode precise relations among lexical items and show how this differs from the standard notion of inheritance. We also show how schemas can play both a generative role, acting as productive rules, and also a relational role, where they codify nonproductive but nevertheless prolific patterns within the lexicon. We then show how this theory of lexical relations can be embedded directly into a theory of lexical access and lexical processing, such that it can make direct contact with experimental findings.

### 1. Overall goals

Looking to the big picture, a cognitive science of language aspires to a unified account of the structure of language, how it is processed, and how it is instantiated in the brain. Roughly following Marr 1982, such an account involves (at least) three subtheories:

– A theory of *representations*: the data structures stored in memory and those built online. This corresponds to Marr's *computational theory* and Chomsky's (1965) theory of *competence*.
– A theory of *processing*: the (virtual) architecture of the language processor, how representations are deployed in real time to create new structures, and how stored representations are acquired. Marr calls this an *algorithmic* theory, though strictly speaking, processing in the brain is most likely not in the nature of a step-by-step algorithm.
– A theory of *neural computation*: how the machinery of the processing theory is instantiated in the brain. Ultimately, this should include not only an account of brain localization, but also fine-scale details such as how neural encoding distinguishes the sound *b* from the sound *p* and the word *cat* from the word *dog*. Marr calls this an *implementational* theory.

Marr stresses (as Chomsky does not) that these theories must integrate with one another. The theories of processing and neural computation must be rich enough to support one's postulated representations; conversely, the representational theory must lend itself to postulated processing mechanisms and neural implementation. If the theories conflict, there is no predetermined outcome as to which theory prevails. For instance, the linguist's theory of competence should not be immune to evidence from performance. But neither should a theory of neural computation be immune to evidence from what an adequate theory of linguistic representation requires.

A theory of linguistic representations and processing should in addition seek integration between the theories of phonology, morphology, syntax, and semantics. And insofar as possible, a theory of the language faculty should seek integration with theories of other mental faculties. However, such integration must respect the details of the individual faculties: it must not result in ignoring much of what we know about language in order to fit it into a procrustean bed of, say, associationism.

The present article attempts some steps toward such an integrated theory. We focus on morphology, which highlights features of linguistic organization that are not so prominent in syntax (unless one thinks to look for them). Our approach, which we call Relational Morphology (RM: Jackendoff and Audring, 2020) grows out of the Parallel Architecture of Jackendoff 2002 and Culicover and Jackendoff 2005; we also draw freely on Construction Morphology (Booij 2010). We will explore some aspects of a theory of morphological processing, raising many questions that we hope might encourage psycholinguists, neurolinguists, and computational modelers to try to fill out the picture, both theoretically and experimentally.

## 2. Representations in relational morphology

### 2.1 *Rules, reconceived as schemas, are in the lexicon*

A prominent construct within most linguistic theories has been the distinction between words and rules, or between the lexicon and the grammar, as though they are in different metaphorical "places" in the mind. In parallel, much of *psycho*linguistics has tended to treat storage of words as distinct from rules of grammar, a position explicitly advocated by e.g. Ullman 2015.[1] This split is rejected by the Parallel Architecture, by Construction Grammar (Goldberg 1995; Hoffman and Trousdale 2013), and by Construction Morphology. These approaches argue that grammatical rules are themselves lexical items – that is, the grammar is part of the lexicon. Here is how the reasoning goes.

1. Other strains of psycholinguistics, e.g. connectionism, deny altogether that there are rules (Rumelhart and McLelland 1986).

In practically every linguistic theory, a word contains pieces of structure on three levels: its semantic structure, its syntactic features, and its phonology.[2] In the Parallel Architecture framework, these levels are in principle independent, each with its own characteristic conditions of well-formedness. But each is also linked to the others: *this* phonological string can be linked to *this* syntax and *this* semantics. We call these connections **interface links**, and we notate them with indices that show which parts of structure on one level correspond with structure on another level (other frameworks use different notations with similar effect). The indices are to be thought of as marking the ends of association lines. In the easiest cases, the mapping is trivial, as in *sheep*.

| | | |
|---|---|---|
| (1) | Semantics: | $\text{SHEEP}_1$ |
| | Morphosyntax: | $\text{N}_1$ |
| | Phonology: | /ʃip/$_1$ |

More interesting is a morphologically complex word, such as *sheepish* in (2).

| | | |
|---|---|---|
| (2) | Semantics: | [$\text{SHEEP}_1$-LIKE; TIMID]$_2$ |
| | Morphosyntax: | $[_A\ \text{N}_1 - \text{aff}_3]_2$ |
| | Phonology: | /ʃip$_1$ ɪʃ$_3$/$_2$ |

Unpacking the notation in (2): The morphosyntax encodes the fact that this word is an adjective composed of a noun plus an affix. Coindex 2 links the semantics, morphosyntax, and phonology of the whole word, just as coindex 1 links the layers in (1). Coindex 1 links the syntactic category Noun with the meaning SHEEP and the phonology /ʃip/. Coindex 3 links the affix with the phonology /ɪʃ/.[3] However, coindex 1 also links the relevant parts of *sheepish* to the word *sheep*. In this role, cutting across words, the coindexation serves as a ***relational link***. If we notate the links within (1) and (2) as association lines rather than coindices, as in (3), we can see that the interface links (solid lines) connect levels within a lexical item, while the relational links (dashed lines) connect parts that are the same in different lexical items:[4]

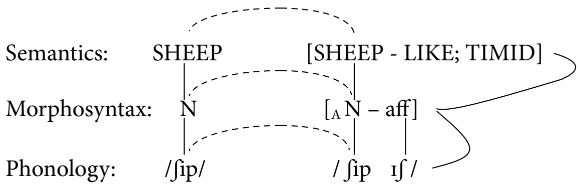
---

**2.** – with limited exceptions. Words like *hello* and *ouch* have no syntactic category: they can constitute utterances on their own but only combine into utterances paratactically (*Hello, Joe*) or in quotes (*"Ouch!" she cried*.) A few other words lack semantics and serve only as "grammatical glue", such as the underlined words in *I think that it's raining* and *I do not like green eggs and ham*. Literate individuals will have a fourth level, orthography.

**3.** One might wonder if coindex 3 should appear also in semantics, indexed on *LIKE*. Our reasons for not doing so are beyond the scope of this article.

**4.** The reader can also see why we have chosen coindexation as our notation rather than association lines: they're a mess!

(3)




The interface links in (2) establish a 1:1 relation between the parts of the semantics, the morphosyntax, and the phonology. However, such a transparent relation is not always present. For example, consider the plural word *sheep*, shown in (4).

(4) Semantics: $[\text{PLUR}\ (\text{SHEEP}_1)]_4$
Morphosyntax: $[_{\text{N}}\ \text{N}_1\ \text{pl}]_4$
Phonology: $/\text{ʃip}/_{1,4}$

(4) has a plural feature in morphosyntax, as is evident from determiner and verb agreement (*those* *sheep* *are*…); but there is no corresponding part in phonology. Rather, the full phonological string has two indices: one for its connection to the stem *sheep* (coindex 1), and one for its connection to the word as a whole (coindex 4). There is no "zero morpheme" in the phonology, as in many theories.

Given this format for representing words, the notation can easily be enriched so as to implement rules. Extracting the contribution of *sheep* (1) to *sheepish* (2), we arrive at (5) as the pattern whose instances include *sheepish, childish, sluggish*, and so forth.

(5) Semantics: $[\text{X}_y\ \text{-LIKE}]_z$
Morphosyntax: $[_{\text{A}}\ \text{N}_y - \text{aff}_3]_z$
Phonology: $/\ldots_y\ \text{ɪʃ}_3/_z$

Following the terminology of Construction Morphology, we call (5) a **schema** rather than a rule, to stress that it is declarative rather than procedural. (The term *schema* goes back in linguistics at least to Bybee and Slobin 1982, with more general antecedents cited in Rumelhart 1980 all the way back to Kant. Construction Grammarians would call (5) a **construction**.) (5) is not a procedure to convert an "input" into an "output," by adding *-ish* to a noun to form an adjective; rather, it *licenses* or *motivates* its instances, a notion to which we return in a moment. This schema is formally like the words that instantiate it, with the same three layers of structure. However, instead of a lexical base, it has open slots or **variables**: *X* in semantics, an unspecified noun in morphosyntax, and an arbitrary phonological string in phonology.

