



THE UNIVERSITY *of* EDINBURGH

This thesis has been submitted in fulfilment of the requirements for a postgraduate degree (e.g. PhD, MPhil, DClinPsychol) at the University of Edinburgh. Please note the following terms and conditions of use:

This work is protected by copyright and other intellectual property rights, which are retained by the thesis author, unless otherwise stated.

A copy can be downloaded for personal non-commercial research or study, without prior permission or charge.

This thesis cannot be reproduced or quoted extensively from without first obtaining permission in writing from the author.

The content must not be changed in any way or sold commercially in any format or medium without the formal permission of the author.

When referring to this work, full bibliographic details including the author, title, awarding institution and date of the thesis must be given.

Word length and the Principle of
Least Effort
Language as an evolving, efficient code for
information transfer

Jasmeen Kanwal

Doctor of Philosophy
School of Philosophy, Psychology, and Language Sciences
University of Edinburgh
2017

Lay Summary

In most languages, words that are used more frequently tend to be shorter than words that are used less frequently. In English, for example, “the”, “it”, and “and” are among the most frequent words, while words like “enhancement”, “hypotenuse”, and “carbuncle” are among the least frequent. The linguist George Zipf hypothesised that this inverse relationship between word length and frequency is the result of an optimisation process. Speakers are lazy, and inclined to produce the shortest utterances possible, even if that means risking ambiguity or potentially being misheard. Hearers, on the other hand, prefer as little ambiguity as possible, to ensure they can understand what is being said, even if it means the speaker has to use longer, less easily confusable words. An optimal solution for balancing these two competing pressures would be to keep the most frequently used words short, while allowing less frequent words to be longer. This minimises the average length of any given utterance, while still mostly keeping distinct words for distinct meanings.

In this thesis, I present the results of experiments that test Zipf’s hypothesis. Participants learn miniature made-up languages and then I observe how they reshape their input as they communicate with other participants. I find that, when under pressures to be both efficient and accurate in their communications, participants tend to use shorter words for more frequent or predictable meanings. When these pressures are removed, optimised languages do not emerge. This suggests that these competing communicative pressures are directly responsible for inciting optimisation behaviour. I then investigate a large data set of text from books published over the last 100 years to see whether these effects accumulate over time, producing the inverse frequency-length relationships observed in so many languages. I find that, indeed, for words whose frequency is increasing over time, a shorter variant of that word gains in popularity faster than it does for other words. Overall, these results support the theory that factors related to the communicative use of language are instrumental in shaping certain universal features of language structure.

Abstract

In 1935 the linguist George Kingsley Zipf made a now classic observation about the relationship between a word’s length and its frequency: the more frequent a word is, the shorter it tends to be. He claimed that this “Law of Abbreviation” is a universal structural property of language. The Law of Abbreviation has since been documented in a wide range of human languages, and extended to animal communication systems and even computer programming languages. Zipf hypothesised that this universal design feature arises as a result of individuals optimising form-meaning mappings under competing pressures to communicate accurately but also efficiently—his famous Principle of Least Effort.

In this thesis, I present a novel set of studies which provide direct experimental evidence for this explanatory hypothesis. Using a miniature artificial language learning paradigm, I show in Chapter 2 that language users optimise form-meaning mappings in line with the Law of Abbreviation only when pressures for accuracy and efficiency both operate during a communicative task. These results are robust across different methods of data collection: one version of the experiment was run in the lab, and another was run online, using a novel method I developed which allows participants to partake in dyadic interaction through a web-based interface.

In Chapter 3, I address the growing body of work suggesting that a word’s predictability in context may be an even stronger determiner of its length than its frequency alone. For instance, Piantadosi et al. (2011) show that shorter words have a lower average surprisal (i.e., tend to appear in more predictive contexts) than longer words, in synchronic corpora across many languages. We hypothesise that the same communicative pressures posited by the Principle of Least Effort, when acting on speakers in situations where context manipulates the information content of words, can give rise to these lexical distributions. Adapting the methodology developed in Chapter 2, I show that participants use shorter words in more predictive contexts only when subject to the competing pressures for accurate and efficient communication. In a second experiment, I show that participants are more likely to use shorter words for meanings with a lower average surprisal. These results suggest that communicative pressures acting on individuals during language use can lead to the re-mapping of a lexicon to align with “Uniform Information Density”, the principle that information

content ought to be evenly spread across an utterance, such that shorter linguistic units carry less information than longer ones.

Over generations, linguistic behaviour such as that observed in the experiments reported here may bring entire lexicons into alignment with the Law of Abbreviation and Uniform Information Density. For this to happen, a diachronic process which leads to permanent lexical change is necessary. However, crucial evidence for this process—decreasing word length as a result of increasing frequency over time—has never before been systematically documented in natural language. In Chapter 4, I conduct the first large-scale diachronic corpus study investigating the relationship between word length and frequency over time, using the Google Books Ngrams corpus and three different word lists covering both English and French. Focusing on words which have both long and short variants (e.g., *info/information*), I show that the frequency of a word lemma may influence the rate at which the shorter variant gains in popularity. This suggests that the lexicon as a whole may indeed be gradually evolving towards greater efficiency.

Taken together, the behavioural and corpus-based evidence presented in this thesis supports the hypothesis that communicative pressures acting on language-users are at least partially responsible for the frequency-length and surprisal-length relationships found universally across lexicons. More generally, the approach taken in this thesis promotes a view of language as, among other things, an evolving, efficient code for information transfer.

Acknowledgements

My journey as a linguistics PhD student began at UCSD, where the wonderful linguistics faculty there welcomed me as a fugitive from the philosophy PhD program. Thanks to Eric Baković, Ben Bergen, Gabriela Caballero, Marc Garellek, Grant Goodall, Andy Kehler, Robert Kleunder, Roger Levy, and John Moore for teaching some especially memorable classes. And to Farrell Ackerman, for discussions that deeply influenced my thinking about linguistics, and for becoming an all-round top-notch mentor and friend. UCSD was as understanding as they were welcoming, when I jumped ship yet again, this time to follow my destiny and migrate to Scotland to join the Centre for Language Evolution at the University of Edinburgh.

My first day at the CLE involved a trip to Dagda. By the end of my first pint of black isle porter, I knew I had arrived home. A more weird and wonderful group of humans surely isn't to be found anywhere in the world. The ethos of the CLE—that our interests outside of science are just as important as, even inextricable from, our interests within science—led by the shining example of acclaimed artist and funk guitarist extraordinaire Simon, buoyed me through the most challenging periods of this PhD and meant that I always felt supported and understood, and certainly never bored. Simon, Jenny, and Kenny: you've each in your own way been the best supervisor ever. The constant sense I got from each of you that you hadn't the slightest flicker of doubt that I would finish this PhD (or your excellent ability to hide any such doubt) is the main thing that kept me from ever reaching a state of sheer panic. I guess what I'm trying to say without bursting into tears is: thanks for believing in me.

Shoutouts to all my fellow PhD pals, both at UCSD and the CLE: Gazza, Yazza, Amanda, Jon, Marieke, and my co-convener of the 5.01 writing+gossip club, Carmen Saldaña. Also to Olga, Marieke, Chrissy, Stella, and Isabelle, for letting me join their elite, super-duper-ladies-only postdoc office for a brief stint.

I'm pretty sure the idea that led to this entire PhD project came from a discussion with Vanessa Ferdinand over deep fried red bean ice cream in Berkeley, California. She also gets the credit for luring me to Edinburgh after we met at the complex systems summer school in Santa Fe and became bffs4lyfe after eating poisoned berries together. To my other co-PhD/co-Amtrak/bandmate/SHEEPmate/bff4lyfe Kevin Stadler: thanks and no thanks for always being around.

Finally, this thesis is dedicated to my parents, whose science PhDs inspired me to decide as a toddler that I would someday pursue one myself. Without their completely unwavering support (both financial and emotional), basically nothing I've ever done in my life would have been possible. To Charlie Murphy, for being a warm lump at my side while writing much of this thesis, and to Cooshka, for being a warm lump on

the other side, and much more.

Declaration

I declare that this thesis was composed by myself and that the work contained therein is my own, except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified. As detailed in the introduction to Chapter 2, part of this work has previously been published as a journal article jointly-authored with my supervisors, in the journal *Cognition*.

(Jasmeen Kanwal)

‘Minimum effort, maximum effect!’

—*Tom Hodgkinson, The Idler*

Contents

Abstract	iv
1 Introduction	1
1.1 Introducing G. K. Zipf	1
1.2 The Law of Abbreviation	3
1.2.1 Evidence from natural language corpora	3
1.2.2 Evidence in a programming language	5
1.2.3 Evidence in animal communication systems	5
1.3 The Principle of Least Effort	7
1.3.1 Information theoretic reformulation	8
1.4 Probability in context	9
1.5 Motivating this thesis project: what is the status of the Least Effort hypothesis?	11
1.5.1 Random typing models	11
1.5.2 Behavioural evidence for optimisation during communication	12
1.5.3 Diachronic evidence for optimisation processes	14
1.6 Artificial language learning as a tool for investigating the evolution of linguistic structure	15
1.7 Language as an efficient code for information transfer	17
2 Word length and frequency	21
2.1 Online experiment (Experiment 1)	21
2.1.1 Introduction	21
2.1.2 Miniature Artificial Language Learning Experiments	25
2.1.3 Results	30
2.1.4 Discussion	36
2.1.5 Conclusions	39
2.2 Lab experiment (Experiment 2)	39
2.2.1 Participants	40
2.2.2 Materials	40
2.2.3 Procedure	40

2.2.4	Results	40
2.2.5	Discussion	42
3	Word length and predictability in context	45
3.1	Introduction	45
3.2	Experiment 3	48
3.2.1	Participants	48
3.2.2	The Training Language	49
3.2.3	Training Procedure	51
3.2.4	Testing Procedures	51
3.2.5	Results	54
3.2.6	Discussion	58
3.3	Experiment 4	59
3.3.1	Participants	60
3.3.2	The Training Language	60
3.3.3	Training Procedure	62
3.3.4	Testing Procedure	62
3.3.5	Results	64
3.3.6	Discussion	68
3.4	General discussion and conclusions	69
4	The diachronic evolution of long/short word pairs	73
4.1	Introduction	73
4.2	The Google Books Ngrams corpus	77
4.3	Info/information: a case study	79
4.3.1	Data	79
4.3.2	Linear regression analysis	81
4.3.3	Granger causality analysis	81
4.3.4	Discussion	84
4.4	Experiment 5: English clipped pairs	85
4.4.1	Data set	85
4.4.2	Linear regression analyses	86
4.4.3	Granger causality analyses	87
4.4.4	Discussion	88
4.5	Experiment 6: English <i>-ic/-ical</i> pairs	90
4.5.1	Data set	90
4.5.2	Linear regression analyses	91
4.5.3	Granger causality analyses	92
4.5.4	Discussion	94
4.6	Experiment 7: French clipped pairs	94

4.6.1	Data set	95
4.6.2	Linear regression analyses	96
4.6.3	Granger Causality analyses	96
4.6.4	Discussion	98
4.7	General Discussion and Conclusions	99
5	Summary and conclusion	103
A	Relevant publications	107
B	Word lists used in Chapter 4 (Experiments 5-7)	123

List of Figures

1.1	Zipf’s original hand-drawn schema for the inverse relationship between frequency and ‘conspicuousness’ of a linguistic unit. Reproduced from Zipf (1929).	2
1.2	The information-theoretic model of communication, as laid out by Shannon in his foundational paper ‘a Mathematical Theory of Communication’ (1948). This figure is reproduced from the article.	8
2.1	The 1000 most frequent words in English. Each point represents an individual word (some points are labeled). The red line marks the mean frequency for the words of each length (here, orthographic length is used, but the same overall pattern would be seen if phonetic length were used instead.) The more frequent a word is, the shorter it tends to be. According to Zipf’s Law of Abbreviation, this is a universal pattern of human languages. Frequency counts used here are from the 450 million word COCA corpus (Davies, 2008).	22
2.2	a) A schematic diagram of the frequencies of the objects and labels presented during the training trials in all four experimental conditions. One object appeared three times more frequently than the other. Each object was labeled half the time with its unique long name, and half the time with its ambiguous short name, which was a clipped version of the long name. b) An example training trial. c) An example of a director trial in the Combined condition (top) and a matcher trial followed by feedback (bottom).	27

2.3	The proportion of trials in which the short name was used to label the frequent object versus the proportion of trials in which it was used to label the infrequent object. For the Combined (a) and Accuracy (b) condition, each data point combines a pair of communicating players, representing the sum of their director trial productions. For the Time (c) and Neither (d) condition, each data point corresponds to an individual player’s productions. The size of the circles is perceptually scaled (Tanimura et al., 2006) to reflect the number of data points coinciding at each value. Data from only the second half of testing trials is shown here, as participants were more likely to have converged on a stable mapping by this time. Data points in the top right quadrant of each graph indicate participants who are mostly using the short name for both objects; participants are clustered in this quadrant only in the Time condition (graph c). Data points in the bottom left quadrant of each graph indicate those who are mostly using the unique long names for both objects; participants are most clustered here in the Accuracy condition (graph b). Data points in the bottom right quadrant of each graph indicate participants who are mostly using the short name for the frequent object and the long name for the infrequent object. This behaviour, consistent with the Law of Abbreviation, only reliably arises in the Combined condition (graph a), where both pressures are present.	31
2.4	The average token length of an individual participant’s ‘language’ (the full set of all their director trial productions) plotted against the expressivity (the mutual information between the forms and meanings) of their language. The size of the circles is perceptually scaled (Tanimura et al., 2006) to reflect the number of data points coinciding at each value. The input language that participants are exposed to in training trials is marked with an asterisk, and the grey points represent possible output languages. (Possible output languages are constrained by the number of different expressivity values that are possible for a language with a given average token length. For example, there is only one possible configuration for both the shortest and longest average token lengths—all objects are either mapped to the short name or the long name, respectively—and thus only one possible expressivity value at the endpoints.) The optimal language—the language with the minimum avg. token length while achieving maximum expressivity—is marked with a target symbol	33
2.5	Timecourse of productions in the critical Combined condition. Each data point shows the average word length taken over all participants’ productions at a given repetition number of an object.	34

2.6	The proportion of trials in which the short name was used to label the frequent object versus the proportion of trials in which it was used to label the infrequent object, for the lab experiment (a), and the online experiment (b). For the Combined and Accuracy conditions, each data point combines a pair of communicating players, representing the sum of their director trial productions. For the Time and Neither conditions, each data point corresponds to an individual player’s productions. The size of the circles is perceptually scaled (Tanimura et al., 2006) to reflect the number of data points coinciding at each value. Data from only the second half of testing trials is shown here, as participants were more likely to have converged on a stable mapping by this time. Data points in the top right quadrant indicate participants who are mostly using the short name for both objects; participants are clustered in this quadrant in the Time condition. Data points in the bottom left quadrant indicate those who are mostly using the unique long names for both objects; participants are clustered here in the Accuracy condition. Data points in the bottom right quadrant indicate participants who are mostly using the short name for the frequent object and the long name for the infrequent object. This behaviour, consistent with the Law of Abbreviation, only reliably arises in the Combined condition, where both pressures are present. Both the lab and online experiments exhibit qualitatively similar behaviour.	41
3.1	(A) The input frequencies of the objects and framing sentences presented during training trials in all four experimental conditions of Experiment 3. (B) A sample training trial from Experiment 3. (C) A sample director trial from the Combined condition of Experiment 3 (top) and a matcher trial followed by feedback (bottom).	50
3.2	The proportion of trials in which the short name was used in predictive contexts versus the proportion of trials in which it was used in surprising contexts in Experiment 3. For the Combined and Accuracy condition, each data point combines a pair of communicating players, representing the sum of their director trial productions. For the Time and Neither condition, each data point corresponds to an individual player’s productions. The size of the circles is perceptually scaled (Tanimura et al., 2006) to reflect the number of data points coinciding at each value. Data from only the second half of testing trials is plotted here, as participants were more likely to have converged on a stable mapping by this time. These results demonstrate that behaviour consistent with the principles of uniform information density and smooth signal redundancy—using short forms in predictive contexts and long forms in surprising contexts, generating systems that fall in the bottom right corner of each graph—only reliably arises in the Combined condition.	55

3.3	The extent to which individual participants' name choices are conditioned on context (lefthand graph) and object (righthand graph) in Experiment 3. The dotted line in the lefthand graph represents the mutual information between name and context (MI_c) associated with the 'optimal' language in UID terms—the language in which the short form is used only in predictive contexts, and the long form only in surprising contexts. $MI_c=0$ for the input language. In the righthand graph, mutual information between name and object (MI_o) can range from 0 (same name fixed for both objects) to 1 (distinct names fixed for each object). $MI_o=0.5$ for the input language, marked by the dotted line. Data from only the second half of testing trials is shown in this figure, as participants were more likely to have converged on a stable mapping by this time.	57
3.4	A schematic diagram showing the frequencies of the object-sentence pairings presented during training trials in Experiment 4. The average surprisal of each object is given in the bottom row. Object A has a lower average surprisal than objects B and C, and thus we predict that, in the output lexicon, it will also have a lower average label length.	61
3.5	(A) A screenshot of a sample training trial from Experiment 4. (B) A sample director trial (top) and a matcher trial followed by feedback (bottom) from Experiment 4.	63
3.6	The proportion of times individual participants used the short name for the lower average surprisal object (object A) versus the higher average surprisal objects (objects B and C), during the second half of testing trials in Experiment 4. The proportion of trials in which each object appeared in its most predictive context is marked with an asterisk. As the asterisks fall close to the medians of each distribution, this suggests that most participants adopted the strategy of using the short name in only the most predictive context for each object, and using the long name elsewhere.	64
3.7	The proportion of times individual participants used the short name in different contexts, during the second half of testing trials in Experiment 4. Contexts are labeled here by the probability of the given object appearing in that context.	67
3.8	The number of participants who are perfectly optimising their form-meaning mappings in accordance with the uniform information density principle (i.e., using the short name only in the most predictive context, and the long name elsewhere) for each object, as a function of trial block. Each block consists of 24 director trials.	68

4.1	Number of words per year in the Google Books Ngrams corpus. The density of tokens is not evenly spread in time, but increases year by year, rising especially rapidly after 1950. Adapted from Michel et al. (2011), Supplementary Online Material p. 62. Copyright 2011 by AAAS. Reprinted with permission.	78
4.2	Meaning frequency (MF , blue) and short form advantage (SFA , red) for the the <i>info/information</i> clipped pair over time. The shape of the short form advantage time series appears to closely follow that of the meaning frequency. Note that the two time series correspond to different y-axes.	80
4.3	The differenced time series for meaning frequency (MF' , blue) and short form advantage (SFA' , red) for the the <i>info/information</i> clipped pair. Note that the two time series correspond to different y-axes. . . .	83
4.4	The meaning frequency (blue lines) and short form advantage (red lines) for the 25 French clipped pairs analysed in Experiment 7. The two time series in each graph correspond to different y-axes, but none of the y-axes are marked here for reasons of space and readability. . . .	97

List of Tables

2.1	Summary of fixed effects for a binomial regression model with short name use as the binary dependent variable, and by-participant random effects for object frequency. Like Figure 2.3, this model is fit using only the second half of each participant’s training trial data, as participants were more likely to have converged on a stable linguistic mapping by then.	32
2.2	Summary of fixed effects for a binomial regression model with short name use as the binary dependent variable, and by-participant random effects for object frequency and trial number. This model is fitted to the data from all participants’ production trials in the Combined condition.	35
2.3	Summary of fixed effects for a binomial regression model with short name use as the binary dependent variable, and by-participant random effects for object frequency, fit to data from only the lab experiment. Like Figure 2.6, this model is fit using only the second half of each participant’s testing trial data, as participants were more likely to have converged on a stable linguistic mapping by then. Significant effects are in bold.	42
2.4	Summary of fixed effects for a binomial regression model with short name use as the binary dependent variable, and by-participant random effects for object frequency, fit to the combined data set of the lab and online studies. The key predicted effects, which were also present in the online data alone (Table 2.1), are in bold.	43
3.1	Summary of fixed effects for a binomial regression model with short name use as the binary dependent variable. Significant effects are in bold.	54
3.2	Summary of fixed effects for a binomial regression model with short name use as the binary dependent variable. “Prob” signifies “probability of the target object occurring in the given context”. Significant and trending effects are in bold.	66

4.1	List of English <i>-ic/-ical</i> pairs that show Granger causality in the predicted direction only—a change in the meaning frequency causes a subsequent corresponding change in the short form advantage (and not also the other way around).	93
4.2	List of English <i>-ic/-ical</i> pairs that show Granger causality in the <i>non-predicted</i> direction only—a change in the short form advantage time series appears to precede a subsequent corresponding change in the meaning frequency time series (and not also the other way around).	93
B.1	List of English clipped pairs used in Experiment 5.	124
B.2	List of English <i>-ic/-ical</i> pairs used in Experiment 6. Only the short form from each pair is included here for reasons of space.	125
B.3	List of French clipped pairs used in Experiment 7.	126

Chapter 1

Introduction

1.1 Introducing G. K. Zipf

One feature that unites communicators of all stripes is that we are lazy. The idea that this tendency towards laziness might shape certain structural properties of our communication systems was perhaps most prominently advanced and developed in the work of linguist George Kingsley Zipf. Zipf first planted the seed of this idea in his PhD thesis, *Relative frequency as a determinant of phonetic change* (1929). In that thesis, submitted to the department of comparative philology at Harvard, Zipf describes a number of examples of historical sound change in Indo-European languages. The changes include loss of accent or stress on syllables, lenition and devoicing of consonants, and vowel shift. Generalising over these observations, Zipf proposes a *principle of frequency*:

The accent, or degree of conspicuousness, of any word, syllable, or sound, is inversely proportionate to the relative frequency of that word, syllable, or sound, among its fellow words, syllables, or sounds, in the stream of spoken language. As usage becomes more frequent, form becomes less accented, or more easily pronounceable, and vice versa (Zipf, 1929; p. 4).

The inverse relationship between frequency and ‘conspicuousness’ is illustrated in a schematic diagram (Figure 1.1). By way of explanation for this inverse relationship, Zipf lays down the following argument. Accented syllables, certain vowels, and consonants with more voicing, aspiration, or “explosiveness”, all require relatively greater effort to produce than other sounds. A speaker, being lazy, will not pronounce “a single sound more forcible than is necessary” for the hearer to understand her (p. 82). Therefore, she should expend extra articulatory effort only on those sounds which the hearer is most likely to be unfamiliar with, and thus potentially misunderstand. Hearers will be most familiar with those sounds which occur most frequently in the language; the sounds which are less frequent are the ones they are less likely to be

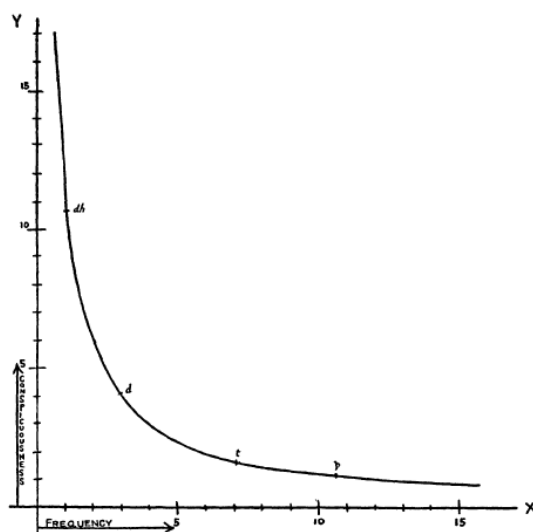


Figure 1.1: Zipf's original hand-drawn schema for the inverse relationship between frequency and 'conspicuousness' of a linguistic unit. Reproduced from Zipf (1929).

familiar with, and thus most likely to mishear. Given this, the speaker should produce less frequent sounds with a higher degree of conspicuousness, and more frequent sounds with a lower degree of conspicuousness. This results in an inverse proportionality between frequency and conspicuousness.

The line of reasoning followed here laid the groundwork for Zipf's famous *Principle of Least Effort* (PLE), which he proposed as an explanation for the inverse relationship he observed between word length and word frequency. In his 1935 book, *the Psycho-biology of language*, Zipf begins by discussing why he wishes to focus on the study of words. Words, he notes, are both composed of less complex linguistic units (phonemes) and can be combined together to form more complex linguistic units (phrases). Thus, they are the 'Goldilocks' choice: ideally sized linguistic units whose behaviour might be representative of what is going on at both higher and lower levels of organisation.

Just as sounds differ along the parameters of voicing, aspiration, etc., one of the most striking parameters along which words differ is their length. This naturally leads Zipf to the question of "what, if any, is the significance of these observable differences in length?" (Zipf, 1935; p. 20). No doubt influenced by his earlier findings on the link between frequency and sound change, Zipf embarks on a meticulous quantitative assessment of the relationship between word length and word frequency in large bodies of texts. First he uses the pre-calculated frequencies from a German frequency dictionary compiled by F. W. Kaeding, based on a sampling of written texts containing over 10 million words total. Zipf tabulates the number of occurrences of each word in the frequency dictionary against the number of syllables it contains.

He finds that monosyllabic words comprise nearly 50% of all tokens, disyllabic words nearly 30%, trisyllabic only 13%, and so on, until words between seven and 15 syllables together comprise less than 0.3% of all tokens. He then examines a corpus of colloquial Chinese, made up of 20 different samples of connected speech, each sample containing one thousand syllables; a corpus of Latin containing four plays by Plautus, totalling $\sim 33,000$ words; and a corpus of American newspaper English, totalling $\sim 44,000$ words. In all three cases, Zipf finds that “the magnitude of words tends, on the whole, to stand in an inverse (not necessarily proportionate) relationship to the number of occurrences” (p. 25). Given that these findings hold across different languages, genres, and measures of length (he looks at phonemes in English and syllables in the other languages), Zipf concludes that this relationship is likely universal. He terms this relationship—wrapped up together with its diachronic corollary, which will be discussed further below—the Law of Abbreviation.

At the time Zipf’s work was published, linguistic generalisations were based on qualitative descriptions of linguistic phenomena. His method of using rigorous quantitative analysis of large bodies of texts in order to derive statistical generalisations about language was groundbreaking. For this, Zipf’s work is often acknowledged as the foundation of modern statistical and corpus linguistics (Prün, 2002; Antieau, 2013).

1.2 The Law of Abbreviation

Today, we would deem a claim of universality based on relatively small (by modern standards) corpora from only four different languages to be somewhat extravagant. However, Zipf’s claim has only been strengthened by time. The inverse relationship between word length and frequency has been observed consistently, even as the size and diversity of available corpora grow. Not only has it been found to hold for a wide range of natural languages, but it has also been observed in an organically-evolving computer programming language, as well as in the communication systems of some non-human animals. I review these findings below.

1.2.1 Evidence from natural language corpora

Using the now-classic ~ 1 million word Brown corpus of English, together with a ~ 1 million word newspaper corpus of Swedish, Sigurd et al. (2004) found an approximate inverse relationship between word length and word frequency in both languages. Using a ~ 1 million word corpus of Mandarin Chinese newspaper articles, Teahan et al. (2000) came across a similar finding: single-character words are the by far the most frequent words in the corpus. As the character-length of words grows, the proportion of the corpus made up of these words goes down.

This same pattern is found using even larger, more recently available corpora. Piantadosi et al. (2011) looked at 10 European language corpora, each containing approximately 100 billion words extracted from publicly available web pages. The languages included were: Czech, Dutch, French, German, Italian, Polish, Portuguese, Romanian, Spanish, and Swedish. A significant correlation between word length and unigram probability (i.e., frequency) was found for all languages.

Traversing orders of magnitude in the opposite direction, it turns out that these results are not restricted to large corpora composed of many texts. Strauss et al. (2007) conducted a cross-linguistic study in which they examine 11 individual texts: two in Russian (one prose and one poetry), and one each in English, German, Slovenian, Czech, Slovak, Croatian, Hungarian, Sundanese, and Indonesian. The inverse relationship between word length and word frequency was found to hold independently in all 11 texts.

The above studies were all carried out on corpora based on written language, albeit of different genres. Ferrer-i-Cancho and Hernández-Fernández (2013) ran a study using the CHILDES corpus, a cross-linguistic corpus based on spoken language (specifically, child-produced and child-directed speech). The languages included in the analysis were US English, Swedish, Russian, Croatian, Greek, Spanish, and Indonesian. Once again, a significant inverse relationship between word length and word frequency was found in all cases.

In addition to frequency bearing a relationship to phonological word length (i.e. the number of graphemes, phonemes, or syllables in a word), a relationship between frequency and phonetic realisation has also been observed. Wright (1979) gave participants a list of monosyllabic words from the Brown corpus to read aloud. Words were split into a high frequency and a low frequency class (using the frequency counts from the corpus), and were matched for grapheme length across the two classes. The authors found that the phonetic duration (in milliseconds) of the high frequency words was significantly shorter than that of low frequency words with the same grapheme length. A later study by Whalen (1991) asked participants to read a word list containing pairs of high-frequency/low-frequency homophones (e.g. *through/threw*). This ensured that the high-frequency/low-frequency pairs were exactly matched for number and type of phonemes, which was a possible confounding factor for the Wright (1979) study. Additionally, the authors ensured that the low-frequency word in each pair contained equal or fewer graphemes than the high-frequency word. Even in these conditions, the high-frequency words were found to have a shorter phonetic length than the the low-frequency words.¹

Finally, even when the domain of interest is limited to certain specific subsets of

¹Relatedly, less frequent words have also been found to contain rarer phonotactic sequences, making them more phonetically distinct on average than frequent words (Meylan and Griffiths, 2017).

the lexicon, evidence of this pervasive inverse frequency-length relationship is found. In Russian, nearly all verbs have two different root forms, one for the perfective aspect and one for the imperfective aspect. For any given verb, these roots will differ in length, but which form—the perfective or imperfective—is shorter is not consistent across all verbs. Fenk-Oczlon (2001) finds that when the usage frequencies of each root in the pair are compared, the more frequently used root is almost always the shorter one. Moreover, she finds that frequency is a better predictor of aspectual root length than measures of semantic markedness. Fenk-Oczlon also argues that frequency is a better predictor of case marker length than semantic markedness is. She cites the example of the genitive plural case for Russian feminine and neuter nouns, which is zero-marked. This zero marking clearly cannot be explained through markedness arguments, but can be explained by the relatively high usage frequency of this case.

In sum, the Law of Abbreviation has been observed in large written language corpora, in individual texts, and in spoken language corpora. Moreover, it has been observed across a wide range of languages from different language families. It is even found when phonetic rather than phonological word length is measured, and when the dataset is restricted to morphological paradigms within a lexicon. We can therefore plausibly conclude that the inverse relationship between word length and word frequency is a robust pattern found universally within the lexicons of human languages.

1.2.2 Evidence in a programming language

Intriguingly, there is some evidence that this inverse relationship is not restricted to natural languages, but to any organically-evolving code used by humans. A study by Ellis and Hitchcock (1986) investigated 10 NASA workers' use of the `alias` function as they programmed using the Unix C-shell scripting language. They found that more experienced users tended to adopt shorter abbreviations for the strings they used most frequently, resulting in a low average string length. Less experienced users were not as efficient in their use of aliasing, and thus the average length of their produced strings was higher. This suggests some movement towards optimisation, as a user becomes more familiar with the language and more aware of how frequently they use different commands.

1.2.3 Evidence in animal communication systems

Beyond human languages, natural or artificial, there is evidence that the Law of Abbreviation may even hold in the communication systems of some non-human animals. Formosan macaques are old world primates endemic to Taiwan. Semple et al. (2010)

collected 375 hours of field recordings of the vocalisations of this species. These vocalisations were then analysed (blindly) for call duration and frequency of occurrence in the data set. A significant negative correlation between call duration and frequency was found, even when the data was subsetting to include only the calls used by members of all age-groups in the population.

However, when the vocalisations of two other primate species—the common marmoset and the golden-backed uakari, both new world primates—were analysed in a study by Bezerra et al. (2011), a significant negative correlation between call duration and frequency was not found. Instead, the authors found that context played an important role in determining signal length. Shorter calls were associated with situations involving vigilance and danger, where rapid signalling is most critical. Longer calls were associated with long-range communication, where noise and distance may lead to signal degradation. In these situations, added redundancy might increase the likelihood of signal detection. These results suggest that, while the call durations of the two species investigated here may not appear optimised when only frequency of use is considered, they may be optimally coded with respect to other considerations. In addition, the authors call attention to other potential measures of effort in primate communication, over and above call duration. For example, brief calls that are also loud may require more effort to produce than longer but quieter calls. When such considerations are taken into account, an inverse relationship between production effort and frequency may yet exist, aside from the contribution of context.

Turning to another class of mammalian species known for their complex vocalisations, Luo et al. (2013) analyse the social and distress calls of four species of echolocating bats. They distinguish between short-range and long-range communication signals within a species' call repertoire, and argue that we should only expect to find a negative length-frequency correlation in short range calls. This is because long-range calls tend to be subject to intense selective pressures for length and complexity, as they are often used to advertise reproductive or territorial information. Incidentally, this might explain the lack of a significant negative correlation found in Bezerra et al.'s primate study above, which included both types of call in the same analysis.

In the short-range social and distress calls recorded in the bats, each instance was broken down into 'syllables'. These were defined as the smallest single sound elements, such as noise bursts or segments of constant frequency. These syllables can be considered analogous to words in human languages, as bats combine them in different ways to form complex calls with distinct referents (Kanwal et al., 1994). In the social calls of all four species of bat tested, a significant negative correlation between syllable length and frequency of use was found. No significant correlation was found, however, in the distress calls. Combining these results with those of the primate

studies, a consistent picture begins to emerge—long-range calls and calls associated with danger or distress are special cases in which call duration is strongly influenced by factors relevant to the context and function of use. However, in the short-range social calls used for ‘normal’ communication in some non-human animals, the inverse relationship between utterance length and frequency is found.

Ferrer-i-Cancho and Lusseau (2009) extend the evidence for this beyond vocalisations, to the surface behavioural patterns of bottlenose dolphins. These physical behaviours, such as ‘side-flop’ and ‘lobtail’, have been shown to have important communicative function within a dolphin school. Ferrer-i-Cancho and Lusseau decompose the surface behaviours into distinct, mutually exclusive behavioural units. E.g., side-flop = jump + side (2 units), and lobtail = stationary + hit + tail (3 units). The number of units per behavioural pattern was compared to its frequency of occurrence in a set of observations of a dolphin population in Doubtful Sound, New Zealand. A significant negative correlation between the number of units in a surface behaviour and its frequency of occurrence was found.

These studies of non-human animal communication, together with the findings in human natural languages as well as a programming language, point to the universality of the Law of Abbreviation across *all* naturally evolved communication systems (see also Ferrer-i-Cancho et al., 2013). This suggests that the explanation for this law is fundamentally tied up with the nature of communication itself. I elaborate on this hypothesis further in section 1.5 below.

1.3 The Principle of Least Effort

Zipf’s proposed explanation for the universality of the Law of Abbreviation followed much the same lines as his explanation of the Principle of Frequency for sound change, given in his 1929 PhD thesis. Speakers are concerned with saving time and effort, which can be achieved by using few, short words. However, there is a trade-off between this and ensuring that the hearer understands what is being said. The hearer’s concerns are best served by using many distinct, longer words, which are less likely to be ambiguous (Zipf, 1935, 1949). Balancing these two competing demands leads to what Zipf refers to as an equilibrium state, in which frequent words are short, and less frequent words are longer. This optimal state is reached through “the accumulated effects of abbreviatory acts of truncation during the long periods of years in which language has slowly evolved” (Zipf, 1935; p. 33). The acts of truncation, he assumes, disproportionately affect frequent long words (this is the diachronic corollary to the Law of Abbreviation mentioned earlier), thus giving rise to the inverse frequency-length relationship observed in language.

This “underlying law of economy”, in which speakers seek out the least effortful way

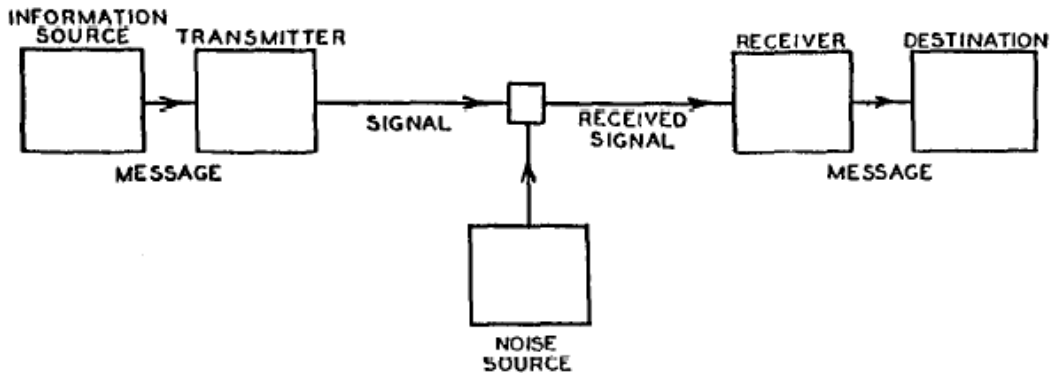


Figure 1.2: The information-theoretic model of communication, as laid out by Shannon in his foundational paper ‘a Mathematical Theory of Communication’ (1948). This figure is reproduced from the article.

to accurately convey information to the hearer, is dubbed the *Principle of Least Effort*. Zipf hypothesised that this principle is the driving force behind the maintenance of an optimal frequency-length mapping in the lexicon.

1.3.1 Information theoretic reformulation

Over a decade after Zipf first published his ideas about equilibrium and the “law of economy”, Shannon published his groundbreaking paper ‘a Mathematical Theory of Communication’ (1948), which established the field of information theory. It turns out that Shannon’s rigorous mathematical formulation of communication provides the tools for a new, more general formulation of Zipf’s Principle of Least Effort, in terms of information theoretic concepts.

According to Shannon’s model of communication, information is encoded in a signal, and then transmitted from the source to the receiver through a noisy channel (see Figure 1.2). For a communication system to be optimal in information-theoretic terms, it must satisfy two requirements. First it must encode the information from the source efficiently, i.e. the source signal should be compressed as much as possible. This encapsulates Zipf’s conjecture that speakers are lazy, and thus prefer low-effort productions. Second, the information must be encoded so as to minimise transmission errors introduced by the noisy channel. This can be done by adding redundancy, or simply maintaining pre-existing redundancy, and this captures Zipf’s proposed conflicting pressure for communicative accuracy—in a communication system, the hearer also imposes a pressure on the speaker, to produce utterances that can be understood. According to information theory, the optimal solution to these conflicting pressures is to assign signal length proportional to the amount of information contained in the signal, and thus maintain a roughly constant flow of information over the channel, at a rate arbitrarily close to the channel capacity.

In information theory, the information content of a signal is quantified in terms of its probability. The less probable a signal is, the more surprising it is, and thus the more information it carries. The information content—also termed the *surprisal*—of a signal is thus given as $\log \frac{1}{P}$, where P is the probability of the signal occurring.

If our focus is on words, and the probability of a word is simply measured in terms of how many times it occurs in the speech stream, then a word’s information content will scale inversely with its frequency of occurrence. As stated above, the optimal way of assigning signal length, according to information theory, is to set signal length proportional to information content. Therefore, the information-theoretically optimal coding system will assign word length inversely proportional to word frequency, giving us Zipf’s Law of Abbreviation.

1.4 Probability in context

This information-theoretic reformulation of Zipf’s Principle of Least Effort not only provides an explanation for the Law of Abbreviation in terms of optimal coding theory, but it opens up the scope to make further predictions. One key prediction—for which there is now well-established evidence in the literature—is that a more salient estimate of the information content of a linguistic unit can be obtained by considering its probability *in context*, rather than simply its marginal probability, or frequency, in the speech stream. In language, words rarely occur in isolation; a consideration of the context surrounding a word’s position can provide a great deal of information about which word is expected in that position. Thus, for example, while the word *beholder* is relatively low-frequency overall, in the carrier phrase “beauty is in the eye of the...”, the probability of *beholder* appearing next is close to 1. Therefore, to capture the intuition that in this instance, the word *beholder* carries little new information, we need to account for its probability *in context* in our measure of information content. This can be done by specifying the measure of information content, or surprisal, as $\log \frac{1}{P(w|c)}$, where $P(w|c)$ is the probability of word w occurring in context c .