Next consider the coindexation in (5). Coindex 3 links the affix in morphosyntax with the phonology *-ish*. It also serves as a relational link to the affix in *sheepish* and all other *-ish* adjectives. However, the layers of the schema as a whole are linked by the **variable coindex** *z*, and the variables within the schema are linked by the variable coindex *y*. These coindices link (5) to any lexical entry with parallel coindexation. For instance, index *y* links to index 1 in *sheepish* and to index 5 in *childish*, shown in (6); and index *z* links to index 2 in *sheepish* and index 6 in *childish*. In other words, the constant coindex 3 in (5) links to what is the same across the instances, and the variable coindices link to what is different but parallel in structure.

(6) Semantics: $[\text{CHILD}_5\text{-LIKE; SILLY}]_6$
Morphosyntax: $[_A\ N_5 - \text{aff}_3]_6$
Phonology: $/\text{tʃaɪld}_5\ \text{ɪʃ}_3/_6$

Space prohibits us from discussing how coindexation applies in more complex morphological phenomena crosslinguistically; see Jackendoff and Audring 2020. For the coindexation notation in phrasal syntax, see Jackendoff 2002 and Culicover and Jackendoff 2005.

### 2.2 *Two functions of schemas*

Replacing rules by schemas may seem like a small move, but there is an important difference. Schemas have the same format as words, differing only in that part of their structure consists of variables and variable coindices. Hence words and schemas can be in the same 'place' in the theory – or in the mind. In other words, this theory does not need two independent constructs of "lexicon" and "grammar," and in that respect it is more constrained.

Another difference between schemas and rules lies in their function in the grammar. Traditional rules build up novel composite expressions from smaller parts, for example "add *-ish* to a noun to form an adjective." Schemas too can be used to build up novel composite expressions, by **unifying** their variables with other pieces of structure. So, for instance, if one encounters a new noun, say *wug*, one can unify it with the variable in the *-ish* schema to construct a new adjective *wuggish* with the meaning 'like a wug.' We will call this the **generative** function of schemas.

But schemas also have a second function: they can capture generalizations among existing lexical items. In this **relational** function of schemas, they encode what is the same among their instances. Consequently, the instances take less work to learn, store, and/or process. Well-known approaches to this function are **motivation** and **inheritance**, which we take up in Section 2.3.

In their relational function, schemas need not exhaustively specify the contents of the items they motivate. For instance, *sheepish* does not just mean 'like a sheep,' it means something like 'timid' with perhaps an overlay of 'embarrassed,' as in *a sheepish grin*. But it is still motivated by the *-ish* schema.

However, the relational use of schemas is not confined to lexical items with idiosyncratic properties. Regular complex forms must be stored when embedded in idioms (such as the plural in *raining cat<u>s</u> and dog<u>s</u>*) and when there is no corresponding singular form (as in *scad<u>s</u> of words*). Moreover, experimental research (e.g. Baayen, Dijkstra, and Schreuder 1997) has shown that speakers store many regular forms, such as Dutch *dieren* 'animals', which could in fact be created generatively.[5] These facts are problematic for a rule-based theory, which decrees that items generated by rules cannot be in the lexicon – what Langacker 1987 has called the "rule-list fallacy." In contrast, a schema for the regular past tense can apply to motivate these stored forms. It is the same schema, the same generalization, but instantiated here through relational rather than generative means.

The two functions of schemas are not confined to morphology. (7) is the phrase structure rule for English transitive VPs, formulated as a schema (or "treelet" in Janet Fodor's (1998) sense).

(7) $[_{\text{VP}}$ V – NP]

This schema obviously functions generatively in constructing novel instances like *swallow the beer*. But it can also function relationally to motivate the syntactic structure of idioms like *chew the fat* ('converse idly') which cannot be built online from their parts, because, although they may share some semantic features, overall they mean the wrong thing.

Crucially, not all schemas have both functions. The suffix *-en* in *harden* has many instances: *whiten, darken, soften*, and so on. However, unlike the *-ish* suffix, which can be used freely to generate new words such as *Trumpish*, the *-en* suffix is not productive: there is *soften* but not **louden, thicken* but not **slicken*, and so on. A great deal of morphology, especially derivational morphology, is like this: there is a generalization, but nevertheless one has to store the existing instances, and one is not free (or at least not *very* free) to make up new instances. The schema for $[_{\text{V}}$ A – *en*] is what has been called a "semiproductive rule" or a "lexical redundancy rule"; it has only a relational function, not a generative one.

Purely relational schemas also turn up in phrasal syntax, although less prominently, in what Culicover 1999 calls 'syntactic nuts' – bits of syntax that only occur with a few listed words. For example, the construction exemplified by *day after day* occurs with only a limited class of prepositions (8a) (Jackendoff 2010); and the little determiner formula in *what a job* occurs only with a restricted set of forms preceding the indefinite article (8b).

5. Here is a point where psycholinguistic research has a crucial bearing on the theory of representations.

(8) a. N-P-N: *day **after** day, week **by** week, face **to** face*
but: **student **before** student, gun **beside** gun*
b. [$_{Det}$ X a]: ***what** a job, **such** a job, **quite** a change*
but: ****who** a professor, ***where** a city*

One has to learn the instances of these patterns one by one. Hence these are syntactic patterns that cannot be encoded by productive generative rules. We conclude that syntax as well as morphology requires schemas functioning relationally.

In addition to schemas that have both the generative and relational functions, and schemas that have only the relational function, we might wonder if there could also be schemas that have only the generative function. The answer would be no. Such schemas cannot exist, because any instance of a schema that can be generated online can then be stored, where it falls under the relational function. One can store items of all sizes – catch phrases, song lyrics, even the entire Koran – without losing the internal structure that connects them to the grammatical patterns in the language.

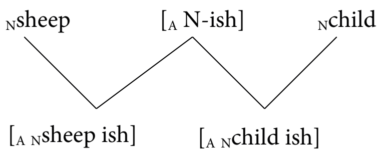
Hence there are basically two kinds of schemas: (a) productive schemas, which both serve the function of traditional generative rules and also capture generalizations over stored lexical items, and (b), nonproductive schemas, which have only the relational function. Thus one might consider productive schemas as relational schemas that have "gone viral"; in effect, the generative function is the innovative add-on.

This conclusion about rules of grammar reorients the perspective of linguistic theory. Modern linguistics has been primarily concerned with the creativity of language, Humboldt's "infinite use of finite means," and therefore it has focused on generative rules that create novel utterances. For instance, the second paragraph of Berwick and Chomsky, 2016: 1 says: "Generative grammar sought … to … explain what we will call the Basic Property of language: that a language is a finite computational system yielding an infinity of expressions…." However, once it is recognized that rules/schemas also serve in a relational function – and especially that many rules/schemas have *only* this function – it becomes apparent that we ought to focus at least as much on the relations within the lexicon, which cannot be described in terms of traditional generative rules. The relational function is the foundation on which the generative function builds.[6]

**6.** Distributed Morphology (Halle and Marantz 1993; Siddiqi 2019) can be seen as an attempt to treat nonproductive schemas in terms of generative rules. See Jackendoff and Audring (2020) for a critique.