Under this more general measure of information content, an optimal coding system would assign word length inversely proportional to a word’s *probability in context*. This prediction, it turns out, is indeed borne out by the data. A study by Piantadosi et al. (2011) compares the inverse relationship between word length and frequency to the relationship between word length and *average surprisal* in 11 different Indo-European languages. The average surprisal of a word w is obtained by averaging over the set C of different contexts it appears in:

$$\sum_{c \in C} P(c|w) \log \frac{1}{P(w|c)} \tag{1.1}$$

Note that in Piantadosi et al.’s study, the relevant context was taken to be the two words preceding the target word. (This measure of probability in context is also referred to as the trigram probability.) In all 11 languages, the authors find that the negative correlation between word length and average surprisal is stronger than that between word length and frequency. A study by Manin (2006) finds a similar result, from a different angle. Manin shows that longer words are harder for participants to guess correctly in a game where words are blotted out at random from a range of different Russian text fragments. This implies that longer words correspond to a higher average surprisal.

Word length is not only modulated by the number of phonemes contained in a word, but can also be modulated by the addition or deletion of entire morphemes. In an artificial language learning study, Fedzechkina et al. (2012) show that participants preferentially use optional case markers when grammatical roles are ambiguous. When other disambiguating information is present, thereby reducing the information content of the target word, case markers are omitted, resulting in a shorter production.

Another way of reducing the length of a word, without changing its underlying phonemic form, is through phonetic reduction. There is ample evidence that this type of word reduction also reliably occurs when a word is more predictable in context. Lieberman (1963) compared the phonetic duration of different tokens of the same word in predictive and surprising contexts, as read aloud by three participants. He found that word duration was shorter in the predictive contexts. Aylett and Turk (2004) investigated a corpus of spoken language, and found that shorter phonetic duration was correlated with higher trigram probability, as well as a higher level of another measure of probability in context—semantic givenness, i.e. how many times that word had already been mentioned in the same discourse context.

Gahl and Garnsey (2004) use yet another measure of probability in context: syntactic probability, i.e. how likely a word is to appear in different syntactic constructions. For example, they look at verbs which differ in how likely they are to either take a direct object, or introduce a complement clause. They find that verbs appearing in more syntactically likely contexts are also more likely to undergo verb-final /t,d/-deletion. Both Tily et al. (2009) and Kuperman and Bresnan (2012) find a similar effect on overall word duration, for a different syntactic construction: the dative alternation. Finally, Seyfarth (2014) shows that a word’s average probability in context, measured in terms of bigram probability, also affects its phonetic duration, over and above effects in each specific context. Interestingly, all these different measures of probability in context may contribute separately to the phonetic duration of a word; a regression study by Bell et al. (2009) finds that frequency, bigram probability, and semantic givenness each have a significant effect on word duration, even when the others are partialled out.

Zipf proposed an explanation for the inverse relationship between word length and frequency in terms of an “equilibrium” reached under competing communicative pressures. The framework provided by information theory allowed us to recast this explanation in terms of an optimisation process, between the drive for maximally compressed yet minimally error-prone signals. It also allowed us to extend the prediction of an inverse relationship between word length and frequency, to a more general prediction of an inverse relationship between a word’s length and its *probability in context*. This quantity arguably provides a more accurate measure of the information content of a word, as it takes into account the specific context surrounding the word, rather than just considering the limiting case where the ‘context’ is the entire corpus of productions. The strong evidence that has subsequently been found for this more general relationship between word length and information content appears to further support the hypothesis that it results from an optimisation process aimed at producing efficient coding systems. However, as I will discuss in the following section, observation of synchronic distributions is not enough to uniquely justify this proposed explanation: an investigation into the mechanistic and evolutionary processes that actually lead to the observed synchronic distributions is needed.

1.5 Motivating this thesis project: what is the status of the Least Effort hypothesis?

1.5.1 Random typing models

Soon after Zipf first published his work on the Principle of Least Effort, a couple of his contemporaries, Benoit B. Mandelbrot and George Miller, attempted to provide a proof that Zipf’s PLE hypothesis was unwarranted, and thus subject to Occam’s razor. Mandelbrot (1954) introduces a *random typing model* of language, where a toy language is created by choosing (with replacement) a sequence of symbols at random. (In other scenarios, the image of a monkey indiscriminately bashing at keys on a typewriter is invoked.) The possible symbols consist of the 26 letters and a space, which delimits individual word tokens. Mandelbrot proves that a toy language created in this way will obey Zipf’s *rank-frequency law*, i.e. that the rank and frequency of a word are related by an inverse power law.

$$frequency \propto rank^{-1} \tag{1.2}$$

Miller (1957) points out that, for a language created using this stochastic symbol-concatenation method, shorter strings have a higher probability of recurring than longer strings. Therefore such languages will, by default, also align with the Law of Abbreviation.

Miller argues that this finding renders the universality of the Law of Abbreviation “not really very amazing”, given that “monkeys typing at random manage to do it [produce an efficiently-coded lexicon] about as well as we do” (313). Thus, he concludes that this universally-found pattern is in need of no further explanation over and above the mathematical laws that govern stochastic systems. More recent work on stochastic modelling has also shown the default emergence of such patterns, including the inverse relationship between word length and probability in context (Ferrer-i-Cancho and Moscoso del Prado, 2012; Moscoso del Prado, 2013).

Miller’s final assessment was that “Zipf’s rule can be derived from simple assumptions that do not strain one’s credulity (unless the random placement of spaces seems incredible), without appeal to least effort” (314). However, I would argue that the random placement of spaces *does* seem incredible, when invoked as an explanation for the formation of a natural language. Lexicons are not produced all at once via a stochastic process which places word delimiters at random, but rather they evolve gradually, shaped by the many factors introduced by human agents, who are subject to cognitive biases and environmental pressures. Random typing models are therefore not realistic models of lexicon formation, and cannot be taken seriously as an alternative explanation to the Principle of Least Effort (see, e.g., Pustet, 2004 and Piantadosi et al., 2013 for similar arguments).

To be clear, the argument put forth by Moscoso del Prado (2013) is that rather than providing an alternative to the Principle of Least Effort, what these stochastic models provide is a necessary *baseline* against which to assess the hypothesis. However, if languages are not actually formed in any way resembling the manner assumed by the stochastic models, then it is not clear that these stochastic models are even suitable as a baseline against which to compare the Least Effort hypothesis. Instead, what arguments such as those made by Miller, Mandelbrot, Ferrer-i-Cancho, and Moscoso del Prado draw attention to, is the fact that the frequency-length correlations observed in synchronic corpora are, by themselves, not sufficient to warrant the PLE. To identify the most plausible explanation, we need also to consider how languages come about, and how they change over time through use. Thus, what is really needed is an investigation of the specific mechanisms that, over time, give rise to the patterns described in Zipf’s Law of Abbreviation. If the PLE hypothesis is correct, then we should observe evidence for optimisation processes acting on the lexicon, both in real time by individuals during communication, and over evolutionary timescales. Providing such evidence is the main goal of this thesis.

1.5.2 Behavioural evidence for optimisation during communication

According to the Principle of Least Effort, communicative pressures acting on language-users cause them to gradually push the lexicon towards an optimally efficient distri-

bution over length-meaning mappings. Specifically, the competing pressures to make less effortful productions and to maximise accurate transmission demand an optimal solution in which word length is assigned inversely proportional to probability. If this is indeed the case, then we should be able to observe language-users actively reshaping the lexicon in this direction, in situations where they are subject to these communicative pressures. Crucially, to determine whether it is the communicative pressures themselves that play a critical causal role, and not some other factor, we should also *not* observe a reshaping of the lexicon in this direction when the communicative pressures are removed.

Some prior studies provide promising evidence to suggest this may be the case. In communication game experiments conducted by Krauss and Weinheimer (1964) and Clark and Wilkes-Gibbs (1986), and even more recently replicated by Hawkins et al. (2017), participants communicated with a partner, taking turns playing ‘director’ and ‘matcher’. The director would use English to describe objects for the matcher to identify from a set. The objects being communicated about were abstract geometrical shapes lacking canonical English names. The director would typically begin by using a long, elaborate phrase to help the matcher identify the correct object. However, on repeat occurrences of the object, as its discourse-based givenness increased, directors would gradually shorten the description produced. For example, an object described as “upside-down martini glass in a wire stand” on its first occurrence ultimately became shortened to just “martini” after several repeat occurrences. The more times an object reoccurred, the shorter its average length by the end of the experiment.

To determine the extent to which this process is due to communicative pressures rather than mere repetition, Krauss and Weinheimer (1966) ran a study in which the timing and quality of matcher feedback was manipulated across conditions. In the ‘concurrent feedback’ condition, the director and matcher were allowed to communicate freely, and the matcher was able to cut the director short as soon as they identified which object was being described. In the control condition, there was no verbal communication between matcher and director. The experimenters found that directors were more likely to shorten their descriptions of objects in the concurrent feedback condition than in the control condition. Furthermore, when directors received feedback that the matcher correctly guessed the objects 100% of the time, they were more likely to shorten the descriptions than when they received feedback that the matcher only guessed correctly 50% of the time.

In addition to receiving real-time, positive feedback, being engaged directly with a partner—as opposed to having someone comprehend from a spatially or temporally distant position—also appears to be an important factor in driving this behaviour (Wilkes-Gibbs and Clark, 1992; Hupet and Chantraine, 1992). Consistent with this, Fowler (1988) found that speakers reduce the phonetic duration of repeated words

more when spontaneously communicating with a listener than when reading aloud into a microphone. These studies suggest that the shortening behaviour observed in the communication game experiments is likely due to more than mere repetition. Rather, it is something about the communicative context which triggers the drive to reduce utterances.

But why are some utterances reduced and not others? In the communication game experiments presented in Krauss and Weinheimer (1964) and Clark and Wilkes-Gibbs (1986), participants were communicating about a small set of meanings using a large space of possible forms—multi-word descriptive phrases in English. All phrases could thus be shortened in this task without risking ambiguity. However, at the level of a lexicon, shorter words are subject to greater confusability for a number of reasons: they have less space for signal redundancy and are thus more likely to be lost in noisy signal transmission; and because there are a limited supply of phonotactically-regular short words available (especially as not all available forms get used—see Dautriche et al., 2017 and Mahowald et al., 2015 for possible explanations for this), word shortening can also result in outright ambiguity of form-meaning mappings. Indeed, shorter words are more likely to be polysemous and homophonous (Piantadosi et al., 2012).

The experiments presented in Chapters 2 and 3 of this thesis adapt the basic communication game paradigm of Krauss and Weinheimer (1964) to account for this lexicon-level effect. In our experiments, communication is restricted to an artificial language with a small lexicon, in which the pressure to achieve successful communication by avoidance of ambiguity is pitted against the pressure to reduce effort in one’s productions. We additionally introduce a 2×2 between-subjects manipulation, such that one or both of these pressures is removed in each of three different control conditions. This allows us to test whether these competing pressures have a direct causal role in driving optimisation behaviour. Chapter 2 investigates whether language-users modulate word length according to frequency only when both pressures are present, and Chapter 3 investigates whether they also modulate word length according to probability in context, under the same pressures. Both sets of studies find that language-users restructure the lexicon to optimise for efficient communication, but only when they are subject to both competing pressures. This establishes a clear causal link between the communicative pressures in question and the optimisation behaviour that leads to lexicons which assign word length inversely proportional to probability.

1.5.3 Diachronic evidence for optimisation processes

As mentioned previously, Zipf’s Law of Abbreviation includes a diachronic corollary: that as the frequency of a word increases with time, its length should correspondingly decrease. The results of the optimisation processes occurring during individual com-

munication acts should accrue over time to reveal a long-term pattern, visible at the timescale of language evolution. This accumulation over time of individual acts of optimisation, Zipf hypothesises, is what is ultimately responsible for the universally observed inverse relationship between word length and probability. In Chapter 4 of this thesis, I review the few pre-existing studies that provide indirect evidence of such an evolutionary process taking place. I then present a diachronic corpus study aimed at directly addressing this question. The results suggest that words across the lexicon which have shorter variants do show an increased preference towards these shorter variants over time. Moreover, long/short word pairs whose overall frequencies are increasing over time, on average, move towards the shorter variant at a faster rate, as compared with word pairs whose frequencies are constant or decreasing on average.

1.6 Artificial language learning as a tool for investigating the evolution of linguistic structure

The use of a miniature artificial language in Chapters 2 and 3 continues a now long tradition of employing this methodology to investigate the evolution of abstract linguistic structures. By taking participants out of the context of their known languages and observing how they learn, process, and use an unfamiliar language, it is possible to focus more sharply on the underlying mechanisms, both cognitive and environmental, that may shape general structural features of language. The ability to construct artificial languages with the specific features one is interested in also allows one to more easily test specific causal hypotheses about which types of conditions lead to the emergence of which types of linguistic structure.

As early as 1925, Esper used this technique to illuminate the mechanisms by which compositional morphological structure arises in language (see also Esper, 1966 and Esper, 1973). Esper constructed a miniature language with a structured semantic space (eight objects: four red and four green, and two each of four different unfamiliar shapes) and a morphologically unstructured lexicon—each object was mapped to a phonologically distinct “suppletive” form which could not be broken down into smaller parts contained in any other labels. This language was taught to the first participant (or learner), who then became the teacher, teaching the language to the next learner, and so on up to 44 iterations. Errors in the languages produced by the teachers were not corrected. What Esper found was that:

...When a miniature linguistic system characterized by semantic but not by morphologic categories is transmitted from person to person in a long series of individuals, morphologic categories tend to develop in correspondence with the semantic categories. (Esper, 1966; p. 579)

In other words, the language learned by the final learner was one in which the various forms could be broken down into phonological chunks which correlated with the features of the object. For example, objects of ‘shape 2’ began with the phoneme /v/ and ended with /ʒ/; objects of ‘shape 3’ began with /p/; most green objects contained the phoneme /e/ while red ones did not. Esper was especially interested in *how* this structure evolved. He found that specific, sometimes predictable, phonetic errors occasionally led to accidental phonetic similarities between semantically related forms. These then led to analogical changes, through which morphological categories emerged. This groundbreaking study provided an insight into how compositional morphological structure might arise through *self-organisation*, without the influence of any domain-specific, pre-encoded linguistic rules. Rather, simply through noisy transmission from person to person over generations, combined with the domain-general cognitive bias for analogical reasoning, morphology-like structure emerges spontaneously.

This experimental set-up, involving “iterated learning” of an artificial language across multiple generations of participants, has seen a fruitful revival in recent years, initiated by Kirby et al. (2008). In this study, the iterated chain was initialised with a language similar to Esper’s: the semantic space was structured, containing objects categorisable by shape, colour, and movement, while the forms in the lexicon were non-compositional. Noisy transmission and linguistic innovation were encouraged by the addition of a transmission bottleneck—participants only learned the names for some of the objects in the semantic space, though they had to produce the names for all of them in the testing portion of the experiment. In just 10 generations, compositional morphological structure emerged, with different phonemic chunks corresponding to different shapes, colours, and movements. Additionally, the language became more faithfully replicated in successive generations. This result suggests that compositional structure emerges in language *because* it makes language more learnable—this feature is selected for by the nature of cultural transmission itself. The authors conclude, “just as biological evolution can deliver the appearance of design without the existence of a designer, so too can cultural evolution” (Kirby et al., 2008; p. 10685).

Others have employed artificial language learning and production tasks in individual participants, rather than across transmission chains, to illuminate the individual learning biases that shape acquisition of linguistic patterns. For example, Hudson Kam and Newport (2005) exposed participants to an artificial language containing some variability (specifically, in whether nouns were preceded by determiners). They found that while adults reproduced this variability in their output, children tended to eliminate the variability and *regularise*, either systematically producing or omitting determiners. This suggests a simple, domain-general mechanism by which creoles are eventually formed from previously inconsistent and unstable linguistic systems.

Using a similar methodology, Culbertson et al. (2012) exposed participants to artificial languages with variable noun phrase word order. They found that participants learning a language with mostly harmonic word orders (both adjective and number either post-nominal or pre-nominal) were more likely to regularise in the direction of using this order systematically, while participants learning a language with mostly non-harmonic word order shifted their productions towards harmonic orders. This learning bias for harmonic word orders may explain why the vast majority of the world’s languages today exhibit harmonic word orders in the noun phrase.

In addition to iterated learning and individual learning tasks, artificial languages have now also been used in communicative tasks involving pairs of participants. For example, in Kirby et al. (2015), participants played a simple signalling game using an artificial language, similar to that of Krauss and Weinheimer (1964), in which players took turns acting as director and matcher. In one condition, the productions from one member of the pair were transmitted to a new pair, forming an iterated chain, while in another, members of the same pair communicated in successive rounds. Only in the iterated condition, when pressures for both expressivity *and* learnability were both present, did compositional structure reliably arise. This suggests that compositional structure emerges in language as a solution for satisfying both of these competing pressures. In other studies, researchers have used artificial language communication games to show that communication can lead to the elimination of unpredictable variation (Fehér et al., 2016) and that languages evolve to encode features relevant to the communicative contexts in which they are used (Winters et al., 2015).

In all the studies reviewed in this section, artificial languages are used as a kind of petri dish for testing how different cognitive and environmental conditions can lead to the evolution of different abstract structural features of language, such as compositionality, systematicity, and harmony. In Chapters 2 and 3 of this thesis, I adopt this methodology to shed light on the emergence of communicatively efficient lexicons.

1.7 Language as an efficient code for information transfer

A central question to the field of linguistics is “why do so many of the world’s languages display the same broad structural features?” Explanations for universal linguistic structures have tended to take one of two general approaches: attributing them to innate, domain-specific cognitive biases (e.g., Chomsky, 1965, 2011; Hauser et al., 2002, 2014); or placing the primary focus on domain-general and/or functional explanations for the observed behaviour (e.g., Christiansen and Chater, 2008; Evans and Levinson, 2009). Under this latter approach, the function of language as a system for communication is considered key in shaping some of its structural properties (see,

e.g., Miller, 1951). In contrast, proponents of the former approach argue that the structure of language is due primarily to the structure of abstract thought—the fact that language is also used to *communicate* thought is not important in determining its structure. For example, Chomsky (2011) states that “the core of language appears to be a system of thought, with externalization a secondary process (including communication, a special case of externalization)” (p. 263). Similar views are espoused by Pinker (1995).

The very basis of the Principle of Least Effort hypothesis lies in considering language as a behaviour whose primary function is communication. This view of language was emphasised by Zipf himself, but is also made formally explicit in our information-theoretic reformulation of the PLE. Only by understanding language as a system for encoding information, to be transmitted from source to receiver, can we clearly express the competing pressures for maximising source signal compression and minimising transmission error. By providing evidence that an optimisation process between these two pressures is indeed taking place, we implicitly provide evidence that the communicative function of language is key in shaping some aspects of its structure. Moreover, the shaping of this structure occurs *through* language use (i.e. during communication with a partner), and relies crucially on factors that are determined purely *by* language use (i.e. frequency and probability in context). Therefore, the studies discussed in Section 1.5.2, as well as the new experimental evidence provided in this thesis, support an even stronger claim: that language cannot be separated from its function as a communication system, as without this, we would not be able to explain core features of its structure. A further aim of this thesis is to show that by applying information-theoretic concepts to the study of language—and thereby conceiving of language as, among other things, an evolving, efficient code for information transfer—we can formulate plausible explanatory hypotheses for patterns of linguistic behaviour, that are upheld by the empirical evidence.

Zipf concluded his 1935 book with the following qualifier: “It remains to be seen whether further empirical studies will completely substantiate our findings, and, if so, how far they will limit, modify, extend, or reinterpret what appears already to have been found” (p. 310). In this introductory chapter, I have reviewed a substantial body of work that has replicated, strengthened, and extended Zipf’s original findings regarding the inverse relationship between word length and frequency. Since his first publication on the topic, the field of information theory was subsequently developed, providing a new framework in which to express his hypothesis about how this inverse relationship arises—via the Principle of Least Effort. By reformulating this PLE hypothesis in terms of an optimisation process between the competing pressures to maximise signal compression and minimise transmission error, I was able to design a series of experiments which explicitly test the causal link between this hypothesis

and the structure of word length in the lexicon. The results of these experiments—reported in Chapters 2 and 3 of this thesis—together with those of the diachronic corpus study presented in Chapter 4, vindicate Zipf’s original hypothesis. At the same time, they reveal complexities in the underlying mechanistic processes which have only just begun to be explored.

Chapter 2

Word length and frequency

The following chapter is organised as follows: the first part comprises the text of a paper published in the journal *Cognition*, reporting the main experiment which was run online. The paper was co-authored with my supervisors, Kenny Smith, Jennifer Culbertson, and Simon Kirby. The text was reformatted within the main body of the thesis to allow for integrated indexing of figures, tables, and references. The published version, using the format of the journal, is included at the end of this thesis in Appendix A. The second part of this chapter details an earlier version of the main experiment, which was run in the lab, and compares these results to those found in the online experiment.

2.1 Online experiment (Experiment 1)

2.1.1 Introduction

In 1935, the linguist George Kingsley Zipf pointed out what he claimed to be a universal property of human language: that “the magnitude of words tends...to stand in an inverse...relationship to the number of occurrences” (Zipf, 1935; pp. 1). In other words, the more frequent a word is, the shorter it tends to be. This “Law of Abbreviation” has now been verified in a wide range of human languages, including: Chinese, Croatian, Czech, Dutch, English, French, German, Greek, Hungarian, Indonesian, Italian, Polish, Portuguese, Romanian, Russian, Slovenian, Slovak, Spanish, Sundanese, and Swedish (Teahan et al., 2000; Sigurd et al., 2004; Strauss et al., 2007; Piantadosi et al., 2011; Ferrer-i-Cancho and Hernández-Fernández, 2013). For example, one can clearly see this relationship for English words in Figure 2.1. Interestingly, there is even evidence for its broader application in animal communication systems (in the vocalisations of common marmosets and formosan macaques, and in the surface behavioural patterns of dolphins; Ferrer-i-Cancho et al., 2013) and in computer programming (e.g., use of the *alias* function in Unix to abbreviate frequent

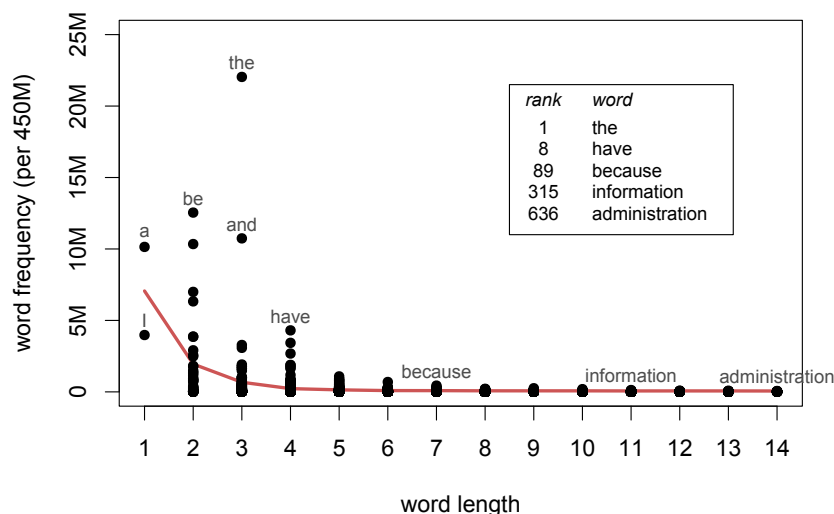


Figure 2.1: The 1000 most frequent words in English. Each point represents an individual word (some points are labeled). The red line marks the mean frequency for the words of each length (here, orthographic length is used, but the same overall pattern would be seen if phonetic length were used instead.) The more frequent a word is, the shorter it tends to be. According to Zipf’s Law of Abbreviation, this is a universal pattern of human languages. Frequency counts used here are from the 450 million word COCA corpus (Davies, 2008).

commands; Ellis and Hitchcock, 1986).

Zipf hypothesised that this universal pattern arises as a result of a tradeoff between two competing pressures: a pressure for accurate (successful) communication and a pressure for efficiency or less effort.¹ The idea is that together, these pressures would shape how forms are mapped to meanings, because languages have a finite inventory of discrete sounds that can be recombined to form words. This results in a lexicon with a limited number of words of a given length. Importantly, the shorter the length, the fewer distinct possible words there will be of that length, and the greater the potential confusability—shorter forms have less space for signal redundancy and thus are more likely to be confused in noisy signal transmission. Therefore, while a pressure for efficiency should favour these short words since they require less effort to produce (all things being equal), this is in direct conflict with the pressure for accurate communication. The latter should instead favour unique form-meaning mappings which minimise potential ambiguity—from this perspective, longer words have the clear advantage. How, then, can a language use the available short forms optimally, while still keeping ambiguity in check? The solution is to assign the shortest words

¹The assumption that information is packaged into repeating words of variable length, and not fixed-length blocks—as in, e.g., block codes such as Hamming codes (Hamming, 1950)—is also necessary to make this prediction. Thanks to an anonymous reviewer for pointing this out.

to the most frequent meanings, leaving longer words for less frequent meanings, as in variable-length, e.g. Huffman, coding (Huffman, 1952). Zipf called this hypothesised tendency to produce short utterances wherever possible the “Principle of Least Effort”.

The Principle of Least Effort offers a functional explanation for the Law of Abbreviation, if we imagine it playing out through incremental changes over time. If language users track frequency differences between meanings (consciously or otherwise), then processes of change may differentially affect words whose frequencies differ. For example, if a word is more frequently used, then it may be more likely to be targeted for reduction or shortening (e.g., ‘information’ becomes ‘info’). Form-meaning mappings would then gradually shift toward more optimal alignment of frequency with length (Zipf, 1935).

While this is an attractive explanatory account, several researchers have raised the possibility that the inverse relationship between word length and word frequency could emerge instead from simple constraints on randomly generated systems. For example, a lexicon generated through a random typing process, in which ‘words’ are produced by pressing keys (including the space bar) at random, has properties that are consistent with the Law of Abbreviation (Moscoso del Prado, 2013; Ferrer-i-Cancho and Moscoso del Prado, 2012). While we know that languages are not actually generated at random in this way, it nevertheless remains a possibility that the Law of Abbreviation could result from some yet-unidentified statistical process, unrelated to optimisation behaviour on the part of language users.

Several studies provide indirect evidence connecting competing pressures for accurate and efficient communication to properties of linguistic systems introduced by language users. For example, previous research has shown that learners restructure case marking systems such that case markers are preferentially used when grammatical roles are ambiguous and omitted when other disambiguating information is present (Fedzechkina et al., 2012). This is consistent with the idea that effort (here, producing case markers) is reduced in a way which preserves communicative function. Language learners have also been shown to capitalise on differences in the length of novel labels to make pragmatic inferences about the communicative intentions of speakers (Degen et al., 2013). A computational model of iterated learning (Kirby, 2001) shows that short, non-compositional morphological forms are more likely to evolve for frequent meanings, while longer, compositional ‘regular’ forms are more likely to persist for infrequent meanings, due to a tradeoff between the pressure for learnability and the pressure for producing shorter, more replicable forms.

A direct link between frequency and utterance-length shortening in actual language users has been shown in studies such as Krauss and Weinheimer (1964) and Clark and Wilkes-Gibbs (1986). In these studies, participants played a dyadic communication game, where ‘directors’ used English to describe objects for their partners

(‘matchers’) to identify from a set. The objects being communicated about were abstract geometrical shapes lacking canonical English names. The director would typically begin by using a long, elaborate phrase to help the matcher identify the correct object. However, on repeat occurrences of the object, the director would take advantage of a growing base of shared knowledge, established through communication, to gradually shorten the descriptive phrase and thereby reduce the effort expended. For example, an object described as “upside-down martini glass in a wire stand” on its first occurrence ultimately became shortened to just “martini” after several repeat occurrences. The more times an object reoccurred, the shorter its average length by the end of the experiment. These results depended on the director receiving positive, real-time feedback from the matcher during the signalling game (Krauss and Weinheimer, 1966; Hupet and Chantraine, 1992), suggesting that it is a communicative context which triggers the drive to reduce effort. Thus, this result suggests one mechanism by which the Law of Abbreviation could arise: if the form associated with a meaning becomes shorter the more times it occurs in conversation, and these mappings are retained and spread across speakers, then in the lexicon overall, more frequent meanings will end up with shorter forms than less frequent meanings.

However, as we mentioned above, there is competition for the short forms in a lexicon. For example ‘info’ refers to ‘information’, and not ‘informality’, ‘infoliation’, or ‘infoedation’. Why is this? In the Krauss and Weinheimer (1964) and Clark and Wilkes-Gibbs (1986) studies, participants were communicating about a small set of meanings using a large space of possible utterances. All labels could thus be shortened in this task without resulting in ambiguity. However, when several meanings are in direct competition for a single short label—a problem that arises at the level of an entire lexicon—the mechanism shown in these studies is not sufficient to account for why one meaning gets mapped to the short form and not the others.²

Thus, while these previous studies are consistent with the idea that something like the Principle of Least Effort operates during language use, they do not explicitly target the hypothesised role of competing communicative pressures—the pressure for reduced effort versus the pressure against ambiguous form-meaning mappings—in modulating word length within the lexicon. In our study, we make use of a miniature artificial language learning paradigm to create a setting in which these two pressures are directly in conflict: a reduction in effort cannot be achieved without also increasing

²Interestingly, not all possible short forms in a language actually get used. This could be a consequence of noisy communication—using short forms sparingly would further minimise potential confusability. However, it has been found that frequent (and by proximity short) forms tend to be tightly clustered together in the phonological space, in seeming opposition to this end (Dautriche et al., 2017). This may be due to the influence of constraints on learning, memory, and production, which favour lexicons with high phonetic regularity. Thus, even though not all possible short forms are used, there will be particularly tough competition for those forms that fall within the more densely-populated regions of the phonological space. Thanks to an anonymous reviewer for raising this topic.

the ambiguity of form-meaning mappings. Crucially, our set-up allows us to isolate these different pressures in order to determine their individual contribution to the overall behaviour of a miniature artificial lexicon. Following Zipf, we hypothesise that only when these pressures are both present—and thus in direct conflict—will language users restructure their input to align shorter forms with more frequent meanings. In this way, our study aims to provide a concrete link between optimisation behaviour at the level of the individual and the global pattern Zipf first observed.

2.1.2 Miniature Artificial Language Learning Experiments

We use a miniature artificial language learning paradigm, which has previously been used to shed light on the cognitive mechanisms and environmental pressures that shape language structure (e.g., Kirby et al., 2008; Fedzechkina et al., 2012; Culbertson et al., 2012). In this paradigm, participants learn a miniature artificial language, and then we observe how they reshape their input as they use the language, in this case to communicate with a partner (see also Winters et al., 2015; Kirby et al., 2015; Fehér et al., 2016).

2.1.2.1 Participants

124 participants (51 females, 64 males; a further 9 chose not to report their gender) were recruited through Amazon Mechanical Turk. 106 of these reported themselves as native English speakers, of which 88 were monolingual. A broad range of other languages were represented across the remaining participants. Ages ranged from 18 to 73 (mean=33).

2.1.2.2 Materials

Participants were trained on two names for each of two plant-like alien objects, by repeatedly being shown pictures of the objects labeled with their names on a computer screen (see also Reali and Griffiths, 2009; Vouloumanos, 2008). Crucially, one of the two objects appeared three times more frequently than the other—specifically, one object appeared 24 times and the other 8, for a total of 32 training trials.

Each object appeared half the time labeled with its long name, a 7-letter word, and half the time with its short name, a 3-letter word derived by clipping the last two syllables off the long name. The process of clipping, or word-truncation, is a common word-shortening device in many languages (e.g. *info* for *information* in both English and French; Antoine, 2000). In natural languages, shorter words are subject to greater confusability for a number of reasons. They have less space for signal redundancy and are therefore more likely to be misinterpreted or lost in noisy transmission. There are also more unique possible 7-letter strings than 3-letter strings, and thus word

shortening can often result in outright ambiguity. Indeed, shorter words are more likely to be polysemous and homophonous (Piantadosi et al., 2012). To model these phenomena in our miniature lexicon, we designed the names such that the short name for both objects was identical (*zop*), while the long names were unique (*zopekil* and *zopudon*). A schematic diagram of the object frequencies and labels is provided in Figure 2.2a.

Which object (the blue fruit or the red stalk) was more frequent, as well as which object was paired with each label, were both counterbalanced between participants, giving a total of 4 possible object-frequency-label pairings which a participant might be trained on. This ensured that potential factors such as sound symbolism, or higher saliency of one of the objects, could not systematically bias our results.

2.1.2.3 Procedure

Participants were assigned to one of four conditions, where we manipulated the presence of pressures to communicate accurately and quickly in a between-subjects 2×2 design. In all conditions, the experiment consisted of two phases: training and testing. The training phase was identical for all four conditions, but the testing phase differed across conditions.

Training phase On each training trial, an object was presented on screen alone for 700ms. The appropriate label then appeared beneath the object for a further 2000ms, yielding a total trial duration of 2700ms. A blank screen showed for 500ms between each trial. The 32 training trials were presented in a different randomised order for each participant.

Testing phase After the training phase, testing procedures varied depending on the experimental condition. In the Combined condition, participants were under a pressure to communicate accurately *and* to communicate efficiently, as according to Zipf’s hypothesis, both of these competing pressures must be present for the Law of Abbreviation to emerge. The remaining three conditions removed one or both of these accuracy and time pressures. In all conditions, the testing trials contained the same frequency ratio over objects as the training trials: the frequent object appeared three times more frequently than the infrequent object.

Condition 1: Combined In the testing phase of this condition (henceforth referred to as the Combined condition), participants were paired with a partner to play a communication game. This was done by putting participants in a virtual queue, managed by a central server script, after completing the training trials. Participants were paired sequentially as they finished training; once a participant entered the

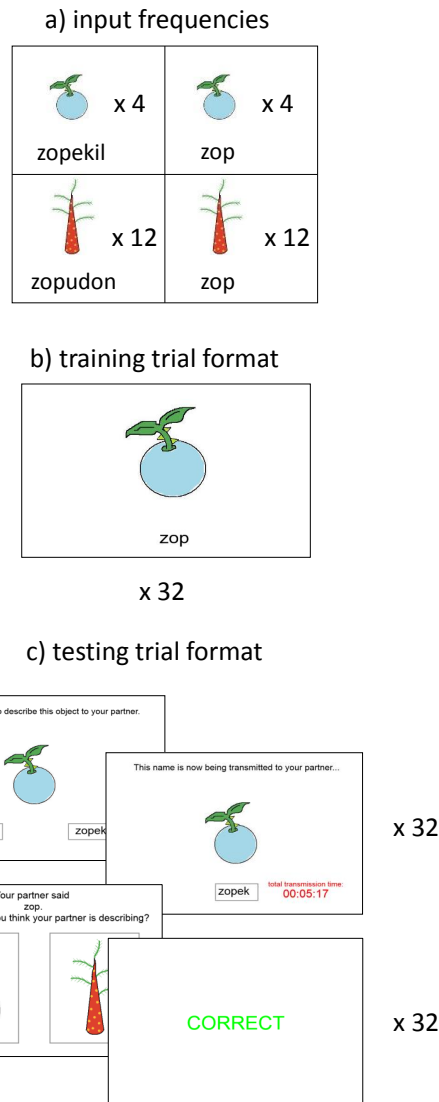


Figure 2.2: a) A schematic diagram of the frequencies of the objects and labels presented during the training trials in all four experimental conditions. One object appeared three times more frequently than the other. Each object was labeled half the time with its unique long name, and half the time with its ambiguous short name, which was a clipped version of the long name. b) An example training trial. c) An example of a director trial in the Combined condition (top) and a matcher trial followed by feedback (bottom).

queue, the server would pair them with the next participant to finish training after them. To encourage participants to wait as long as possible in the queue without leaving the game, they were shown a humorous cat video while they waited. However, if participants had still not been paired with a partner after 5 minutes, they were removed from the queue and paid for their time. This method allowed us to successfully run a dyadic artificial language communication experiment online using a crowdsourcing platform. We were therefore able to relatively quickly and easily collect data from a more culturally and linguistically diverse group of participants than is usually possible with traditional lab-based experiments that draw mainly from a university's undergraduate population.

Once paired with a partner, participants began the communication game. On each trial, the 'director' was shown an object on the screen and told to transmit its name to the 'matcher'. The director always had two options for which name to send: the long name for the object or the (ambiguous) short name. The director chose a name by clicking on it, and was then given instructions for how to actually transmit the name to the matcher. This was done by pressing and holding the mouse in a central transmission box in which each letter in the name appeared one by one, at 1200 ms intervals. Note that participants never had to type the names or necessarily remember their correct spelling; once they chose a name from the two options on the screen, the letters would appear sequentially in the transmission box as they held down the mouse. Only once all the letters had appeared in the box was the name transmitted to the matcher. If the mouse was released before all letters had been transmitted, the participant would have to start again from the first letter (but the total transmission time was only counted for the successful transmission). This belaboured method of transmission, in which the long name was significantly slower to transmit than the short name, introduced an element of effort into communication, modelling the difference in effort in spoken communication associated with producing long versus short utterances.

Once the matcher received the name from the director, the matcher was asked to choose which of the two objects they thought the director was referring to. Both players were then given feedback as to whether the matcher chose the correct object.

The players alternated roles after every trial, with the matcher becoming the director and the director becoming the matcher, until both completed 32 director trials and 32 matcher trials. The frequency with which each object appeared in each player's director trials matched those of the training frequencies: 24 occurrences of the frequent object, and 8 of the infrequent object. The order of these 32 director trials was randomly shuffled for each participant. The member of the pair who entered the queue first was the first player to direct.

To model the pressures in spoken communication to be both efficient and accurate,

pairs were told at the beginning of the game that they would be rewarded a bonus payment if they were the pair to complete the game in the quickest time with the highest number of correct match trials. Time was only counted during name transmission, and the time count was displayed next to the transmission box as the participant was transmitting a name, to underline the time pressure. Example screenshots of a director trial and matcher trial are shown in Figure 2.2c.

In order to tease apart the influence of the two pressures on the participants' patterns of behaviour, we included three further experimental conditions, described below, for a full 2×2 manipulation of the pressures for accuracy and efficiency.

Condition 2: Accuracy In this condition, participants were paired to play a communication game as described above, but in the director trials, there was no intermediate step between the director choosing a name to send and the matcher receiving the name; the names were sent instantaneously, thus removing any difference in effort between transmitting long or short names. Pairs were told that the goal of the game was to have their partner make as many correct guesses as possible. There was no bonus reward given for the most accurate pair, as the task was extremely easy and we predicted that most pairs would achieve maximum accuracy, which turned out to be the case.

Condition 3: Time In this condition, communication was taken out of the game entirely; participants played a one-player game consisting of 64 director trials only. In each director trial, participants were told to choose a name to describe the object shown on the screen, but there was no subsequent communicative task. As in the previous conditions, the choice was always between the long name and the short name. Once chosen, the name had to be entered as in the Combined condition, by pressing and holding the mouse in a transmission box, with each letter appearing at 1200 ms intervals. The next trial began only when all the letters had appeared in the box. Thus, the long name was significantly slower to produce than the short name. The transmission process was also timed with an on-screen timer as in the Combined condition, and participants were told at the beginning of the game that they would be rewarded a bonus payment if they were the player with the shortest overall transmission time.

Condition 4: Neither The fourth and last condition contained neither a pressure for efficiency nor a pressure for accuracy. As in the Time condition, participants played a one-player game with no explicit communicative element, but additionally there was no time difference associated with transmission; once a label was chosen to describe an object, long or short, it was instantaneously recorded and the player was advanced to the next trial. We included this condition in order to provide a baseline

for participants' behaviour from which to assess the effects of the accuracy and time pressures in the other three conditions.

Payment Participants were paid depending on the condition they were in, commensurate with the average time it took to complete that condition. Participants in the Combined condition, the longest to complete due to both the slow transmission process and having to wait for the partner's response after each trial, were paid \$2; participants in the Accuracy and Time conditions were paid \$1, and participants in the Neither condition, the shortest to complete, were paid \$0.50.

2.1.2.4 Predictions

Our predictions for the Neither condition were that participants would either probability-match—i.e. use the long and short forms for both objects with equal frequency, as in the training trials (see Hudson Kam and Newport, 2005)—or their behaviour would reveal prior biases language users bring to the task, such as a preference against using ambiguous forms.