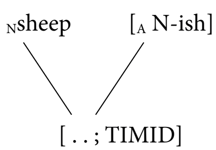
### 2.3 *Inheritance: Spelling out the relational function*

We now return to the question of what it means for a schema to "motivate" its instances, in the sense widely invoked in morphology (e.g. Daelemans, De Smet, and Gazdar 1992; Booij 2010, 2017; Lakoff 1987; Goldberg 1995; Radden and Panther 2004; and indeed de Saussure 1916/1977 (133)). A common position in these sources is that families of words like *sheep, sheepish*, and [$_A$ N -*ish*] motivate each other through an **inheritance hierarchy**, of the sort illustrated in (9). The items lower in the hierarchy are taken to inherit structure from items they are connected to higher in the hierarchy. Thus *sheepish* and *childish* inherit from *sheep* and *child* respectively, and both inherit from the [$_A$ N-*ish*] schema.

(9)



Inheritance is especially attractive because it has also been frequently invoked as a factor in the organization of concepts (Murphy 2002): the concept *poodle* inherits from *dog*, which inherits from *animal*, and so on. Hence inheritance is a domain-general theoretical construct that requires no special machinery for morphology or even for language per se. But what does inheritance mean? What do the lines signify?

A common interpretation of inheritance (e.g. Collins and Quillian 1969; Pollard and Sag 1994; Riehemann 1988) is that it serves to keep the lexicon maximally economical: any information present in a higher node of the hierarchy does not have to be specified in a lower node. For instance, *sheepish* inherits almost everything from the two higher nodes and therefore can be listed something like this:

(10)



Following Jackendoff 1975, we call this the **impoverished entry theory**: the idea is that lexical items contain only information that cannot be inherited from elsewhere.

Despite its intuitive appeal to elegance and minimum description length, there are good reasons to reject this position. Consider a case from conceptual inheritance: RJ's concept of his late cat Peanut can't specify the color of her paws without linking it to her paws. But by hypothesis, the fact that she has paws is inherited from his concept of cats in general, and is not a direct part of his concept of Peanut. In other words, according to this theory, the concept of Peanut per se contains nothing that the color can be the color *of*. As Bybee (2001: 7) puts it (citing Langacker 1987), "if predictable properties are taken away from objects, they become unrecognizable."

Another argument comes from the theory of acquisition. In order to construct a schema such as $[_A$ N-*ish*], one must generalize over existing lexical items whose details are present in memory. The impoverished entry theory forces one to claim that once the schema is established, the redundant details of all the instances used to establish it are immediately erased from memory, thereby optimizing the lexicon. We find this implausible (though psycholinguistic evidence might prove us wrong).[7] Similarly, in the course of acquiring a new complex word, one must first discover its details, and only then determine what schemas it falls under and what its base is. The impoverished entry theory suggests that as soon as one establishes the new word's relation to a base and to one or more schemas, all its redundant features are immediately expunged. Again, we find this implausible. (Similar arguments can be found in Langacker 1987 and Booij 2017.)

A third argument against impoverished entries is that inheritance is taken to be asymmetric: one lexical entry is the ancestor of the other. But consider pairs like *assassin/assassinate* or *linguist/linguistics*. On the basis of phonology, the second member of each pair should inherit from the first, just like *sheep/sheepish*. However, the semantic dependency goes the other way: an assassin is 'someone who assassinates people' and a linguist is 'someone who does linguistics.' This mixed dependency cannot be expressed in terms of standard inheritance with impoverished entries: Which is the ancestor and which is the dependent? What gets omitted from which entry?

On the basis of these and other arguments (Jackendoff and Audring 2020), we conclude that economy is the wrong criterion for a theory of lexical storage in the brain. An alternative that we find plausible is that the brain *embraces* redundancy, at least up to a point. For instance, languages seem to have no problem marking thematic roles redundantly, through word order, case marking, and verb agreement. Further afield, the visual system has at least five partially redundant mechanisms for depth perception: lens accommodation, eye convergence, stereopsis, occlusion, and perspective (Marr 1982). Redundancy appears to have the effect of making mental computation more robust (Libben 2006: 6).

7. This does not preclude the possibility that redundant information, lie any stored information, may decay and be forgotten. The point is that until this is the case, it is redundant; the impoverished entry theory claims this is impossible.

These precedents encourage us to adopt a **full-entry theory**, in which lexical items are encoded in their entirety, even where redundant.[8] In this approach, the lexical entry of *sheepish* is like (2), with its full structure, rather than like (10), in which it has been evacuated of content. We explicate motivation in terms of shared structure: the *-ish* schema (3) and the word *sheep* (1) motivate (2) by virtue of the structure they have in common. The relational links, encoded by coindexation, show precisely what parts of their respective structures the items share.

This formulation avoids the difficulties for the impoverished entry theory. First, the fact that Peanut has paws is directly present within her conceptual structure, coindexed with the paws of the more general schema for cats. Hence the idiosyncratic color of Peanut's paws has something to be attributed to. Second, there is no need to erase redundant information in instances of newly constructed schemas or in newly encountered instances of existing schemas. Third, unlike standard inheritance, coindexation is not inherently directional, and it can apply to the three levels of representation independently where necessary. This permits the mixed connections between *assassin* and *assassinate* to be expressed by coindexation as in (11): the semantics of *assassinate* is shared with part of the semantics of *assassin* (coindex 8), while the phonology of *assassin* is shared with part of the phonology of *assassinate* (coindex 7).

(11) a. Semantics: $[\text{PERSON WHO [MURDERS POLITICIAN]}_8]_7$
Morphosyntax: $N_7$
Phonology: /əsæsən/$_7$
b. Semantics: $[\text{MURDER POLITICIAN}]_8$
Morphoyntax: $[_V N_7 – \text{aff}_9]_8$
Phonology: /əsæsən$_7$ eɪt$_9$ /$_8$

We call such relations "sister" relations, in contrast to the mother-daughter relations of standard inheritance.

## 2.4 *Summary so far and prospectus*

To sum up the theory of representation:

- The lexicon contains words and schemas in the same format: pieces of semantic, morphosyntactic, and phonological structure, connected by interface links.
- Schemas differ from words only in that they contain variables.

8. We acknowledge that there are important questions about what it means to code an item "in its entirety." Does that include phonetic detail? Does it include semantic detail that might be termed "real-world knowledge"? We must leave these questions open. However, we resist the other extreme, the position that one stores all and only fully detailed exemplars.

- Words and schemas explicitly motivate each other through relational links, which designate pieces of structure shared among fully specified lexical entries, whether words or schemas.
- Both productive and nonproductive schemas can be used to motivate the structure of listed lexical items. Productive schemas can also be used to generate novel expressions on line. Both morphology and syntax have both kinds of schemas.

## 3. Acquisition of productive schemas relies on the relational function

In order to be able to generate an indefinitely large number of novel utterances, we obviously need productive schemas, functioning generatively. But the status of schemas in their *relational* function is less clear – especially nonproductive schemas, which function only relationally and never generate anything. One might legitimately wonder if the relational function is necessary at all. For instance, Pinker 1999 takes the position that there are rules for regular morphology – in our terms, productive schemas used generatively – but that there are only associations and analogies, not rules, for nonproductive patterns. (He does not reflect on productive schemas used relationally.)

There do exist some small families of words that do not fit any general pattern. For instance, the pairs *bomb/bombard, laugh/ laughter, hate/hatred*, and *compare/ comparison* have apparent suffixes that are at best exceedingly rare in modern English. Such pairs certainly do not warrant an abstract schema with variables. Nor is there a general schema that connects pairs like *assassin* and *assassinate* in the precise way they are related in (11). It is reasonable to think such cases are connected only as sisters.

On the other hand, fairly general – though nonproductive – schemas are intuitively more appealing for the hundreds of adjectives that end in *-ous* (*joyous, glorious, adventurous, avaricious* and so on) and the thousands of nouns that end in *-ion*. Nevertheless, it is worthwhile to insist: Are schemas really needed for these? Wouldn't analogy and association be enough?

Here is an argument from acquisition. Everyone who believes in rules of grammar agrees that children are trying to construct the productive rules of their language – here, the productive schemas – on the basis of primary linguistic input. The input of course contains no rules: rather, children must be discovering (or inferring) patterns in the input. This requires appropriate evidence to be stored in memory, in the form of stored words and fixed phrases. In other words, roughly following Tomasello 2003; Culicover and Nowak 2003; Bybee 2010, and many others, the acquisition of a schema has to be item-based.

However, just storing items is not enough: there must be some process that ranges over long-term memory (LTM), seeking commonalities among items.[9] In terms of the present framework, this process creates sister links among items that appear to share structure, for instance between *sheep* and *sheepish*, between *assassin* and *assassinate*, and between *chew* and *chew the fat*.

On the basis of such sister links, the learner must (unconsciously) create hypotheses for putative rules. In the present framework, such hypotheses take the form of tentative schemas, which are in the same format as words. The parts of the linked sisters that are alike form the constants in the schema (e.g. *ish*), and the parts that differ (e.g. *sheep, child, slug*) are replaced by variables in the schema. The acquisition process does not need to to shift from the format of words to an entirely different format for rules (a criticism mounted by Bybee 1995 of Pinker's "words and rules" dichotomy) – a formal advantage of the schema approach.

Now suppose that a learner has posited a tentative schema on the basis of some collection of sister items in LTM. Its function vis-à-vis these items is by necessity relational from the outset: the instances are already in memory. Crucially, though, the learner has no way of knowing in advance whether or not this schema can be used generatively as well – that is, whether it is productive. If in the end it is not – as will be the case most of the time, especially in (derivational) morphology[10] – the learner is left with a nonproductive schema that can only motivate lexical items that have been heard. In other words, nonproductive schemas are a logical steppingstone

---

**9.** We might conceptualize this process in various ways. First, there might be a global "web crawler" that (unconsciously) trolls LTM for similarities, and establishes links among similar items. Alternatively, individual lexical items might send out signals, looking for partners, and linking when they find a match. This might take place on first presentation of a word, or later, once the word is stored in memory. None of these possibilities really lends itself to neural implementation as it is currently understood (we think); one does not normally think of neural networks as adding connections.

Individuals undoubtedly differ in how thoroughly they search for shared structure. In particular, we should not exclude the possibility that children try out hypotheses that adults would find rather far-fetched. For instance, at about three years old, RJ's grandson Ezra responded to someone's remark about Saturday, saying "it makes you sad." For another case, Lila Gleitman admits (p.c.) that until well into adulthood, she was convinced that *robinson* meant an adventure involving a sea voyage, on the strength of *Swiss Family Robinson* and *Robinson Crusoe*.

**10.** What criteria does the learner use to determine whether a schema is productive? Goldberg 2005 offers some possibilities, as does Yang 2016. Here we observe only that the obvious factor of frequency cannot be the only factor in play. For instance, English has a construction *(all) X-ed out* 'worn out from too much X', as in *I'm coffeed/Olympic'd/historied/knitted out*. This construction is completely productive in choice of noun or verb for *X*. Yet it is quite infrequent; one hears it perhaps three times per year.

to open schemas, and the relational function of schemas is a *prerequisite* to their taking on a generative function.

A nonproductive schema can also aid in learning new words. If the child has the *-ish* schema and encounters a new instance, say *wolfish*, it is possible to decompose the word and at least partly interpret it, perhaps with the help of context. This differs from simply interpreting the new word on the basis of analogy with known words. By invoking a schema, the learner does not have to compare the new word unselectively with all potential sisters, whatever their degree and dimensions of similarity. The schema in effect offers a prestructured or "precompiled" basis for analogy.

## 4. Theory of processing: Basic assumptions

We now turn to the question of how nonproductive schemas might play a role in morphological processing. Our intuition is that schemas should facilitate the processing of morphologically complex words. This intuition seems to be confirmed at least for compounds: Fiorentino and Poeppel (2007) find that compounds are processed faster than comparable monomorphemic words. We wish to show how such a difference could come about in the case of nonproductive suffixation. In order to do so, we now briefly set out our thinking on language processing in general (building on Jackendoff 2002, 2007), and in particular on the processing of spoken words. Our approach shares a great deal with other theories of the lexicon. However, following Marr's desideratum of integrating representational and processing theories, we build features of our linguistic theory directly into the account of processing.

As in many approaches, we treat knowledge of language as a network consisting of linked nodes. Like some versions of lexical networks (e.g. Levelt, Roelofs, and Meyer 1999), its nodes fall into three layers, corresponding to semantic, morphosyntactic, and phonological structures. Such a network has no independent level of "lexical nodes." Rather, a "lexical node" is a complex of semantic, morphosyntactic, and phonological nodes connected by interface links.

A difference from most network models (including neural networks) is that we do not regard nodes in the network as simplex monads. Rather, they are *structures*, as in Booij 2010; Bybee 2010, and Kapatsinski 2007. For instance, the lexical entry for *sheepish* comprises semantic, morphosyntactic, and phonological nodes, connected by interface links, and each node has the internal structure shown in (2) above. In other words, the contents of the network mirror the representational theory's description of morphologically complex words.

Another important feature of our conception of the lexical network is that the relations among words are not instantiated by simple links among nodes. Rather, because the nodes have internal structure, their connections can be encoded by the relational links posited in Section 2, which pinpoint the regions of similarity among structures. So again, the representational theory determines details of the processing theory.

Yet another difference from traditional networks, again following the representational theory, is that the network includes not only all the words, but also all the rules, in the form of schemas; these too are lexical items, with all the privileges and responsibilities thereof.

We conceive of the lexicon as supporting spreading activation, in much the sense of standard approaches going back at least to Collins and Loftus 1975. However, again we propose a refinement arising from the representational theory: since nodes are now connected by interface and relational links, these are what form the paths for spreading activation. The strength of spreading activation can then be modulated in part by how extensive the linkage is.

We make some further assumptions about processing that are not linked directly to the representational theory. Departing from pure network models, but in concurrence with many other approaches, working memory (WM) is to be regarded as a functional component distinct from long-term memory (LTM); it does not consist simply of the parts of LTM that are active. LTM contains the lexical network of "knowledge of language" that we have just described; it corresponds roughly to the "memory" component in Hagoort's (2005) MUC ("memory, unification, control") approach. It is in LTM that schemas fulfill their relational function, through their links to more fully specified items. In contrast, working memory (WM) is the functional component in which pieces of lexical items are assembled into larger structures, either to create an utterance (in production), or to analyze and parse an input (in comprehension). WM is where productive schemas perform their generative function; it corresponds roughly to Hagoort's "unification" component. (For arguments for this division of labor between LTM and WM, including differences from Baddeley's (1986) conception of working memory, see Jackendoff 2002, 2007; Marcus 1998, 2001, and Gallistel and King 2009.)

Following much contemporary thought, tracing back to Swinney 1979; Tanenhaus, Leiman, and Seidenberg 1979, and Woods 1980, as well as to cohort theory and its descendants (Marslen-Wilson and Tyler 1980; Marslen-Wilson 1987; Marslen-Wilson and Zwitserlood 1989), we assume that processing is "promiscuous": in language comprehension, everything in LTM sufficiently similar to the current input is activated and retrieved into WM, where it serves as a candidate for "what is being heard," in competition with other candidates. The degree to which any particular candidate is activated in WM depends on a number of factors, including its current level of activation in LTM (see below), the relative strength of its competitors, and how well it fits the current context.

The competition among candidates need not be resolved immediately, as can be seen from minimal pairs such as (12) and (13), involving word identification and syntactic structure respectively. An ambiguity arises at the italicized words, but it cannot be resolved until the semantic processing of the underlined words. (We know of no experiments testing these particular situations.)

(12) a. That's not *a parent*; it's actually a teacher.
b. That's not *apparent*; it's actually quite obscure.