In the Accuracy condition, we predicted that participants would be more likely to use the long names for both objects compared to the baseline condition, given the potential loss of accuracy from using the ambiguous short name, and with no time considerations to favour the use of short but ambiguous labels. Given the task demands, this would therefore be the best strategy to use in this condition.

In contrast, in the Time condition, we predicted that participants would use the short name for both objects: with no communicative purpose attached to the transmissions, and an incentive to be as quick as possible, using the short name in every trial is the best strategy in this condition.

In the critical Combined condition, with both a time and an accuracy pressure, we predicted that participants would converge on the optimal strategy, in which the frequent object is consistently mapped to the ambiguous short name, and the infrequent object to its unique long name, in line with Zipf's Law of Abbreviation. Using this strategy, transmission time is minimised as much as possible while still maintaining one-to-one form-meaning mappings, thereby also ensuring accurate communication.

2.1.3 Results

Figure 2.3 shows the proportion of trials on which the short (ambiguous) label was selected by the director, for high- and low-frequency objects. As predicted, in the Accuracy condition, most participants retained the unique long names for both objects, while in the Time condition, most participants mapped both objects to the ambiguous short name. Crucially, in the Combined condition, where participants were under pressure to communicate both accurately and efficiently, most pairs converged on

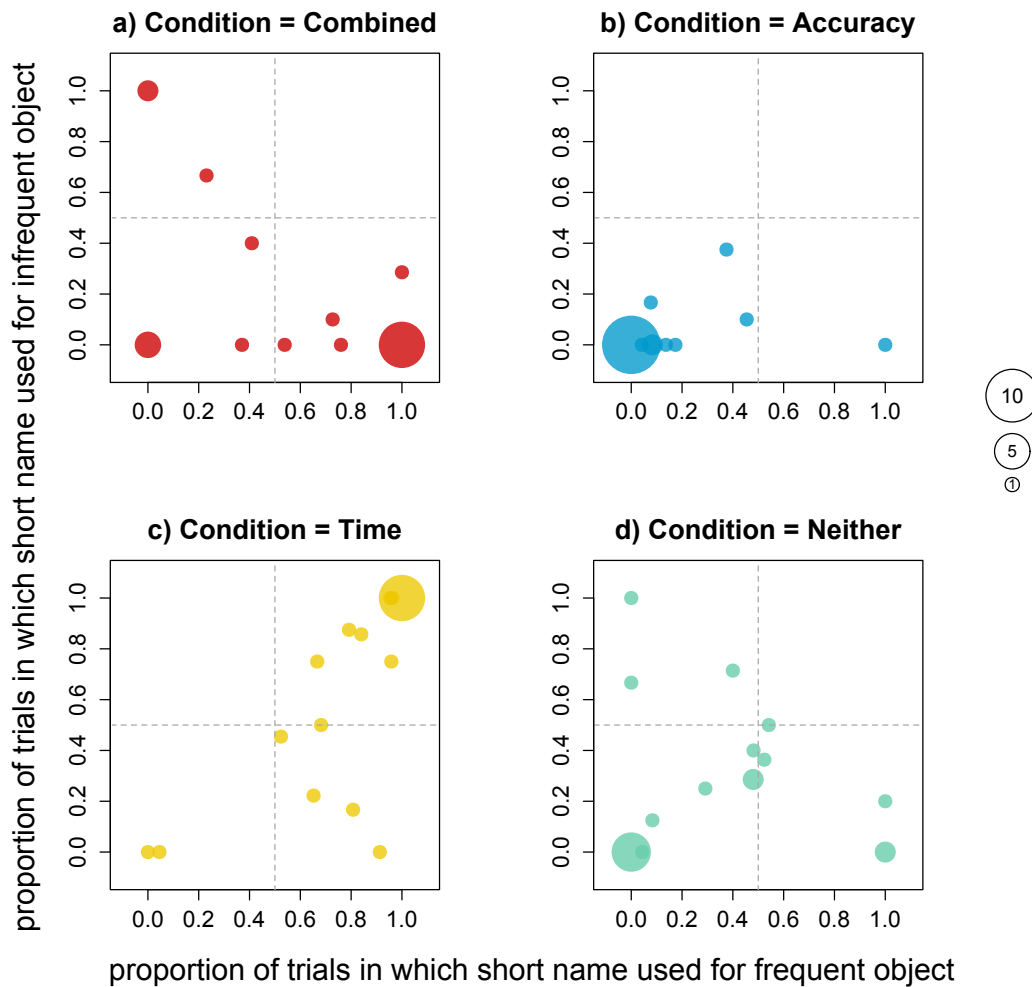


Figure 2.3: The proportion of trials in which the short name was used to label the frequent object versus the proportion of trials in which it was used to label the infrequent object. For the Combined (a) and Accuracy (b) condition, each data point combines a pair of communicating players, representing the sum of their director trial productions. For the Time (c) and Neither (d) condition, each data point corresponds to an individual player’s productions. The size of the circles is perceptually scaled (Tanimura et al., 2006) to reflect the number of data points coinciding at each value. Data from only the second half of testing trials is shown here, as participants were more likely to have converged on a stable mapping by this time. Data points in the top right quadrant of each graph indicate participants who are mostly using the short name for both objects; participants are clustered in this quadrant only in the Time condition (graph c). Data points in the bottom left quadrant of each graph indicate those who are mostly using the unique long names for both objects; participants are most clustered here in the Accuracy condition (graph b). Data points in the bottom right quadrant of each graph indicate participants who are mostly using the short name for the frequent object and the long name for the infrequent object. This behaviour, consistent with the Law of Abbreviation, only reliably arises in the Combined condition (graph a), where both pressures are present.

Table 2.1: Summary of fixed effects for a binomial regression model with short name use as the binary dependent variable, and by-participant random effects for object frequency. Like Figure 2.3, this model is fit using only the second half of each participant’s training trial data, as participants were more likely to have converged on a stable linguistic mapping by then.

	β	SE	p
intercept	-2.225	0.501	<0.001
object=frequent	1.392	0.484	0.004
condition=Accuracy	-5.149	0.781	<0.001
condition=Time	6.031	1.207	<0.001
condition=Combined	0.343	0.746	0.645
object=frequent & condition=Accuracy	-0.722	0.751	0.337
object=frequent & condition=Time	-1.079	1.180	0.360
object=frequent & condition=Combined	2.573	0.709	<0.001

the optimal strategy wherein the most frequent object was mapped to the ambiguous short name, and the infrequent object to its unique long name. This made the participants’ lexicon both efficient and expressive, in line with Zipf’s Law of Abbreviation. Finally, the Neither condition revealed an underlying bias towards avoiding ambiguity.³

A logistic regression model was fit in R (R Core Team, 2015) using the lme4 package (Bates et al., 2015), with short name use (as contrasted with long name use) as the binary dependent variable, object frequency, experimental condition, and their interaction as fixed effects, and by-participant random intercepts and random slopes for object frequency. The model was sum coded, setting the grand mean as the intercept, to which each level was then compared. This model yielded a significant positive interaction for the frequent object in the critical Combined condition. Thus, in this condition, participants were significantly more likely to assign the short name to the frequent object than in any other. Participants were significantly *less* likely to assign the short name to either object in the Accuracy condition, and significantly more likely to assign it to both objects in the Time condition, as reflected by the large negative coefficient for the former condition, and the large positive coefficient for the latter. Finally, the intercept is significantly negative, indicating that there is a baseline preference for avoiding the short form (see Table 2.1 for a full list of model coefficients).

In Figure 2.4 we plot participants’ ‘languages’ (the collection of form-meaning mappings produced in their director trials) according to their average token length and the mutual information between their forms f and meanings m : $\sum_f \sum_m p(f, m) \log \frac{p(f, m)}{p(f)p(m)}$.⁴

³The complete set of raw data from this experiment can be accessed using the following link: <http://datashare.is.ed.ac.uk/handle/10283/2702>.

⁴We computed the mutual information directly from the empirical distributions, rather than

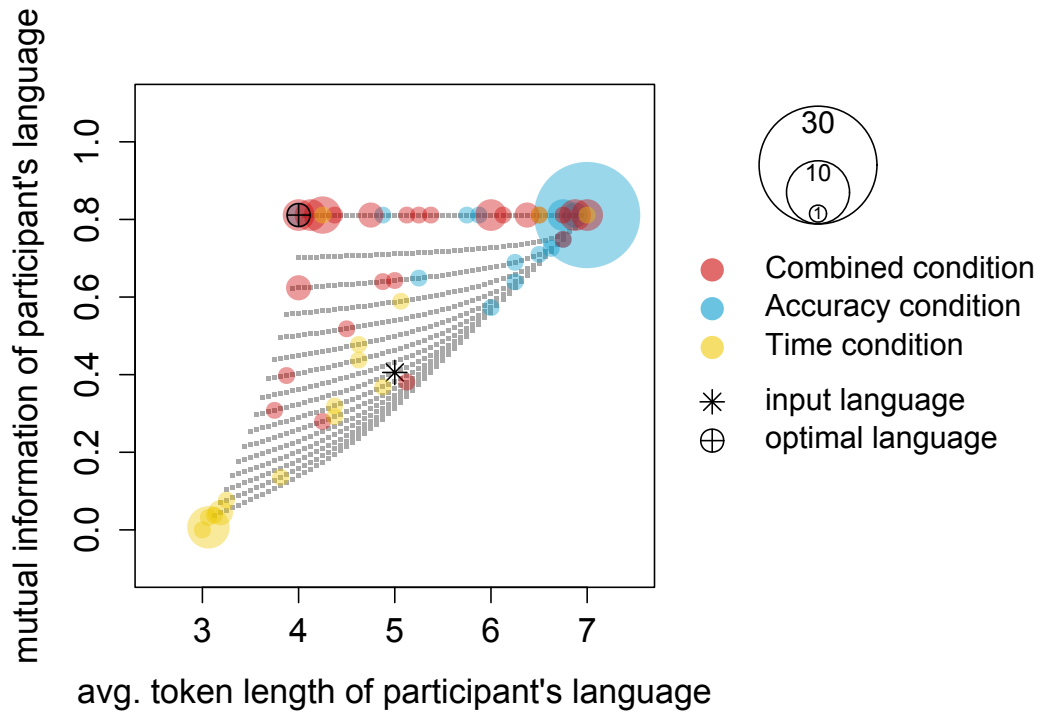


Figure 2.4: The average token length of an individual participant’s ‘language’ (the full set of all their director trial productions) plotted against the expressivity (the mutual information between the forms and meanings) of their language. The size of the circles is perceptually scaled (Tanimura et al., 2006) to reflect the number of data points coinciding at each value. The input language that participants are exposed to in training trials is marked with an asterisk, and the grey points represent possible output languages. (Possible output languages are constrained by the number of different expressivity values that are possible for a language with a given average token length. For example, there is only one possible configuration for both the shortest and longest average token lengths—all objects are either mapped to the short name or the long name, respectively—and thus only one possible expressivity value at the endpoints.) The optimal language—the language with the minimum avg. token length while achieving maximum expressivity—is marked with a target symbol

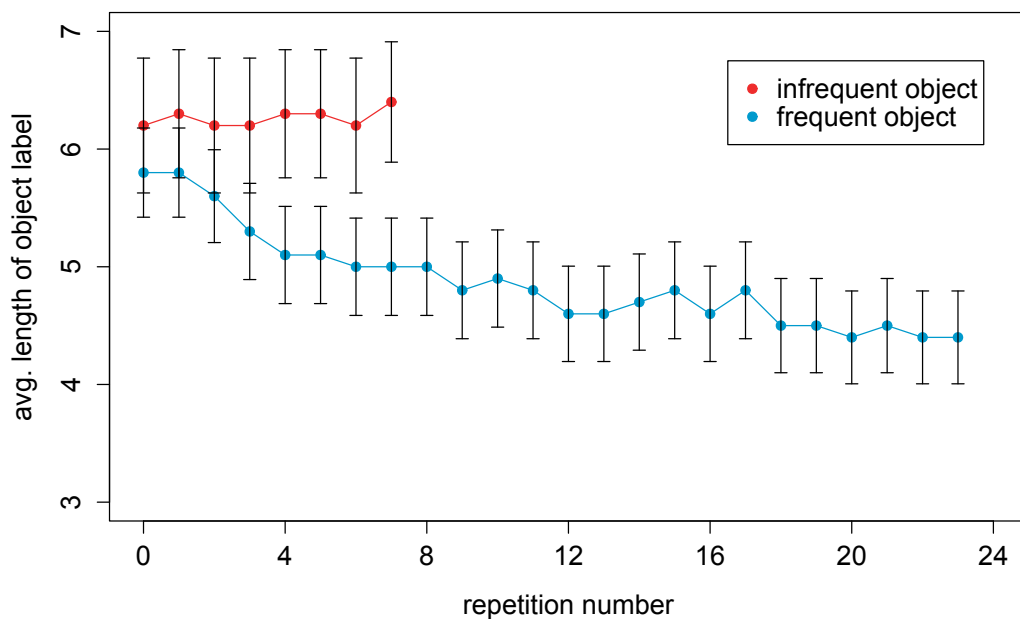


Figure 2.5: Timecourse of productions in the critical Combined condition. Each data point shows the average word length taken over all participants’ productions at a given repetition number of an object.

The mutual information between the forms and meanings in a participant’s lexicon gives us a measure of how predictable the meanings are given the forms and vice versa, and thus tells us how *expressive* a language is, i.e. how much information is expressed by the forms in the lexicon. The average token length of director trial productions serves as a measure for the effort expended. According to the Principle of Least Effort, an optimal language would maximise expressivity while minimising effort. Only participants in the critical Combined condition produce languages which are optimal in this way. Participants in the Accuracy condition gravitate overwhelmingly towards the strategy that maximises expressivity *and* average token length, and participants in the Time condition maintain minimal average token length but sacrifice expressivity to do so; these were the optimal strategies to use in these respective conditions, given the different task demands.

In Figure 2.5, we take a closer look at the possible mechanisms behind participants’ trial-by-trial production choices in the Combined condition, by measuring the average length of each object’s label over successive repetitions. (Note that participants’ frequent and infrequent object production trials are randomly shuffled, and

using a bias-corrected estimate; since our use of this measure is for purposes of comparison between participants, we are not concerned with the absolute values, which would be lowered by roughly the same factor across all participants using a bias-correction method such as the Miller-Madow method.

Table 2.2: Summary of fixed effects for a binomial regression model with short name use as the binary dependent variable, and by-participant random effects for object frequency and trial number. This model is fitted to the data from all participants’ production trials in the Combined condition.

	β	SE	p
intercept (object=infrequent)	-7.115	2.067	0.001
object=frequent	3.949	2.251	0.079
trial number	0.064	0.059	0.279
object=frequent x trial number	0.137	0.046	0.003

thus repetition number does not correspond with a specific spacing of trial numbers.) As discussed in §2.1.1, earlier studies by Krauss and Weinheimer (1964) and Clark and Wilkes-Gibbs (1986) show that object descriptions tend to shorten with repetition, and that more frequent objects end up with shorter descriptions simply because they go through more repetitions. In these studies, because the meaning space was small compared to the large descriptive space available (i.e., English phrases with no length restriction), all descriptions could be shortened somewhat without producing ambiguous form-meaning mappings. In our study, we investigated the case where a pressure to use shorter forms comes into direct conflict with the pressure to avoid ambiguity: in this miniature lexicon, shortening yields the same, ambiguous label for the two objects in the meaning space.

If participants are simply more likely to use a shorter form for an object the more times they communicate about that object, then we would expect the average label length for both the frequent object and the infrequent object to decrease at a similar rate as the number of repetitions increases. However, as Figure 2.5 shows, this is not what we find. Only the average label length of the *frequent* object decreases with successive repetitions; the average label length of the infrequent object remains roughly constant over the course of the trials. A logistic regression model fit to just the data from the Combined condition, with short name use as the binary dependent variable, object frequency, trial number and their interaction as fixed effects, and by-participant random intercepts and slopes for object frequency and trial number, confirms this. The model results (Table 2.2) show an overall significant positive effect of trial number on short form use only when the object is frequent. Note that there is also a marginal difference between the two objects at repetition number 0. Thus, in the critical Combined condition, while most participants switch to using the short form for the more frequent object at some point during production trials, most also maintain the long form for the infrequent object throughout the trials—the threat of ambiguity appears to block shortening altogether for this object. This suggests that, in cases where the pressure to decrease effort and the pressure to avoid ambiguity come into direct conflict, language-users’ production choices result in systems which maximise

expressivity while minimising effort, optimising across the lexicon as a whole.

Interestingly, there were a small number of participants (for example in the Combined condition) who consistently mapped the short form to the *infrequent object*. While shortening the label for either object does satisfy the time pressure to some extent, why might this sub-optimal strategy be used? One possibility is that a participant’s strategy is not to optimise based on overall frequency distributions within the signalling game, but simply to shorten the first object they are presented with in production trials, which then blocks shortening of the other object. However, of the 10 participants who were presented with the infrequent object first, 30% converged on a ‘reversed’ or other non-optimal strategy as opposed to the optimal strategy. Of the remaining 30 participants who saw the frequent object first, 37% converged on a reversed or other non-optimal strategy. Thus, which object appeared in the first production trial (or even the first several trials, which we also checked) is not predictive of which strategy (optimal or otherwise) the participants converged on in the critical condition. We believe these occasional reversed lexicons are thus more likely due to an effect of the cost of switching an incipient convention during the task. For example, if a participant starts out producing labels probabilistically, following the language they were trained on, they will sometimes produce a short name for the infrequent object. If this results in successful communication, and is picked up by a communicative partner, then this pattern may become conventionalised. However, once such a pattern is established, the cost of switching to a different mapping becomes an obstacle. The pressure to maximise the number of correct guesses in the testing trials means the cost of switching labels would further penalise participants who attempted to abandon an incipient sub-optimal convention midway through the task.

2.1.4 Discussion

More than 80 years ago, Zipf hypothesised that the inverse relationship between word length and word frequency was a universal feature of human language, resulting from language users optimising form-meaning mappings for efficient communication. Our study provides direct experimental evidence linking pressures that operate at the level of the individual during communication to the Law of Abbreviation, an emergent structural feature of languages. In particular, language users converge on an optimally-configured lexicon, preferentially using short but potentially ambiguous labels for frequent objects and long labels for infrequent objects. Importantly, this holds only when both a pressure to communicate accurately and a pressure to communicate efficiently are present.

When these pressures were isolated, the Law of Abbreviation did not emerge; an accuracy pressure alone led participants to use the longer non-ambiguous forms

regardless of frequency, while a time pressure alone led them to use the short forms. Some participants mapped the short form to the more frequent object in the Neither condition, however the effect was much weaker. Thus, while biases towards accuracy and efficiency might be implicitly present in any linguistic task, emphasising these pressures significantly amplified the effect, as predicted. Even though this experiment involved a miniature lexicon consisting of three possible forms, our result is a proof of concept that such pressures can push a lexicon to align with the Law of Abbreviation. We expect the results to scale up to lexicons with more forms and meanings; with the groundwork in place we can now test this in future studies.

It is important to note, however, that there is a distinction between a language-user's *mental representation* of the lexicon, and the form-meaning mappings they actually produce in communication. Participants using the short form for the frequent object and the long form for the infrequent object may still retain associations of both forms with both objects in their mental lexicon—however, the nature of the communicative task in this experiment may have caused them to produce only the short form for one object and the long form for the other based on purely *pragmatic* considerations (see, e.g., Franke, 2017). Given that our experiment only recorded participants' actual productions, we cannot with certainty distinguish between these two possible explanations for the observed behaviour. However, we did include an exit survey which asked participants to explain their strategies during the production stage. Some of the language used in the responses suggested that some participants *had* remapped their mental lexicons. E.g., “I waited until my partner sent Zop twice for the blue round object and then we had a mutual understanding that that’s what the Zop was” and “the small round object was Zop, and the orange tall figure was the longer word.” However, some other participants indicated that they interpreted the short form as either a prefix or convenient shortening—e.g., “one of the objects had to use the long name, as the short Zop was the same prefix for both” and “[I] used just Zop when transmitting Zopekil [as] the other needed more transmission time”—suggesting that they still retained the long form in their mental lexicon even if they stopped using it.⁵

Our interpretation of such cases is that, while this pragmatics-driven asymmetry in usage may or may not lead to an immediate shift in lexical representations, it may be an important first step in such a change. In English, many words exist that initially began as convenient shortenings of longer forms, which are now either no longer in use, or no longer associated with the same meaning as the short forms. Some examples are: *bus* (from *omnibus*); *wig* (from *periwig*); *pram* (from *perambulator*); *pub* (from *public house*); and *pants* (from *pantaloons*). In all these cases, the clipped form has

⁵All the exit survey responses are available along with the full dataset at: <http://datashare.is.ed.ac.uk/handle/10283/2702>.

undergone “opacification”, i.e. it is no longer widely recognised as a derivation of the full form, and exists autonomously in the lexicon as an unmarked, standard form (Jamet, 2009). Likewise, even if participants in our experiment are retaining the long form in their mental lexicon, the rapid decrease in its frequency of use over successive generations of learners would likely lead the long form to eventually drop out of the lexicon, with the short form becoming lexicalised as the standard form. Indeed, studies in the iterated learning paradigm show that, in the lexicons produced by successive generations of participants, those in which two labels map to the same meaning are dispreferred (e.g., Reali and Griffiths, 2009; Smith and Wonnacott, 2010). In short, permanent lexical changes often begin life as pragmatics-driven asymmetries in usage (Bybee, 2010). Thus, even if the alignment with the Law of Abbreviation that we observe in participants’ usage is not yet accompanied by a corresponding shift in their mental lexicons, it is an important intermediary stage on the way to this outcome.

It is also worth noting that across conditions we found evidence for a baseline preference against ambiguity: when no pressures were present, participants tended towards retaining the unique long forms for *both* objects, and no participants used the ambiguous short names for both objects simultaneously. Indeed, in both conditions featuring a time pressure, a few participants nevertheless used the long names across the board. These results suggest that for some participants, the framing of the task as one of learning a language carries with it some expectation of communicative utility.

Returning to the issue of the explanation for the widespread application of the Law of Abbreviation, our results demonstrate that optimisation behaviour on the part of language-users can lead to the production of lexicons which align with this law. Our study expands on previous work that investigates the relationship between frequency and utterance length, by setting up a small lexicon in which the pressures for efficiency and expressivity in a communicative task come sharply head-to-head. We find that these conflicting pressures do indeed lead language-users to map shorter forms to more frequent meanings, as Zipf hypothesised. However, this result does not rule out that additional processes are involved in shaping this global linguistic pattern as well. Indeed, we expect there are many other factors that come into play as the size of the lexicon is scaled up and the conditions become closer to actual language-use: for example the bottlenecks of learning and memory; the influence of predictability in context; constraints of speech production; and the propagation of errors. There may be a role for random statistical processes to play as well. Future work should focus on how the pressures involved in this task interact with these and other factors, and especially on how the behaviour of individuals communicating in a pair spreads outside this context to the level of an entire population.

2.1.5 Conclusions

Zipf’s proposal—that the inverse relationship between a word’s length and its frequency is a universal design feature of language—has been borne out repeatedly in observations of the world’s languages (Teahan et al., 2000; Sigurd et al., 2004; Strauss et al., 2007; Piantadosi et al., 2011; Ferrer-i-Cancho and Hernández-Fernández, 2013). The long-standing explanation for this phenomenon appeals to the idea that language users want to communicate as efficiently as possible. However, the critical link between this Principle of Least Effort and the emergence of an optimal lexicon has remained largely untested. Our study explored the hypothesis that the mechanisms operating in individual language users during online language production can result in the active restructuring of a lexicon. Our findings reveal that when pressures to communicate accurately and efficiently are both present and in conflict, language users exploit information in the input about the frequency of meanings to converge on an optimally-configured lexicon. When only one of these pressures is present, the effect does not emerge. This result provides evidence that the universal pattern Zipf observed can indeed arise through individual-level optimisation of form-meaning mappings. More generally, this method provides a model for future work showing how explanations of population-level properties of languages can be grounded in the moment-to-moment behaviours of individuals.

2.2 Lab experiment (Experiment 2)

The experiment described in the preceding part of this chapter was first run in the Language Evolution Lab at the University of Edinburgh, recruiting participants through the university’s experiment advertisement system. Because two of the four conditions require pairs of interacting participants, and it was often the case that some participants who signed up would not show on the day, running these conditions was slow and time-consuming. This motivated us to switch to online data collection, using the crowdsourcing platform Amazon Mechanical Turk. However, before we made the switch and ran the experiment described in Section 2.1 above, we were able to collect 8 data points for each of the four conditions in the lab setting. Here, I describe the results of this earlier mini-experiment, and compare them to the results of the online experiment. The conclusion is that the results of the two experiments are qualitatively similar, confirming the robustness of our findings across different populations and data collection methods. However, the lab participants’ behaviour appears to gather more strongly around the extremes, while the online participants’ behaviour is a bit “messier”.

2.2.1 Participants

48 participants (37 females, 11 males) were recruited through the University of Edinburgh’s paid work advertisement hub. Half of these reported themselves as native English speakers, of which 7 were monolingual. A broad range of other languages were represented across the remaining participants. Ages ranged from 18 to 38 (mean=23, SD=4.56).

2.2.2 Materials

The materials were identical to those used in the online experiment (Section 2.1).

2.2.3 Procedure

The experimental procedure was also identical to that of the online experiment, except participants were seated inside isolated booths in the Language Evolution Lab, where they could be monitored through a small window by the experimenter. Each booth contained a computer, from which the experiment was run in a browser window. Participants in the paired conditions (the Combined and Accuracy conditions) were aware of their partner sitting in a nearby (but non-adjacent, to minimise any potential noise disturbances) booth.

The payment protocols also differed from that of the online experiment, as we established from past experience that a minimum payment of £5 was necessary to incite people to come into the lab. All participants were thus paid £5, regardless of experimental condition. Including reading and signing of consent forms, wait time for partners to arrive, and debriefing (which was offered upon request), participants in the slowest condition to complete (the Combined condition) spent approximately 30 minutes in the lab, while participants in the fastest condition to complete (the Neither condition) usually spent less than 10 minutes. In the Combined and Time conditions, where a bonus was offered to encourage quick (and in the Combined condition, also accurate) communication, the bonus was also set to £5 per winning player.

2.2.4 Results

Figure 2.6a shows the proportion of trials on which the short (ambiguous) label was selected by the director, for high- and low-frequency objects. As predicted, in the Accuracy condition, most pairs retained the unique long names for both objects, while in the Time condition, most participants mapped both objects to the ambiguous short name. Crucially, in the Combined condition, where participants were under pressure to communicate both accurately and efficiently, most pairs converged on the optimal strategy wherein the most frequent object was mapped to the ambiguous short name, and the infrequent object to its unique long name. Finally, the Neither

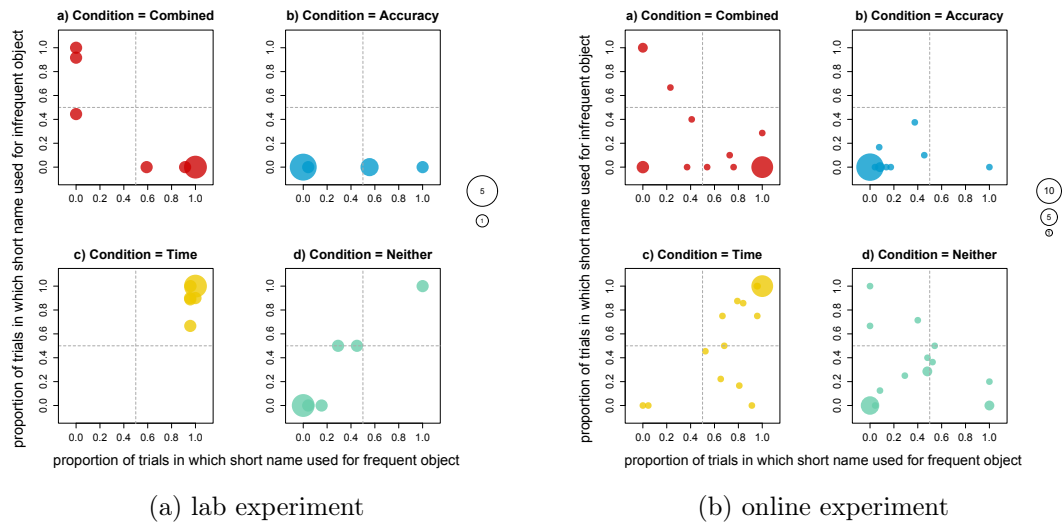


Figure 2.6: The proportion of trials in which the short name was used to label the frequent object versus the proportion of trials in which it was used to label the infrequent object, for the lab experiment (a), and the online experiment (b). For the Combined and Accuracy conditions, each data point combines a pair of communicating players, representing the sum of their director trial productions. For the Time and Neither conditions, each data point corresponds to an individual player’s productions. The size of the circles is perceptually scaled (Tanimura et al., 2006) to reflect the number of data points coinciding at each value. Data from only the second half of testing trials is shown here, as participants were more likely to have converged on a stable mapping by this time. Data points in the top right quadrant indicate participants who are mostly using the short name for both objects; participants are clustered in this quadrant in the Time condition. Data points in the bottom left quadrant indicate those who are mostly using the unique long names for both objects; participants are clustered here in the Accuracy condition. Data points in the bottom right quadrant indicate participants who are mostly using the short name for the frequent object and the long name for the infrequent object. This behaviour, consistent with the Law of Abbreviation, only reliably arises in the Combined condition, where both pressures are present. Both the lab and online experiments exhibit qualitatively similar behaviour.

condition revealed an underlying bias towards avoiding ambiguity. These results therefore display qualitatively the same behaviour that was observed in the online version of the experiment (the results of the online experiment are reproduced in Figure 2.6b for comparison).

However, fitting a linear mixed effects regression model to the data did not yield significance for any of the predicted effects (which were an overall negative effect for the Accuracy condition, an overall positive effect for the Time condition, and a positive interaction effect for the frequent object in the Combined condition). The model (Table 2.3) was fit using the same parameters as that in the online study (Table 2.1): short name use as the binary dependent variable; fixed effects for object frequency, condition, and their interaction; and by-participant random effects for object frequency. As in the online study, the model was also fit to only the second half of participants’ testing trials, as they were more likely to have converged on a

stable mapping by this point, and it was contrast-coded, setting the intercept to the grand mean. The only significant effects found were for the intercept—revealing a baseline preference against using the ambiguous short form, as found in the online study—and the Combined condition, where an overall preference to use the short form as compared to the mean was found. We suspect that the model’s inability to find any other significant effects, and particularly interaction effects, is due to the low number of data points used in this model, which included data from 8 or 16 participants per condition, as compared to 20 or 40 per condition in the online study.

Table 2.3: Summary of fixed effects for a binomial regression model with short name use as the binary dependent variable, and by-participant random effects for object frequency, fit to data from only the lab experiment. Like Figure 2.6, this model is fit using only the second half of each participant’s testing trial data, as participants were more likely to have converged on a stable linguistic mapping by then. Significant effects are in bold.

	β	SE	p
intercept	-1.977	0.433	<0.001
object=frequent	0.050	0.217	0.820
condition=Accuracy	1.081	0.594	0.069
condition=Time	0.241	0.729	0.741
condition=Combined	1.238	0.594	0.037
object=frequent & condition=Accuracy	0.063	0.226	0.780
object=frequent & condition=Time	-0.008	0.210	0.970
object=frequent & condition=Combined	-0.334	0.226	0.140

When we combine the data from both the lab and online experiment together, and fit them using the same model parameters as in the previous models, we find that the results yield the same significant effects, in the same direction, as the online model alone (Table 2.1), including the three main predicted effects. These results are shown in Table 2.4. Adding the source of the data (lab or online) as a fixed effect does not change these results, or significantly improve the model’s fit to the data, as confirmed by a model comparison test ($\chi^2(1) = 0.056, p = 0.813$). Thus our models did not reveal any significantly different behaviour in the lab data as compared with the online data.

2.2.5 Discussion

While the qualitative resemblance between the lab and online data is apparent when visually comparing Figures 2.6a and 2.6b, one striking difference that can also be seen is the lack of any data points falling near the centre of the graphs in Figure 2.6a, in all conditions but the Neither condition. It seems that in the lab experiment, participants were more likely to fall at the extreme corners of the graphs, meaning that they had a fully consistent or near-consistent mapping between forms and meanings by the

Table 2.4: Summary of fixed effects for a binomial regression model with short name use as the binary dependent variable, and by-participant random effects for object frequency, fit to the combined data set of the lab and online studies. The key predicted effects, which were also present in the online data alone (Table 2.1), are in bold.

	β	SE	p
intercept	-2.193	0.380	< 0.001
object=frequent	0.813	0.313	0.009
condition=Accuracy	-2.754	0.529	<0.001
condition=Time	3.643	0.678	<0.001
condition=Combined	0.726	0.501	0.147
object=frequent & condition=Accuracy	-0.152	0.404	0.707
object=frequent & condition=Time	-0.617	0.545	0.257
object=frequent & condition=Combined	1.087	0.381	0.004

second half of testing trials. In the online experiment, participants’ performance is a bit messier, in the sense that participants show varying degrees of consistency in their mappings during the second half of testing trials. Even if many of these participants ultimately converge on a stable mapping, they do not all reach it as quickly as those in the lab experiment do.

This may be due to several factors: in the lab, participants are more likely to give the task their full concentration, as there are no distractions and they can be observed through a window. However, the online participants have potentially many distractions from both the internet and from whichever place they are accessing the internet. The payment was also higher in the lab condition, and this may have provided a stronger motivation for participants to perform “optimally”. Another factor that might have strengthened the motivation to perform is the fact that the lab participants had direct contact with the experimenter, whereas for the online participants, the experimenter was simply a theoretical entity, thus perhaps providing less reason for them to behave in a conscientious or responsible manner as compared with the lab participants. Finally, the lab participants were all students at the University of Edinburgh, and therefore representative of a highly motivated and educated demographic. We do not have any information on the levels of motivation and education represented in our online participant pool, but we can assume that most are unlikely to be current university students at a prestigious British university, and thus represent a more diverse cross-section of society. We do know that in the online experiment, the age range was much wider, and the participants were located all over the world.

Of course, the higher proportions of consistent behaviour observed in the lab experiment may also simply be due to the smaller sample of data points available in this study, as compared with the online study, for which we collected nearly three times as much data.

Despite the slight difference discussed above, the key point to take away from this direct comparison of two different data collection methods is how *similar* the findings were across both experiments. Specifically, comparing the graphs of each condition across the two experiments reveals remarkably consistent patterns of behaviour: in the Accuracy condition, participants tend to use the the unique long forms for both objects; in the Time condition, they tend to use the short form both objects; and in the critical Combined condition, a majority of participants map the short form to the frequent object and the long form to the infrequent object, in line with Zipf’s Law of Abbreviation. We even find the same baseline preference against ambiguity in the Neither condition of both experiments. What this points to is the general robustness of the effects we observed, independent of the precise method of data collection used, the size of the monetary incentive offered, the location or environmental conditions of the participants, and the demographic makeup of participants.

This conclusion is important for the following chapter—in which we discuss results from both online and lab studies using a similar experimental setup—and for any future work done using this setup. What our direct comparison of lab and online data collection methods in this experiment shows is that future results obtained from either a lab experiment or an online experiment using this framework are unlikely to be mere artifacts of the data collection method used. Moreover, the results found using one method are very likely to be replicable using the other method. This is not the case for all experimental designs—some are extremely sensitive to the conditions in which the experiment is run and to the demographic makeup of the participants. Therefore, an additional achievement of our study is that we have hit upon a robust and versatile experimental design, one that can be run in the lab or online depending on the available resources and timing considerations, and which allows us to observe reliable effects that are replicable across demographics, locations, and environmental conditions. We hope that others can tweak this design to investigate different linguistic behaviours, as we do in the following chapter.

Chapter 3

Word length and predictability in context

The text of this chapter comprises the body of a manuscript in preparation for journal submission. The manuscript was prepared with feedback from my supervisors—Kenny Smith, Simon Kirby, and Jennifer Culbertson—who will also be co-authors on the journal submission.

3.1 Introduction

Zipf (1935) observed that word length tends to be inversely proportional to word frequency in the lexicon. He hypothesised that this widespread cross-linguistic pattern was due to the *Principle of Least Effort*: language users align form-meaning mappings in such a way that effort is minimised while expressivity is still maintained. However, word frequency is not the only reliable predictor of word length. Using corpora from 11 different languages, Piantadosi et al. (2011) show that a word’s predictability in context (where context is defined as the two words preceding the target word) is even more strongly correlated with word length than frequency is: words that are, on average, more predictable in context tend to be shorter.

Measuring how predictable or unpredictable a word is in a particular context gives us a way of defining the *information content* or *surprisal* of a word. For example, consider the two sentences:

- (1) The early bird catches the worm.
- (2) Our early bird special today is worm.

In sentence (1), a well-known proverb, the word *worm* is entirely predicted by the preceding words. The word itself thus gives us practically no new information, and so we say it has *low* information content or surprisal. In sentence (2), the same word

is highly unlikely given the preceding words, and thus we find it surprising. This element of surprise is what we associate with *high* information content or surprisal.

Using these concepts, we can apply Zipf’s Principle of Least Effort to hypothesise that a speaker’s drive to reduce effort will be directed towards words that are highly predictable given the context, as the context already does much of the work in helping the hearer deduce the intended word. Words that are more surprising in a particular context will be less likely to be reduced, or more likely to be lengthened, to maximise the hearer’s chances of accurate comprehension. The resulting state is one in which low-information words are shorter on average than high-information words, and thus the length of a word is roughly proportional to the amount of information it carries. This is consistent with the uniform information density (UID) principle—also known as the smooth signal redundancy (SSR) hypothesis—which states that information is distributed roughly evenly across units of time in an utterance, and therefore longer words should carry more information than shorter words. This hypothesis was introduced at least as early as 1980 by Gertraud and August Fenk, but has since been put forth in Fenk-Oczlon (2001); Genzel and Charniak (2002); Aylett and Turk (2004); Levy and Jaeger (2007) and Jaeger (2010), among others.

There are many ways to operationalise the information content of a word. One way is to use the *N-gram probability* of a word, i.e. its probability conditioned on a window of N preceding or following words. This is the method used by Piantadosi et al. (2011), and the method we will use here. Zipf’s word frequency measure is in fact just a limiting case of this N-gram probability, where $N=0$. Other measures include *syntactic probability*, a word’s probability of appearing in a particular syntactic structure (e.g., Tily et al., 2009), and *givenness*, a word’s likelihood of mention given the semantic context (Aylett and Turk, 2004).

Both corpus studies and controlled language production experiments have linked low information content, operationalised in these different ways, to various types of linguistic reduction. Lieberman (1963); Aylett and Turk (2004); Gahl and Garnsey (2004); Tily et al. (2009); Kuperman and Bresnan (2012), and Seyfarth (2014) show that words with low information content are more likely to undergo different types of phonetic reduction, such as shortened word duration and word-final /t,d/-deletion. Bell et al. (2009) show that some of the different measures of information content mentioned above (frequency, N-gram probability, and semantic givenness) in fact contribute separately to the phonetic duration of a word. Jaeger (2010) shows that *that*-complementisers are more often dropped when the following word is less surprising in context. Fedzechkina et al. (2012) show that case markers are more likely to be omitted on nouns in more probable syntactic roles.

If predictability in context can lead to phonetic reduction and deletion in real-time, then these effects might make their way to the lexicon, producing, for example, the

inverse relationship between predictability and word length observed by Piantadosi et al. (2011). However, there is relatively little work directed at understanding the causal mechanisms that could ultimately give rise to this widely observed pattern.

One way of approaching this issue is by tracking language users' online choices when producing words that are part of a 'clipped pair', i.e. when both a long form and an abbreviated or 'clipped' form exist that have the same or very similar meanings (Mahowald et al., 2013). For example, in both English and French, *info/information* is a clipped pair. Mahowald et al. presented participants with English sentences containing a blank and asked them to complete the sentence with either the long or the clipped form. They found that participants were more likely to choose the clipped form in predictive contexts. This is consistent with the hypothesis that the lexicon-level patterns observed by Piantadosi et al. (2011) may be due in part to a *least-effort* mechanism, in which speakers balance communicative success with efficiency to assign word length proportionally to information content. Indeed, the authors suggest that there may be a mechanistic link between this type of online speaker behaviour and the large-scale proportional relationship observed between word length and information content, via processes of lexical change: "if a word's surprisal decreases, one should expect that word to shorten over time" (p. 317).