(13) a. *The* more attractive and intuitively plausible theory isn't always right. [*the* = determiner]
b. *The* more attractive and intuitively plausible a theory is, the better chance it has of becoming popular. [*the* = mark of comparative correlative construction] (Culicover and Jackendoff 2005)

A final assumption of our approach concerns the correlation between the speed of lexical access and the corpus frequency of the word being identified (Oldfield and Wingfield 1965). This actually involves two correlations. First, reaction time correlates with resting activation: the higher an item's level of resting activation, the faster it responds (Baayen, Dijkstra, and Schreuder 1997; Bybee 1985, 1995 uses the term "lexical strength"; see also Plag 2003). In turn, concurring with practically every theory of lexical acquisition, an item's level of resting activation is incrementally driven up by repeated use, along the familiar lines of Hebbian learning. Thus there is no direct representation of frequency in the brain. Rather, frequency in a written corpus stands proxy for frequency of use, which affects resting activation – which in turn affects reaction time. (Notice here the interplay between representational and processing theories. The structure of an item is determined by the representational theory, but its resting activation belongs to the theory of processing.)

Activation of a node in the LTM network can be raised above resting level through a number of familiar mechanisms.

- After an item is called by WM, we find it reasonable to assume that it does not return to resting activation immediately, but takes a little while to settle down. This provides an account of *identity priming*: if the item recurs in the input soon enough, its level is still above resting activation, and thus it is easier to summon it back to WM, resulting in more rapid identification.
- A node's activation may also be raised by activation that has spread along a relational link from another activated lexical item in LTM. This again makes it quicker to respond to a call from WM. This is *neighborhood priming*.
- A third possible source of heightened activation is semantic priming from the understanding of context. Such priming need not be lexical: for instance (12a) might be favored over (12b) because the speaker accompanied the utterance of *that* with a point to a person in a school.

– Finally, among neighborhood primes are fixed expressions containing the word in question. For a trivial instance, the input *baa baa black*… activates the fixed expression whose continuation is *sheep*, which in turn primes *sheep*. Hence hearing /ʃ/ activates the candidate *sheep* more strongly than the candidate *shoe*.

The Parallel Architecture's continuity between words and rules has an interesting consequence here. Since schemas are lexical entries, they too have a resting activation and can be called into WM, with a strength based on their frequency. This immediately yields an account of *syntactic priming*: if, for instance, one has heard a ditransitive dative structure (e.g. *Kate handed Sam a banana*), one is more likely to favor producing a ditransitive (e.g. *Max sent Anne a message*) over a prepositional dative (*Max sent a message to Anne*). If words and rules are entirely different mental entities, syntactic priming requires a separate mechanism, as noted by Bock and Loebell 1990 in an early description of the phenomenon. On the present account, syntactic priming is just another form of identity priming. The ditransitive dative construction is a lexical item in LTM, and in this case it has not entirely returned to resting activation, so at the moment it is easier to reactivate than usual (Wittenberg and Snedeker 2014).

This conception of WM processing offers a connection with probabilistic/predictive approaches to syntactic parsing (for syntax: Hale 2003, 2011; Levy 2008; Kuperberg and Jaeger 2015; for morphology, O'Donnell 2015). The basic idea of this approach is that at any particular moment, the processor is making predictions about what is to come in the input; these predictions are not all-or-nothing, but probabilistic. As new input comes in, its degree of surprisal – basically its deviation from prediction – affects processing time and electrophysiological measures of processing effort. Surprisal is measured via corpus frequency and cloze probability. Continuations with low surprisal (*baa baa black … <u>sheep</u>*) are processed faster and elicit a lower processing cost than continuations with high surprisal (*baa baa black … <u>shoe</u>*).

To translate this approach into present terms: A candidate interpretation of the input typically includes structure that continues beyond the input. This extra structure can be considered a prediction. For instance, an initial /ʃi…/ activates the cohort *she, sheep, sheen, sheet, sheath*, and so on, each of which can be considered a prediction about the continuation. The probability assigned to each prediction can be reframed as the relative strength of activation of each of the candidates, which, as we have seen, varies with frequency and priming. A low-surprisal continuation is one that conforms to a high-strength candidate, such as the continuation …*eep* after *baa baa black sh*… A high-surprisal continuation is one that conforms to a low-strength candidate, such as the continuation …*oe* in the same context. A high-surprisal continuation has the effect of dethroning a leading candidate and elevating a lowly one; this seems a reasonable cause for increased processing effort, in proportion to the disparity between the two competitors. We are not aware of any attempts to adapt the mathematics of the surprisal perspective to a framing in terms of competitive activation, but it should be possible.

On the other hand, the architecture we have proposed presents severe challenges to a theory of neural implementation. For one thing, it is unknown how to neurally implement any sort of structured representations such as the three levels of *sheepish* or *chew the fat*. Still more problematic is how material is retrieved from LTM into WM, and how WM unifies these retrieved pieces by instantiating variables. A third problem is how new words or phrases are learned: this requires adding new representations to the LTM network by copying them from WM, where they have been constructed in response to input. We are not aware of any proposals about neural computation that can fulfill these functions (with the possible exception of Smolensky and Legendre 2006). We take it that one of the Big Questions for cognitive neuroscience is how to reconcile these apparently incompatible demands from the representational theory and the theory of neural implementation.

## 5. Schemas in processing

We now return to the question of how morphological schemas in the sense of Section 2 play a role in processing.

Much of the literature on morphological processing is concerned with the question of lookup versus computation: When processing morphologically complex words, does one process them as whole words, or does one decompose them into constituents (Nooteboom, Weerman, and Wijnen 2002)? Various positions have been proposed:

– All decomposition, a.k.a. full parsing (Clahsen 1999; Taft 2004)
– All whole-word, a.k.a. full listing (Butterworth 1983, and in a way Rumelhart & McClelland 1986)
– Simultaneous decomposition and whole-word retrieval, in competition, a.k.a. Race model (Baayen 1993)
– Simultaneous decomposition and whole-word retrieval, potentially reinforcing each other (Schreuder & Baayen 1995; Kuperman et al., 2008))

Given the non-compositional semantics of many morphologically complex words, we can immediately reject the all-decomposition theory. Given speakers' ability to understand novel morphologically complex words – particularly in languages with exuberantly productive morphological systems, such as Turkish (Hankamer 1979), we can immediately reject the all-whole-word theory. The choice thus comes down to the latter two theories, both in the spirit of promiscuous access. Of these two, we favor the latter: when the decompositional and whole-word strategies result in incompatible candidates, they compete. But when they result in compatible candidates, they reinforce each other, and their redundancy creates a more robust outcome and potentially a faster reaction time.

A crucial part of the analysis is Relational Morphology's hypothesis that, in the representational theory, schemas are lexical items right alongside words. Hence in processing they should behave like just like words. In particular, they should have a resting activation correlated with the frequency with which they are used; and they should be able to be primed by related lexical items or schemas, thereby raising their activation above resting level.

In general, a schema ought to be easier to access than its instances, for two reasons. First, the frequency of a schema is the sum of the frequencies of its instances, so (at least for a first approximation) its resting activation ought to be higher than any of them. Second, an instance of a schema (say *sheepish*) might be expected to activate the schema itself (here, $[_{A}$ N – *ish*]) more strongly than it activates sister instances (say *childish*). The reason is that the schema's variable does not conflict with the *sheep* in *sheepish*, whereas the *child* in *childish* does, reducing the intensity of spreading activation.

To see how schemas might play a role in processing, we look step by step at the activation of four types of words: (a) as a baseline, a monomorphemic word such as *hurricane*; (b) a novel bimorphemic word such as *purpleness*; (c) a stored word with a legitimate suffix but whose base is not a lexical item, such as *scrumptious*; (d) a stored bimorphemic word such as *sheepish*.

We idealize to the situation where the word is being heard in isolation. External factors such as priming and syntactic or semantic context will affect all four cases comparably. For example, if previous context has set up a syntactic expectation of an adjective, say *the very…*, this bias will equally boost the activation of *scrumptious* and *sheepish*, and inhibit the activation of *hurricane* and *purpleness*. We also note that visually presented words will have a different timecourse, because the word can be perceived all at once rather than sequentially.

The monomorphemic word *hurricane* sets the background for the other three cases.