However, because Mahowald et al.'s study uses English sentence frames and target words, we cannot rule out potentially confounding contributions from register, prosody, and participants' learned preferences to their word choice in particular instances. For example, English-speakers may have learned to associate clipped forms like "math" with contexts that suggest an informal register, or they may choose longer forms when it improves the phrasal rhythm. More importantly, we cannot assess whether the effect is really driven by the competing pressures for communicative accuracy and efficiency without manipulating the presence or absence of these different communicative pressures. For instance, in Mahowald et al.'s task, participants clicked on a word rather than typing it, and thus there was no difference in effort associated with choosing the long or short form. In addition, participants were told to choose a word based on "which sounded more natural", rather than being directly engaged in a task requiring successful communication.

Here we use an artificial language learning paradigm to investigate the effect of communicative pressures on word length in context. In our setup, participants learn a miniature lexicon, then use it to either communicate with another participant, or—in a control condition—to simply describe objects. The use of an artificial language allows us to test participants' behaviour in the absence of any specific prior learned associations, which will exist in the languages they already speak. Kanwal et al. (2017b) used this setup to show that competing pressures to communicate both accurately and efficiently led participants to restructure their input so that shorter words

were mapped to more frequent meanings. Here, in Experiment 1, we test whether language users shorten words that are more *predictable in context* when subject to these same pressures. In three control conditions, we switch these pressures off one by one. The use of a design that implements this structure of control conditions allows us to directly assess the causal contribution of these interacting communicative pressures on participants' behaviour. We find that participants only shorten words in predictive contexts in the critical condition and not in any of the control conditions. This result supports the hypothesis that optimisation behaviour by language users is a likely contributing factor to the proportionality between word length and information content observed across languages.

In Experiment 3 (experiments have been renumbered relative to the thesis as a whole), objects differ in their probability of occurrence given a particular context, but their *average surprisal* is the same. Experiment 4 introduces objects that differ in their overall average surprisal in the artificial language. We then test the prediction that lower average surprisal within the language corresponds to lower average word length, and find evidence that it does. This strengthens the plausibility of one proposed mechanism for converting context-specific word length alternation into permanent lexical change.

3.2 Experiment 3

Artificial language learning paradigms have previously been used to shed light on the cognitive mechanisms and environmental pressures that shape language structure (e.g., Kirby et al., 2008; Fedzechkina et al., 2012; Culbertson et al., 2012). Participants learn a miniature artificial language, and then we observe how they reshape their input as they use the language, in this case to communicate with a partner (see also Winters et al., 2015; Kirby et al., 2015; Fehér et al., 2016).

The specific experimental design used in this paper is closely based on that of Kanwal et al. (2017b)—we implement a 2×2 between-subjects manipulation of communicative pressures, such that pressures to communicate accurately *and* efficiently are present in the critical condition, and one or both of these are switched off in three different control conditions, allowing us to assess the causal role of these pressures on the observed behaviour. All four conditions use the same initial training language, but their procedures differ in the testing phase. The experiment was run online through the crowdsourcing platform Amazon Mechanical Turk.

3.2.1 Participants

120 participants (53 females, 66 males, 1 other) took part. These were grouped into 20 pairs each in the Combined and Accuracy conditions, and 20 individuals each in

the Time and Neither conditions. Of the 40 participants in the Combined condition, 36 reported themselves as native English speakers, of which 34 were monolingual. The remaining 4 participants were native speakers of Hindi, Tamil, Telugu, and Spanish, respectively. Ages ranged from 19 to 61 (mean=34, SD=10). Of the 40 participants in the Accuracy condition, 38 reported themselves as native English speakers, of which 32 were monolingual. The other two participants were a native Spanish speaker and a native Telugu speaker. Ages ranged from 22 to 70 (mean=34, SD=10.4). Of the 20 participants in the Time condition, 17 reported themselves as native English speakers, of which 15 were monolingual. The remaining participants were native speakers of Tamil, Russian, and Ukrainian, respectively. Ages ranged from 18 to 50 (mean=31.5, SD=9.1). Finally, of the 20 participants in the Neither condition, 17 reported themselves as monolingual native English speakers; the remaining participants were two native Tamil speakers and a native Spanish speaker. Ages ranged from 21 to 50 (mean=30.1, SD=6.2).

3.2.2 The Training Language

Participants were trained on two names for each of two plant-like alien objects, by repeatedly being shown pictures of the objects labeled with a simple sentence. The sentence consisted of a framing word followed by the object’s name. There were two possible frames, *bix* and *gat*. Overall there were 64 training trials, with each object appearing 32 times and each frame appearing 32 times. Crucially, one object appeared seven times more frequently with the frame *bix* than *gat* (28 and 4 times, respectively), while the other object appeared seven times more frequently with the frame *gat* than *bix* (again, 28 and 4 times, respectively). This meant that each object appeared in both a predictive context and a surprising context, but which frame was predictive and which was surprising was flipped between the two objects.

Furthermore, the object name appeared half the time in its full form, a 7-letter word, and half the time in shortened form, a 3-letter word derived by clipping the last two syllables off the long name. These short and long forms were evenly distributed across both predictive and surprising contexts, ensuring that the input language contained no bias towards using one form in any particular context. Specifically, short forms were not more likely to appear in predictive contexts, as they would if the language were adhering to the principle of uniform information density. A schematic diagram of the object-label frequencies in the training language is provided in Fig. 3.1A.

We designed the names such that the short name for both objects was identical (*zop*), while the long names were unique (*zopekil* and *zopudon*). We did this in order to simulate the fact that in natural languages, shorter words are subject to greater confusability, for a number of reasons. Shorter forms have less space for

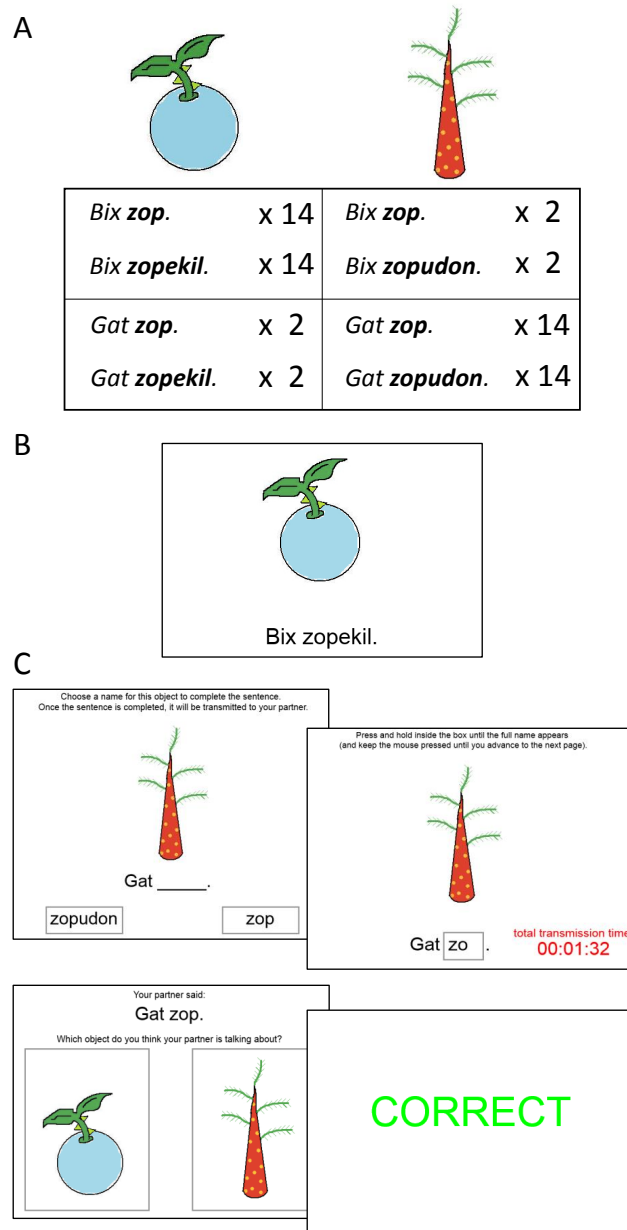


Figure 3.1: (A) The input frequencies of the objects and framing sentences presented during training trials in all four experimental conditions of Experiment 3. (B) A sample training trial from Experiment 3. (C) A sample director trial from the Combined condition of Experiment 3 (top) and a matcher trial followed by feedback (bottom).

signal redundancy and thus are more likely to be lost in noisy signal transmission. Because languages have a finite phoneme inventory, there are more unique possible long strings than short strings, and thus word shortening can also result in outright ambiguity of form-meaning mappings. Indeed, shorter words are more likely to be polysemous and homophonous (Piantadosi et al., 2012).

Which object (the blue fruit or the red stalk) was more predictable given which frame, as well as which object was paired with which long name, were both counter-balanced between participants, giving a total of 4 possible object-frame-name pairings which a participant might be trained on. This ensured that potential factors such as sound symbolism, or higher saliency or learnability of any specific object-word pairing, could not systematically bias our results.

3.2.3 Training Procedure

On each training trial, an object was presented on screen alone for 700ms. The appropriate sentence then appeared beneath the object for a further 3000ms, yielding a total trial duration of 3700ms. A blank screen showed for 500ms between trials. The 64 training trials were presented in a different randomised order for each participant.

3.2.4 Testing Procedures

After the training phase, the testing procedures varied depending on the condition. In the Combined condition, participants were under a pressure to communicate accurately *and* efficiently. According to the Principle of Least Effort, it is balancing these competing pressures that would lead language users to distribute word length inversely to word predictability. The remaining three conditions removed one or both of these pressures, resulting in a between-subjects 2×2 manipulation of the accuracy and time pressures.

3.2.4.1 Combined condition

In the testing phase of this condition, participants were paired with a partner to play a communication game, using the method developed for running two-player online experiments in Kanwal et al. (2017b). On each trial, the ‘director’ was shown an object on the screen with a framing word followed by a blank. The director was instructed to choose a name for the object to complete the sentence, and once the name was entered, the sentence would be transmitted to the ‘matcher’. The director could choose one of two options to complete the sentence: the unique long name for the object or the (ambiguous) short name. Once the chosen name was selected by clicking on the appropriately labeled button, it had to be entered into the blank space by pressing and holding the mouse as each letter appeared one after the other

at 1200 ms intervals. Only after all the letters in the name had appeared in the box was the completed sentence transmitted to the matcher. This belaboured method of production, in which the long name was significantly slower to produce than the short name, was introduced to model the difference in effort and speed associated with producing long versus short utterances.

Once the director completed their description, the entire sentence (frame plus chosen word) was transmitted to the matcher, who was asked to choose which of the two objects they thought the director was referring to. Both players were then given feedback as to whether the matcher’s choice was correct.

The players alternated roles after every trial, with the matcher becoming the director and the director becoming the matcher, until both completed 32 director trials and 32 matcher trials. The proportion of times each object appeared with each frame in each player’s director trials matched those of the training proportions: one object appeared seven times more frequently with the frame *gat* than *bix*, and the other appeared seven times more frequently with *bix* than *gat*. The order of each participant’s 32 director trials was randomly shuffled.

To model the pressures in spoken communication to be both efficient and accurate, pairs were told at the beginning of the communication phase that they would be rewarded a bonus payment of \$1 if they were the pair to complete the game in the quickest time with the highest number of correct match trials. Time was only counted when the director was entering a name into the blank, and the total time count was displayed next to the blank during this process, to emphasise the time pressure. Screenshots of sample director and matcher trials are shown in Fig. 3.1C.

In this condition, with pressures to be speedy yet accurate, we expected participants to converge on an optimal strategy in which the short name is used for an object when it appears in its predictive context, and the long name otherwise. In predictive contexts, the framing word already provides a lot of information to the matcher about which object is likely under discussion, and thus participants can minimise effort by using the short form. Conversely, in surprising contexts, the full object name is required to ensure disambiguation.

In order to establish a causal link between these purported mechanisms and the behaviour we observe, we included three further experimental conditions, described below, for a full 2×2 manipulation of the pressures for accuracy and efficiency.

3.2.4.2 Accuracy condition

In this condition, participants were paired to play a communication game as described above, but in the director trials, there was no intermediate step between the director choosing a name to complete the sentence and the matcher receiving the sentence; the names were entered instantaneously, thus removing any difference in effort between

producing long or short names. Pairs were told that the goal of the game was simply to have their partner make as many correct guesses as possible. No bonus prize was offered in this condition, as we expected most pairs to hit ceiling as they did in Kanwal et al. (2017b).

We predicted that participants would be more likely to use the long names for both objects across all contexts in this condition. This is because the long names are less confusable, and without a pressure to be efficient, there is little reason to use the shorter name.

3.2.4.3 Time condition

In this condition, communication was taken out of the game entirely; participants played a one-player game consisting of 64 director trials. In each trial, participants were simply asked to choose either the long or short name for the object shown to complete the sentence. The name was then entered as in the Combined condition, by pressing and holding the mouse in the blank space, with each letter appearing at 1200 ms intervals, while a timer displayed the total time count. The next trial began once all the letters had appeared in the box. Participants were told at the beginning of the game that they would be rewarded a bonus payment of \$1 if they were the player with the shortest total time.

Here, we expected participants to use the short name for both objects across all contexts: with no communicative purpose attached to the transmissions, and an incentive to be as quick as possible, using the short name in every trial is the best strategy.

3.2.4.4 Neither condition

The fourth and last condition contained neither a pressure for efficiency nor a pressure for accuracy. As in the Time condition, participants played a one-player game with no explicit communicative element. Additionally, there was no time difference associated with transmission; once a label was chosen to complete a sentence, it was instantaneously entered and the player advanced to the next trial. We included this condition to provide a baseline for participants' behaviour from which to assess the effects of the accuracy and time pressures in the other three conditions.

In this condition we expected that participants might probability-match—i.e. use the long and short forms for both objects with equal frequency, as in the training trials (Hudson Kam and Newport, 2005)—or that their behaviour might reveal additional prior biases language users bring to the task, such as a preference against using ambiguous forms, as observed in Kanwal et al. (2017b).

Table 3.1: Summary of fixed effects for a binomial regression model with short name use as the binary dependent variable. Significant effects are in bold.

Fixed Effect	β	SE	p
Intercept	0.139	0.149	0.351
Context=Predictive	0.118	0.100	0.238
Condition=Accuracy	-1.470	0.232	<0.001
Condition=Time	2.187	0.292	<0.001
Condition=Combined	-0.392	0.230	0.089
Context=Predictive, Condition=Accuracy	-0.490	0.161	0.002
Context=Predictive, Condition=Time	0.224	0.202	0.269
Context=Predictive, Condition=Combined	0.619	0.158	<0.001

3.2.5 Results

Fig. 3.2 shows the proportion of trials in which the short name was produced by each participant or pair of participants in predictive versus surprising contexts. Our predictions were borne out by the results in all four conditions. In the critical Combined condition, in which participants were subject to the combined pressures for accuracy and efficiency, pairs of communicating participants produced systems in which the short name was used in predictive contexts and the long name in surprising contexts. Crucially, only when both pressures were present did participants reliably produce systems where word length was conditioned on context in this way. In the Accuracy condition, participants tended to use the long name for both objects regardless of context, and in the Time condition, they used the short name for both objects regardless of context. In the Neither condition, some participants stuck with the long name or the short name throughout the trials regardless of context, as in the Accuracy or Time conditions; however, most participants probability-matched.

A logistic regression model was fit to the full dataset in R (R Core Team, 2015) using the lme4 package (Bates et al., 2015), with short name use (as contrasted with long name use) as the binary dependent variable; context (predictive or surprising), experimental condition, and their interaction as fixed effects; and by-participant random slopes and intercepts for context.¹ All fixed effects were sum coded, setting the grand mean as the intercept. The results, shown in Table 3.1, yielded a significant positive interaction between context and the Combined condition, indicating that in this condition, participants were significantly more likely to use the short name in predictive contexts. The only other significant effects found were as follows: a positive overall effect in the Time condition, indicating that participants were more likely to

¹The regression models reported in this paper are fit to the full set of training trials, rather than the second half of training trials—used in the figures shown here and in our earlier frequency study (Kanwal et al., 2017b)—because data for the less frequent contexts would otherwise be too sparse to fit reliable models.

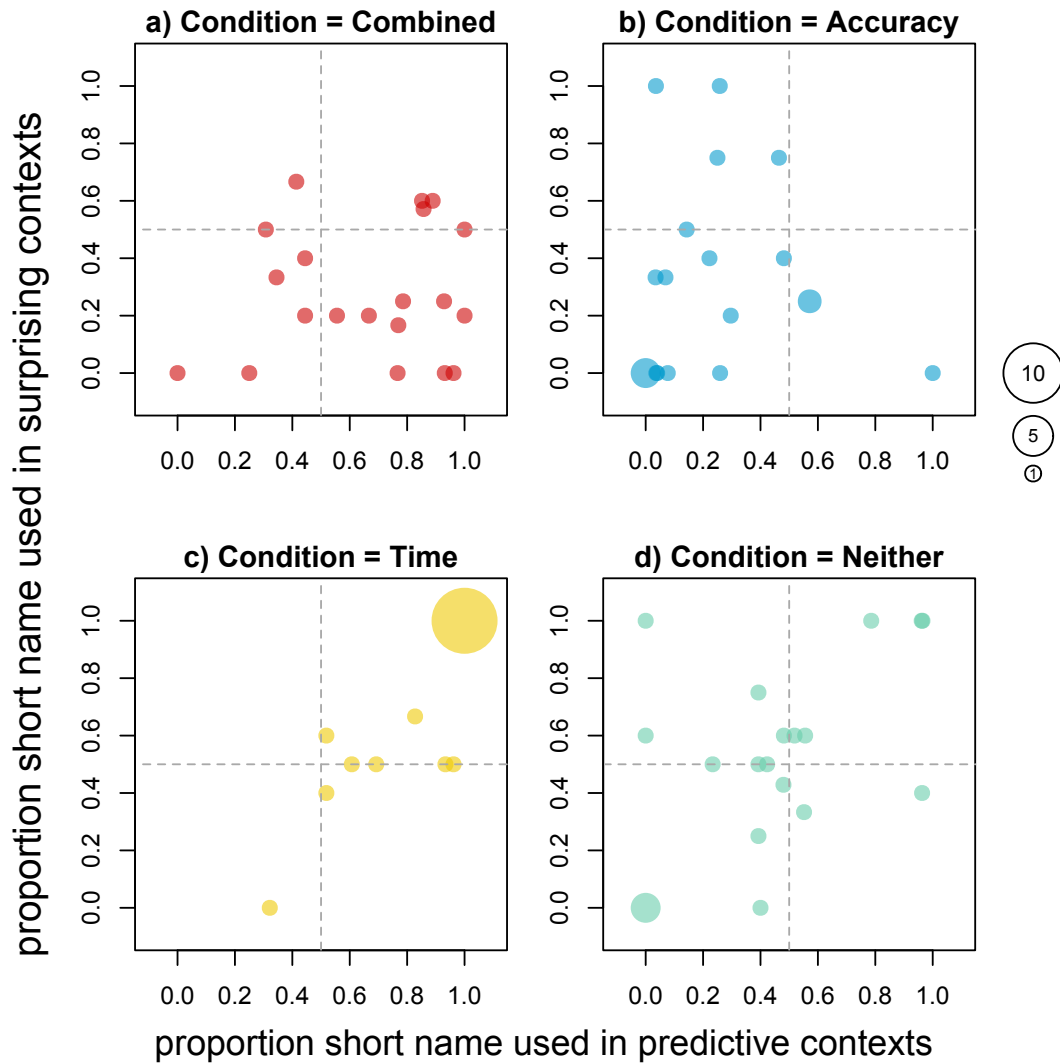


Figure 3.2: The proportion of trials in which the short name was used in predictive contexts versus the proportion of trials in which it was used in surprising contexts in Experiment 3. For the Combined and Accuracy condition, each data point combines a pair of communicating players, representing the sum of their director trial productions. For the Time and Neither condition, each data point corresponds to an individual player’s productions. The size of the circles is perceptually scaled (Tanimura et al., 2006) to reflect the number of data points coinciding at each value. Data from only the second half of testing trials is plotted here, as participants were more likely to have converged on a stable mapping by this time. These results demonstrate that behaviour consistent with the principles of uniform information density and smooth signal redundancy—using short forms in predictive contexts and long forms in surprising contexts, generating systems that fall in the bottom right corner of each graph—only reliably arises in the Combined condition.

use the short form in this condition regardless of context; a negative overall effect in the Accuracy condition, indicating that participants were *less* likely to use the short form in this condition regardless of context; and finally a negative interaction effect of context in the Accuracy condition, indicating that in fact participants were *even* less likely to use the short form in the predictive context in this condition.

We also calculated the average mutual information between name produced and context (predictive or surprising) in each participant’s output language (MI_c). The mutual information between two variables X and Y is given by

$$\begin{aligned} MI(X;Y) &= H(X) - H(X|Y) \\ &= - \sum_{x \in X} P(x) \log P(x) + \sum_{x \in X, y \in Y} P(x,y) \log P(x|y) \end{aligned} \quad (3.1)$$

where $H(X)$ is the entropy of variable X , defined in terms of its probability distribution, $P(X)$. Here, the two variables are name and context, and thus MI_c is defined as

$$MI_c = - \sum_{c \in C} P(c) \log P(c) + \sum_{c \in C, n \in N} P(c,n) \log P(c|n) \quad (3.2)$$

where C is the set of context types {predictive, surprising} and N is the set of names {*zop*, *zopekil*, *zopudon*}.² The more reliably participants are conditioning their use of the long and short names on context, the higher we would expect the value of MI_c to be. The distributions for all four conditions are plotted on the lefthand graph of Fig. 3.3. We find that participants’ MI_c is significantly higher in the Combined condition than in any other condition. A linear regression was run on the set of all participants’ MI_c values, using the second half of their testing trial data. The predictor variable was experimental condition, and the intercept was set at the Combined condition. A significant negative effect of the Accuracy ($\beta = -0.081, SE = 0.033, p = 0.016$), Time ($\beta = -0.184, SE = 0.041, p < 0.001$), and Neither ($\beta = -0.128, SE = 0.041, p = 0.002$) conditions was found. Note that the bulk of MI_c values in the Combined condition are still well below the value for the optimal language. This is driven by the many participants whose output languages were suboptimal in this condition, falling outside of the bottom right corner of Figure 3.2a. However, there are still many more participants optimising in this condition than in the other three conditions, where almost no data points fall in this bottom right region, explaining why the MI_c values are even lower in these conditions.

We additionally calculated the average mutual information between name produced and *object* (the blue fruit or the red stalk) in each participant’s output language

²All values of mutual information in this paper are computed directly from the empirical distributions, without applying any bias-correction methods.

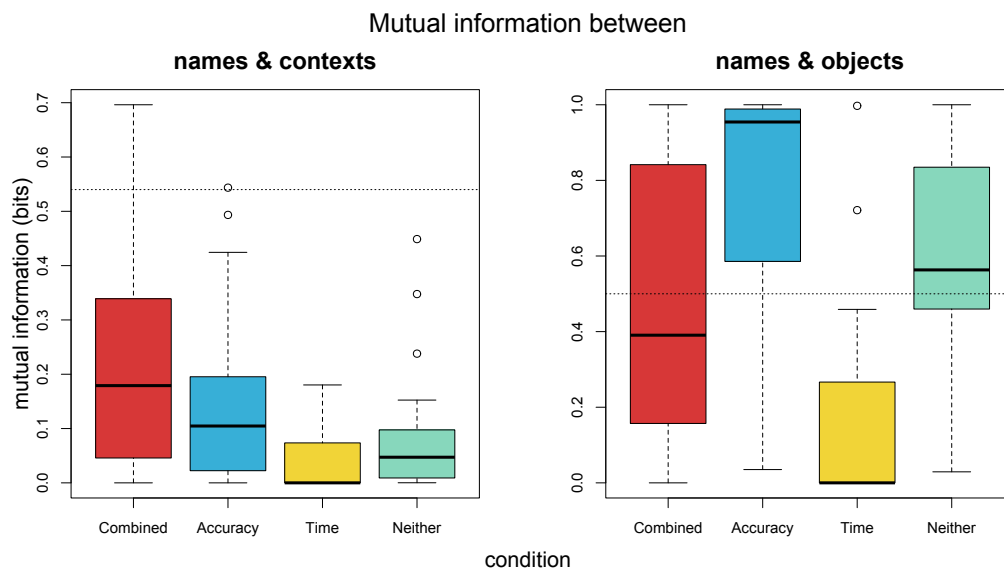


Figure 3.3: The extent to which individual participants’ name choices are conditioned on context (lefthand graph) and object (righthand graph) in Experiment 3. The dotted line in the lefthand graph represents the mutual information between name and context (MI_c) associated with the ‘optimal’ language in UID terms—the language in which the short form is used only in predictive contexts, and the long form only in surprising contexts. $MI_c=0$ for the input language. In the righthand graph, mutual information between name and object (MI_o) can range from 0 (same name fixed for both objects) to 1 (distinct names fixed for each object). $MI_o=0.5$ for the input language, marked by the dotted line. Data from only the second half of testing trials is shown in this figure, as participants were more likely to have converged on a stable mapping by this time.

(MI_o). This is given by

$$MI_o = - \sum_{o \in O} P(o) \log P(o) + \sum_{o \in O, n \in N} P(o, n) \log P(o|n) \quad (3.3)$$

where O is the set of objects {blue fruit, red stalk} and N is again the set of names {*zop*, *zopekil*, *zopudon*}. This measure allows us to determine whether some participants are using fixed names for each object, regardless of context, and so provides an additional check that the context-based optimisation behaviour observed in the Combined condition is not driven in any way by factors related to the object. The results are plotted in the righthand graph of Fig. 3.3. If participants are using a distinct name for each object, MI_o will be close to 1; if they are using the same name for both objects, MI_o will be close to 0. The former pattern is what we see in the Accuracy condition: most participants use the unique long name for each object, regardless of context. The latter pattern is what we see in the Time condition: most participants use the ambiguous short form for both objects, regardless of context. In the Combined and Neither conditions, MI_o hovers around that of the input language, indicating that, here, participants are not reliably conditioning their name choice on object. Looking back at the lefthand graph then clarifies that context conditions name choice to some extent in the Combined condition, but not the Neither condition.

3.2.6 Discussion

Experiment 3 shows that language users reliably produce shortened words in predictive contexts only when the competing pressures to communicate accurately and efficiently are both present in a communicative task. When one or both of these pressures are removed, participants fail to reliably condition their word choices on context. This provides clear evidence that language users actively modulate phonological word length to align with information content during communication. But how is this online behaviour linked to the large-scale lexicon-level effect observed by Piantadosi et al. (2011), in which shorter words tend to have a lower average surprisal than longer words?

One mechanism that has been suggested in the literature (e.g., Mahowald et al., 2013) is that as a word’s average surprisal evolves over time through shifts in its usage, its average length will also evolve accordingly. Words that become on the whole more predictable across their different contexts of use will become shorter, and conversely, words that become less predictable may be lengthened. One known mechanism of word shortening is clipping, discussed above.

Experiment 3 shows that, for words that already have both a full form and clipped form in active use, the Principle of Least Effort drives language users to produce the short form only in predictive contexts. We can then hypothesise that, 1) if the

proportion of predictive contexts increases (and thus the average surprisal of the word decreases), the proportion of short form use will correspondingly increase; and 2) as new generations of language users learn from these distributions, processes of regularisation will magnify these proportional asymmetries, eventually causing the long form to either drop out of the lexicon or shift its meaning, and the short form to then be considered an autonomous, unmarked, standard form in the lexicon (Jamet, 2009).

In the next section, we describe a follow-up experiment, that, combined with Experiment 3, provides a test of part 1 of this hypothesised causal story. Specifically, we created a new training language consisting of three meanings, each mapped to a unique long name as well as a shared ambiguous short name. The distributions of object-name-context pairings in this language are such that one of the meanings (and its corresponding long name) has a *lower average surprisal* than the other two. We investigate whether this meaning is consequently associated with a shorter average word length than the other two meanings.

3.3 Experiment 4

Experiment 3 comprised a between-subjects manipulation of the pressures to communicate accurately and efficiently, and investigated the effect of these pressures on the resulting lexicon. Experiment 4 implements the testing procedure of only the Combined condition of Experiment 3, in which both a pressure to communicate accurately and a pressure to communicate efficiently were present, and sets up a within-subjects/between-items manipulation of *average surprisal*. In the optimal language in Experiment 3, the short label is used for each object in its respective predictive context, but neither object is more strongly associated with the short label than the other. Thus, in the optimal language of Experiment 3, both meanings have the same average word length. In Experiment 4, we attempt to break this symmetry, and observe the consequences: there are three objects instead of two, and one of these objects has a lower average surprisal than the other two. Our prediction is that the object with the lowest average surprisal will also have the shortest average word length in the resulting lexicon.

Because there were now three rather than two objects to learn names for in the training phase, and each object required 32 instead of 16 testing trials (to ensure enough instances of each object-context pairing were present to carry out meaningful statistical analysis on the results), the length of the experiment nearly doubled, bringing the total task time to approximately one hour. As we were concerned that this longer duration would be problematic when running the experiment online, we moved it to the lab. Earlier research by the authors has shown that the current experimental

paradigm is robust to changes in the method of data collection: similar results are found whether the experiment is run online or in the lab (Chapter 2). Therefore, we do not expect the commensurability of Experiments 3 and 4 to be significantly affected by this difference in the data collection method.

For data collection in the lab, participants were seated in individual booths containing a computer, from which the experiment was run in a browser window. Participants were aware that their partner was sitting in a nearby but non-adjacent booth.

3.3.1 Participants


40 participants (28 females, 12 males) were recruited through the University of Edinburgh. Each was paid £7 for the approximately 1 hour long session. 19 participants reported themselves as native English speakers, of which 5 were monolingual. A range of other languages were represented across the remaining participants. Ages ranged from 18 to 32 (mean=22.55, SD=3.32).

3.3.2 The Training Language

As in Experiment 3, participants were trained on two names (a long name and short name) for each alien object by repeatedly being shown pictures of the objects labeled with a simple sentence, consisting of a framing word followed by the object’s name. This time, however, there were three different plant-like alien objects, and three possible framing words (i.e. contexts), *bix*, *gat*, and *lum*.

There were a total of 96 training trials, with each object appearing 32 times and each context (framing word) appearing 32 times. Crucially, the frequency with which each object occurred in each context was manipulated such that one object had a lower average surprisal than the other two. This was done by unevenly distributing the frequencies with which objects and framing words appeared together, in the manner shown in Figure 3.4. Object A appears almost always with the frame *bix*, and thus is, on average, highly predictable in context. Objects B and C are more spread out across the possible contexts, and thus are both, on average, less predictable in context than object A. Objects B and C, however, do still each have one context which is most predictive for them: *gat* for B and *lum* for C. In this way none of the objects differ in their overall frequencies of appearance, nor in the fact that they are each most predictable in one particular context. The critical difference is in their precise distributions across the three contexts, and thus in their average surprisal. The average surprisal (i.e. *information entropy*) of an object o is given by

$$-\sum_{c \in C} P(o|c) \log P(o|c) \tag{3.4}$$



	A	B	C
	<i>Bix zop.</i> x 14	<i>Bix zop.</i> x 1	<i>Bix zop.</i> x 1
	<i>Bix zopekil.</i> x 14	<i>Bix zopudon.</i> x 1	<i>Bix zopirax.</i> x 1
	<i>Gat zop.</i> x 1	<i>Gat zop.</i> x 10	<i>Gat zop.</i> x 5
	<i>Gat zopekil.</i> x 1	<i>Gat zopudon.</i> x 10	<i>Gat zopirax.</i> x 5
	<i>Lum zop.</i> x 1	<i>Lum zop.</i> x 5	<i>Lum zop.</i> x 10
	<i>Lum zopekil.</i> x 1	<i>Lum zopudon.</i> x 5	<i>Lum zopirax.</i> x 10
average surprisal:	0.67	1.20	1.20

Figure 3.4: A schematic diagram showing the frequencies of the object-sentence pairings presented during training trials in Experiment 4. The average surprisal of each object is given in the bottom row. Object A has a lower average surprisal than objects B and C, and thus we predict that, in the output lexicon, it will also have a lower average label length.

where C is the set of contexts $\{bix, gat, lum\}$ and $P(o|c)$ is the probability of object o occurring in context c .

As in Experiment 3, each object name appeared half the time in its full form, a unique 7-letter word, and half the time in its shortened form, a 3-letter word derived by clipping the last two syllables off the long name. This shortened form was the same for all three objects, for the reasons stated earlier: it is intended to model the fact that short words in the lexicon are more susceptible to noise and outright ambiguity than are long words.

These short and long names were evenly distributed across all contexts, ensuring that the input language contained no inbuilt bias towards using the short form in any particular context. Which object (red, yellow, or blue) appeared in which distribution among the three contexts (columns A, B, or C in Figure 3.4), as well as which object was paired with which long name, were all randomly assigned for each new pair of participants. This ensured that any potential influence from factors such as sound symbolism, or higher saliency or learnability of any specific object-word pairing, could

not systematically bias our results. A schematic diagram showing the frequencies of the object-sentence pairings in the training language, as well as the average surprisal for each object, are shown in Figure 3.4.

3.3.3 Training Procedure

The timing of training trials was identical to that of Experiment 3: on each training trial, an object was presented on screen alone for 700ms. The appropriate sentence then appeared beneath the object for a further 3000ms, yielding a total trial duration of 3700ms. A blank screen showed for 500ms between trials. The 96 training trials were presented in a different randomised order for each participant.

3.3.4 Testing Procedure

The testing procedures were identical to that of the Combined condition in Experiment 3, described in Section 3.2.4.1, except for the addition of a third object to the miniature language, and a different pay and bonus rate.

After training, participants moved on to the communication phase, where they took turns playing director and matcher, until both players completed 96 director trials and 96 matcher trials. The director transmitted sentences to the matcher exactly as in the Combined condition of Experiment 3, with a time cost to produce longer names. The matcher then chose which of the three objects they thought the director was referring to, followed by feedback to both players.

The proportion of times each object appeared with each frame in each player's director trials exactly matched those of the training proportions (shown in Figure 3.4), with one object thus having a lower average surprisal than the other two. The order of each participant's 96 director trials was randomly shuffled. Screenshots of sample director and matcher trials are shown in Figure 3.5b.

As in the Combined condition of Experiment 3, the total transmission time was displayed to the director during the process of entering in a name, next to the blank space, in order to emphasise the time pressure. As before, participants were also offered a bonus payment if they were the pair to complete the game in the quickest time with the highest number of correct match trials. The bonus payment was set to £5 per player.³

³Note that this is a higher bonus than that in Experiment 3. In Experiment 3 we were able to pay the bonus automatically through Amazon Mechanical Turk; in Experiment 4 participants had to physically come to the lab to collect it, so we felt a higher bonus was necessary to offset the cost of collection.

A



B

Choose a name for this object to complete the sentence.
Once the sentence is completed, it will be transmitted to your partner.

Gat _____.

zopudon zop

Press and hold inside the box until the full name appears
(and keep the mouse pressed until you advance to the next page).

Gat zo . total transmission time: 00:01:32

Your partner said:
Gat zop.

Which object do you think your partner is talking about?

CORRECT

Figure 3.5: (A) A screenshot of a sample training trial from Experiment 4. (B) A sample director trial (top) and a matcher trial followed by feedback (bottom) from Experiment 4.

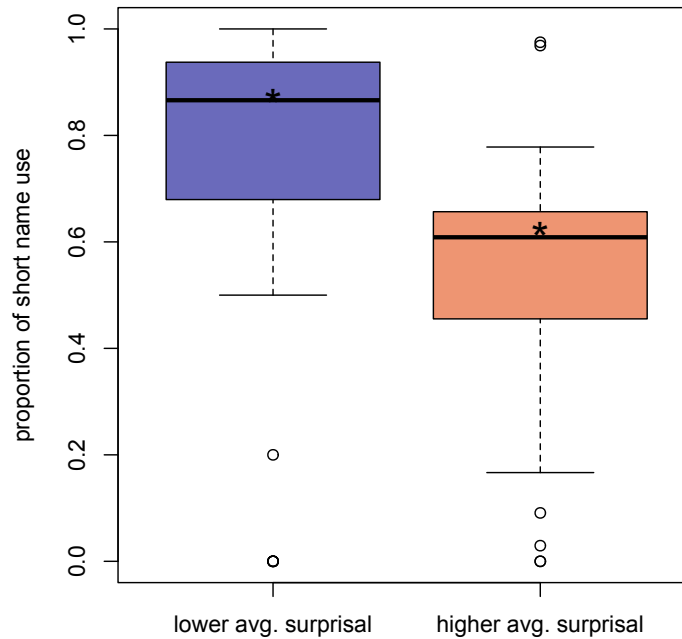


Figure 3.6: The proportion of times individual participants used the short name for the lower average surprisal object (object A) versus the higher average surprisal objects (objects B and C), during the second half of testing trials in Experiment 4. The proportion of trials in which each object appeared in its most predictive context is marked with an asterisk. As the asterisks fall close to the medians of each distribution, this suggests that most participants adopted the strategy of using the short name in only the most predictive context for each object, and using the long name elsewhere.

3.3.5 Results

Figure 3.6 shows the proportion of trials in which participants used the short name to refer to the low average surprisal object (object A) versus the higher average surprisal objects (objects B and C). Our main prediction was that the lower surprisal object would have a shorter average length in participants' output lexicons than the higher surprisal objects. As can be seen in the figure, the lower surprisal object corresponds to a higher proportion of short name use, and thus a shorter average length, than the higher surprisal objects.

A binomial regression model on the full dataset shows this difference to be significant. Short name use was set as the binary dependent variable, with object (lower or higher average surprisal) as a fixed effect. By-participant random slopes and intercepts for object were included. With the intercept set at the mean for the higher

average surprisal objects, a significant positive effect for lower average surprisal was found ($\beta = 0.767$, $SE = 0.214$, $p < 0.001$), indicating that participants were significantly more likely to use the short name for this object. In this overarching measure, then, our predictions were borne out.

However, to understand how these name choices are distributed across the different contexts, we need to take a more fine-grained look at the data. Recall that object A has the lowest average surprisal because it appears in its most predictive context (i.e. with framing word *bix*) in the highest proportion of cases (the probability is 0.875). Objects B and C also have one context in which they are the most predictable object (namely with framing word *gat* and framing word *lum*, respectively), but they appear less frequently in each of these contexts than Object A appears with *bix* (specifically, with probability 0.625). Given this, one possible strategy a participant might adopt would be to use the short name in the most predictive context for each object, relying on the fact that the framing word would then provide enough information for the matcher to guess the correct object. When the object appears with a framing word that it is less usually seen with, the long name would be used to provide disambiguating information about the object in question.

If participants used this strategy, then the proportion of times the short name was used for each object would be equivalent to the proportion of trials in which the object appeared in its most predictive context. This proportion—0.875 for the lower surprisal object and 0.625 for the higher surprisal objects—is marked with an asterisk in each column of Figure 3.6. As can be seen, the median proportions of short name use for each object are indeed close to the proportions they appear in their most predictive contexts. This suggests that participants may be adopting the strategy of using the short name in only the most predictive context for each object, and using the long name elsewhere.

A more fine-grained examination of the data in fact reveals a somewhat richer pattern than this. Figure 3.7 shows the proportion of times participants used the short name for an object, depending on the probability of that object appearing in the given context. We see that for both types of most predictive context—those in which the object appears with probability 0.875, which applies only to object A, and those in which the object appears with probability 0.625, which applies only to objects B and C—participants use the short name most of the time. For somewhat surprising contexts (which only objects B and C occur in) and very surprising contexts (which all three objects occur in), participants rarely use the short name. However, there is a difference between how often participants use the short name in the most predictive context for object A (the lower surprisal object), and how often they use the short name in the most predictive contexts for objects B and C (the higher surprisal objects). We ran a binomial regression model on the full dataset with short name

Table 3.2: Summary of fixed effects for a binomial regression model with short name use as the binary dependent variable. “Prob” signifies “probability of the target object occurring in the given context”. Significant and trending effects are in bold.