- By the time /hərɪk/ is heard, the uniqueness point has been reached, and the only viable candidate remaining is indeed *hurricane*.
- Hence the continuation of the input is predicted to be /eyn/, and when that comes, the prediction is confirmed.

Next consider *purpleness*, a novel word with a lexical base. This is an altogether plausible word, given *blackness, whiteness*, and *redness*, but it is not a stored lexical item (at least for the authors, before they made it up as an example).

– By the time /pɹ̩pl̩/ is heard, the cohort has been reduced to *purple*, whose semantics primes that of other color words to some extent.
– But this time there is no hint that *purpleness* is coming, since (by assumption) there is no such word in the lexicon that could be activated.
– When /nəs/ is heard, the lexical item *purple* is now incompatible with the input and it may start to decay in WM.[11]
– However, /nəs/ activates the [$_N$ A-*ness*] schema, which calls for a preceding adjective to instantiate its variable.
– *Purple*, still (somewhat) active in WM, satisfies the variable. It becomes reactivated and unifies with the [$_N$ A-*ness*] schema, accounting for the input.
– Meanwhile, [$_N$ A-*ness*] might prime all the listed *-ness* words to some degree.

Hence this is a pure case of what the literature calls decomposition or computation.

The third case is *scrumptious*. This has the common suffix *-ous*, but it is attached to a nonce base *scrum(p)(t)* (speakers differ on the exact pronunciation). The literature calls such bases "bound roots." They are ubiquitous in English morphology, for instance *deleterious, generous, egregious; commotion, contraption, compassion; deprecate, insulate*, and *vindicate*. To our knowledge, neither morphological theory nor psycholinguistic research has had much to say about such words. However, they are no marginal phenomenon. For instance, at least a third of the hundreds of *-ous* words in English are of this character, and the proportion of verbs with *-ate* is if anything higher.[12]

Our intuition is that such items ought to take less effort to identify than equally long and comparably frequent monomorphemic words such as *orchestra* and *hurricane*. Walking through the processing of *scrumptious*:

– By the time /skrʌmptʃ/ is heard, the cohort in WM has been reduced to the single word *scrumptious*; unlike with *purpleness*, there is no word **scrumpt* in the lexicon that could be activated.
– The word *scrumptious* in WM activates the [$_A$ X-*ous*] schema.

**11.** If the word were not heard in isolation, the syllable /nəs/ might be part of a following word, e.g. *purple nasturtium*, leading to further hypotheses in competition.

**12.** These estimates are based on an informal count by RJ of the words he recognizes in a reverse dictionary created by Mark Lindsay and brought to our attention by Mark Aronoff. (The exact numbers are of no consequence.)

- It also primes all the other *-ous* words such as *joyous* and *glamorous*. But the schema is primed more than its instances, because its variable does not conflict with /skrʌmptʃ/, whereas /dʒɔj/ and /glæməɹ/ do conflict with it. (Non-adjectival words such as *artifice* may be primed too, but even less, as they are morpho-syntactically unlike *scrumptious*.)
- The resting activation of [$_A$ X-*ous*] is higher than that of all its instances, as discussed above. Hence already at the point where only /skrʌmptʃ/ has been heard, the [$_A$ X-*ous*] schema is ready to go.
- When the whole word /skrʌmptʃəs/ has been heard, the lexical item *scrumptious* has already achieved a presence in WM and can satisfy the input.
- But in addition, now the [$_A$ X-*ous*] schema is called to WM, again priming all the *-ous* words a little.
- [$_A$ X-*ous*] then partially satisfies the input, redundantly.

We conjecture that this redundancy should add strength and/or robustness and/or speed to the identification of the input as the word *scrumptious*, compared to a monomorphemic word, which does not get the extra boost. We eagerly await the experiment. (Schreuder and Baayen 1995 argue that redundancy speeds processing in a different set of cases.)

Finally, we return to *sheepish*, which combines the conditions of the previous two cases: the whole word is a lexical item, and the base is also a lexical item.

- At the point where /ʃip/ has been heard, it activates the lexical items *sheep, sheepish, sheepherding*, and perhaps others, in proportion to their resting activation.
- All these lexical items prime each other, along with other neighbors.
- In addition, once *sheepish* is primed, it primes all the other *-ish* words, but especially the [$_A$ N-*ish*] schema (just as *scrumptious* primes the [$_A$ X-*ous*] schema above).
- When the whole word /ʃipɪʃ/ has been heard, the lexical items *sheep* and *sheepherding* are incompatible with the input and start to decay; *sheepish* becomes activated still more.
- But now [$_A$ N-*ish*] is called to WM, and because it has been primed, it comes relatively quickly.
- It satisfies the input, redundantly supporting the affix of *sheepish* (parallel to *-ous* in *scrumptious*).
- In addition, its variable can be unified with *sheep*, creating a structure that is compatible with the input (parallel to the case of *purplish*). This revives the activation of *sheep* (but does not revive *sheepherding*). Hence the interpretation of the input as the word *sheepish* is supported by both the lexical item *sheepish* and the assembly of the lexical items *sheep* and [$_A$ N-*ish*].

We therefore expect *sheepish* to be identified more robustly and faster than monomorphemic *hurricane*, purely compositional *purpleness*, and partially compositional *scrumptious*, all else being equal.

The compositional route naturally takes more steps than the whole-word route: the base has to be retrieved, the affix schema has to be retrieved, and the base has to be unified with the affix. Each step takes time. So the contribution of the whole word will tend to dominate in determining the speed of processing. However, if the whole word is infrequent, hence relatively slow, but the base and the affix are both frequent, hence relatively fast, the compositional route may be faster and/or stronger. This is in accord with Baayen and Lieber 1991, who show that the compositional route takes precedence if the affix in question has a large number of low-frequency instances, and particularly if the base is of relatively high frequency compared to the derived word. (There may in addition be individual differences in this balance. For instance, Reifegerste 2014 shows that older speakers tend to rely on storage to a greater extent than younger speakers, who make relatively greater use of composition.)

Hay and Baayen 2005 discuss such gradient results and conclude that the compositionality of morphologically complex items is not discrete but graded: words are divided into morphemes *to such-and-such a degree*. We suggest that the gradient they observe is a result of the relative strength and/or speed of whole-word access vs. compositional access. In turn, compositional access depends on (a) the resting activation of the affix (which depends on how numerous and how frequent its instances are), (b) how similar the remaining part of the word is to a lexical base (in *sheepish*, a lot; in *malicious*, less, because of the phonological differences from *malice*; in *scrumptious*, not at all), and (c) if there is a lexical base, its own resting activation.

## 6. Conclusions

We have sketched here the theory of Relational Morphology, based on many antecedents in the literature. It is important that this is not simply a theory of morphology: it serves as an additional component of the Parallel Architecture, helping to fulfill the aspiration for an integrated theory of linguistic description. In particular, we have taken care to show that many innovations we have proposed for morphology apply to syntax as well.

We have furthermore tried to show that RM can be embedded directly into a processing theory of some sophistication. Our reasoning rests on a large number of premises that we have tried to lay out here.

From the representational theory of Relational Morphology:

- Words and rules (now schemas) are stated in the same format, namely pieces of linguistic structure connected by interface links.
- Words and schemas are stored together in a network of lexical relations, whose connections are formalized as relational links.
- Productive rules/schemas are only one part of the grammar; lexical relations are at least equally as important.

From the processing theory in which Relational Morphology is embedded:

- Schemas are stored in memory with a resting activation related to their frequency, just like words.
- Therefore schemas can prime their instances and be primed by them.
- Schemas unify in WM with their bases, just like verbs unify with their arguments.

Admittedly, we have only scratched the surface here. For instance, similar analyses need to be constructed for visual word recognition (both in ordinary reading and in eye-tracking), for production in picture naming and continuous speech, for phoneme and word monitoring, for masked priming, and other natural and experimental tasks. Nor have we said anything about the many other complex effects (e.g. neighborhood density) cited by e.g. deJong 2004, Kuperman et al. 2008, Moscoso del Prado Martín et al. 2004, Bien, Baayen, and Levelt 2010, Kapatsinski 2007, and Amenta and Crepaldi 2012. Nevertheless, we hope our approach offers an incentive to revisit the experimental literature and to design new experiments. We also hope that it might invite computational modeling that allows one to fine-tune the parameters of the theory such as activation strength, the effects of spreading activation, and the relative timing of the many events taking place in lexical access, in order to best simulate the experimental results.

## Acknowledgements

This article is largely excerpted from Jackendoff and Audring (2020), which deals with these issues in much greater detail. For important suggestions, we are grateful to Geert Booij, Peter Culicover, Gina Kuperberg, Pim Levelt, Peter Hagoort, Ari Goldberg, three reviewers, and audiences at talks, workshops, and seminars at Tufts, Harvard, MIT, and the Max Planck Institute for Psycholinguistics.

## References

Amenta, S., and Crepaldi, D. (2012) Morphological processing as we know it: An analytical review of morphological efffecs in visual word identification. *Frontiers in Psychology*. https://doi.org/10.3389/fpsyg.2012.00232

Baayen, R. H., Dijkstra, T., and Schreuder, R. (1997) Singulars and Plurals in Dutch: Evidence for a Parallel Dual-Route Model, *Journal of Memory and Language* 37, 94–117. https://doi.org/10.1006/jmla.1997.2509

Baayen, R. H., and Lieber, R. (1991) Productivity and English derivation: a corpus-based study. *Linguistics* 29, 801–844. https://doi.org/10.1515/ling.1991.29.5.801

Baddeley, A. (1986) *Working Memory*. Oxford: Clarendon Press.

Berwick, R., and Chomsky, N. (2016) *Why Only Us?* Cambridge, MA: MIT Press. https://doi.org/10.7551/mitpress/9780262034241.001.0001

Bien, H., Baayen, R. H., and Levelt, W. J. M. (2011) Frequency effects in the production of Dutch deverbal adjectives and inflected verbs. *Language and Cognitive Processes* 26, 683–715. https://doi.org/10.1080/01690965.2010.511475

Bock, K., and Loebell, H. (1990) Framing sentences. *Cognition* 35, 1–39. https://doi.org/10.1016/0010-0277(90)90035-I

Booij, G. (2010) *Construction Morphology*. Oxford: Oxford University Press.

Booij, G. (2017) Inheritance and motivation in Construction Morphology. In Nikolas Gisborne and Andrew Hippisley (eds.), *Default inheritance*. Oxford: Oxford University Press. https://doi.org/10.1093/oso/9780198712329.003.0002

Butterworth, B. (1983) Lexical representation. In B. Butterworth (ed.), *Language Production: Vol. 2*, 257–294. San Diego, CA: Academic Press.

Bybee, J. (1985) *Morphology: a Study of the Relation between Meaning and Form*. Amsterdam: John Benjamins. https://doi.org/10.1075/tsl.9

Bybee, J. (1995) Regular morphology and the lexicon. *Language and Cognitive Processes* 10: 425–455. https://doi.org/10.1080/01690969508407111

Bybee, J. (2001) *Phonology and Language Use*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511612886

Bybee, J. (2010) *Language, Usage and Cognition*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511750526

Bybee, J., and D. Slobin. (1982) Rules and schemas in the development and use of the English past tense. *Language* 58: 265–289. https://doi.org/10.1353/lan.1982.0021

Chomsky, N. (1965) *Aspects of the Theory of Syntax*. Cambridge, MA: MIT Press.

Clahsen, H. (1999) Lexical entries and rules of language: A multidisciplinary study of German inflection. *Behavioral and Brain Sciences* 22, 991–1060. https://doi.org/10.1017/S0140525X99002228

Collins, A., and Loftus, E. (1975) A *spreading-activation* theory of semantic processing. *Psychological Review* 82, 407–428. https://doi.org/10.1037/0033-295X.82.6.407

Collins, A., and Quillian, M. (1969) Retrieval time from semantic memory. *Journal of Verbal Learning and Verbal Behavior* 9, 240–247. https://doi.org/10.1016/S0022-5371(69)80069-1

Culicover, P. W. (1999) *Syntactic Nuts*. Oxford: Oxford University Press.

Culicover, P. W., and Jackendoff, R. (2005) *Simpler Syntax*. Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199271092.001.0001

Culicover, P. W., and Nowak, A. (2003) *Dynamical Grammar*. Oxford: Oxford University Press.

Daelemans, W., De Smet, K., and Gazdar, G. (1992) Inheritance in natural language processing. *Computational Linguistics* 18, 205–218.
deJong, N. (2002) *Morphological Families in the Mental Lexicon*. Nijmegen: Max Planck Institute for Psycholinguistics.
de Saussure, F. (1916) *Cours de linguistique générale*, ed. C. Bally and A. Sechehaye, with the collaboration of A. Riedlinger, Lausanne and Paris: Payot; trans. W. Baskin, *Course in General Linguistics*, Glasgow: Fontana/Collins, 1977.
Fiorentino, R. and Poeppel, D. (2007) Compound words and structure in the lexicon. *Language and Cognitive Processes* 22, 953–1000. https://doi.org/10.1080/01690960701190215
Fodor, J. D. (1998) Learning to parse? In D. Swinney (ed.) *Anniversary issue of Journal of Psycholinguistic Research*, 27.2, 285–318. https://doi.org/10.1023/A:1023258301588
Gallistel, C. R., and King, P. A. (2009) *Memory and the Computational Brain*. Malden, MA; Wiley-Blackwell. https://doi.org/10.1002/9781444310498
Goldberg, A. (1995) *Constructions: A Construction Grammar approach to argument structure*, University of Chicago Press, Chicago.
Goldberg, A. (2005) *Constructions at work*, Oxford University Press, New York. https://doi.org/10.1093/acprof:oso/9780199268511.001.0001
Hagoort, P. (2005) On Broca, brain, and binding: a new framework. *Trends in Cognitive Sciences* 9, 416–423. https://doi.org/10.1016/j.tics.2005.07.004
Hale, J. (2003) 2003. The Information Conveyed by Words in Sentences. *Journal of Psycholinguistic Research* 32, 101–123. https://doi.org/10.1023/A:1022492123056
Hale, J. (2011) What a rational parser would do. *Cognitive Science* 35, 399–443. https://doi.org/10.1111/j.1551-6709.2010.01145.x
Halle, M., and Marantz, A. (1993) Distributed Morphology and the pieces of inflection. In K. Hale and S. J. Keyser (eds.), *The View from Building 20*, 111–176. Cambridge MA: MIT Press.
Hankamer, J. (1989) Morphological parsing and the lexicon. In W. Marslen-Wilson (ed.), *Lexical Representation and Process*, 392–408. Cambridge, MA: MIT Press.
Hay, J., and Baayen, H. (2005) Shifting paradigms: Gradient structure in morphology, *Trends in Cognitive Science* 9.7, 342–348. https://doi.org/10.1016/j.tics.2005.04.002
Hoffman, T. and Trousdale, G. (eds.) (2013) *The Oxford Handbook of Construction Grammar*. Oxford: Oxford University Press. https://doi.org/10.1093/oxfordhb/9780195396683.001.0001
Jackendoff, R. (1975) Morphological and Semantic Regularities in the Lexicon, *Language* 51.3, 639–671. https://doi.org/10.2307/412891
Jackendoff, R. (2002) *Foundations of Language*. Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198270126.001.0001
Jackendoff, R. (2007) A Parallel Architecture Perspective on Language Processing, *Brain Research* 1146, 2–22. https://doi.org/10.1016/j.brainres.2006.08.111
Jackendoff, R. (2010) *Meaning and the Lexicon*. Oxford: Oxford University Press.
Jackendoff, R. and Audring, J. (2020) *The Texture of the Lexicon: Relational Morphology and the Parallel Architecture*. Oxford: Oxford University Press.
Kapatsinski, V. (2007) Frequency, neighborhood density, age-of-acquisition, lexicon size, neighborhood density and speed of processing: Towards a domain-general, single-mechanism account. In S. Buescher, K. Holley, E. Ashworth, C. Beckner, B. Jones, and C. Shank. *Proceedings of the 6th Annual High Desert Linguistics Society Conference*, 121–40. Albuquerque, NM: High Desert Linguistics Society.