Fixed Effect	β	SE	p
Intercept (prob=0.0625)	-0.313	0.201	0.119
prob=0.3125	0.132	0.290	0.648
prob=0.625	3.181	0.472	<0.001
prob=0.875	0.558	0.311	0.072

use as the binary dependent variable and context type (i.e. how probable the object is in the given context) as the predictor variable (Table 3.2). By-participant random intercepts and slopes for context type were included. The model was backwards difference coded, meaning that the mean of each factor is compared only to the mean of the previous factor. We find that short name use in the somewhat surprising contexts (prob=0.3125) is not significantly different from that in the most surprising contexts (prob=0.0625). Short name use in the most predictive contexts for objects B and C (prob=0.625) is significantly greater than that in the somewhat surprising contexts. Finally, short name use in the most predictive context for object A (prob=0.875) is notably trending above that in the most predictive contexts for objects B and C.⁴

Why might this effect come about? Given that the most predictive context for object A appears in a higher proportion of object A’s trials than the most predictive contexts for objects B and C appear in their respective trials, it may be clearer to participants which of the contexts is the most predictive one for object A; there is a sharp difference between how often we see object A with *bix* than with *gat* or *lum*. For objects B and C, however, there is a shallower gradient, spread out over a predictive context, a somewhat surprising context, and an even more surprising context. Given this, it may be easier to confuse the predictive and surprising contexts for objects B and C, or it may take longer to learn which ones are which. This increased uncertainty may cause participants to choose the longer name for these objects in more trials, regardless of context, resulting in the greater average word length associated with these higher average surprisal objects.

To view this in another way, we plot in Figure 3.8 the number of participants who are perfectly optimising in accordance with the UID principle (i.e., using the short name *only* in the most predictive context of an object, and the long name elsewhere)

⁴Running the model on only the second half of testing trials, where participants are more likely to have converged on stable mappings, brings this effect to significance at $p < 0.001$. However, this model fails to adequately converge, most likely due to data sparsity in the lower frequency contexts. Simplifying the random effects structure (i.e. including intercepts but not slopes) or using a coding scheme other than backwards difference coding results in a converging model, yielding a significant difference between the prob=0.875 and prob=0.625 contexts. Therefore, we can be fairly confident about the trend reported in Table 3.2.

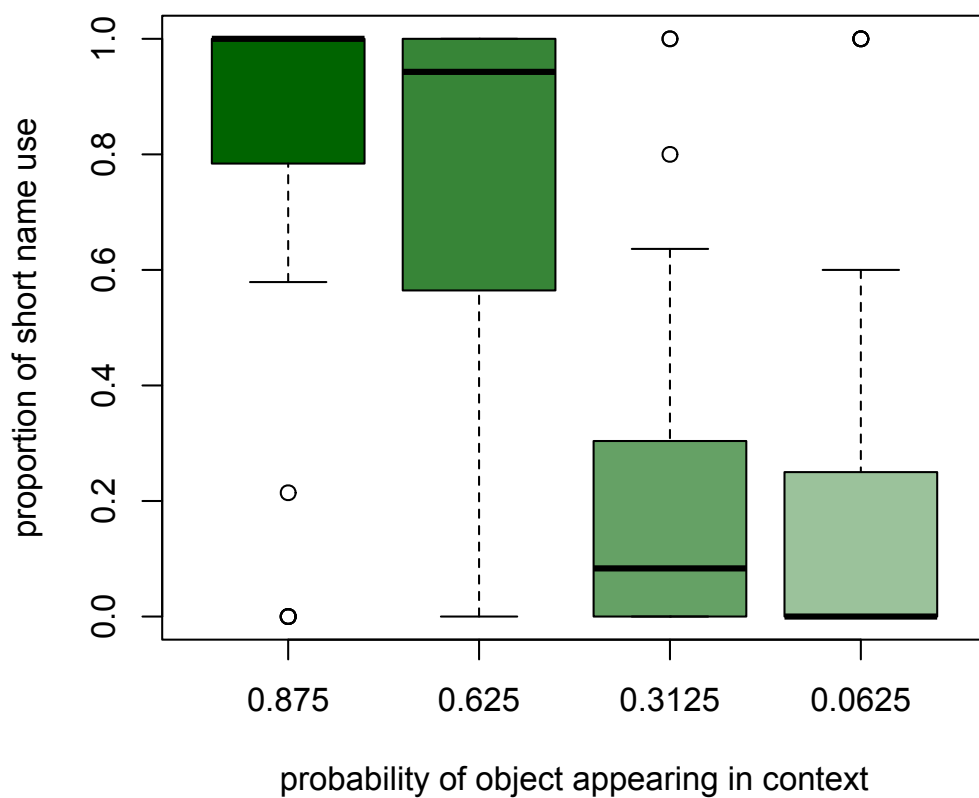


Figure 3.7: The proportion of times individual participants used the short name in different contexts, during the second half of testing trials in Experiment 4. Contexts are labeled here by the probability of the given object appearing in that context.

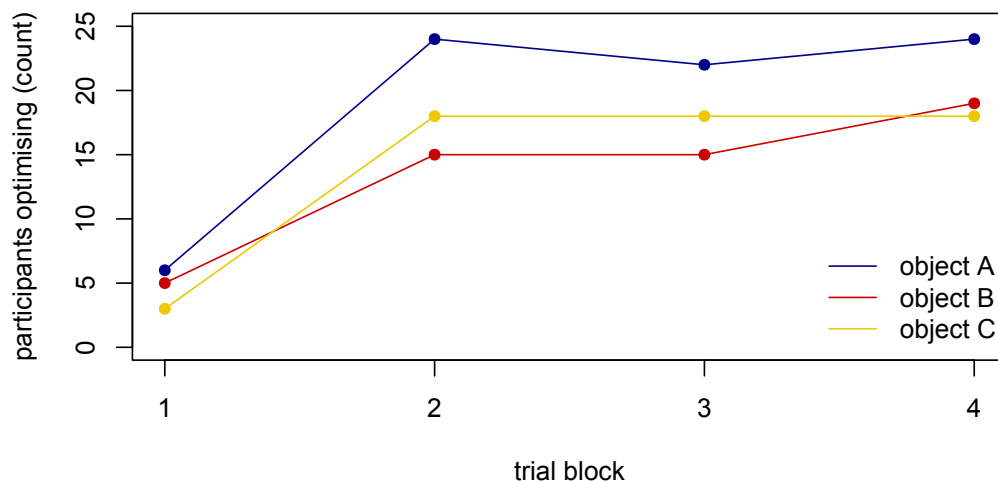


Figure 3.8: The number of participants who are perfectly optimising their form-meaning mappings in accordance with the uniform information density principle (i.e., using the short name only in the most predictive context, and the long name elsewhere) for each object, as a function of trial block. Each block consists of 24 director trials.

for each object, as a function of time. What we see is that, by the end of the second block of testing trials, more participants are perfectly adhering to this strategy for object A than for objects B and C, and this difference persists thereafter.

3.3.6 Discussion

In the critical Combined condition of Experiment 3, we found that participants tended to use the ambiguous short name for objects when they appeared in predictive contexts, and the unique long name when they appeared in surprising contexts. This strategy, which assigns word length proportionally to information content, in line with the uniform information density principle, was also observed in Experiment 4. The average surprisal differential introduced in this experiment between one object and the other two also corresponded to an observed difference in the *extent* to which this strategy was applied to each object. Namely, the object with a lower average surprisal also had a lower average length in participants' output lexicons, in that it was referred to using the short name in a higher proportion of trials.

Digging deeper into the details showed that this difference was due to a combination of several factors. First, object A appears in predictive contexts a higher proportion of the time than objects B and C. For participants who perfectly adhere to the strategy of using the short name only in the most predictive context for each

object, this fact alone would explain why object A ends up with a higher proportion of short name use. However, a more fine-grained analysis showed that while most participants optimally condition word length on context for object A, they do this less so for objects B and C. This may be because it is more difficult to learn which are the predictive and which are the surprising contexts for these objects. This increased uncertainty may cause participants to choose the longer name in more trials, regardless of context, resulting in the greater average word length associated with these higher average surprisal objects.

This overall finding, that meanings with a lower average surprisal also correspond to a lower average word length, accords with the large-scale lexicon-level findings of Piantadosi et al. (2011). In this experiment, we find that language users assign word length proportionally to *average* information content, resulting in a more uniform information density across the lexicon as a whole. Our results also lend plausibility to the diachronic hypothesis suggested by Mahowald et al. (2013) for clipped pairs: as a meaning’s average surprisal decreases over time, its shortened form should, accordingly, become more prevalent.

3.4 General discussion and conclusions

There is mounting evidence that utterance length is linked to information content (Lieberman, 1963; Aylett and Turk, 2004; Gahl and Garnsey, 2004; Tily et al., 2009; Bell et al., 2009; Jaeger, 2010; Piantadosi et al., 2011; Kuperman and Bresnan, 2012; Fedzechkina et al., 2012; Seyfarth, 2014). The explanation put forth in much of this previous work is that speakers are driven by pressures like those outlined in Zipf’s Principle of Least Effort: the competing demands for accurate and efficient communication lead speakers to converge on an optimal system in which information content is spread roughly uniformly across the utterance, resulting in low-information words being shorter than high-information words. This effect appears to have made its way into the structure of the lexicon as a whole: shorter words occur on average in more predictive contexts than longer words (Piantadosi et al., 2011). Previous work thus reveals a relationship between information content and word length in the synchronic grammars of natural languages. However, these studies provide only indirect evidence for the mechanisms underlying the causal link between information (e.g., surprisal) and changes in the lexicon over time.

Here we use artificial language learning to explicitly manipulate the causal factors hypothesized to underlie this cross-linguistic pattern. In Experiment 3, by observing participants’ online behaviour in tasks in which the pressures to communicate accurately and efficiently were manipulated across four experimental conditions, we show that participants use shorter words in more predictive contexts *only* when both

competing pressures were at play. When one or both of these pressures were removed, participants failed to reliably condition their word choices on context. This echoes our previous finding regarding the conditions under which participants map shorter word length to more *frequent* objects (Kanwal et al., 2017b). In Experiment 4, by manipulating the average surprisal across three objects during a communicative task, we show that meanings with a lower average surprisal correspond to a shorter average word length.

Our results serve as a proof of concept that the large-scale lexicon-level effect observed by Piantadosi et al. (2011) could be driven in part by a least effort principle, in which language users balance the competing pressures for communicative accuracy and efficiency to reshape the lexicon into one where word length is roughly proportional to average information content, as defined by usage. However, there is a crucial step between what we have observed here—language users alternating between long and short variants for a single meaning depending on context—and what Piantadosi et al. (2011) observed in the lexicon of different languages, where most meanings don’t correspond to both a long and a clipped variant, but rather map to a single fixed form. For these cases, which make up the majority of the lexicon, the length of the form is strongly correlated with the *average* predictability-in-context of the meaning, across all its different occurrences.

We can, however, hypothesise a causal link between these two phenomena. One common mechanism of adjusting word length is through truncation, or ‘clipping’. In this process, part of a word is truncated to give rise to a shortened, or ‘clipped’, form. As a word appears in increasingly more predictive contexts, this may instigate the coining of a clipped variant. Often, both the full and the clipped form remain simultaneously in active usage for a time. In many cases, however, the full form eventually falls out of usage or shifts its meaning, leaving the shortened form to then be considered an autonomous, unmarked, standard form in the lexicon (Jamet, 2009). This leaves us with a lexicon in which words with a lower average surprisal tend to be shorter, as observed by Piantadosi et al. (2011).

Our experiments do not speak to the conditions that give rise to clipping initially, as we already present participants with both a full and a clipped form for each meaning during training. However, they do provide insight into the conditions that contribute to the clipped form overtaking the full form. Experiment 3 shows that, for words that already have both a full form and clipped form in simultaneous active usage, the Principle of Least Effort drives language users to produce the short form in predictive contexts, and the long form in surprising contexts. If language users tend to produce the short form only in predictive contexts, then it follows that, as the relative proportion of predictive contexts increases, the relative proportion of the short form will also increase. This was confirmed in Experiment 4, where lower av-

erage surprisal corresponded with a relative increase in short form use, and thus a lower average word length.

As the average surprisal of a meaning continues to decrease, the instances of long form use will become rarer and rarer. At this stage, we hypothesise that factors related to learning and memory will play an important role in amplifying these asymmetries in the direction of ultimately eliminating the long form. As new generations of language users learn from exposure to these highly skewed input distributions, processes of regularisation will take over. Indeed, previous work on regularisation (Hudson Kam and Newport, 2005) and iterated learning (Real and Griffiths, 2009; Smith and Wonnacott, 2010) shows that, in the output lexicons produced by participants, those in which two labels map to the same meaning are dispreferred.

As our experimental paradigm is designed to eliminate the load of memory and learning pressures on participants, and instead focus on the interaction of communicative pressures with information content, we cannot here provide direct support for this final stage in our causal hypothesis. For this, a significantly amended experimental design would be necessary, including production of forms from memory (rather than a 2AFC task), a larger lexicon and corresponding meaning space, and possibly iteration over multiple generations of language users. We therefore leave this endeavor for future work.

In conclusion, our results contribute to a growing body of literature that shows how language is shaped through usage by the influence of communicative pressures. The experiments presented here allowed us to test the plausibility of a proposed causal hypothesis—that active speaker choice, subject to pressures to communicate both accurately and efficiently, can lead to reshaping of the lexicon to align with the principles of uniform information density and smooth signal redundancy.

Chapter 4

The diachronic evolution of long/short word pairs

4.1 Introduction

The experiments described in the previous two chapters showed that meanings that are more frequent or predictable in context are more likely to be mapped to shorter forms, when both a long name and a clipped name are available. This supports the hypothesis that a pressure for communicative efficiency operates on language-users, influencing their word choice during communication. This hypothesis is bolstered by the fact that, in synchronic corpora across a wide range of languages, an inverse relationship is found between word frequency and word length (Teahan et al., 2000; Sigurd et al., 2004; Strauss et al., 2007; Ferrer-i-Cancho and Hernández-Fernández, 2013). An even stronger inverse relationship is found between a word’s predictability in context and its length (Piantadosi et al., 2011).

In this chapter, I will focus on frequency alone, as this quantity is more straightforwardly operationalised and measured in pre-existing linguistic datasets. The result in Chapter 2—that more frequent meanings are more likely to be mapped to shorter forms during communication—naturally gives rise to a diachronic hypothesis: as a single meaning itself becomes more frequent over time, it will be more likely to be mapped to a shorter form over time. This diachronic process would serve as a link between the online behaviour observed in the communication games, and the synchronic patterns found at the level of the lexicon.

Here I target clipped pairs—pairs of words consisting of a long form and a short form derived by truncating the long form. These pairs are directly analogous to the long/short pairs used in the experiments presented in Chapters 2 and 3. Crucially, both forms in a clipped pair either refer to the same meaning, or are extremely close in meaning. Consider, for example, the English clipped pair *info/information*.

One can approximate the overall frequency of the meaning in a corpus by summing the frequencies of the long form and the short form. One can also track the *relative* frequency of the two forms, and thereby obtain a measure of the “average word length” ascribed to the meaning over time (see, e.g., Mahowald et al., 2013). The hypothesis is as follows: *as the overall frequency of a meaning increases over time, its associated short form will increase in frequency relative to its long form.* This relative increase in short form use equates to a decrease in the “average word length” ascribed to the meaning.

Conversely, one might expect the frequency of the short form to *decrease* relative to the long form in cases where the overall frequency of the clipped pair is decreasing. However, it is also possible that there is no such analogous pressure to lengthen less frequent meanings. Instead, clipped pairs whose frequency decreases over time may simply maintain a level ratio of short:long form frequency. Another alternative possibility is that *all* words are subject to the pressure to be shortened, but this pressure is strengthened for words whose frequency is trending upwards. In this case, the short form in a clipped pair would tend to increase in frequency relative to the long form *regardless* of whether the meaning is becoming more or less frequent over time. However, for those pairs where the meaning *is* becoming more frequent over time, the short form might gain on the long form *at a faster rate.*

As a short form increases in frequency relative to its corresponding long form, conditions are created for regularisation processes to take hold. In the lexicons produced by speakers, those in which two forms map to the same meaning are dispreferred (e.g., Reali and Griffiths, 2009; Smith and Wonnacott, 2010). If one of these forms is significantly more frequent than the other in the lexicon that a speaker is exposed to, then the less frequent form is likely to be dropped entirely from their subsequent productions—an instance of regularisation (Hudson Kam and Newport, 2005). In some cases, after enough generations have passed, the long form may even pass out of lexical knowledge. The clipped form would then become the sole label associated with a given meaning, existing autonomously in the lexicon as an unmarked, standard form. When the short form is no longer widely recognised as a derivation of the long form, we say it has undergone “opacification” (Jamet, 2009). An example is the word *bus*: this was originally a clipped version of the word *omnibus*, which is now obsolete.

When such processes occur, the resulting lexicon is one in which more frequent meanings tend to be mapped to shorter forms. Less frequent meanings are less likely to undergo this entire process—truncation, subsequent increase in clipped form usage, regularisation, and finally opacification—and are therefore more likely to remain mapped to longer forms in the lexicon. This hypothesis thus provides one plausible mechanism by which languages could gradually align or maintain alignment with the Law of Abbreviation, giving rise to the large-scale, cross-linguistic patterns observed

in synchronic corpora.

The outlines of such a hypothesis were suggested by Zipf himself:

The accumulated effects of abbreviatory acts of truncation during the long periods of years in which the language has slowly evolved are probably responsible for the shortness of many of the frequently occurring words in speech today (Zipf, 1935; p. 33).

He cites such examples as *movies* from *moving pictures*, *bus* from *omnibus*, and *gas* from *gasoline* (Zipf, 1935, 1949). These shortened forms, he argues, arose from the rapid increase in use of these concepts in our language, which itself was due to their rapidly growing presence in our daily lives. He then contrasts these with words such as *constitutionality*, *quintessentially* and *idiosyncrasy*, which, he claims, are never shortened because they are not frequently used. Zipf concludes this discussion by proposing a general rule governing lexical change: “as the relative frequency of a word increases, it tends to diminish in magnitude” (Zipf, 1935; p. 38).

More recently, Joan Bybee has provided indirect support for this hypothesis through a body of work linking processes of reduction to higher frequency of use. In a series of case studies—word-final t/d deletion in high-frequency English words like *just*, *perfect*, *child*, and *grand*); historical [ð] deletion in Spanish; and different length realisations of the contraction *don't*—she finds that higher frequency constructions are disproportionately targeted for reduction (Bybee, 2006). Bybee essentially reiterates the causal hypotheses outlined above, to explain how these changes become cemented in the lexicon. The quote below is in reference to a current sound change in progress in English: the deletion of unstressed schwa.

After the initial stages, the language acquisition process begins to play a role, for it is probably in the transmission of the language that restructuring takes place. A very frequent word such as *every* may be variable for an adult speaker, but a new learner may take the schwa-less pronunciation to be the norm or base form. The younger speaker will have a schwa-less underlying form, and the process will be complete for that word. (Bybee, 2006; p. 31)

This causal hypothesis relies on the assumption that increasing frequency will correspond to an increasing proportion of the use of shorter variants over time. However, to my knowledge there is currently no direct quantitative evidence for this process occurring *diachronically*. In Bybee’s study of unstressed schwa deletion, she provides a snapshot in time of English words containing unstressed schwa: the frequency count for a given word in a synchronic corpus is compared to its incidence of schwa-deletion (deleted *usually*, *sometimes*, or *rarely*, as judged by eight native speakers). She finds that the words that are usually produced without the schwa also have a higher average

frequency. However, we do not know whether, as the frequency of these words increases over time, their incidence of schwa deletion will correspondingly increase. Nor do we know what might happen if their frequency begins to subsequently decrease.

A recent study by Pate (2017) gets closer to providing a temporal picture. If a word's length exactly corresponds to the amount of information contained in it—no more, no less—then it should be given by the Shannon entropy calculated over the word's probability of occurrence. Pate calculated this optimal word length for words found in synchronic corpora of English, Spanish, and Mandarin, using each word's unigram probability (i.e., its frequency in the corpus). For each word, he took the distance between its actual length and this optimal length. He then obtained an approximate measure of each word's age, by finding the first year in which it appeared in the Google Books Ngrams corpus (which I will describe in more detail below). The result was that older words tended to have fewer extra letters above their optimal length than newer words—the lexicon appears to be becoming more efficient over time. Similar to the synchronic findings of Piantadosi et al. (2011), this effect is even stronger when the trigram probability of a word (its probability given the context of the two preceding words) is used to calculate its optimal length, rather than its unigram probability. This 'backward-looking' corpus study is complemented by a 'forward-looking' study, in which Pate shows that words with more extra letters above their optimal length are also more likely to disappear from the lexicon in subsequent years. These backward- and forward-looking views suggest that a diachronic optimisation process is indeed occurring. However, a direct causal link between a word's changing frequency and its length has yet to be demonstrated using diachronic data.

One reason for the lack of direct support for this diachronic hypothesis is the relative dearth of suitable corpora available. To compare frequency counts across time spans, there must be a sufficient amount of data at each point in time to allow reliable frequency counts at that time slice. In addition, the corpus must have sufficient time depth to allow for reliable analysis of diachronic trends. Thus, for a corpus to be able to provide diachronic support for the hypothesis that increasing meaning frequency leads to increasing short form use over time, it must be very large. Until recently, no adequately sized corpora were available for this purpose. However, with the creation of the Google Books Ngrams corpus (Michel et al., 2011), an opportunity for reliably investigating diachronic trends in word frequency at a large scale became newly available.

In this chapter I conduct a large-scale diachronic test of the causal hypothesis for long/short word pairs—that as a meaning becomes more frequent over time, its short form will become more frequent relative to its long form. The results fail to provide decisive quantitative support for this hypothesis. Nevertheless, they do suggest some intriguing trends that are consistent with the hypothesis that increasing frequency

may boost the rate of short form use.

4.2 The Google Books Ngrams corpus

The Google Books Ngrams corpus is based on the texts of approximately 5 million digitized books ($\sim 4\%$ of all books ever published). It contains 500 billion tokens, across seven languages: English, French, Spanish, German, Chinese, Russian, and Hebrew (Michel et al., 2011). The English portion is the largest, with over 300 billion words. The texts themselves are not included whole in the corpus; rather, the corpus consists of frequency counts per year of individual word tokens, as well as of bigrams, trigrams, 4-grams, and 5-grams. Here, I will focus on just the unigram frequency counts.

In terms of its diachronic coverage, the oldest books in the corpus date from the 1500s, and the most recent are from 2008. While the size and breadth of this corpus is impressive, one important point to note is that the data is not evenly spread out in time—the early decades contain only a few books per year, and then the density of tokens gradually increases to approximately 100 million words per year by 1800, 1.8 billion by 1900, and then sharply rising to 11 billion by 2000 (see Figure 4.1).

The method of book digitization—first sheet-scanning or stereo-scanning, then optical character recognition (OCR)—is such that Google estimates that over 98% of words in modern English books are accurately identified. Another important potential source of error, however, is in the date ascribed to each word. These were taken from the year-of-publication metadata for each book, but this automatically-retrieved metadata is susceptible to error. A filtering process was thus applied to eliminate texts which were more likely to be mis-dated, such as serials and periodicals. Even after this filtering process, 5.8% of books from a random sample of 1000 were found to have metadata dates that were more than 5 years from the actual date of publication, as determined by a human checker (Michel et al., 2011; Supplementary Online Material). The problems of OCR quality and metadata accuracy are more pervasive for the non-English portions of the corpus, to which more careful filtering and checking processes have not yet been applied. These issues are important to keep a note of as I discuss the results of the analyses in the following sections.

I begin my investigation with a detailed case study of one clipped pair (*info/information*), and then proceed to large-scale analyses of three different datasets: English clipped pairs, English *-ic/-ical* pairs (e.g. *electric/electrical*), and French clipped pairs.

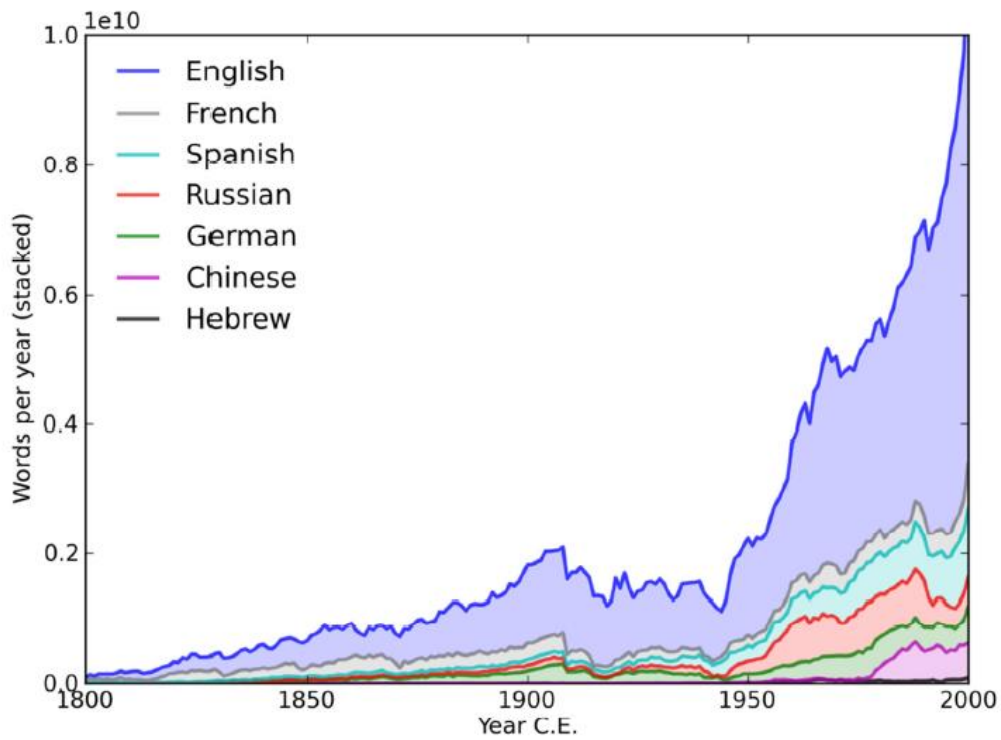


Figure 4.1: Number of words per year in the Google Books Ngrams corpus. The density of tokens is not evenly spread in time, but increases year by year, rising especially rapidly after 1950. Adapted from Michel et al. (2011), Supplementary Online Material p. 62. Copyright 2011 by AAAS. Reprinted with permission.

4.3 Info/information: a case study

4.3.1 Data

From the English portion of the Google Books Ngrams corpus, I extracted the frequency counts of the words *info* and *information*, in all years where both words occurred at least 40 times each. The search was case-insensitive, including all differently-cased variations of a word. This threshold of 40 occurrences per year was chosen to echo the threshold used by Google of 40 occurrences over *all* years to determine whether to include the data for a word in the final corpus. Setting a threshold for each year reduces the likelihood that a word's occurrences in a given year are not merely due to either a digitization error or a mis-dating in the metadata.

Due to the uneven size of the corpus across time (see Figure 4.1), I normalised the frequency count for each word in each year by the total frequency of all tokens in that year. Thus, I obtained the proportion of times *info* and *information* occurred in each year, respectively. I call these normalised frequencies F_s (frequency of short form) and F_l (frequency of long form).

I then summed these two values to obtain the (normalised) total meaning frequency (MF) in each year.

$$MF = F_s + F_l \quad (4.1)$$

Next, I calculated the *short form advantage* (SFA) per year. This was done by taking the difference of the short form frequency and the long form frequency in a given year, then dividing it by the total meaning frequency in that year. This provides a measure of how frequent the short form is relative to the long form, as a proportion of the total meaning frequency. Dividing the difference by the total meaning frequency gives us a more generalisable quantity that we can later use to compare between different word pairs across different years.

$$SFA = \frac{F_s - F_l}{MF} \quad (4.2)$$

If the short form is more frequent than the long form in a given year, the SFA will be positive in that year, and if the short form is less frequent than the long form, the SFA will be negative. The slope of the curve representing the SFA over time signifies whether the short form is becoming more or less frequent relative to the long form, over and above any change in difference merely due to a change in the overall meaning frequency itself.

Finally, I discarded data for all years before 1900, due to increasingly sparse data in the corpus further back in time, and hence potentially unrepresentative frequency counts.

Figure 4.2 shows the resulting time series of the meaning frequency MF and the

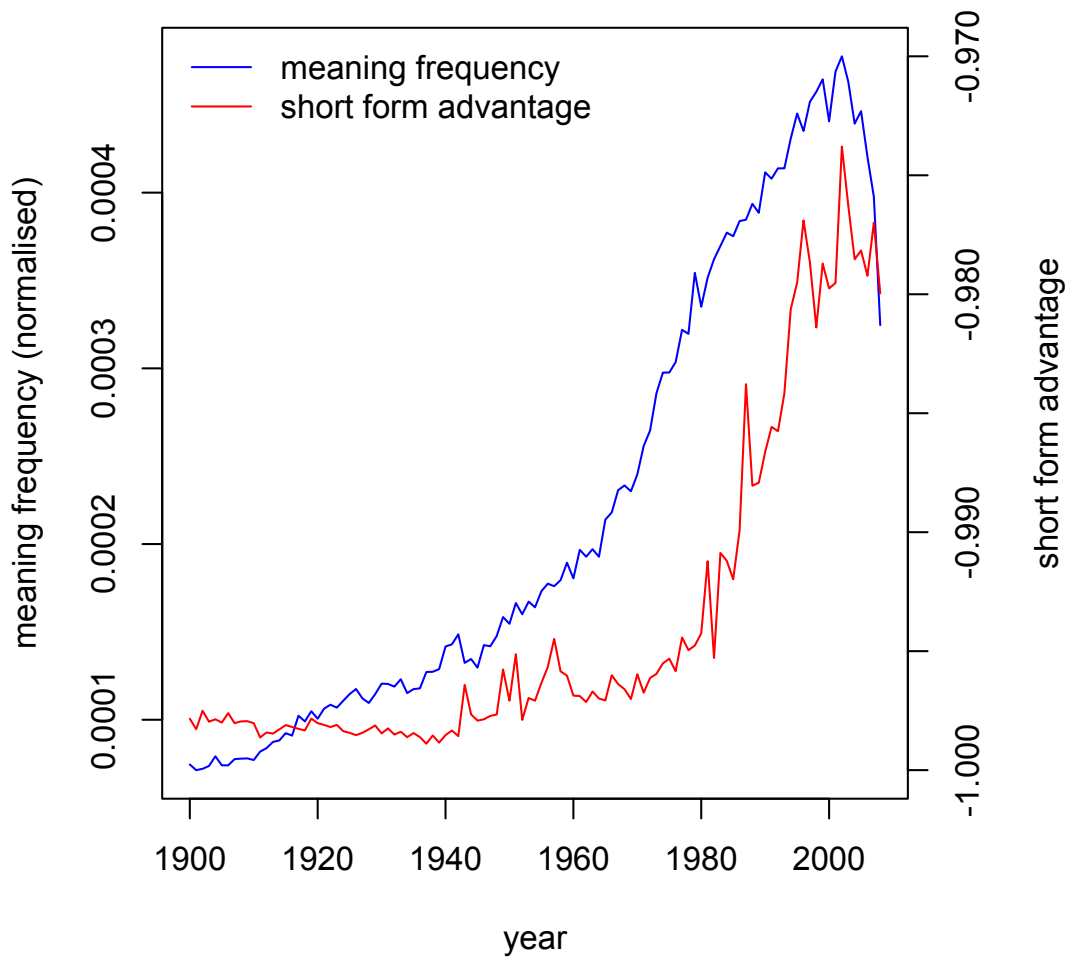


Figure 4.2: Meaning frequency (MF , blue) and short form advantage (SFA , red) for the the *info/information* clipped pair over time. The shape of the short form advantage time series appears to closely follow that of the meaning frequency. Note that the two time series correspond to different y-axes.

short form advantage *SFA* for the *info/information* clipped pair. Through visual inspection it is clear that the shape of the short form advantage time series closely follows that of the meaning frequency. As the meaning frequency increases, the *SFA* also begins to increase, after some time delay and, at first, less steeply. When the meaning frequency begins to drop around the year 2000, the *SFA* appears to begin dropping too or at least leveling out. This result is precisely what is predicted by the diachronic causal hypothesis: as the frequency of the meaning increases over time, the relative frequency of *info* to *information* appears to increase accordingly, possibly with some time lag. Note, however, that the long form, *information*, is always more frequent than the short form, *info*, evidenced by the negative values of the *SFA*. These values, however, become less negative during the period roughly between 1950 and 2000, when the frequency of the meaning is also increasing.

4.3.2 Linear regression analysis

Although visual inspection alone reveals a striking result in the case of this particular clipped pair, it is important to evaluate these findings with statistical tests. These same tests will be used when we expand our analysis to a larger set of word pairs.

The first, somewhat crude, test is simply to check whether the meaning frequency of *info/information* shows an *average* increase or decrease over the time span of our corpus data, and then whether the *SFA* shows an average trend in the same direction. This can be done by first fitting a generalised linear model, with meaning frequency as the predicted variable and year as the predictor variable. This model yields a significant positive effect of year on the meaning frequency ($\beta = 3.943e - 06, SE = 1.265e - 07, p < 0.001$), as expected based on visual inspection of Figure 4.2.¹ Next, I fit a linear model to the *SFA* data, again with year as the predictor variable. This yields a significant positive effect of year on the *SFA* ($\beta = 1.750e - 04, SE = 1.273e - 05, p < 0.001$), again as expected based on Figure 4.2. Thus, the tests confirm that the meaning frequency and the short form advantage are both increasing with time, on average. While this does not capture whether the dips and rises in one curve follow those of the other, it provides a first rough measure of whether the curves are moving in the same direction, on average. This heuristic test will be useful when dealing with a much larger set of words, to assess whether the meaning frequency and *SFA* are correlated in general.

4.3.3 Granger causality analysis

While the linear models provide a rough assessment of the alignment between the *average* tendencies of the two time series, a more precise test would check whether

¹All *p*-values reported in this chapter for generalised linear models are obtained from χ^2 model comparison tests.

changes in the meaning frequency *predict* subsequent changes in the *SFA*. A statistical test which is ideally-suited to address this question is the Granger test for causality (Granger, 1969).

The Granger causality test was originally developed for application to econometric time series, but it has been applied successfully in the context of diachronic linguistics in at least one instance: to show that, historically, changes in morphology tend to trigger changes in syntax in Icelandic (Moscoso del Prado, 2014). Assuming there are two stationary time series, $X(t)$ and $Y(t)$, and one wants to test whether the behaviour of $Y(t)$ is predicted of the behaviour of $X(t)$, the Granger test proceeds by first fitting an autoregressive model to $Y(t)$. An autoregressive model is one in which the current value of $Y(t)$ is predicted by earlier values of $Y(t)$, up to some lag value m .

$$\begin{aligned} Y(t) &= a_0 + a_1Y(t-1) + a_2Y(t-2) + \dots + a_mY(t-m) + \epsilon(t) \\ &= a_0 + \sum_{i=1}^m a_iY(t-i) + \epsilon(t) \end{aligned} \tag{4.3}$$

Then, a second model of $Y(t)$ is fitted, in which lagged values of the time series $X(t)$ are also included.

$$Y(t) = a_0 + \sum_{i=1}^m a_iY(t-i) + \sum_{i=1}^m b_iX(t-i) + \epsilon(t) \tag{4.4}$$

The $\epsilon(t)$ terms denote (uncorrelated) Gaussian white noise series. If the second model is found by an F-test to be a significant improvement on the first model, then one can conclude that $X(t)$ Granger-causes $Y(t)$ —in other words, $Y(t)$ is better predicted by taking into account the past values of $Y(t)$ *and* $X(t)$, than by just taking into account the past values of $Y(t)$ alone. The null hypothesis in this case is that the second model is not an improvement on the first, i.e. that adding the past values of $X(t)$ does not add any additional predictive power to the model of $Y(t)$.

It is also important to run a Granger test in the opposite direction—testing whether lagged values of $Y(t)$ predict values of $X(t)$ —to check that the causal relationship exists only in the hypothesised direction, and not in both directions, which would imply some sort of feedback relationship.

4.3.3.1 Transforming the time series to achieve stationarity

In the present case, the two time series $X(t)$ and $Y(t)$ are the meaning frequency of the *info/information* clipped pair, and the short form advantage, respectively. Since time

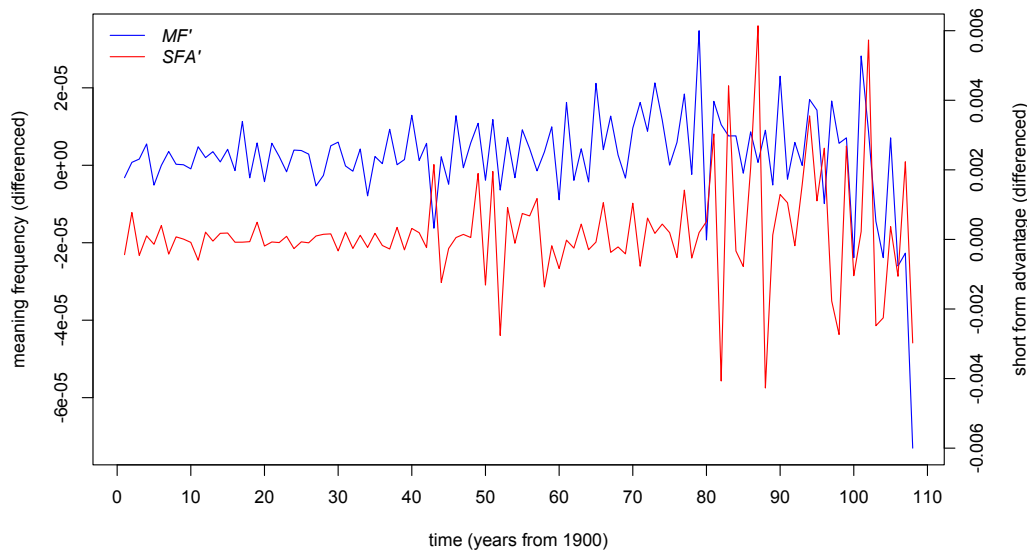


Figure 4.3: The differenced time series for meaning frequency (MF' , blue) and short form advantage (SFA' , red) for the the *info/information* clipped pair. Note that the two time series correspond to different y-axes.

series must be stationary (i.e. have a roughly constant mean and variance over time) in order to apply a Granger causality test, the first step is to transform these two time series to render them approximately stationary. This can be done by *differencing* the time series, i.e. taking the year-to-year differences between each data point, yielding a new time series with one fewer data point. This transformation can be applied any number of times to successive differenced time series, until stationarity is achieved.

To determine the number of differences required for the time series to achieve stationarity, I applied a unit root test, using the `ndiffs` function in the `forecast` package (Hyndman and Khandakar, 2008) in R (R Core Team, 2015). The function returns the least number of differences required to pass the unit root test at the level $\alpha = 0.05$. Applying this yielded a value of one for both time series, meaning that one difference would be enough to achieve stationarity. I thus took the first-order differences of the meaning frequency and SFA series, giving the *differenced meaning frequency* (MF') and the *differenced short form advantage* (SFA'). These transformed time series are plotted in Figure 4.3.

4.3.3.2 Determining the optimal lag and applying the Granger test

It is now possible to apply the Granger causality test to these stationary time series. One parameter that is still not fixed, however, is the lag value m , which tells us the

date in the past ($t - m$) before which we can ignore data in our predictive model for $Y(t)$. This lag value is usually determined by fitting models for multiple different lag values, then selecting the model that minimises the AIC (Akaike Information Criterion)—see, e.g., Moscoso del Prado (2014).

I decided to try lag values between 1 and 10 years, assuming this to be a plausible window for frequency shifts to show up in a written diachronic corpus. This range is also supported by visual inspection of the changes in both time series (see Figures 4.2 and 4.3).² I first fitted autoregressive models of SFA' given MF' , and found the AIC to be minimised for $m = 8$. I then applied this in the other direction—fitting autoregressive models of MF' given SFA' —and found in this case an optimal lag of $m = 10$.

To test whether the meaning frequency Granger-causes the short form advantage, I used the `grangertest` function in the `lmtest` R-package (Zeileis and Hothorn, 2002), inputting the lag value $m = 8$. The test found that the model for SFA' which included lagged values of MF' was a significant improvement on the model without these values ($p = 0.003$) and thus we can conclude that the meaning frequency Granger-causes the short form advantage.

I then ran the test in the opposite direction—fitting autoregressive models for MF' with and without lagged values of SFA' , inputting the lag value $m = 10$. This time, the augmented model including lagged values of SFA' was *not* found to be a significant improvement over the first model ($p = 0.182$). Thus, we can conclude that the short form advantage does *not* Granger-cause the meaning frequency.

As an extra check, I re-ran the Granger tests (in both directions) for all lag values between 1 and 10, and found that the results were qualitatively similar for all lag values greater than 1.

This result, in which the meaning frequency of the *info/information* clipped pair is found to predict changes in the short form advantage, but not the converse, is exactly in line with the predictions of the diachronic hypothesis outlined at the beginning of this chapter.

4.3.4 Discussion

A detailed analysis of the *info/information* clipped pair showed that: 1) the meaning frequency of the clipped pair is increasing over time, on average; 2) the relative frequency of the short form over the long form (the ‘short form advantage’ or *SFA*) is also increasing over time, on average; and 3) changes in the meaning frequency predict changes in the short form advantage, according to a Granger test for causality. These results support the diachronic hypothesis, originally laid out by Zipf, that “as

²Furthermore, a sanity check testing lag values greater than 10 did not yield consistently improved model fits.

the relative frequency of a word increases, it tends to diminish in magnitude” (Zipf, 1935; p. 38).

The case study presented here provides rigorous quantitative support for a causal link propagating through time between the meaning frequency of the *info/information* clipped pair and its corresponding short form advantage. However, to assess how widely this diachronic causal link can be observed across the lexicon, we need to test a comprehensive set of word pairs. This brings me to Experiment 5, below.

4.4 Experiment 5: English clipped pairs

4.4.1 Data set

In this experiment, I ran the tests described above—first, a set of linear regressions to check whether the average tendencies of the meaning frequency and short form advantage of clipped pairs are aligned; and second, a set of Granger causality tests to check whether the *MF*’ time series predicts changes in the *SFA*’ time series—on a comprehensive list of English clipped pairs. I compiled this list using *an English-French dictionary of clipped words* (Antoine, 2000), an etymological and slang dictionary containing hundreds of clipped pairs in both English and French.

First, I extracted the full list of English clipped pairs from the dictionary. I then discarded any clipped pairs with ambiguous short or long forms, as these could render the frequency counts inaccurate by conflating homonyms which refer to different meanings. This resulted in a list of approximately 300 clipped pairs.

For this list, I then extracted frequency counts per year for both the short form and the long form from the Google Books Ngrams corpus. As before, I only included years in which each token had at least 40 occurrences (see Section 4.3.1). Frequency counts from each year were normalised by the total number of tokens in that year to give the normalised frequencies F_l and F_s for each clipped pair. These were summed to give the total meaning frequency per year for each pair (see Equation 4.1). They were then also differenced, and divided by the meaning frequency, to give the short form advantage per year (see Equation 4.2).

As with the *info/information* clipped pair, I discarded data from all years before 1900 due to increasing data sparsity further back in time in the corpus. I then performed one final filtering step to remove obscure, low frequency words (many of which were even opaque to native speakers, often because they were slang words with very specific contexts of use). By restricting the analysis to relatively high frequency words, we can ensure that frequency counts in the corpus—which is compiled from published books—reliably reflect actual usage, and not just which books happened to be published in a given year. Thus, I obtained synchronic frequency counts from the Google 1T corpus (Brants and Franz, 2006), which are based on one trillion tokens

sourced from publicly accessible web pages. I then removed all clipped pairs from the data set whose short form fell below the threshold of half a million occurrences in the 1T corpus. Applying this threshold succeeded in removing nearly all of the more obscure short forms, and resulted in a final list of 92 clipped pairs. The list is provided in Table B.1 of Appendix B.

4.4.2 Linear regression analyses

For each clipped pair in the list, I fitted a generalised linear model of the meaning frequency, with year as the predictor variable, to check whether the meaning frequency was rising or falling over time on average. Of the 92 pairs, 69 showed a significant positive effect of year on the meaning frequency, and 20 showed a significant negative effect of year on the meaning frequency. Three pairs showed no significant effect.

According to the diachronic hypothesis, the clipped pairs whose meaning frequency is increasing over time on average should also show an average increase in the short form advantage. To test this, I ran a linear mixed effects regression using the `lme4` package in R (Bates et al., 2015), on just the set of 69 pairs with a positive meaning frequency slope. The predicted variable was the short form advantage and the predictor variable was the year. A random intercept for clipped pair was also included.³ This model yielded a positive significant effect of year on the short form advantage ($\beta = 1.607e - 03$, $SE = 1.076e - 04$, $p < 0.001$). This result confirms the prediction: the clipped pairs that show an average increase in their meaning frequency also show an average increase in their short form advantage.

However, also crucial to the hypothesis is the prediction that the behaviour should be different for clipped pairs whose meaning frequency is *decreasing* over time, on average. The short form advantage of these pairs, according to the predictions outlined in Section 4.1, should either be decreasing, level, or increasing at a slower rate than that of the clipped pairs in the first set. To test this, I fitted another linear mixed effects model, this time to the set of 20 pairs with an average negative meaning frequency slope. The results yielded a positive significant effect of year on the short form advantage ($\beta = 1.175e - 03$, $SE = 5.833e - 05$, $p < 0.001$), implying that, for these pairs, the short form is also becoming increasingly preferred, as it is for the pairs with a positive meaning frequency slope. However, the effect size is slightly smaller than that found for the first set of pairs, suggesting that the short form advantage may be increasing at a slower rate for most of these pairs.

To test whether this difference in the rates of short form increase was significant,

³A model using the full random effects structure—i.e. including random slopes for year—produced qualitatively similar results. However, it failed to adequately converge and therefore we only report models with the reduced random effects structure here. Convergence issues were flagged as being due to overly large differences in the scales between different word pairs. Adequately rescaling this frequency data turns out to be a complex issue, however, and beyond the scope of the current project.

I ran a third linear mixed effects regression. This time the model was fitted to the combined set of clipped pairs—both those with a significant increasing and those with a significant decreasing meaning frequency. The predicted variable was the short form advantage, and the predictor variables were year, sign of meaning frequency slope (positive or negative), and their interaction. A random intercept was included for clipped pair. As before, the model yielded a significant positive effect of year on the *SFA* ($\beta = 1.177e-03$, $SE = 1.524e-04$, $p < 0.001$). However, there was an additional significant positive interaction effect of year for clipped pairs whose meaning frequency slope was positive ($\beta = 4.298e-04$, $SE = 1.795e-04$, $p = 0.017$). A χ^2 test confirmed that this model was a significantly better fit to the data than models without the interaction term included, and models without the meaning frequency slope included as a predictor at all. This indicates that the short form indeed increases at a significantly faster rate for clipped pairs whose meaning frequency is increasing, than for those whose meaning frequency is decreasing.

It is important to remember, however, that these linear regressions capture only the average trend of each time series, and do not account for the subtleties of changing upward and downward movement in the frequencies of each clipped pair. Therefore, I turn next to the Granger causality analysis (introduced in Section 4.3.3), to investigate whether changes in the meaning frequency of each pair *predict* subsequent corresponding changes in the short form advantage.

4.4.3 Granger causality analyses

Before it is possible to run any Granger tests, the time series in question must be made stationary, as described in Section 4.3.3. To do this, I ran the `ndiffs` function on the meaning frequency and short form advantage time series of each clipped pair, to determine the number of differences required for each of them to achieve stationarity. Because differencing reduces the number of data points in a time series by one every time it is applied, it is important that the *MF* and *SFA* series of each clipped pair are differenced the same number of times; this ensures that both the resulting time series MF' and SFA' are the same length when the Granger test is applied. For some pairs, however, one of the time series required more differences than the other to achieve stationarity. In these cases I chose the larger number of the two to apply to both time series. No time series required more than two differences to achieve stationarity; most only required one.

Once all the time series had been transformed, the next step was to determine the optimal lag to use in the Granger tests. This was done, as in Section 4.3.3, by fitting autoregressive models using different lag values between 1 and 10, and choosing the lag value m which minimised the AIC. For each clipped pair, I then ran two Granger tests, using the optimal lag values found for that pair in each respective direction—

one in which the short form advantage is modeled with and without lagged values of the meaning frequency, and one in which the meaning frequency is modeled with and without lagged values of the short form advantage.

Of the 92 clipped pairs tested, there were 15 where the meaning frequency was found to Granger-cause the short form advantage at significance level below $\alpha = 0.05$. These were: *aqua/aquamarine*, *boobs/boobies*, *bro/brother*, *carbs/carbohydrates*, *decal/decalomania*, *fax/facsimile*, *info/information* (which we already determined in Section 4.3.3), *intro/introduction*, *memo/memorandum*, *neg/negatives*, *piano/pianoforte*, *porn/pornography*, *ref/referee*, *sax/saxophone*, and *servo/servomechanism*. Of these, 8 also showed no significant Granger-causality in the opposite direction: *aqua/aquamarine*, *carbs/carbohydrates*, *decal/decalomania*, *info/information*, *intro/introduction*, *neg/negatives*, *piano/pianoforte* and *sax/saxophone*. For these clipped pairs, we therefore see changes in the meaning frequency causing corresponding changes in the short form advantage, but not the other way around, which is the result predicted by our causal diachronic hypothesis.

However, for the other 7 pairs, a significant causal effect of short form advantage on the meaning frequency was also found, indicating some kind of feedback relationship between the two variables. Interestingly, a further 14 pairs were found to show *only* a significant causal effect of short form advantage on the meaning frequency, rather than in the opposite, predicted direction: *boob/booby*, *bot/robot*, *champ/champion*, *combo/combination*, *comfy/comfortable*, *comm/communication*, *limo/limousine*, *mis/miserable*, *nav/navigator*, *pol/politician*, *prof/professor*, *rehab/rehabilitation*, *stats/statistics* and *undies/underwear*. For the remaining 63 pairs, no significant Granger causality effect was found in either direction.

4.4.4 Discussion

The explanatory hypothesis that an increase in frequency would often lead to a decrease in length “for the purpose of saving time and effort” was one that Zipf deemed “too self-evident to require demonstration” (Zipf, 1935; p. 30). As for the converse possibility, that the brevity of a word might cause it to be used more frequently, he felt there was “no cogent reason for believing” it, because speakers primarily select their words based on the meanings they wish to convey, rather than on their lengths (Zipf, 1935; p. 29).

A Granger causality analysis over a list of 92 clipped pairs, however, did not provide large-scale quantitative support for these claims. Only 15 pairs showed evidence of the predicted causal influence of frequency changes on average length, and 21 pairs in fact appeared to show a causal relationship in the opposite direction, either in addition to the predicted direction, or solely in this direction.

Interestingly, however, the linear regression analysis did confirm that meanings whose frequency increased over time on average, also showed an average increase in their short form advantage. Meanings whose frequency *decreased* over time on average showed an increase in their short form advantage as well, but at a slower rate. This result lends support to a hypothesis made by Bybee in reference to ongoing phonological change: “changes often affect high-frequency words to a greater extent or at a faster rate than they do low-frequency words” (Bybee, 2006; p. 296). Perhaps similarly for lexical change through truncation, all words may be susceptible to it to some extent, but words whose frequency is increasing over time will be faster to show a disproportionate shift in usage towards the shorter variant. This general tendency for all words in the lexicon to shorten over time accords with the findings of Pate (2017), who showed that newer words tend to contain more extra letters above their optimal length than older words. For the types of words I focus on here—clipped pairs—there is the additional advantage that a shortened form has already been coined for these meanings. Once these shortened forms exist, it may be that they eventually become preferred over the long form in almost all cases, as long as the meaning is above a certain threshold frequency—recall that we restricted our analysis here to relatively high-frequency pairs. This would be evidence of an overall general increase in the efficiency of language over time.

Taken together, what do these results imply for Zipf’s causal hypothesis? The nature of the corpus used for these analyses means one must take care in interpreting the results. First, there are a host of potential factors other than frequency that might cause speakers to use a short form in a particular instance. For example in English, short forms often come with a more casual connotation—*info/information* is a good example of this. This means that the level of formality demanded by the context could sometimes take precedence over frequency considerations in determining whether a speaker chooses to use *info* or *information* in a particular instance. If frequency is only one factor among many determining short form use, it may be that, when viewed from a high level, its causal influence is dwarfed by other factors. This could explain why a Granger causality analysis would fail to find a significant effect for some clipped pairs.

Second, Google Books Ngrams is a written corpus based on published books. This is a format in which concerns about “saving time and effort” when using frequent meanings may be of minimal importance—such factors are primarily of relevance to spoken language. Therefore, while changes in the overall frequency of a meaning over time might be reflected in the corpus with reasonable accuracy, the actual change in usage of short forms in speech may not be as accurately reflected here, since what is considered acceptable in speech is not always considered to be appropriate in published written work.

Third, and continuing in the same vein, the composition of the corpus itself might have the largest influence on the short form usage measured. There is no data available on how different written genres are spread across years, and thus any significant shift in the genres represented from year to year could have a strong influence on the short form usage measured. Specifically, more books from informal genres being included in a given year could raise the value of the short form advantage measured in that year, irrespective of its actual usage in speech. The fact that I found the short form advantage to be generally increasing over time regardless of whether the meaning frequency was also generally increasing or instead generally decreasing (Section 4.4.2) could simply be due to an increasing proportion of informal genres appearing in the corpus over time. It could also, however, be reflective of a general increase in the use of informal language in published work.

Fourth and finally, the OCR methods used to digitize the books themselves is error prone, as mentioned in Section 4.2. As pointed out by the creators of the corpus, “for simple statistical reasons, short words are much more likely to arise due to OCR mistakes (for instance, the word ‘hat’ can be the result of an OCR error recognizing an ‘a’ instead of an ‘o’ in ‘hot’) than longer words (such as ‘outstanding’)”. They therefore warn that “results for short 1-gram must also be taken with caution” (Michel et al., 2011; Supplementary Online Material p. 17).

4.5 Experiment 6: English *-ic*/*-ical* pairs

In this section, I investigate the extent to which the results of Experiment 5 were due to corpus- and word-specific factors like those discussed above, by running the same set of analyses on a different set of word pairs. Namely, I look at word pairs that share the same root, but one word ends with the suffix *-ic* and the other with the longer suffix *-ical*. For example: *diabolic* and *diabological*, or *symmetric* and *symmetrical*. Such word pairs are similar to the clipped pairs used in Experiment 5 in that they differ in length, but are very close in meaning. Crucially, however, the two words in an *-ic*/*-ical* pair rarely differ in terms of their register—neither is likely to be construed as more casual or formal than the other. This means one can investigate their frequency shifts over time without worrying about the genre composition of the corpus or the register associated with the spoken or written context as potential confounds. The short forms in these word pairs also tend to be on the longer side, minimising the potential role of OCR digitization errors.

4.5.1 Data set

To compile a comprehensive list of English *-ic*/*-ical* pairs, I used Moby Words II (Ward, 2002), a large electronic English word list available in the public domain. I

extracted all words of the form **-ic* for which I could also find a word of the form **-ical*. This resulted in a list of 556 pairs. I then removed all pairs with the endings *-ologic/-ological*, as these have been shown to have unusual morphological behaviour. For example, the *-olog* suffix strongly prefers the *-ical* ending over the *-ic* ending, most likely for historical reasons (Lindsay and Aronoff, 2013). This reduced the list to 459 pairs.

I then combed through these pairs and removed any where the two forms had clearly different meanings. For example, in some cases one of the words in the pair had at least one sense which was a different part of speech from the other word in the pair, e.g., *music/musical* and *mechanic/mechanical*. This brought the list to 402 pairs. Among the remaining word pairs, some were close in meaning but with slightly different senses or connotations, e.g. *historic/historical* and *classic/classical*, while others were completely indistinguishable in meaning, e.g. *prototypic/prototypical* and *diabolic/diabolical*.

I did not exclude any words according to their synchronic frequency counts in the Google 1T corpus, as I did with the clipped pairs, because all words in this list were relatively low frequency words. I did however remove data from all years prior to 1900, as before, and I then extracted the total meaning frequency and short form advantage for each pair in each year in which both forms occurred at least 40 times in the Google Books Ngrams corpus. For some pairs, this meant there was data left from very few years. Because I am conducting diachronic analyses, and therefore need a reasonable time span over which to analyse diachronic trends, I removed any pairs which did not occur across at least 40 different years. I chose this number because it was 4 times the maximum lag value I would be looking at in the Granger causality analyses, and thus a reasonable amount of time over which causal changes might be observed propagating from meaning frequency to short form usage. This filtering step led to a removal of a further 141 pairs, bringing the final list to 261 *-ic/-ical* pairs. The full list is provided in Table B.2 in the Appendix B.

4.5.2 Linear regression analyses

I used the same methods described in Section 4.4.2: a generalised linear regression was run on the total meaning frequency of each pair, with year as the predictor variable, to assess whether the meaning frequency was increasing or decreasing on average over time. 132 pairs showed a significant positive effect of year, 82 pairs showed a significant negative effect of year, and 47 pairs showed no significant effect.

I then ran a linear mixed effects regression on the short form advantage of all pairs with an average increasing meaning frequency, to determine whether, for these pairs, the *SFA* was correspondingly increasing over time. Year was set as the predictor variable, and a random intercept for pair was included. A significant positive effect of

year on the short form advantage was found ($\beta = 2.624e - 03, SE = 6.412e - 05, p < 0.001$), signifying that for most of these pairs, the *SFA* was increasing over time, in line with our predictions.

For the pairs with an average *decreasing* meaning frequency, I ran another linear mixed effects regression on the short form advantage, again with year as the predictor variable and a random intercept for pair included. Again a significant positive effect of year on the short form advantage was found ($\beta = 1.132e - 03, SE = 5.705e - 05, p < 0.001$), but with a smaller effect size than that found for the set of pairs with an average increasing meaning frequency. As in the case of the clipped pairs in Experiment 5, this suggests that, although shorter forms are becoming increasingly preferred across the board, they rise at a faster rate for pairs whose overall meaning frequency is also increasing.

To test whether this difference in rate was significant, I fitted a third model. This was fitted to the *SFA* values of the combined data sets, including both pairs whose meaning frequency was increasing on average, and those whose meaning frequency was decreasing. The predictor variables were year, slope of the meaning frequency (positive or negative), and their interaction. A random intercept for word pair was included. As expected, a significant positive effect of year was found ($\beta = 1.132e - 03, SE = 7.245e - 05, p < 0.001$). In addition, a significant positive interaction effect of year was found for pairs with a positive meaning frequency slope ($\beta = 1.492e - 03, SE = 9.263e - 05, p < 0.001$). As in Experiment 5, this indicates that the short form usage does indeed increase at a faster rate for pairs whose meaning frequency is also increasing.

4.5.3 Granger causality analyses

The regression models reported above provide only a rough measure of the average trends for each word pair. Thus, a Granger causality analysis was also conducted, in order to provide a more fine-grained insight into whether changes in the meaning frequency of a pair influenced corresponding changes in the short form advantage. The method used was the same as that for the clipped pairs, described in Section 4.4.3. As before, no time series required more than two differences to achieve stationarity; most only required one.

For each pair of differenced time series, the optimal lag values were determined by fitting autoregressive models in both directions and choosing those which minimised the AIC. I then ran Granger causality tests using the chosen lag value in each direction. I found 46 word pairs with a significant causal effect of meaning frequency on the short form advantage. 35 of these did not also show a significant causal effect of the short form advantage on the meaning frequency, and thus these word pairs exhibit behaviour consistent with Zipf's causal hypothesis—that a change in the meaning

frequency will cause a subsequent corresponding change in the short form advantage, but not the other way around. These 35 word pairs are listed in Table 4.1.

Interestingly, a further 44 words showed only a causal relationship in the non-predicted direction—the *SFA'* time series was found to have a significant causal effect on the *MF'* time series, but not the other way around. These word pairs are given in Table 4.2. For the remaining 171 words, no significant causal relationship was found in either direction.

Table 4.1: List of English *-ic/-ical* pairs that show Granger causality in the predicted direction only—a change in the meaning frequency causes a subsequent corresponding change in the short form advantage (and not also the other way around).

short form/long form	short form/long form	short form/long form
amic/amical	dogmatic/dogmatical	nonanalytic/nonanalytical
angelic/angelical	domic/domical	paleogeographic/paleogeographical
aristocratic/aristocratical	dramatic/dramatical	periodic/periodical
arithmetic/arithmetical	ecumenic/ecumenical	phonetic/phonetical
bibliographic/bibliographical	egotistic/egotistical	schismatic/schismatical
brahminic/brahminical	electric/electrical	scientific/scientifical
classic/classical	elliptic/elliptical	subtropic/subtropical
climatic/climatical	emphatic/emphatical	synodic/synodical
conic/conical	hermetic/hermetical	theatric/theatrical
demagogic/demagogical	homiletic/homiletical	trigonometric/trigonometrical
despotic/despotical	hypothetic/hypothetical	typic/typical
didactic/didactical	microanalytic/microanalytical	

Table 4.2: List of English *-ic/-ical* pairs that show Granger causality in the *non-predicted* direction only—a change in the short form advantage time series appears to precede a subsequent corresponding change in the meaning frequency time series (and not also the other way around).

short form/long form	short form/long form	short form/long form
acoustic/acoustical	epic/epical	philosophic/philosophical
aeronautic/aeronautical	genealogic/genealogical	phylogenetic/phylogenetical
aesthetic/aesthetical	hagiographic/hagiographical	physiognomic/physiognomical
anarchic/anarchical	identic/identical	piratic/piratical
apologetic/apologetical	metallurgic/metallurgical	radiographic/radiographical
apostolic/apostolical	mineralogic/mineralogical	rhythmic/rhythmical
artistic/artistical	noncyclic/noncyclical	specific/specifical
ascetic/ascetical	obstetric/obstetrical	spheric/spherical
biochemic/biochemical	oceanographic/oceanographical	stoic/stoical
brahmanic/brahmanical	palaeogeographic/palaeogeographical	symmetric/symmetrical
casuistic/casuistical	parabolic/parabolical	topographic/topographical
chronic/chronical	parenthetic/parenthetical	tragic/tragical
cytogenetic/cytogenetical	pharisaic/pharisaical	tyrannic/tyrannical
democratic/democratical	phenotypic/phenotypical	unpoetic/unpoetical
enigmatic/enigmatical	philanthropic/philanthropical	

4.5.4 Discussion

The behaviour observed for these *-ic/-ical* pairs is similar to that observed for the clipped pairs in Section 4.4. Only 17.6% of the pairs in our list showed a significant effect of Granger causality in the predicted direction, similar to the 16.3% found in our list of clipped pairs in Experiment 5. Though this effect is not pervasive, it may still be greater than that expected by chance. As before, it may also simply be that the effect is dwarfed by other factors which play a larger role in determining short form use at any given time, making the influence from frequency alone undetectable as a significant cause. Or, it may be linked to the nature of the corpus itself, as a written rather than spoken corpus; frequency changes occurring in spoken language may not necessarily make their way into the written medium.

Similar to the results of the linear regression analyses in Experiment 5, we did also observe here a general increase across the board of short forms relative to their longer counterparts. Moreover, the rate of increase for the pairs with an average increasing meaning frequency was significantly higher than that for the pairs with an average decreasing meaning frequency, just as we found for the clipped pairs in Experiment 5. In the case of the clipped pairs, one possible explanation for the overall increase of short forms was that there may have been an increase in informal genres being represented in the corpus in more recent years. Another possible explanation was that published writing is simply becoming more informal across the board. These explanations, however, become less likely in light of the results observed here for *-ic/-ical* pairs. Since the *-ic* forms are not generally considered any more informal than their *-ical* counterparts, something other than a trend towards informality must explain their gradual proliferation.

One explanation discussed above in Section 4.4.4 remains a possibility: that once a clipped or shorter form exists for a meaning, then that form tends to gradually win out, although its rate of proliferation will depend on whether the meaning frequency is on the rise or on the decline. In this way, the English lexicon as a whole would be gradually becoming more efficient over time. There is already some evidence that speakers utilise the differentiating role of context to maximise use of ambiguous shorter words (Piantadosi et al., 2012). It could be that an average decreasing word length across the entire lexicon is also responsible for the higher-than-expected levels of phonetic clustering observed in English and some other languages (Dautriche et al., 2017).

4.6 Experiment 7: French clipped pairs

The dictionary from which we extracted our English clipped pairs (Antoine, 2000) also contains entries for French. The Google Books Ngrams corpus includes frequency

counts from published books in six additional languages, one of them being French. I therefore ran one further experiment, investigating the relationship between the meaning frequency and short form use of a small set of clipped pairs in French. The purpose of this experiment is to provide some indication of the generality of our findings concerning English word pairs. Is the behaviour observed for clipped pairs and *-ic/-ical* pairs in Experiments 5 and 6 restricted to English, or is it suggestive of a cross-linguistic diachronic pattern?

4.6.1 Data set

While word formation through clipping is a common morphological process in French, many of these clipped forms occur only in specific slang contexts. To obtain a uniform word list where all clippings were known to be in more general usage, and derived using a common morphological process, I extracted all clipped pairs from the the Antoine (2000) dictionary whose long form was a noun ending in the suffix *-ion*. Nouns ending in *-ion* are commonly clipped words in French, and most of their clipped forms are widely recognised.⁴ This extraction returned 41 clipped pairs. As an additional filter, I removed any form which either did not occur or was a hapax legomenon (occurred only once) in the Frantext corpus (ATILF - CNRS and Université de Lorraine, 2016), a historical corpus of French containing approximately 300 million tokens. This brought the list to 30 pairs.

I then extracted the total meaning frequency and short form advantage values for each pair from the French portion of the Google Books Ngrams corpus, for all years post-1900 in which each word occurred at least 40 times. As in Section 4.5.1, for some pairs this returned data for very few years. Because I am conducting a diachronic analysis, a sufficient number of data points at different time values is necessary in order to reliably assess any diachronic trends. In Section 4.5.1, I set the threshold to at least 40 different years worth of data, to be on the safe side. In this case, setting the threshold to 40 years would eliminate too many pairs, so I decided to lower the threshold to only 20 years, to include more word pairs in the analysis (though I later checked that using the higher threshold and thus a smaller word list did not qualitatively change the overall results). The final word list contained 25 clipped pairs. These are given in Table B.3 in Appendix B.

It is worth noting that the French portion of the Google Books Ngrams corpus is much smaller than the English portion, containing approximately 45 billion words in total, as compared with more than 300 billion in the English portion. Moreover, the filtering and error correction methods which were fine-tuned to the English sub-corpus have not yet been applied with the same rigor to any of the non-English sub-corpora. For instance, the metadata, including year-of-publication information, is not as accu-

⁴This information was provided by my native French speaking collaborator, Olivier Bonami.

rate in these sub-corpora. The authors of the corpus therefore recommend that one should “use extra caution in interpreting the results of culturomic analyses...based on the various non-English corpora” (Michel et al., 2011; Supplementary Online Material p. 17). We must keep this in mind as we interpret the results of our analyses of French clipped pairs below.

Because the list of pairs being analysed in this experiment is relatively small (only 25 pairs), I include here individual graphs showing the meaning frequency and short form advantage of each pair, plotted over time (Figure 4.4).

4.6.2 Linear regression analyses

I applied the same methods used in Sections 4.4.2 and 4.5.2 to determine the average trends of the meaning frequency and short form advantage for each word pair. An initial generalised linear model was fitted to the total meaning frequency of each pair, with year as the predictor variable. Of the 25 pairs, 18 were found to have a significant positive effect of year on the meaning frequency, indicating that the meaning was becoming more frequent over time, on average, for these pairs. Only four pairs were found to have a significant negative effect of year on the meaning frequency, and three pairs showed no significant effect.

I then fitted a linear mixed effects model to the set of 18 pairs with an increasing meaning frequency, with short form advantage as the predicted variable, year as the predictor variable, and a random intercept for word pair. In contrast to the results observed in Experiments 5 and 6, a small but significant *negative* effect of year on the short form advantage was found ($\beta = -2.677e - 04$, $SE = 1.012e - 04$, $p = 0.008$), indicating that in fact the short forms of these clipped pairs are becoming less frequent relative to the long form over time.

Fitting a similar linear mixed effects model to the set of 4 pairs with an average decreasing meaning frequency—with short form advantage as the predicted variable, year as the predictor variable, and a random intercept for word pair—yielded no significant effect, most likely due to lack of power.

4.6.3 Granger Causality analyses

Again, the same methods were applied as those described in Sections 4.4.3 and 4.5.3. The maximum number of differences required for stationarity was two, and most clipped pairs only required one. Of the 25 pairs, four showed a significant effect of Granger causality from the meaning frequency to the short form advantage, and no effect in the opposite direction. These four pairs were therefore in alignment with the predictions of Zipf’s causal hypothesis. The pairs were: *asso/association*, *conso/consommation*, *intro/introduction*, and *sono/sonorisation*.

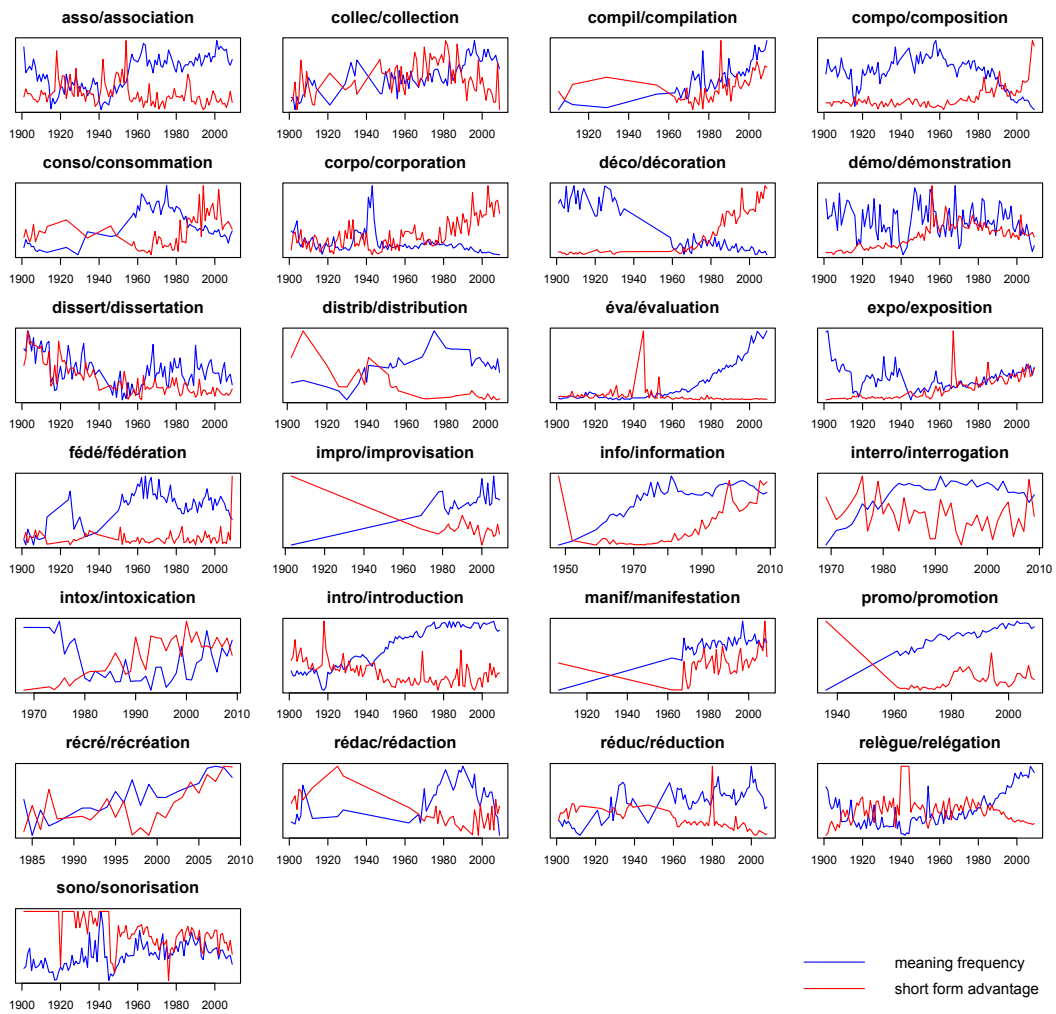


Figure 4.4: The meaning frequency (blue lines) and short form advantage (red lines) for the 25 French clipped pairs analysed in Experiment 7. The two time series in each graph correspond to different y-axes, but none of the y-axes are marked here for reasons of space and readability.

One pair—*relègue/relégation*—showed an effect in the opposite direction: lagged values of the short form advantage were found to be significant predictors of the meaning frequency, but not the other way around. For the remaining 20 pairs, no significant Granger causality effect was found.

4.6.4 Discussion

The small set of French clipped pairs investigated in this experiment did not replicate the behaviour observed in Experiment 5 for the English clipped pairs or in Experiment 6 for the *-ic/-ical* pairs, in terms of the linear regression analyses. For both these previous data sets, a small but significant positive effect of year on the short form advantage was found for pairs where the meaning frequency was increasing over time, and a smaller positive effect of year on the short form advantage for pairs where the meaning frequency was decreasing over time. This indicated that the short form was gradually gaining in use over the long form for the majority of English long/short pairs, but at a faster rate for those whose meaning frequency was also on the rise. In contrast, the French clipped forms, which were all derived from nouns ending in *-ion*, showed a small but significant *negative* effect of year on the short form advantage, for the pairs which had an average increasing meaning frequency (which was most pairs). This indicates that, for these pairs, the short form is in fact becoming less popular over time, relative to the long form.

At face value, the results suggest that something interesting may be changing about this particular class of clipped pairs in French. It could be that these types of shortenings are becoming less acceptable, if not across the board then at least in published writing in French. However, we must also acknowledge the warning given by the authors of the Google Books corpus, to use caution in interpreting the results obtained from any of the non-English sub-corpora. The metadata, which includes the crucial information of year of publication of each book, is less reliable for these corpora. Additionally—and this is a problem for both the English and non-English portions of the corpus—we know little about the changing composition of the corpus from year-to-year. Thus, the frequencies we observe for these types of clipped words could be heavily influenced by different genres and types of books represented in any given year.

The Granger causality analysis on this data set yielded somewhat similar results to those found for the English word pairs: a small proportion of the pairs (16%, as compared with 16.3% and 17.6% in Experiments 5 and 6, respectively) showed a significant causal relationship in the predicted direction. However, large-scale support for Zipf’s causal hypothesis across the lexicon was not found. I discuss possible reasons for this in the general discussion, below.

4.7 General Discussion and Conclusions

Analyses of large synchronic corpora continue to confirm that Zipf’s Law of Abbreviation is a universal design feature of human languages—more frequent words tend to be shorter than less frequent words. In Chapter 2 of this thesis, we investigated one hypothesis about the causal mechanisms that give rise to this universal pattern: that when presented with both a long and clipped form for different meanings, language-users tend to use the short form for more frequent meanings, and the long form for less frequent meanings, when subject to competing pressures for accurate and efficient communication. For this tendency on the part of speakers to accumulate into long-term lexical change and thus be responsible for shaping lexicons that demonstrate the Law of Abbreviation, an additional series of linking mechanisms are needed. First, as the frequency of a word increases over time, its short variant should also become increasingly frequent relative to its longer variant. This gradual skewing of the frequency distribution towards the short variant would ultimately lead to the conditions necessary for processes such as regularisation and opacification to take place. Over time, this will result in permanent lexical change, and hence a lexicon in which more frequent meanings are mapped to shorter words.

The lack of availability of sufficiently large diachronic corpora has precluded direct tests of this crucial diachronic hypothesis. Here, I attempted the first such direct test, using the relatively new Google Books Ngrams corpus, an extremely large diachronic corpus with cross-linguistic coverage. The aim was to provide large-scale quantitative support for the hypothesis, first outlined by Zipf, that “as the relative frequency of a word increases, it tends to diminish in magnitude” (Zipf, 1935). More specifically, since my focus here is on long/short word pairs, I hypothesised that tracking the frequency of a meaning over time would provide predictive information about the frequency of its short form relative to its long form.

I investigated this hypothesis by running a series of statistical tests on diachronic frequency counts extracted from the Google Books Ngrams corpus, first in a case study on the English clipped pair *info/information*, then on three comprehensive word lists: English clipped pairs, English *-ic/-ical* pairs, and French clipped pairs.

Tests of Granger causality between the meaning frequency and short form advantage in these pairs did not produce large-scale quantitative support for Zipf’s causal hypothesis. In all three of the word lists tested, only 16-18% of words displayed a significant Granger-causal effect of the meaning frequency on the short form advantage. However, this may still be significantly more than is expected due to chance—further work, perhaps including the running of simulations, is needed to assess this.

The results obtained by linear regression analysis were clearer. These showed that, for English long/short word pairs, those which have an average increasing meaning frequency also show an average increase over time in their short form advantage.

Those which have an average *decreasing* meaning frequency also showed an average increase over time in their short form advantage, but *at a slower rate*. These analyses support our hypothesis that the diachronic trajectory of the meaning frequency influences the short form advantage.

Why would our hypotheses be supported in the regression analyses but not the Granger causality analyses? The explanation is likely related to the fact that the Granger test is critically sensitive to time at a much finer-grained level. If specific changes in the chosen lagged time span of the meaning frequency do not strictly precede the corresponding changes in the short form advantage, then Granger causality will not be found. However, the fact that the corpus is comprised of published books means it is unlikely to reflect frequency changes in spoken language with the accuracy and time-resolution necessary. This may be compounded by the fact that the use of English clipped forms is likely also affected by factors such as context and register, which could drown out some of the effects of meaning frequency from year to year. The frequency effects observed in the experiment reported in Chapter 2 may have only been clearly visible there because other factors were minimised in this controlled setting. In a corpus of natural language, where these other factors cannot be controlled for, we should not expect as strong of an effect to be observable.

As highlighted previously, the corpus also contains numerous potential sources of error that might confound our results, particularly in the case of the French data. The accuracy of the OCR digitisation, as well as of the metadata containing year-of-publication information, are both potentially serious contributors to the measurements extracted. Additionally, the shifting composition of the corpus over time could play a nontrivial role.

Nevertheless, there was one pattern in the English long/short word pairs that seemed to be robust across word pairs of different classes and frequencies: short forms increase in frequency relative to long forms across the board, though at a faster rate for meanings whose overall frequency is also increasing. This same pattern was not observed for the French clipped pairs, potentially due to the small size of the word list used, its restriction to only one morphological class of words (nouns ending in *-ion*), or the poorer accuracy of the frequency data in the French sub-corpus of Google Books Ngrams. The fact that this pattern is replicated across the two, larger, English word lists (for which there is also more accurate frequency data) is promising. It is consistent with Zipf's hypothesis—that as a meaning's frequency increases, its short form also becomes increasingly frequent—and also points to a more nuanced hypothesis. Namely, that the frequency of a meaning over time affects the *rate* at which its short form gains in use. Specifically, the short form advantage will increase *faster* for meanings whose overall frequency is also on the rise.

In addition, the fact that the short form tends to be gaining in relative popularity

for *all* English long/short word pairs, whether their overall meaning frequency is increasing or decreasing, is a new and surprising result. It suggests that language will increase in efficiency where it can: once a shorter form for a meaning exists, it will tend to become preferred over time, all else being equal. This is in line with the Principle of Least Effort posited by Zipf (1935), which claims that humans are driven by the “purpose of saving time and effort”, wherever possible. More generally this result supports a view of language as a communication system susceptible to evolution under external selection pressures, among them the pressure for communicative efficiency.

One possibility is that, had we looked at a word’s predictability in context rather than its frequency, we would have found results more closely in line with our predictions. Namely, meanings whose average surprisal is *increasing* over time may increasingly tend to be mapped to shorter forms, while meanings whose average surprisal is *decreasing* over time may increasingly tend to be mapped to longer forms. Indeed, synchronic data (Piantadosi et al., 2011) and data incorporating the age of words (Pate, 2017) have shown that a word’s length is more strongly predicted by its log trigram probability than with its log unigram probability. However, the size and structure of the Google Books Ngrams corpus make it difficult to obtain average trigram probabilities for words across many years, and thus I was unable to investigate this possibility here. Developing a more efficient indexing structure over the corpus could solve this issue, and I hope that this challenge is taken up in the future.