Kuperberg, G., and Jaeger, F. (2016) What do we mean by prediction in language comprehension? *Language, Cognition, and Neuroscience* 31: 32–59. https://doi.org/10.1080/23273798.2015.1102299

Kuperman, V., Bertram, R. and Baayen, R. H. (2008) Morphological Dynamics in Compound Processing. *Language and Cognitive Processes*, 23, 1089–1132. https://doi.org/10.1080/01690960802193688

Lakoff, G. (1987) *Women, Fire, and Dangerous Things*. Chicago: University of Chicago Press. https://doi.org/10.7208/chicago/9780226471013.001.0001

Langacker, R. (1987) *Foundations of Cognitive Grammar, vol. 1*, Stanford University Press, Stanford, CA.

Levelt, W., Roelofs, A., and Meyer, A. (1999) A theory of lexical access in speech production. *Behavioral and Brain Sciences* 22, 1–75. https://doi.org/10.1017/S0140525X99001776

Levy, R. (2008) Expectation-Based Syntactic Comprehension. *Cognition* 106(3):1126–1177. https://doi.org/10.1016/j.cognition.2007.05.006

Libben, G. (2006) Why study compounds? An overview of the issues. In G. Libben & G. Jarema, (Eds.), *The representation and processing of compound words*. Oxford: Oxford University Press (pp. 1–21).

Marcus, G. (1998) Rethinking eliminative connectionism. *Cognitive Psychology* 37, 243–282. https://doi.org/10.1006/cogp.1998.0694

Marcus, G. (2001) *The Algebraic Mind*. Cambridge MA: MIT Press. https://doi.org/10.7551/mitpress/1187.001.0001

Marr, D. (1982) *Vision*. San Francisco: Freeman.

Marslen-Wilson, W., & Tyler, L. K. (1980). The temporal structure of spoken language understanding. *Cognition* 8, 1–71. https://doi.org/10.1016/0010-0277(80)90015-3

Marslen-Wilson, W. (1987) Functional parallelism in spoken language understanding. *Cognition* 25, 71–102. https://doi.org/10.1016/0010-0277(87)90005-9

Marslen-Wilson, W., and Zwitserlood, P. (1989) Accessing spoken words: The importance of word onsets. *Journal of Experimental Psychology: Human Perception and Performance* 15, 576–585.

Murphy, G. (2002) *The Big Book of Concepts*. Cambridge, MA: MIT Press. https://doi.org/10.7551/mitpress/1602.001.0001

Nooteboom, S., Weeman, F., and Wijnen, F. (eds.) (2002) *Storage and Computation in the Language Faculty*. Dordrecht: Kluwer. https://doi.org/10.1007/978-94-010-0355-1

O'Donnell, T. J. (2015) *Productivity and Reuse in Language: A Theory of Linguistic Computation and Storage*. Cambridge, MA: MIT Press. https://doi.org/10.7551/mitpress/9780262028844.001.0001

Oldfield, R., and Wingfield, A. (1965) Response latencies in naming objects. *Quarterly Journal of Experimental Psychology* 17, 273–281. https://doi.org/10.1080/17470216508416445

Pinker, S. (1999) *Words and rules*, Basic Books, New York.

Plag, I. (2003). *Word-Formation in English*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511841323

Pollard, C. and Sag, I. (1994) *Head-driven Phrase Structure Grammar*, University of Chicago Press, Chicago.

Radden, G., and Panther, K.-U. (2004) Introduction: Reflections on motivation. In G. Radden and K.-U. Panther (eds.), *Studies in Linguistic Motivation*, 1–46. Berlin, New York: Mouton de Gruyter.

Reifegerste, J. (2014) *Morphological Processing in Younger and Older People: Evidence for Flexible Dual-Route Access*. Nijmegen: Max Planck Institute for Psycholinguistics.

Riehemann, S. (1988) Type-based derivational morphology. *Journal of Comparative Germanic Linguistics* 2, 49–77. https://doi.org/10.1023/A:1009746617055

Rumelhart, D. (1980) Schemata: The building blocks of cognition. In R. Spiro, B. Bruce, and W. Brewer, *Theoretical Issues in Reading Comprehension*: 33–58. Hillsdale, NJ: Erlbaum.

Rumelhart, D., and McClelland, J. (1986) On learning the past tense of English verbs. In J. McClelland, D. Rumelhart, and the PDP Research Group, *Parallel Distributed Processing*, vol. ii, 216–71. Cambridge, MA: MIT Press. https://doi.org/10.7551/mitpress/5236.003.0008

Schreuder, R., and R. H. Baayen. (1995) Modeling morphological processing. In L. Feldman (ed.), *Morphological aspects of language processing*, 131–154. Hove: Erlbaum.

Siddiqi, D. (2019) Distributed Morphology. In J. Audring and F. Masini (eds.), *The Oxford Handbook of Morphological Theory*. Oxford: Oxford University Press, 143–165.

Smolensky, P., and Legendre, G. (2006) *The Harmonic Mind*. Cambridge, MA: MIT Press.

Swinney, D. (1979) Lexical access during sentence comprehension: (Re)consideration of context effects. *Journal of Verbal Learning and Verbal Behavior* 18, 645–659. https://doi.org/10.1016/S0022-5371(79)90355-4

Taft, M. (2004) Morphological decomposition and the reverse base frequency effect. *Quarterly Journal of Experimental Psychology*, 57A, 745–765. https://doi.org/10.1080/02724980343000477

Tanenhaus, M., Leiman, J. M., Seidenberg, M. (1979) Evidence for multiple stages in the processing of ambiguous words in syntactic contexts. *Journal of Verbal Learning and Verbal Behavior* 18, 427–440. https://doi.org/10.1016/S0022-5371(79)90237-8

Tomasello, M. (2003) *Constructing a Language: A Usage-Based Theory of Language Acquisition*. Cambridge MA: Harvard University Press.

Ullman, M. T. (2015) The declarative/procedural model: A neurobiologically motivated theory of first and second language. In B. VanPatten and J. Williams (Eds.), *Theories in second language acquisition: An introduction* (2nd ed.). New York: Routledge. pp. 135–158.

Wittenberg, E. and Snedeker, J. (2014) It takes two to kiss, but does it take three to give a kiss? Categorization based on thematic roles. *Language, Cognition and Neuroscience* 29.5, 635–641. https://doi.org/10.1080/01690965.2013.831918

Woods, W. (1980) Multiple theory formation in speech and reading. In R. Spiro, B. Bruce, and W. Brewer, *Theoretical Issues in Reading Comprehension*: 59–82. Hillsdale, NJ: Erlbaum.

Yang, Charles. (2016) *The Price of Linguistic Productivity*. Cambridge, MA: MIT Press. https://doi.org/10.7551/mitpress/9780262035323.001.0001

# Author index

# Subject index

From its beginnings, the study of the mental lexicon has been at the crossroads of research and scholarship. This volume presents a polylogue – a textual conversation of many voices. It is designed to capture the excitement within the field and generate a deeper understanding of key issues and debates for established researchers, students, and readers interested in language and cognition. The first chapter examines how the mental lexicon itself can be seen as a polylogue. In the following six chapters, authors tackle the fundamental questions concerning future research on lexical representation and processing in an interactive structure that presents new perspectives and captures the excitement of the field. The themes include the value of cross-linguistic megastudies, the nature of meaning, how to capture truly natural language, what can be learned from lexical acquisition, the advantages of a functionalist perspective, and the role of schemas in understanding morphology and the lexicon.

ISBN 978 90 272 0925 2

9 789027 209252

John Benjamins Publishing Company