This chapter endeavored a large-scale diachronic investigation of the relationship between frequency and word length. While it has produced some intriguing results, further work is clearly warranted. In particular, using data which are better controlled for genre, or better reflect spoken usage, would yield more reliable results. Though no sufficiently large diachronic spoken corpus currently exists (to our knowledge), the rapid improvement in the state of the art for data storage and automatic speech-to-text parsing make this a possibility in the very near future.

Chapter 5

Summary and conclusion

Explanations for universal structural features of language have tended to take one of two broad approaches. Either the primary focus is placed on domain-specific constraints that have evolved biologically (e.g., Chomsky, 1965, 2011; Hauser et al., 2002, 2014), or it is placed on functional and domain-general constraints, that interact over the course of cultural evolution (e.g., Christiansen and Chater, 2008; Evans and Levinson, 2009). Zipf’s proposed explanation for the universal structural feature considered in this thesis—the Law of Abbreviation—takes the latter of these approaches. The Principle of Least Effort hypothesises that the communicative function of language is crucial in shaping how word lengths are distributed across meanings in a lexicon. Specifically, the hypothesis is that language users optimise the lexicon for efficient communication, aligning shorter word lengths with more frequent or predictable meanings, and that these changes accumulate over generations to give rise to the inverse frequency-length and predictability-length relationships observed across languages today. The resulting state is one in which the greater effort associated with producing longer words is only expended on words that are used rarely, or that tend to be less predicable given the context. Thus, efficiency is maximised without sacrificing communicative accuracy.

There is now a great deal of evidence that lexicons are indeed structured optimally in this way (e.g., Piantadosi et al., 2011 and Ferrer-i-Cancho and Hernández-Fernández, 2013). However, there is less clear evidence regarding the specific causal mechanisms that give rise to this structure. In Chapter 1 I discussed random typing models (Miller, 1957; Ferrer-i-Cancho and Moscoso del Prado, 2012), which show that lexicons which exhibit an inverse frequency-length or predictability-length relationship can arise from a purely stochastic generative process (e.g. monkeys bashing random keys on a typewriter). What such models highlight is that the Least Effort hypothesis cannot be taken for granted (Moscoso del Prado, 2013)—if language users are actively optimising the lexicon for efficient communication, as opposed to this structure arising through some random process, then we must be able to show

concrete evidence of the mechanisms of optimisation at work.

This gap was perhaps first addressed by Krauss and Weinheimer's 1964 study in which participants communicated in pairs to identify unfamiliar objects to one another. Some of these objects appeared more frequently than others. The authors found that by the end of testing trials, more frequent objects were associated with shorter descriptions. A series of follow-up studies (Krauss and Weinheimer, 1966; Wilkes-Gibbs and Clark, 1992; Hupet and Chantraine, 1992) showed that this inverse relationship between frequency and description length is crucially tied up with the inclusion of direct, real-time communication in the task. This suggests that the observed behaviour may indeed be due to participants optimising their form-meaning mappings in response to communicative pressures. In Chapters 2 and 3 of this thesis, I built on these findings by manipulating the presence or absence of specific communicative pressures in order to directly test their causal role in shaping participants' lexicons.

In Chapter 2, I adapted the original communication game setup of Krauss and Weinheimer (1964), where partners take turns playing director and matcher to identify unfamiliar objects. I incorporated an artificial language paradigm, in which participants learn and communicate using a novel miniature lexicon with which they have no prior learned associations. The miniature lexicon was organised such that each of two objects could be referred to by either a long or clipped label, with equal probability. To simulate the fact that in natural languages, shorter forms are more likely to be ambiguous, the short label for both objects was the same, while the long labels were distinct. Crucially, one object appeared more frequently than the other in both training and testing trials. Another novel feature of the design was a 2×2 manipulation of the pressures to communicate accurately and efficiently, such that there was one condition (the Combined condition) in which both pressures were present and competing with one another, and three different control conditions in which one or both of these pressures was removed. Only in the critical Combined condition did participants reliably reshape their input to align with the Law of Abbreviation, mapping the infrequent object to its long form, and the frequent object to the ambiguous short form. These results were robust across different methods of data collection: one version of the experiment was run in the lab, and another was run online using a novel framework I developed, allowing workers on Amazon Mechanical Turk to communicate with one another in real-time.

Experiment 3 (described in Chapter 3) used a similar experimental setup to the above, but instead of a difference in frequency between the two meanings in the miniature lexicon, there was a difference in their probability given different contexts. Object labels were now presented following one of two framing words, *gat* and *bix*. One object was much more likely to appear with *gat*, and the other much more likely

to appear with *bix*. Given the growing body of evidence for an even stronger inverse relationship between a word's length and its probability in context, the goal was to test whether language users do indeed track these probabilities during communication, and shift the distribution of word lengths in the lexicon towards this optimal state. Again, in the critical condition, participants played a communication game in which the competing pressures to communicate accurately and efficiently were both present, and then in three control conditions one or both of these pressures was switched off. Again, only in the critical Combined condition did participants reliably optimise the lexicon for efficient communication, using the short, ambiguous label in predictive contexts, and the longer labels in surprising contexts. Together with the results reported in Chapter 2, this indicates that language users actively adjust word length according to usage-based parameters such as frequency and probability in context, and that they do this in such a way as to optimise the language for efficient information transfer. This supports the original hypothesis made by Zipf in his proposal of the Principle of Least Effort.

Experiment 3 showed that participants alternate which label (long or short) they use for an object based on the context. In Experiment 4, I investigated whether words with a lower *average surprisal* correspond to a lower average word length. Using the setup of the Combined condition of Experiment 3, and adding a third object to the miniature language which had a lower average surprisal than the other two, I found that participants used the short label for the lower surprisal object significantly more than they used it for the higher surprisal objects. This speaks to the question of how the alternation behaviour observed in Experiment 3 ultimately leads to permanent lexical change. Words that appear in more low surprisal contexts than other words will also tend to be used in their shorter variants more often than other words. This leads to the diachronic hypothesis that as the average surprisal of a word decreases over time, the instances of its long variant will become rarer and rarer. At this stage, factors related to learning and memory might play an important role in amplifying these asymmetries. As new generations of language users learn from exposure to these highly skewed input distributions, regularisation may occur, ultimately eliminating the long variant from the lexicon. This would result in a lexicon where shorter words have a lower average surprisal than longer words, as has been observed in a range of different languages (Piantadosi et al., 2011; Manin, 2006).

In Chapter 4, I turned to focus on the diachronic hypothesis outlined above—that as a word becomes more frequent or predictable over time, its shorter variant will become more prevalent relative to its longer variant. I tested this hypothesis using the Google Books Ngrams corpus, a large diachronic corpus of written language. Using the English and French portions of this corpus, I looked at several different types of long/short word pairs. I found that for both types of English pairs investigated

(clipped pairs and *-ic/-ical* pairs), the frequency of short variants relative to long variants tended to increase over time. However, for pairs whose overall meaning frequency was also increasing over time, the *rate* at which the short form gained over the long form was significantly faster than that of the other pairs. This suggests that changes in the frequency of a word over time do indeed influence changes in its average length. Specifically, for words that have both a long and short variant in active use simultaneously, the short form will become preferred over the long form at a significantly faster rate, if the overall frequency of the word pair is increasing.

The findings presented here provide strong support for the theory that the communicative function of language shapes some fundamental properties of its structure. In particular, language users actively shift the distribution of word lengths in a lexicon in such a way that it aligns with the information-theoretic description of an optimal code for efficient communication. This adds to a recent surge of evidence that linguistic features other than length also appear to be optimised for efficient communication, such as phonological distinctiveness (Meylan and Griffiths, 2017), syntactic dependency lengths (Futrell et al., 2015), and semantic category structure (Kemp and Regier, 2012; Regier et al., 2015, 2016). These converging observations make for a convincing case that all these structures are the result of the same type of optimisation process. However, a detailed investigation into the actual mechanisms responsible for each pattern is necessary in order to confirm this, as the case of random typing models highlighted with regard to the Law of Abbreviation. If such investigations do show that these patterns result from the same basic principles of optimal information transfer, then we will have found a simple and powerful unifying explanation for many of the structural features observed in human languages—an explanation with the core assumption that language is, among other things, an evolving, efficient code for information transfer.

Appendix A

Relevant publications

Reproduced in this section are Kanwal et al. (2017b)—which formed the text of Chapter 2.1—and Kanwal et al. (2017a), from which parts of Chapter 3 were adapted.



Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/COGNIT

Original Articles

Zipf's Law of Abbreviation and the Principle of Least Effort: Language users optimise a miniature lexicon for efficient communication



Jasmeen Kanwal*, Kenny Smith, Jennifer Culbertson, Simon Kirby

Centre for Language Evolution, University of Edinburgh, Edinburgh, Scotland, United Kingdom

ARTICLE INFO

Article history:
Received 25 October 2016
Revised 24 March 2017
Accepted 1 May 2017

Keywords:
Zipf's Law of Abbreviation
Principle of Least Effort
Language universals
Efficient communication
Information theory
Artificial language learning

ABSTRACT

The linguist George Kingsley Zipf made a now classic observation about the relationship between a word's length and its frequency; the more frequent a word is, the shorter it tends to be. He claimed that this "Law of Abbreviation" is a universal structural property of language. The Law of Abbreviation has since been documented in a wide range of human languages, and extended to animal communication systems and even computer programming languages. Zipf hypothesised that this universal design feature arises as a result of individuals optimising form–meaning mappings under competing pressures to communicate accurately but also efficiently—his famous Principle of Least Effort. In this study, we use a miniature artificial language learning paradigm to provide direct experimental evidence for this explanatory hypothesis. We show that language users optimise form–meaning mappings only when pressures for accuracy and efficiency *both* operate during a communicative task, supporting Zipf's conjecture that the Principle of Least Effort can explain this universal feature of word length distributions.

© 2017 Elsevier B.V. All rights reserved.

1. Introduction

In 1935, the linguist George Kingsley Zipf pointed out what he claimed to be a universal property of human language: that "the magnitude of words stands in an inverse (not necessarily proportionate) relationship to the number of occurrences" (Zipf, 1935; p. 23). In other words, the more frequent a word is, the shorter it tends to be. This "Law of Abbreviation" has now been verified in a wide range of human languages, including: Chinese, Croatian, Czech, Dutch, English, French, German, Greek, Hungarian, Indonesian, Italian, Polish, Portuguese, Romanian, Russian, Slovenian, Slovak, Spanish, Sundanese, and Swedish (Ferrer-i-Cancho & Hernández-Fernández, 2013; Piantadosi, Tily, & Gibson, 2011; Sigurd, Eeg-Olofsson, & Van Weijer, 2004; Strauss, Grzybek, & Altmann, 2007; Teahan, Wen, McNab, & Witten, 2000). For example, one can clearly see this relationship for English words in Fig. 1. Interestingly, there is even evidence for its broader application in animal communication systems (in the vocalisations of common marmosets and formosan macaques, and in the surface behavioural patterns of dolphins; Ferrer-i Cancho et al., 2013) and in computer programming (e.g., use of the *alias* function in Unix to abbreviate frequent commands; Ellis & Hitchcock, 1986).

Zipf hypothesised that this universal pattern arises as a result of a tradeoff between two competing pressures: a pressure for accurate (successful) communication and a pressure for efficiency or less effort.¹ The idea is that together, these pressures would shape how forms are mapped to meanings, because languages have a finite inventory of discrete sounds that can be recombined to form words. This results in a lexicon with a limited number of words of a given length. Importantly, the shorter the length, the fewer distinct possible words there will be of that length, and the greater the potential confusability—shorter forms have less space for signal redundancy and thus are more likely to be confused in noisy signal transmission. Therefore, while a pressure for efficiency should favour these short words since they require less effort to produce (all things being equal), this is in direct conflict with the pressure for accurate communication. The latter should instead favour unique form–meaning mappings which minimise potential ambiguity—from this perspective, longer words have the clear advantage. How, then, can a language use the available short forms optimally, while still keeping ambiguity in check? The solution is to assign the shortest words to the most frequent meanings, leaving longer words for less frequent meanings, as in variable-length, e.g. Huffman, coding (Huffman,

* Corresponding author.

E-mail addresses: jasmeen.kanwal@ed.ac.uk (J. Kanwal), kenny.smith@ed.ac.uk (K. Smith), jennifer.culbertson@ed.ac.uk (J. Culbertson), simon@ling.ed.ac.uk (S. Kirby).¹ The assumption that information is packaged into repeating words of variable length, and not fixed-length blocks—as in, e.g., block codes such as Hamming codes (Hamming, 1950)—is also necessary to make this prediction. Thanks to an anonymous reviewer for pointing this out.

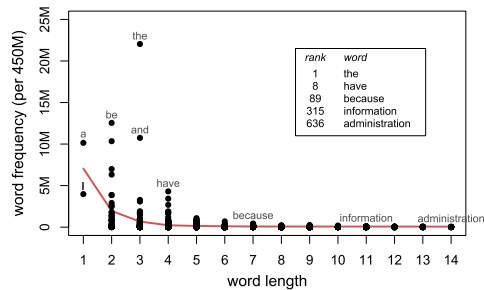


Fig. 1. The 1000 most frequent words in English. Each point represents an individual word (some points are labeled). The red line marks the mean frequency for the words of each length (here, orthographic length is used, but the same overall pattern would be seen if phonetic length were used instead.) The more frequent a word is, the shorter it tends to be. According to Zipf's Law of Abbreviation, this is a universal pattern of human languages. Frequency counts used here are from the 450 million word COCA corpus (Davies, 2008). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

1952). Zipf called this hypothesised tendency to produce short utterances wherever possible the “Principle of Least Effort”.

The Principle of Least Effort offers a functional explanation for the Law of Abbreviation, if we imagine it playing out through incremental changes over time. If language users track frequency differences between meanings (consciously or otherwise), then processes of change may differentially affect words whose frequencies differ. For example, if a word is more frequently used, then it may be more likely to be targeted for reduction or shortening (e.g., ‘information’ becomes ‘info’). Form-meaning mappings would then gradually shift toward more optimal alignment of frequency with length (Zipf, 1935).

While this is an attractive explanatory account, several researchers have raised the possibility that the inverse relationship between word length and word frequency could emerge instead from simple constraints on randomly generated systems. For example, a lexicon generated through a random typing process, in which ‘words’ are produced by pressing keys (including the space bar) at random, has properties that are consistent with the Law of Abbreviation (Ferrer-i-Cancho & Moscoso del Prado, 2012; Moscoso del Prado, 2013). While we know that languages are not actually generated at random in this way, it nevertheless remains a possibility that the Law of Abbreviation could result from some yet-unidentified statistical process, unrelated to optimisation behaviour on the part of language users.

Several studies provide indirect evidence connecting competing pressures for accurate and efficient communication to properties of linguistic systems introduced by language users. For example, previous research has shown that learners restructure case marking systems such that case markers are preferentially used when grammatical roles are ambiguous and omitted when other disambiguating information is present (Fedzechkina, Jaeger, & Newport, 2012). This is consistent with the idea that effort (here, producing case markers) is reduced in a way which preserves communicative function. Language learners have also been shown to capitalise on differences in the length of novel labels to make pragmatic inferences about the communicative intentions of speakers (Degen, Franke, & Jäger, 2013). A computational model of iterated learning (Kirby, 2001) shows that short, non-compositional morphological forms are more likely to evolve for frequent meanings, while longer, compositional ‘regular’ forms are more likely to persist for infrequent meanings, due to a tradeoff between the pressure

for learnability and the pressure for producing shorter, more replicable forms.

A direct link between frequency and utterance-length shortening in actual language users has been shown in studies such as Krauss and Weinheimer (1964) and Clark and Wilkes-Gibbs (1986). In these studies, participants played a dyadic communication game, where ‘directors’ used English to describe objects for their partners (‘matchers’) to identify from a set. The objects being communicated about were abstract geometrical shapes lacking canonical English names. The director would typically begin by using a long, elaborate phrase to help the matcher identify the correct object. However, on repeat occurrences of the object, the director would take advantage of a growing base of shared knowledge, established through communication, to gradually shorten the descriptive phrase and thereby reduce the effort expended. For example, an object described as “upside-down martini glass in a wire stand” on its first occurrence ultimately became shortened to just “martini” after several repeat occurrences. The more times an object reoccurred, the shorter its average length by the end of the experiment. These results depended on the director receiving positive, real-time feedback from the matcher during the signalling game (Hupet & Chantraine, 1992; Krauss & Weinheimer, 1966), suggesting that it is a communicative context which triggers the drive to reduce effort. Thus, this result suggests one mechanism by which the Law of Abbreviation could arise: if the form associated with a meaning becomes shorter the more times it occurs in conversation, and these mappings are retained and spread across speakers, then in the lexicon overall, more frequent meanings will end up with shorter forms than less frequent meanings.

However, as we mentioned above, there is competition for the short forms in a lexicon. For example ‘info’ refers to ‘information’, and not ‘informality’, ‘infoliation’, or ‘infoedation’. Why is this? In the Krauss and Weinheimer (1964) and Clark and Wilkes-Gibbs (1986) studies, participants were communicating about a small set of meanings using a large space of possible utterances. All labels could thus be shortened in this task without resulting in ambiguity. However, when several meanings are in direct competition for a single short label—a problem that arises at the level of an entire lexicon—the mechanism shown in these studies is not sufficient to account for why one meaning gets mapped to the short form and not the others.²

Thus, while these previous studies are consistent with the idea that something like the Principle of Least Effort operates during language use, they do not explicitly target the hypothesised role of competing communicative pressures—the pressure for reduced effort versus the pressure against ambiguous form-meaning mappings—in modulating word length within the lexicon. In our study, we make use of a miniature artificial language learning paradigm to create a setting in which these two pressures are directly in conflict: a reduction in effort cannot be achieved without also increasing the ambiguity of form-meaning mappings. Crucially, our set-up allows us to isolate these different pressures in order to determine their individual contribution to the overall behaviour of a miniature artificial lexicon. Following Zipf, we hypothesise that only when these pressures are both present—and thus in direct con-

² Interestingly, not all possible short forms in a language actually get used. This could be a consequence of noisy communication—using short forms sparingly would further minimise potential confusability. However, it has been found that frequent (and by proximity short) forms tend to be tightly clustered together in the phonological space, in seeming opposition to this end (Dautriche, Mahowald, Gibson, Christophe, & Piantadosi, 2017). This may be due to the influence of constraints on learning, memory, and production, which favour lexicons with high phonetic regularity. Thus, even though not all possible short forms are used, there will be particularly tough competition for those forms that fall within the more densely-populated regions of the phonological space. Thanks to an anonymous reviewer for raising this topic.

flict–will language users restructure their input to align shorter forms with more frequent meanings. In this way, our study aims to provide a concrete link between optimisation behaviour at the level of the individual and the global pattern Zipf first observed.

2. Miniature artificial language learning experiments

We use a miniature artificial language learning paradigm, which has previously been used to shed light on the cognitive mechanisms and environmental pressures that shape language structure (e.g., Culbertson, Smolensky, & Legendre, 2012; Fedzechkina et al., 2012; Kirby, Cornish, & Smith, 2008). In this paradigm, participants learn a miniature artificial language, and then we observe how they reshape their input as they use the language, in this case to communicate with a partner (see also Fehér, Wonnacott, & Smith, 2016; Kirby, Tamariz, Cornish, & Smith, 2015; Winters, Kirby, & Smith, 2015).

2.1. Participants

124 participants (51 females, 64 males; a further 9 chose not to report their gender) were recruited through Amazon Mechanical Turk. 106 of these reported themselves as native English speakers, of which 88 were monolingual. A broad range of other languages were represented across the remaining participants. Ages ranged from 18 to 73 (mean = 33).

2.2. Materials

Participants were trained on two names for each of two plant-like alien objects, by repeatedly being shown pictures of the objects labeled with their names on a computer screen (see also Reali & Griffiths, 2009; Vouloumanos, 2008). Crucially, one of the two objects appeared three times more frequently than the other—specifically, one object appeared 24 times and the other 8, for a total of 32 training trials.

Each object appeared half the time labeled with its long name, a 7-letter word, and half the time with its short name, a 3-letter word derived by clipping the last two syllables off the long name. The process of clipping, or word-truncation, is a common word-shortening device in many languages (e.g. *info* for *information* in both English and French; Antoine, 2000). In natural languages, shorter words are subject to greater confusability for a number of reasons. They have less space for signal redundancy and are therefore more likely to be misinterpreted or lost in noisy transmission. There are also more unique possible 7-letter strings than 3-letter strings, and thus word shortening can often result in outright ambiguity. Indeed, shorter words are more likely to be polysemous and homophonous (Piantadosi, Tily, & Gibson, 2012). To model these phenomena in our miniature lexicon, we designed the names such that the short name for both objects was identical (*zop*), while the long names were unique (*zopekil* and *zopudon*). A schematic diagram of the object frequencies and labels is provided in Fig. 2a.

Which object (the blue fruit or the red stalk) was more frequent, as well as which object was paired with each label, were both counterbalanced between participants, giving a total of 4 possible object-frequency-label pairings which a participant might be trained on. This ensured that potential factors such as sound symbolism, or higher saliency of one of the objects, could not systematically bias our results.

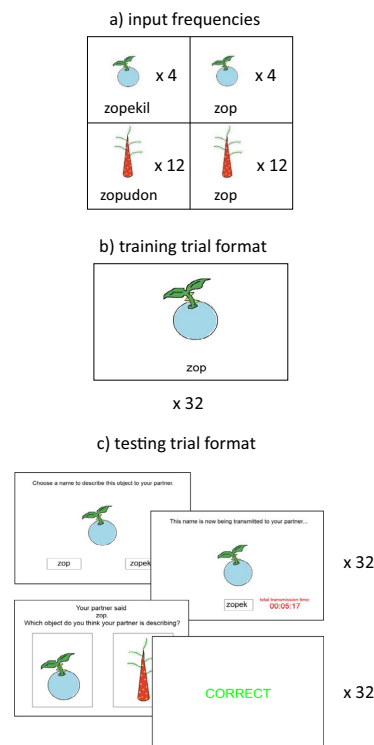


Fig. 2. (a) A schematic diagram of the frequencies of the objects and labels presented during the training trials in all four experimental conditions. One object appeared three times more frequently than the other. Each object appeared three times more frequently than the other. Each object was labeled half the time with its unique long name, and half the time with its ambiguous short name, which was a clipped version of the long name. (b) An example training trial. (c) An example of a director trial in the Combined condition (top) and a matcher trial followed by feedback (bottom).

2.3. Procedure

Participants were assigned to one of four conditions, where we manipulated the presence of pressures to communicate accurately and quickly in a between-subjects 2×2 design. In all conditions, the experiment consisted of two phases: training and testing. The training phase was identical for all four conditions, but the testing phase differed across conditions.

2.3.1. Training phase

On each training trial, an object was presented on screen alone for 700 ms. The appropriate label then appeared beneath the object for a further 2000 ms, yielding a total trial duration of 2700 ms. A blank screen showed for 500 ms between each trial. The 32 training trials were presented in a different randomised order for each participant.

2.3.2. Testing phase

After the training phase, testing procedures varied depending on the experimental condition. In the Combined condition, participants were under a pressure to communicate accurately *and* to communicate efficiently, as according to Zipf's hypothesis, both

of these competing pressures must be present for the Law of Abbreviation to emerge. The remaining three conditions removed one or both of these accuracy and time pressures. In all conditions, the testing trials contained the same frequency ratio over objects as the training trials: the frequent object appeared three times more frequently than the infrequent object.

2.3.2.1. Condition 1: Combined. In the testing phase of this condition (henceforth referred to as the Combined condition), participants were paired with a partner to play a communication game. This was done by putting participants in a virtual queue, managed by a central server script, after completing the training trials. Participants were paired sequentially as they finished training; once a participant entered the queue, the server would pair them with the next participant to finish training after them. To encourage participants to wait as long as possible in the queue without leaving the game, they were shown a humorous cat video while they waited. However, if participants had still not been paired with a partner after 5 min, they were removed from the queue and paid for their time. This method allowed us to successfully run a dyadic artificial language communication experiment online using a crowdsourcing platform. We were therefore able to relatively quickly and easily collect data from a more culturally and linguistically diverse group of participants than is usually possible with traditional lab-based experiments that draw mainly from a university's undergraduate population.

Once paired with a partner, participants began the communication game. On each trial, the 'director' was shown an object on the screen and told to transmit its name to the 'matcher'. The director always had two options for which name to send: the long name for the object or the (ambiguous) short name. The director chose a name by clicking on it, and was then given instructions for how to actually transmit the name to the matcher. This was done by pressing and holding the mouse in a central transmission box in which each letter in the name appeared one by one, at 1200 ms intervals. Note that participants never had to type the names or necessarily remember their correct spelling; once they chose a name from the two options on the screen, the letters would appear sequentially in the transmission box as they held down the mouse. Only once all the letters had appeared in the box was the name transmitted to the matcher. If the mouse was released before all letters had been transmitted, the participant would have to start again from the first letter (but the total transmission time was only counted for the successful transmission). This belaboured method of transmission, in which the long name was significantly slower to transmit than the short name, introduced an element of effort into communication, modelling the difference in effort in spoken communication associated with producing long versus short utterances.

Once the matcher received the name from the director, the matcher was asked to choose which of the two objects they thought the director was referring to. Both players were then given feedback as to whether the matcher chose the correct object.

The players alternated roles after every trial, with the matcher becoming the director and the director becoming the matcher, until both completed 32 director trials and 32 matcher trials. The frequency with which each object appeared in each player's director trials matched those of the training frequencies: 24 occurrences of the frequent object, and 8 of the infrequent object. The order of these 32 director trials was randomly shuffled for each participant. The member of the pair who entered the queue first was the first player to direct.

To model the pressures in spoken communication to be both efficient and accurate, pairs were told at the beginning of the game that they would be rewarded a bonus payment if they were the pair to complete the game in the quickest time with the highest

number of correct match trials. Time was only counted during name transmission, and the time count was displayed next to the transmission box as the participant was transmitting a name, to underline the time pressure. Example screenshots of a director trial and matcher trial are shown in Fig. 2c.

In order to tease apart the influence of the two pressures on the participants' patterns of behaviour, we included three further experimental conditions, described below, for a full 2×2 manipulation of the pressures for accuracy and efficiency.

2.3.2.2. Condition 2: Accuracy. In this condition, participants were paired to play a communication game as described above, but in the director trials, there was no intermediate step between the director choosing a name to send and the matcher receiving the name; the names were sent instantaneously, thus removing any difference in effort between transmitting long or short names. Pairs were told that the goal of the game was to have their partner make as many correct guesses as possible. There was no bonus reward given for the most accurate pair, as the task was extremely easy and we predicted that most pairs would achieve maximum accuracy, which turned out to be the case.

2.3.2.3. Condition 3: Time. In this condition, communication was taken out of the game entirely; participants played a one-player game consisting of 64 director trials only. In each director trial, participants were told to choose a name to describe the object shown on the screen, but there was no subsequent communicative task. As in the previous conditions, the choice was always between the long name and the short name. Once chosen, the name had to be entered as in the Combined condition, by pressing and holding the mouse in a transmission box, with each letter appearing at 1200 ms intervals. The next trial began only when all the letters had appeared in the box. Thus, the long name was significantly slower to produce than the short name. The transmission process was also timed with an on-screen timer as in the Combined condition, and participants were told at the beginning of the game that they would be rewarded a bonus payment if they were the player with the shortest overall transmission time.

2.3.2.4. Condition 4: Neither. The fourth and last condition contained neither a pressure for efficiency nor a pressure for accuracy. As in the Time condition, participants played a one-player game with no explicit communicative element, but additionally there was no time difference associated with transmission; once a label was chosen to describe an object, long or short, it was instantaneously recorded and the player was advanced to the next trial. We included this condition in order to provide a baseline for participants' behaviour from which to assess the effects of the accuracy and time pressures in the other three conditions.

2.3.3. Payment

Participants were paid depending on the condition they were in, commensurate with the average time it took to complete that condition. Participants in the Combined condition, the longest to complete due to both the slow transmission process and having to wait for the partner's response after each trial, were paid \$2; participants in the Accuracy and Time conditions were paid \$1, and participants in the Neither condition, the shortest to complete, were paid \$0.50.

2.4. Predictions

Our predictions for the Neither condition were that participants would either probability-match—i.e. use the long and short forms for both objects with equal frequency, as in the training trials (see Hudson Kam & Newport, 2005)—or their behaviour would

reveal prior biases language users bring to the task, such as a preference against using ambiguous forms.

In the Accuracy condition, we predicted that participants would be more likely to use the long names for both objects compared to the baseline condition, given the potential loss of accuracy from using the ambiguous short name, and with no time considerations to favour the use of short but ambiguous labels. Given the task demands, this would therefore be the best strategy to use in this condition.

In contrast, in the Time condition, we predicted that participants would use the short name for both objects: with no communicative purpose attached to the transmissions, and an incentive to be as quick as possible, using the short name in every trial is the best strategy in this condition.

In the critical Combined condition, with both a time and an accuracy pressure, we predicted that participants would converge on the optimal strategy, in which the frequent object is consistently mapped to the ambiguous short name, and the infrequent object to its unique long name, in line with Zipf's Law of Abbreviation. Using this strategy, transmission time is minimised as much as possible while still maintaining one-to-one form-meaning mappings, thereby also ensuring accurate communication.

3. Results

Fig. 3 shows the proportion of trials on which the short (ambiguous) label was selected by the director, for high- and low-frequency objects. As predicted, in the Accuracy condition, most participants retained the unique long names for both objects, while in the Time condition, most participants mapped both objects to the ambiguous short name. Crucially, in the Combined condition, where participants were under pressure to communicate both accurately and efficiently, most pairs converged on the optimal strategy wherein the most frequent object was mapped to the ambiguous short name, and the infrequent object to its unique long name. This made the participants' lexicon both efficient and expressive, in line with Zipf's Law of Abbreviation. Finally, the Neither condition revealed an underlying bias towards avoiding ambiguity.³

A logistic regression model was fit in R (R Core Team, 2015) using the lme4 package (Bates, Mächler, Bolker, & Walker, 2015), with short name use (as contrasted with long name use) as the binary dependent variable, object frequency, experimental condition, and their interaction as fixed effects, and by-participant random intercepts and random slopes for object frequency. This model yielded a significant positive interaction for the frequent object in the critical Combined condition. Thus, in this condition, participants were significantly more likely to assign the short name to the frequent object than in the baseline condition. Participants were significantly less likely to assign the short name to either object in the Accuracy condition, and significantly more likely to assign it to both objects in the Time condition, as reflected by the large negative coefficient for the former condition, and the large positive coefficient for the latter. Finally, the intercept is significantly negative, indicating that in the Neither condition, there is a baseline preference for avoiding the short form (see Table 1 for a full list of model coefficients).

³ The complete set of raw data from this experiment can be accessed using the following link: <http://datashare.is.ed.ac.uk/handle/10283/2702>.

⁴ We computed the mutual information directly from the empirical distributions, rather than using a bias-corrected estimate; since our use of this measure is for purposes of comparison between participants, we are not concerned with the absolute values, which would be lowered by roughly the same factor across all participants using a bias-correction method such as the Miller-Madow method.

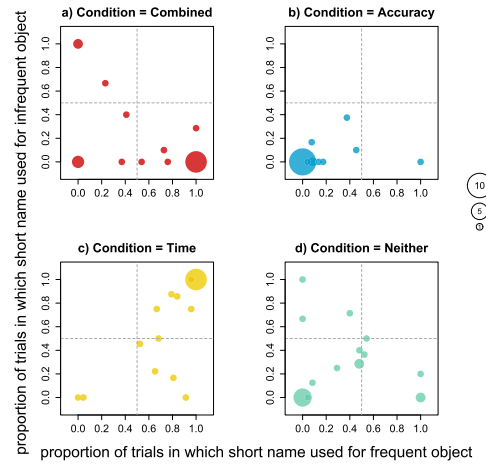


Fig. 3. The proportion of trials in which the short name was used to label the frequent object versus the proportion of trials in which it was used to label the infrequent object. For the Combined (a) and Accuracy (b) condition, each data point combines a pair of communicating players, representing the sum of their director trial productions. For the Time (c) and Neither (d) condition, each data point corresponds to an individual player's productions. The size of the circles is perceptually scaled (Tanimura, Kuroiwa, & Mizota, 2006) to reflect the number of data points coinciding at each value. Data from only the second half of testing trials is shown here, as participants were more likely to have converged on a stable mapping by this time. Data points in the top right quadrant indicate participants who are mostly using the short name for both objects; participants are clustered in this quadrant in the Time condition. Data points in the bottom left quadrant indicate those who are mostly using the unique long names for both objects; participants are clustered here in the Accuracy condition. Data points in the bottom right quadrant indicate participants who are mostly using the short name for the frequent object and the long name for the infrequent object. This behaviour, consistent with the Law of Abbreviation, only reliably arises in the Combined condition, where both pressures are present.

Table 1

Summary of fixed effects for a binomial regression model with short name use as the binary dependent variable, and by-participant random effects for object frequency. The predicted effects are shown in bold. Like Fig. 3, this model is fit using only the second half of each participant's training trial data, as participants were more likely to have converged on a stable linguistic mapping by then.

	β	SE	<i>p</i>
intercept (object = infrequent, condition = Neither)	-2.225	0.501	<0.001
object = frequent	1.392	0.484	0.004
condition = Accuracy	-5.149	0.781	<0.001
condition = Time	6.031	1.207	<0.001
condition = Combined	0.343	0.746	0.645
object = frequent & condition = Accuracy	-0.722	0.751	0.337
object = frequent & condition = Time	-1.079	1.180	0.360
object = frequent & condition = Combined	2.573	0.709	<0.001

In Fig. 4 we plot participants' 'languages' (the collection of form-meaning mappings produced in their director trials) according to their average token length and the mutual information between their forms *f* and meanings *m*: $\sum_f \sum_m p(f, m) \log \frac{p(f, m)}{p(f)p(m)}$.⁴ The mutual information between the forms and meanings in a participant's lexicon gives us a measure of how predictable the meanings are given the forms and vice versa, and thus tells us how expressive a language is, i.e. how much information is expressed by the forms in the lexicon. The average token length of director trial productions serves as a measure for the effort expended. According to the Principle of

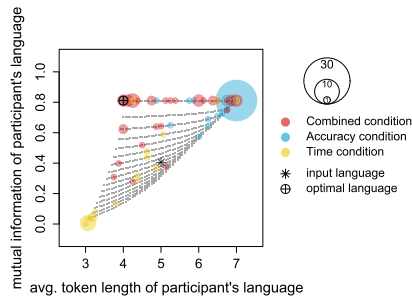


Fig. 4. The average token length of an individual participant's 'language' (the full set of all their director trial productions) plotted against the expressivity (the mutual information between the forms and meanings) of their language. The size of the circles is perceptually scaled (Tanimura et al., 2006) to reflect the number of data points coinciding at each value. The input language that participants are exposed to in training trials is marked with an asterisk, and the grey points represent possible output languages. (Possible output languages are constrained by the number of different expressivity values that are possible for a language with a given average token length. For example, there is only one possible configuration for both the shortest and longest average token lengths—all objects are either mapped to the short name or the long name, respectively—and thus only one possible expressivity value at the endpoints.) The optimal language—the language with the minimum avg. token length while achieving maximum expressivity—is marked with a target symbol.

Least Effort, an optimal language would maximise expressivity while minimising effort. Only participants in the critical Combined condition produce languages which are optimal in this way. Participants in the Accuracy condition gravitate overwhelmingly towards the strategy that maximises expressivity *and* average token length, and participants in the Time condition maintain minimal average token length but sacrifice expressivity to do so; these were the optimal strategies to use in these respective conditions, given the different task demands.

In Fig. 5, we take a closer look at the possible mechanisms behind participants' trial-by-trial production choices in the Combined condition, by measuring the average length of each object's label over successive repetitions. (Note that participants' frequent and infrequent object production trials are randomly shuffled, and thus repetition number does not correspond with a specific spacing of trial numbers.) As discussed in Section 1, earlier studies by Krauss and Weinheimer (1964) and Clark and Wilkes-Gibbs (1986) show that object descriptions tend to shorten with repetition, and that more frequent objects end up with shorter descriptions simply because they go through more repetitions. In these studies, because the meaning space was small compared to the large descriptive space available (i.e., English phrases with no length restriction), all descriptions could be shortened somewhat without producing ambiguous form-meaning mappings. In our study, we investigated the case where a pressure to use shorter forms comes into direct conflict with the pressure to avoid ambiguity: in this miniature lexicon, shortening yields the same, ambiguous label for the two objects in the meaning space.

If participants are simply more likely to use a shorter form for an object the more times they communicate about that object, then we would expect the average label length for both the frequent object and the infrequent object to decrease at a similar rate as the number of repetitions increases. However, as Fig. 5 shows, this is not what we find. Only the average label length of the frequent object decreases with successive repetitions; the average label length of the infrequent object remains roughly constant over the course of the trials. A logistic regression model fit to just the

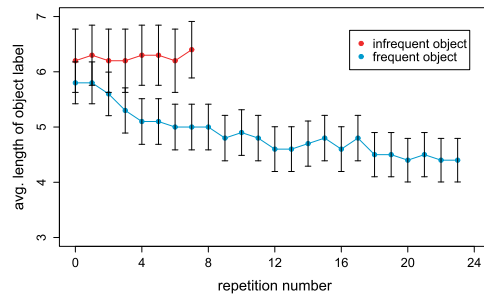


Fig. 5. Timecourse of productions in the critical Combined condition. Each data point shows the average word length taken over all participants' productions at a given repetition number of an object.

data from the Combined condition, with short name use as the binary dependent variable, object frequency, trial number and their interaction as fixed effects, and by-participant random intercepts and slopes for object frequency and trial number, confirms this. The model results (Table 2) show an overall significant positive effect of trial number on short form use only when the object is frequent. Note that there is also a marginal difference between the two objects at repetition number 0. Thus, in the critical Combined condition, while most participants switch to using the short form for the more frequent object at some point during production trials, most also maintain the long form for the infrequent object throughout the trials—the threat of ambiguity appears to block shortening altogether for this object. This suggests that, in cases where the pressure to decrease effort and the pressure to avoid ambiguity come into direct conflict, language-users' production choices result in systems which maximise expressivity while minimising effort, optimising across the lexicon as a whole.

Interestingly, there were a small number of participants (for example in the Combined condition) who consistently mapped the short form to the *infrequent object*. While shortening the label for either object does satisfy the time pressure to some extent, why might this sub-optimal strategy be used? One possibility is that a participant's strategy is not to optimise based on overall frequency distributions within the signalling game, but simply to shorten the first object they are presented with in production trials, which then blocks shortening of the other object. However, of the 10 participants who were presented with the infrequent object first, 30% converged on a 'reversed' or other non-optimal strategy as opposed to the optimal strategy. Of the remaining 30 participants who saw the frequent object first, 37% converged on a reversed or other non-optimal strategy. Thus, which object appeared in the first production trial (or even the first several trials, which we also checked) is not predictive of which strategy (optimal or otherwise) the participants converged on in the critical condition. We believe these occasional reversed lexicons are thus more likely due to an effect of the cost of switching an incipient convention during the task. For example, if a participant starts out producing labels probabilistically, following the language they were trained on, they will sometimes produce a short name for the infrequent object. If this results in successful communication, and is picked up by a communicative partner, then this pattern may become conventionalised. However, once such a pattern is established, the cost of switching to a different mapping becomes an obstacle. The pressure to maximise the number of correct guesses in the testing trials means the cost of switching labels will further penalise participants who attempted to abandon an incipient sub-optimal convention midway through the task.

Table 2

Summary of fixed effects for a binomial regression model with short name use as the binary dependent variable, and by-participant random effects for object frequency and trial number. This model is fitted to the data from all participants' production trials in the Combined condition.

	β	SE	<i>p</i>
intercept (object = infrequent)	-7.115	2.067	0.001
object = frequent	3.949	2.251	0.079
trial number	0.064	0.059	0.279
object = frequent \times trial number	0.137	0.046	0.003

4. Discussion

More than 80 years ago, Zipf hypothesised that the inverse relationship between word length and word frequency was a universal feature of human language, resulting from language users optimising form-meaning mappings for efficient communication. Our study provides direct experimental evidence linking pressures that operate at the level of the individual during communication to the Law of Abbreviation, an emergent structural feature of languages. In particular, language users converge on an optimally-configured lexicon, preferentially using short but potentially ambiguous labels for frequent objects and long labels for infrequent objects. Importantly, this holds only when both a pressure to communicate accurately and a pressure to communicate efficiently are present.

When these pressures were isolated, the Law of Abbreviation did not emerge; an accuracy pressure alone led participants to use the longer non-ambiguous forms regardless of frequency, while a time pressure alone led them to use the short forms. Some participants mapped the short form to the more frequent object in the Neither condition, however the effect was much weaker. Thus, while biases towards accuracy and efficiency might be implicitly present in any linguistic task, emphasising these pressures significantly amplified the effect, as predicted. Even though this experiment involved a miniature lexicon consisting of three possible forms, our result is a proof of concept that such pressures can push a lexicon to align with the Law of Abbreviation. We expect the results to scale up to lexicons with more forms and meanings; with the groundwork in place we can now test this in future studies.

It is important to note, however, that there is a distinction between a language-user's *mental representation* of the lexicon, and the form-meaning mappings they actually produce in communication. Participants using the short form for the frequent object and the long form for the infrequent object may still retain associations of both forms with both objects in their mental lexicon—however, the nature of the communicative task in this experiment may have caused them to produce only the short form for one object and the long form for the other based on purely *pragmatic* considerations (see, e.g., Franke, *in press*). Given that our experiment only recorded participants' actual productions, we cannot with certainty distinguish between these two possible explanations for the observed behaviour. However, we did include an exit survey which asked participants to explain their strategies during the production stage. Some of the language used in the responses suggested that some participants *had* remapped their mental lexicons. E.g., "I waited until my partner sent Zop twice for the blue round object and then we had a mutual understanding that that's what the Zop was" and "the small round object was Zop, and the orange tall figure was the longer word." However, some other participants indicated that they interpreted the short form as either a prefix or convenient shortening—e.g., "one of the objects had to use the long name, as the short Zop was the same prefix for both" and "[I] used just Zop when transmitting Zopekil [as] the other needed

more transmission time"—suggesting that they still retained the long form in their mental lexicon even if they stopped using it.⁵

Our interpretation of such cases is that, while this pragmatics-driven asymmetry in usage may or may not lead to an immediate shift in lexical representations, it may be an important first step in such a change. In English, many words exist that initially began as convenient shortenings of longer forms, which are now either no longer in use, or no longer associated with the same meaning as the short forms. Some examples are: *bus* (from *omnibus*); *wig* (from *periwig*); *pram* (from *perambulator*); *pub* (from *public house*); and *pants* (from *pantaloons*). In all these cases, the clipped form has undergone "opacification", i.e. it is no longer widely recognised as a derivation of the full form, and exists autonomously in the lexicon as an unmarked, standard form (Jamet, 2009). Likewise, even if participants in our experiment are retaining the long form in their mental lexicon, the rapid decrease in its frequency of use over successive generations of learners would likely lead the long form to eventually drop out of the lexicon, with the short form becoming lexicalised as the standard form. Indeed, studies in the iterated learning paradigm show that, in the lexicons produced by successive generations of participants, those in which two labels map to the same meaning are dispreferred (e.g., Reali & Griffiths, 2009; Smith & Wonnacott, 2010). In short, permanent lexical changes often begin life as pragmatics-driven asymmetries in usage (Bybee, 2010). Thus, even if the alignment with the Law of Abbreviation that we observe in participants' usage is not yet accompanied by a corresponding shift in their mental lexicons, it is an important intermediary stage on the way to this outcome.

It is also worth noting that across conditions we found evidence for a baseline preference against ambiguity: when no pressures were present, participants tended towards retaining the unique long forms for *both* objects, and no participants used the ambiguous short names for both objects simultaneously. Indeed, in both conditions featuring a time pressure, a few participants nevertheless used the long names across the board. These results suggest that for some participants, the framing of the task as one of learning a language carries with it some expectation of communicative utility.

Returning to the issue of the explanation for the widespread application of the Law of Abbreviation, our results demonstrate that optimisation behaviour on the part of language-users can lead to the production of lexicons which align with this law. Our study expands on previous work that investigates the relationship between frequency and utterance length, by setting up a small lexicon in which the pressures for efficiency and expressivity in a communicative task come sharply head-to-head. We find that these conflicting pressures do indeed lead language-users to map shorter forms to more frequent meanings, as Zipf hypothesised. However, this result does not rule out that additional processes are involved in shaping this global linguistic pattern as well. Indeed, we expect there are many other factors that come into play as the size of the lexicon is scaled up and the conditions become closer to actual language-use: for example the bottlenecks of learning and memory; the influence of predictability in context; constraints of speech production; and the propagation of errors. There may be a role for random statistical processes to play as well. Future work should focus on how the pressures involved in this task interact with these and other factors, and especially on how the behaviour of individuals communicating in a pair spreads outside this context to the level of an entire population.

⁵ All the exit survey responses are available along with the full dataset at: <http://datashare.is.ed.ac.uk/handle/10283/2702>.

5. Conclusions

Zipf's proposal—that the inverse relationship between a word's length and its frequency is a universal design feature of language—has been borne out repeatedly in observations of the world's languages (Ferrer-i-Cancho & Hernández-Fernández, 2013; Piantadosi et al., 2011; Sigurd et al., 2004; Strauss et al., 2007; Teahan et al., 2000). The long-standing explanation for this phenomenon appeals to the idea that language users want to communicate as efficiently as possible. However, the critical link between this Principle of Least Effort and the emergence of an optimal lexicon has remained largely untested. Our study explored the hypothesis that the mechanisms operating in individual language users during online language production can result in the active restructuring of a lexicon. Our findings reveal that when pressures to communicate accurately and efficiently are both present and in conflict, language users exploit information in the input about the frequency of meanings to converge on an optimally-configured lexicon. When only one of these pressures is present, the effect does not emerge. This result provides evidence that the universal pattern Zipf observed can indeed arise through individual-level optimisation of form-meaning mappings. More generally, this method provides a model for future work showing how explanations of population-level properties of languages can be grounded in the moment-to-moment behaviours of individuals.

Acknowledgments

Thanks to Michael Franke, an anonymous reviewer, Vanessa Ferdinand, Kevin Stadler, and other members of the Centre for Language Evolution for useful feedback and discussion on this work.

References

- Antoine, F. (2000). *An English-French dictionary of clipped words* (Vol. 106). Peeters Publishers.
- Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67, 1–48. <http://dx.doi.org/10.18637/jss.v067.i01>.
- Bybee, J. (2010). *Language, usage and cognition*. Cambridge University Press.
- Clark, H. H., & Wilkes-Gibbs, D. (1986). Referring as a collaborative process. *Cognition*, 22, 1–39.
- Culbertson, J., Smolensky, P., & Legendre, G. (2012). Learning biases predict a word order universal. *Cognition*, 122, 306–329.
- Dautriche, I., Mahowald, K., Gibson, E., Christophe, A., & Piantadosi, S. (in press). Words cluster phonetically beyond phonotactic regularities. *Cognition*, 163, 128–145.
- Davies, M. (2008). *The corpus of contemporary american english: 520 million words, 1990-present*. Available online at <<http://corpus.byu.edu/coca/>>.
- Degen, J., Franke, M., & Jäger, G. (2013). Cost-based pragmatic inference about referential expressions. In *Proceedings of the 35th annual conference of the cognitive science society* (pp. 376–381).
- Ellis, S. R., & Hitchcock, R. J. (1986). The emergence of zipf's law: Spontaneous encoding optimization by users of a command language. *IEEE Transactions on Systems, Man and Cybernetics*, 16, 423–427.
- Fedzechkina, M., Jaeger, T. F., & Newport, E. L. (2012). Language learners restructure their input to facilitate efficient communication. *Proceedings of the National Academy of Sciences*, 109, 17897–17902. <http://dx.doi.org/10.1073/pnas.1215776109>.
- Fehér, O., Wonnacott, E., & Smith, K. (2016). Structural priming in artificial languages and the regularization of unpredictable variation. *Journal of Memory and Language*, 91, 158–180.
- Ferrer-i-Cancho, R., & Hernández-Fernández, A. (2013). The failure of the law of brevity in two new world primates, statistical caveats. *Glottology International Journal of Theoretical Linguistics*, 4, 45–55 <<http://www.degruyter.com/view/j/jglot.2013.4.issue-1/jglot.2013.0004/jglot.2013.0004.xml>>.
- Ferrer-i-Cancho, R., Hernández-Fernández, A., Lusseau, D., Agoramoorthy, G., Hsu, M. J., & Semple, S. (2013). Compression as a universal principle of animal behavior. *Cognitive Science*, 37, 1565–1578.
- Ferrer-i-Cancho, R., & Moscoso del Prado, F. (2012). Information content versus word length in random typing. *JSTAT*.
- Franke, M. (2017). Game theory in pragmatics: Evolution, rationality & reasoning. In *Oxford Research Encyclopedia of Linguistics*, in press.
- Hamming, R. W. (1950). Error detecting and error correcting codes. *Bell Labs Technical Journal*, 29, 147–160.
- Hudson Kam, C. L., & Newport, E. L. (2005). Regularizing unpredictable variation: The roles of adult and child learners in language formation and change. *Language Learning and Development*, 1, 151–195.
- Huffman, D. A. (1952). A method for the construction of minimum-redundancy codes. *Proceedings of the IRE*, 40, 1098–1101.
- Hupet, M., & Chantraine, Y. (1992). Changes in repeated references: Collaboration or repetition effects? *Journal of Psycholinguistic Research*, 21, 485–496.
- Jamet, D. (2009). A morphophonological approach to clipping in English. Can the study of clipping be formalized? *Lexis: Journal in English Lexicology*, HS 1.
- Kirby, S. (2001). Spontaneous evolution of linguistic structure—an iterated learning model of the emergence of regularity and irregularity. *IEEE Transactions on Evolutionary Computation*, 5, 102–110.
- 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*, 105, 10681–10686.
- Kirby, S., Tamariz, M., Cornish, H., & Smith, K. (2015). Compression and communication in the cultural evolution of linguistic structure. *Cognition*, 141, 87–102.
- Krauss, R. M., & Weinheimer, S. (1964). Changes in reference phrases as a function of frequency of usage in social interaction: A preliminary study. *Psychonomic Science*, 1, 113–114.
- Krauss, R. M., & Weinheimer, S. (1966). Concurrent feedback, confirmation, and the encoding of referents in verbal communication. *Journal of Personality and Social Psychology*, 4, 343.
- Moscoso del Prado, F. (2013). The missing baselines in arguments for the optimal efficiency of languages. In *Proceedings of the 35th annual conference of the cognitive science society* (pp. 1032–1037) <<http://csjarchive.cogsci.rpi.edu/Proceedings/2013/papers/0203/paper0203.pdf>>.
- Piantadosi, S. T., Tily, H., & Gibson, E. (2011). Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences*, 108, 3526–3529. <http://dx.doi.org/10.1073/pnas.1012551108>.
- Piantadosi, S. T., Tily, H., & Gibson, E. (2012). The communicative function of ambiguity in language. *Cognition*, 122, 280–291. <http://dx.doi.org/10.1016/j.cognition.2011.10.004>.
- R Core Team (2015). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing <<https://www.R-project.org/>>.
- Real, F., & Griffiths, T. L. (2009). The evolution of frequency distributions: Relating regularization to inductive biases through iterated learning. *Cognition*, 111, 317–328.
- Sigurd, B., Eeg-Olofsson, M., & Van Weijer, J. (2004). Word length, sentence length and frequency-Zipf revisited. *Studia Linguistica*, 58, 37–52.
- Smith, K., & Wonnacott, E. (2010). Eliminating unpredictable variation through iterated learning. *Cognition*, 116, 444–449.
- Strauss, U., Crzybek, P., & Altmann, G. (2007). Word length and word frequency. In *Contributions to the science of text and language* (pp. 277–294). Netherlands: Springer.
- Tanimura, S., Kuroiwa, C., & Mizota, T. (2006). Proportional symbol mapping in R. *Journal of Statistical Software*, 15.
- Teahan, W. J., Wen, Y., McNab, R., & Witten, I. H. (2000). A compression-based algorithm for chinese word segmentation. *Computational Linguistics*, 26, 375–393.
- Vouloumanos, A. (2008). Fine-grained sensitivity to statistical information in adult word learning. *Cognition*, 107, 729–742.
- Winters, J., Kirby, S., & Smith, K. (2015). Languages adapt to their contextual niche. *Language and Cognition*, 7, 415–449.
- Zipf, G. K. (1935). *The psycho-biology of language* (Vol. ix) Oxford, England: Houghton Mifflin.

Language-users choose short words in predictive contexts in an artificial language task

Jasmeen Kanwal (jasmeen.kanwal@ed.ac.uk), Kenny Smith (kenny.smith@ed.ac.uk), Jennifer Culbertson (jennifer.culbertson@ed.ac.uk), and Simon Kirby (simon@ling.ed.ac.uk)

Centre for Language Evolution, University of Edinburgh
3 Charles St., Edinburgh, Scotland EH8 9AD

Abstract

Zipf (1935) observed that word length is inversely proportional to word frequency in the lexicon. He hypothesised that this cross-linguistically universal feature was due to the *Principle of Least Effort*: language-users align form-meaning mappings in such a way that the lexicon is optimally coded for efficient information transfer. However, word frequency is not the only reliable predictor of word length: Piantadosi, Tily, and Gibson (2011) show that a word's predictability in context is in fact more strongly correlated with word length than word frequency. Here, we present an artificial language learning study aimed at investigating the mechanisms that could give rise to such a distribution at the level of the lexicon. We find that participants are more likely to use an ambiguous short form in predictive contexts, and distinct long forms in surprising contexts, only when they are subject to the competing pressures to communicate accurately and efficiently. These results support the hypothesis that language-users are driven by a least-effort principle to restructure their input in order to align word length with information content, and this mechanism could therefore explain the global pattern observed at the level of the lexicon.

Keywords: Information theory; Efficient communication; Artificial language learning; Uniform Information Density

Introduction

Zipf (1935) observed that word length tends to be inversely proportional to word frequency in the lexicon. He hypothesised that this widespread cross-linguistic pattern was due to the *Principle of Least Effort*: language-users align form-meaning mappings in such a way that effort is minimised while expressivity is still maintained. However, word frequency is not the only reliable predictor of word length. Using corpora from 11 different languages, Piantadosi et al. (2011) show that a word's predictability in context (where they define context as the two words preceding the target word) is even more strongly correlated with word length than frequency is: words that are, on average, more predictable in context tend to be shorter.

Measuring how predictable or unpredictable a word is in a particular context gives us a way of defining the *information content* of a word. For example, consider the two sentences:

- (1) The early bird catches the worm.
- (2) Our early bird special today is a baked-apple worm.

In sentence (1), a well-known proverb, the word *worm* is entirely predicted by the preceding words. The word itself thus gives us practically no new information, and so it has *low information content*. In sentence (2), the same word is highly unlikely given the preceding words, and thus we find it surprising. This element of surprise is associated with *high information content*.

Using these concepts, we can apply Zipf's Principle of Least Effort to hypothesise that a speaker's drive to reduce effort will be directed towards words that are already highly predictable given the context, i.e. have low information content. Words that are more surprising in a particular context will be less likely to be reduced, or more likely to be lengthened. The resulting state in which low-information words are shorter than high-information words, and thus the length of a word is roughly proportional to the amount of information associated with the word, is consistent with the Uniform Information Density (UID) principle (Jaeger, 2010) or the Smooth Signal Redundancy (SSR) hypothesis (Aylett & Turk, 2004), which state that information is distributed roughly evenly across words in an utterance.

There are many ways to operationalise the information content of a word. One way is to use the *N-gram probability* of a word, i.e. its probability conditioned on a window of N preceding or following words. This is the method used by Piantadosi et al. (2011). Zipf's word frequency measure is in fact just a limiting case of this N-gram probability, where N=0. Other measures include *syntactic probability*, a word's probability of appearing in a particular syntactic structure (Jaeger, 2010, e.g.), and *givenness*, a word's predictability given the semantic context (Aylett & Turk, 2004).

Both corpus studies and controlled behavioural experiments have linked low information content, operationalised in these different ways, to various types of linguistic reduction. Lieberman (1963); Aylett and Turk (2004); Gahl and Garnsey (2004); Tily et al. (2009); Kuperman and Bresnan (2012), and Seyfarth (2014) show that words with low information content are more likely to undergo various types of phonetic reduction. Bell, Brenier, Gregory, Girand, and Jurafsky (2009) show that each of the four different measures of information content mentioned above may in fact contribute separately to the phonetic duration of a word. Fedzechkina, Jaeger, and Newport (2012) show that case markers are more likely to be omitted on nouns in more probable syntactic roles. Jaeger (2010) shows that *that*-complementisers are more often dropped when the following word is less surprising in context.

If predictability in context can lead to phonetic reduction, as well as deletion of morphemes and entire words, then these effects might make their way to the overall distribution of form-meaning mappings in the lexicon. However, there is relatively little work directed at understanding how predictability affects this widely observed pattern at the level of the lex-

icon.

One way of investigating the issue is by tracking language-users' online choices when producing words that are part of a 'clipped pair', i.e. when both a long form and an abbreviated or 'clipped' form exist that have the same or very similar meanings (Mahowald, Fedorenko, Piantadosi, & Gibson, 2013). E.g. in English, *info/information* is a clipped pair. Mahowald et al. presented participants with sentences containing a blank and asked them to complete the sentence with either the long or the clipped form corresponding to the relevant meaning. They found that participants were more likely to choose the short form in predictive contexts, which is consistent with the hypothesis that the lexicon-level patterns observed by Piantadosi et al. (2011) may be due in part to a *least-effort* mechanism, in which speakers balance communicative efficacy with efficiency.

However, because this study uses English sentence frames and target words, we cannot rule out potentially confounding contributions from register, prosody, and participants' learned preferences to their word choice in particular instances. Moreover, we cannot assess whether the effect is really driven by the competing pressures for communicative accuracy and efficiency without manipulating the presence or absence of these different communicative pressures. For instance, in Mahowald et al.'s task, it seems participants clicked on a word rather than typing it in, and thus there was no difference in effort between choosing the long or short form. In addition, participants were told to choose a word based on "which sounded more natural", rather than being directly engaged in a task requiring successful communication.

Here, we present a new artificial language learning study investigating the question of whether language-users align word length with information content when communicating. Our results are consistent with previous findings that language-users tend to use shorter forms in more predictive contexts. Furthermore, the behaviour we observe across different experimental conditions supports the hypothesis that this effect is driven at least in part by a least-effort principle, in which language-users balance the competing pressures for communicative accuracy and efficiency to reshape the lexicon into one where word length is roughly proportional to average information content.

Method

Artificial language learning studies have previously been used to shed light on the cognitive mechanisms and environmental pressures that shape large-scale linguistic structure. In this paradigm, participants learn an artificial language, and then we observe how they reshape their input as they use the language, in this case to communicate with a partner (e.g., Winters, Kirby, & Smith, 2015; Kirby, Tamariz, Cornish, & Smith, 2015; Fehér, Wonnacott, & Smith, 2016).

Participants

120 participants (53 females, 66 males; one did not report their gender) were recruited and remunerated via Amazon

Mechanical Turk. 108 of these reported themselves as native English speakers, of which 96 were monolingual. A range of other languages were represented across the remaining participants. Ages ranged from 18 to 70 (mean=32.9, SD=9.5).

The Training Language

The study was run online. Participants were trained on two names for each of two plant-like alien objects, by repeatedly being shown pictures of the objects labeled with a simple sentence. The sentence consisted of a framing word followed by the object's name. There were two possible frames, *bix* and *gat*. Overall there were 64 training trials, with each object appearing 32 times and each frame appearing 32 times. Crucially, one object appeared seven times more frequently with the frame *bix* than *gat* (28 and 4 times, respectively), while the other object appeared seven times more frequently with the frame *gat* than *bix* (again, 28 and 4 times, respectively). This meant that each object appeared in both a predictive context and a surprising context; which frame signified which of these contexts was flipped between the two objects.

Furthermore, the object name appeared half the time in its full form, a 7-letter word, and half the time in shortened form, a 3-letter word derived by clipping the last two syllables off the long name. These short and long forms were evenly distributed across both predictive and surprising contexts, ensuring that the input language contained no inbuilt bias towards using one form in any particular context.¹ A schematic diagram of the object frequencies and labels is provided in Fig. 1A.

In natural languages, shorter words are subject to greater confusability for a number of reasons: shorter forms have less space for signal redundancy and thus are more likely to be completely lost in noisy signal transmission; and because languages have a finite phoneme inventory, there are more unique possible long strings than short strings, and thus word shortening often results in ambiguity. Indeed, shorter words are more likely to be polysemous and homophonous (Piantadosi, Tily, & Gibson, 2012). To model this fact in our miniature lexicon, we designed the names such that the short name for both objects was identical (*zop*), while the long names were unique (*zopekil* and *zopudon*).

Procedure

Participants were assigned to one of four conditions, where we manipulated the presence of pressures to communicate accurately and quickly in a between-subjects 2x2 design (Kanwal, Smith, Culbertson, & Kirby, 2017). Each experiment consisted of two phases: training and testing. The training phase was uniform across conditions, while the testing phase varied by condition.

¹Which object (the blue fruit or the red stalk) appeared more frequently with which frame, as well as which object was paired with which long name, were both counterbalanced between participants, giving a total of 4 possible object-frame-name pairings which a participant might be trained on. This ensured that potential factors such as sound symbolism, or higher saliency or learnability of any specific object-word pairing, could not systematically bias our results.

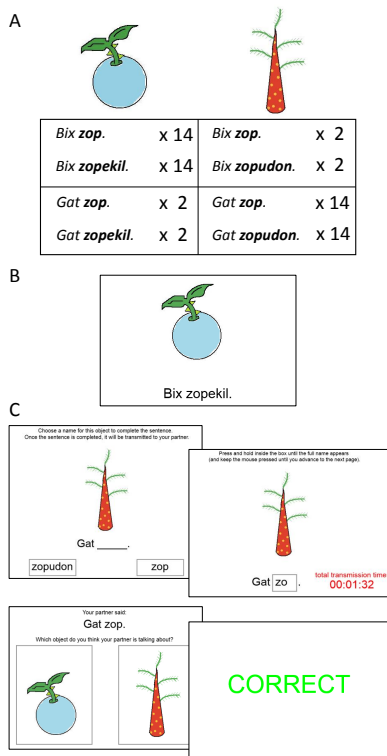


Figure 1: (A) The input frequencies of the objects and framing sentences presented during training trials in all four experimental conditions. (B) A sample training trial. (C) A sample director trial in the Combined condition (top) and a matcher trial followed by feedback (bottom).

Training phase On each training trial, an object was presented on screen alone for 700ms. The appropriate sentence then appeared beneath the object for a further 3000ms, yielding a total trial duration of 3700ms. A blank screen showed for 500ms between trials. The 64 training trials were presented in a different randomised order for each participant.

Testing phase After the training phase, testing procedures varied depending on the experimental condition. In the Combined condition, participants were under a pressure to communicate accurately *and* efficiently, as according to the Principle of Least Effort, it is balancing these competing pressures that leads language-users to distribute word length inversely to word predictability. The remaining three conditions removed one or both of these accuracy and time pressures.

Condition 1: Combined In the testing phase of this condition, participants were paired with a partner to play a communication game, using the method developed for running two-player online experiments in Kanwal et al. (2017). On

each trial, the ‘director’ was shown an object on the screen with a framing word followed by a blank. The director was instructed to choose a name for the object to complete the sentence, and once the name was entered, the sentence would be transmitted to the ‘matcher’. The director could choose one of two options to complete the sentence: the unique long name for the object or the (ambiguous) short name. Once the chosen name was selected by clicking on the appropriately labeled button, it had to be entered into the blank space by pressing and holding the mouse as each letter appeared one after the other at 1200 ms intervals. Only after all the letters in the name had appeared in the box was the completed sentence transmitted to the matcher. This belaboured method of production, in which the long name was significantly slower to produce than the short name, was introduced to model the difference in effort and speed associated with producing long versus short utterances.

Once the director completed their description, it was transmitted to the matcher, who was asked to choose which of the two objects they thought the director was referring to. Both players were then given feedback as to whether the matcher’s choice was correct.

The players alternated roles after every trial, with the matcher becoming the director and the director becoming the matcher, until both completed 32 director trials and 32 matcher trials. The proportion of times each object appeared with each frame in each player’s director trials matched those of the training proportions: one object appeared seven times more frequently with the frame *gat* than *bix*, and the other appeared seven times more frequently with *bix* than *gat*. The order of each participant’s 32 director trials was randomly shuffled.

To model the pressures in spoken communication to be both efficient and accurate, pairs were told at the beginning that they would be rewarded a bonus payment of \$1 if they were the pair to complete the game in the quickest time with the highest number of correct match trials. Time was only counted when the director was entering a name into the blank, and the total time count was displayed next to the blank during this process, to emphasise the time pressure. Screenshots of sample director and matcher trials are shown in Fig. 1C.

In this condition, with pressures to be speedy yet accurate, we expected participants to converge on an optimal strategy in which the short name is used for an object when it appears in its predictive context, and the long name otherwise. In predictive contexts, the framing word already provides a lot of information to the matcher about which object is likely under discussion, and thus participants can minimise effort by using the short form. Conversely, in surprising contexts, the full object name is required to ensure disambiguation.

In order to establish a causal link between these purported mechanisms and the behaviour we observe, we included three further experimental conditions, described below, for a full 2x2 manipulation of the pressures for accuracy and efficiency.

Condition 2: Accuracy In this condition, participants were paired to play a communication game as described above, but in the director trials, there was no intermediate step between the director choosing a name to complete the sentence and the matcher receiving the sentence; the names were entered instantaneously, thus removing any difference in effort between producing long or short names. Pairs were told that the goal of the game was simply to have their partner make as many correct guesses as possible. No bonus prize was offered in this condition, as we expected many pairs to hit ceiling as they did in Kanwal et al. (2017)—however, fewer than expected actually did so here.

In this condition, we predicted that participants would be more likely to use the long names for both objects across all contexts, as the long names are less confusable, and without a pressure to be efficient, there is little reason to shorten.

Condition 3: Time In this condition, communication was taken out of the game entirely; participants played a one-player game consisting of 64 director trials. In each trial, participants completed the sentence with either the long or short name for the object shown, but there was no subsequent communicative task. The name was simply entered as in the Combined condition, by pressing and holding the mouse in the blank space, with each letter appearing at 1200 ms intervals, while a timer displayed the total time count. The next trial began once all the letters had appeared in the box. Participants were told at the beginning of the game that they would be rewarded a bonus payment of \$1 if they were the player with the shortest total time count.

Here, we expected participants to use the short name for both objects across all contexts: with no communicative purpose attached to the transmissions, and an incentive to be as quick as possible, using the short name in every trial is the best strategy.

Condition 4: Neither The fourth and last condition contained neither a pressure for efficiency nor a pressure for accuracy. As in the Time condition, participants played a one-player game with no explicit communicative element. Additionally, there was no time difference associated with transmission; once a label was chosen to complete a sentence, it was instantaneously recorded and the player advanced to the next trial. We included this condition to provide a baseline for participants' behaviour from which to assess the effects of the accuracy and time pressures in the other three conditions.

In this condition we expected that participants would either probability-match—i.e. use the long and short forms for both objects with equal frequency, as in the training trials (Hudson Kam & Newport, 2005)—or their behaviour would reveal prior biases language users bring to the task, such as a preference against using ambiguous forms, as observed in Kanwal et al. (2017).

Results

Fig. 2 shows the proportion of trials in which the short name was produced by each participant or pair of participants in

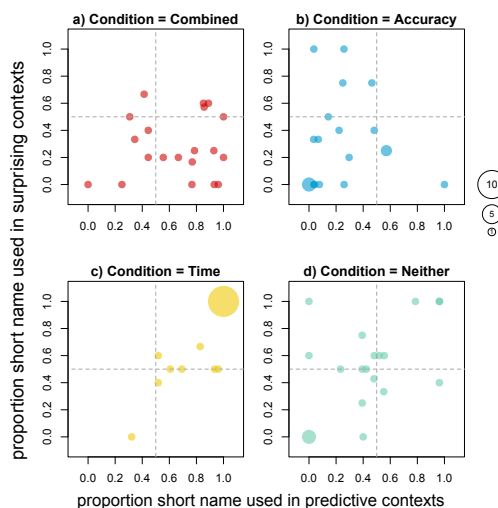


Figure 2: The proportion of trials in which the short name was used in predictive contexts versus the proportion of trials in which it was used in surprising contexts. For the Combined and Accuracy condition, each data point combines a pair of communicating players, representing the sum of their director trial productions. For the Time and Neither condition, each data point corresponds to an individual player's productions. The size of the circles is perceptually scaled (Tanimura et al., 2006) to reflect the number of data points coinciding at each value. Data from only the second half of testing trials is shown here, as participants were more likely to have converged on a stable mapping by this time. These results demonstrate that behaviour consistent with the principles of UID or SSR—using short forms in predictive contexts and long forms in surprising contexts, generating systems that fall in the bottom right corner of each graph—only reliably arises in the Combined condition.

predictive versus surprising contexts. Our predictions were borne out by the results in all four conditions. In the critical Combined condition, in which participants were subject to the combined pressures for accuracy and efficiency, pairs of communicating participants produced systems in which the short name was used in predictive contexts and the long name in surprising contexts. Crucially, only when both pressures were present did participants reliably produce systems where word length was conditioned on context in this way. In the Accuracy condition, participants tended to use the long name for both objects regardless of context, and in the Time condition, they used the short name for both objects regardless of context. In the Neither condition, some participants stuck with the long name or the short name throughout the trials regardless of context, as in the Accuracy or Time conditions; however, most participants probability-matched.

A logistic regression model was fit to the full dataset in R using the lme4 package, with short name use (as contrasted with long name use) as the binary dependent variable; context (predictive or not), experimental condition, and their interaction as fixed effects; and by-participant random slopes

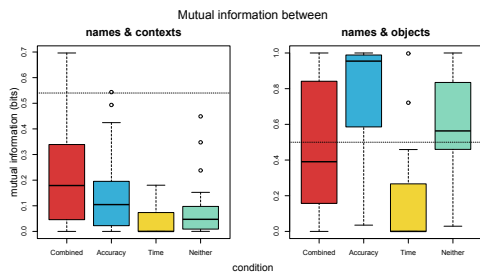


Figure 3: This figure shows the extent to which participants’ name choices are conditioned on context (lefthand graph) and object (righthand graph). The dotted line in the lefthand graph represents the mutual information (MI_C) associated with the ‘optimal’ language in least-effort terms—the language in which the short form is used only in predictive contexts, and the long form only in surprising contexts. $MI_C=0$ for the input language. In the righthand graph, mutual information (MI_O) can range from 0 (same name fixed for both objects) to 1 (distinct names fixed for each object). $MI_O=0.5$ for the input language, marked by the dotted line. Data from only the second half of testing trials is shown in this figure, as participants were more likely to have converged on a stable mapping by this time.

and intercepts for context. The model was sum coded, setting the grand mean as the intercept, to which each level was then compared. The results yielded a significant positive interaction of context in the critical Combined condition ($\beta = 0.619, SE = 0.158, p < 0.001$), indicating that in this condition, participants were significantly more likely to use the short name in predictive contexts. The only other significant effects found were as follows: a positive overall effect in the Time condition ($\beta = 2.187, SE = 0.292, p < 0.001$), indicating that participants were more likely to use the short form in this condition regardless of context; a negative overall effect in the Accuracy condition ($\beta = -1.470, SE = 0.233, p < 0.001$), indicating that participants were *less* likely to use the short form in this condition regardless of context; and finally a negative interaction effect of context in the Accuracy condition ($\beta = -0.490, SE = 0.161, p = 0.002$), indicating that in fact participants were *even* less likely to use the short form in the predictive context in this condition.

An analysis of how participants conditioned the variation in their name usage sheds further light on the differing patterns of behaviour seen across conditions. We calculated the average mutual information between name produced and context (predictive or not) in each participant’s output language (MI_C). The more reliably participants are conditioning their use of the long and short names on context, the higher we would expect the value of MI_C to be. The distributions for all four conditions are plotted on the lefthand graph of Fig. 3.

We also calculated the average mutual information between name produced and *object* (the blue fruit or the red stalk) in each participant’s output language (MI_O). This measure allows us to determine whether some participants are us-

ing fixed names for each object, regardless of context. The results are plotted by condition in the righthand graph of Fig. 3. If participants are using a distinct name for each object, MI_O will be close to 1; if they are using the same name for both objects, MI_O will be close to 0. The former pattern is what we see in the Accuracy condition: most participants use the unique long name for each object, regardless of context. The latter pattern is what we see in the Time condition: most participants use the ambiguous short form for both objects, regardless of context.

In the Combined and Neither conditions, MI_O hovers around that of the input language. Based on this graph alone, participants may be probability matching in both these conditions, or perhaps reliably conditioning their output on other factors. Looking back at MI_C disambiguates: it is significantly higher in the Combined condition than in any other condition. A linear regression on MI_C with condition as predictor variable (fit to the second half of testing trials, as in Fig. 3) yielded a significant negative effect of the Accuracy ($\beta = -0.081, SE = 0.033, p = 0.016$), Time ($\beta = -0.184, SE = 0.041, p < 0.001$), and Neither ($\beta = -0.128, SE = 0.041, p = 0.002$) conditions, with the Combined condition set as the intercept. This result is consistent with what we saw in Fig. 2: in the Combined condition, many participants are optimally conditioning their responses on context, generating systems that fall in the bottom right corner of the graph; in the other conditions, almost no data points fall in this region.

Discussion

There is mounting evidence that utterance length is linked to information content (Lieberman, 1963; Aylett & Turk, 2004; Gahl & Garnsey, 2004; Tily et al., 2009; Bell et al., 2009; Jaeger, 2010; Piantadosi et al., 2011; Kuperman & Bresnan, 2012; Fedzechkina et al., 2012; Seyfarth, 2014). The explanation put forth in much of this previous work is that speakers are driven by pressures much like those outlined in Zipf’s Principle of Least Effort: the competing demands for accurate and efficient communication lead speakers to converge on an optimal system in which information content is spread roughly uniformly across the utterance, resulting in low-information units being shorter than high-information units. This resultant effect appears to have made its way into the structure of the lexicon as a whole: shorter words appear on average in more predictive contexts than longer words (Piantadosi et al., 2011). But is this effect really due to the proposed mechanism? Can speaker choice lead to the reshaping of a lexicon to align it with the principles of Uniform Information Density and Smooth Signal Redundancy?

Here, we presented the first study that concretely addresses these questions. Previous studies either lacked a manipulation of the communicative pressures operating in the task, or lacked a communicative element entirely. In our study, by observing participants’ online behaviour in a task in which the pressures to communicate accurately and efficiently were manipulated across four experimental conditions, we have

shown that participants use shorter words in more predictive contexts *only* when both competing pressures were acting on them. When these pressures were isolated or removed entirely, participants failed to reliably condition their word choices on context.

Furthermore, because our study employed an artificial language learning paradigm, our findings avoid potential confounds from factors such as register, prosody, and participants' learned preferences in their native or second languages. Our results are nevertheless consistent with previous findings that language-users tend to use shorter forms in more predictive contexts when using their native language.

Our results serve as a proof of concept that the lexicon-level effect observed by Piantadosi et al. (2011) could be driven at least in part by a least-effort principle in which language-users balance the competing pressures for communicative accuracy and efficiency to reshape the lexicon into one where word length is roughly proportional to information content. However, there is a crucial step between what we have observed here—language-users alternating between long and short variants for a single meaning depending on context—and what Piantadosi et al. (2011) observed in the lexicon of different languages, where most meanings don't correspond to both a long and a clipped variant, but rather map to a single fixed form. For these cases, which make up the majority of the lexicon, the length of the form is strongly correlated with the *average* predictability-in-context of the meaning, across all its different occurrences. We can hypothesise a link between these two phenomena: as a word appears in increasingly more predictive contexts, a reduced variant may come into use. If speakers use the reduced variant in predictive contexts, then this reduced form will consequently become much more frequent than the long form, leading to the long form eventually dying out altogether. This would end in a scenario where a short word, with no alternative variants currently in use, appears on average in a high number of predictive contexts, and thus has a low average information content. Though this story sounds reasonable, a precise mechanistic explanation of how this preference for short forms in more predictive contexts leads to permanent shifts in form-meaning mappings has yet to be thoroughly investigated. We hope this topic is given more attention in future work.

References

- Aylett, M., & Turk, A. (2004). The smooth signal redundancy hypothesis: A functional explanation for relationships between redundancy, prosodic prominence, and duration in spontaneous speech. *Language and speech*, 47(1), 31–56.
- Bell, A., Brenier, J. M., Gregory, M., Girand, C., & Jurafsky, D. (2009). Predictability effects on durations of content and function words in conversational english. *Journal of Memory and Language*, 60(1), 92–111.
- Fedzechkina, M., Jaeger, T. F., & Newport, E. L. (2012, October). Language learners restructure their input to facilitate efficient communication. *Proceedings of the National Academy of Sciences*, 109(44), 17897–17902.
- Fehér, O., Wonnacott, E., & Smith, K. (2016). Structural priming in artificial languages and the regularisation of unpredictable variation. *Journal of Memory and Language*, 91, 158–180.
- Gahl, S., & Garnsey, S. M. (2004). Knowledge of grammar, knowledge of usage: Syntactic probabilities affect pronunciation variation. *Language*, 748775.
- Hudson Kam, C. L., & Newport, E. L. (2005). Regularizing unpredictable variation: The roles of adult and child learners in language formation and change. *Language learning and development*, 1(2), 151–195.
- Jaeger, T. F. (2010, August). Redundancy and reduction: Speakers manage syntactic information density. *Cognitive Psychology*, 61(1), 23–62.
- Kanwal, J., Smith, K., Culbertson, J., & Kirby, S. (2017). Zipf's law of abbreviation and the principle of least effort: Language users optimise a miniature lexicon for efficient communication. *Cognition*. (In press.)
- Kirby, S., Tamariz, M., Cornish, H., & Smith, K. (2015). Compression and communication in the cultural evolution of linguistic structure. *Cognition*, 141, 87–102.
- Kuperman, V., & Bresnan, J. (2012). The effects of construction probability on word durations during spontaneous incremental sentence production. *Journal of Memory and Language*, 66(4), 588–611.
- Lieberman, P. (1963). Some effects of semantic and grammatical context on the production and perception of speech. *Language and speech*, 6(3), 172–187.
- Mahowald, K., Fedorenko, E., Piantadosi, S. T., & Gibson, E. (2013, February). Info/information theory: Speakers choose shorter words in predictive contexts. *Cognition*, 126(2), 313–318.
- Piantadosi, S. T., Tily, H., & Gibson, E. (2011). Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences*, 108(9), 3526–3529.
- Piantadosi, S. T., Tily, H., & Gibson, E. (2012). The communicative function of ambiguity in language. *Cognition*, 122(3), 280–291.
- Seyfarth, S. (2014). Word informativity influences acoustic duration: Effects of contextual predictability on lexical representation. *Cognition*, 133(1), 140–155.
- Tanimura, S., Kuroiwa, C., Mizota, T., et al. (2006). Proportional symbol mapping in R. *Journal of Statistical Software*, 15(i05).
- Tily, H., Gahl, S., Arnon, I., Snider, N., Kothari, A., & Bresnan, J. (2009). Syntactic probabilities affect pronunciation variation in spontaneous speech. *Language and Cognition*, 1(2), 147–165.
- Winters, J., Kirby, S., & Smith, K. (2015). Languages adapt to their contextual niche. *Language and Cognition*, 7(3), 415–449.
- Zipf, G. K. (1935). *The psycho-biology of language* (Vol. ix). Oxford, England: Houghton Mifflin.

Appendix B

Word lists used in Chapter 4 (Experiments 5-7)

Table B.1: List of English clipped pairs used in Experiment 5.

short form	long form	short form	long form	short form	long form
admin	administration	dorm	dormitory	piano	pianoforte
advert	advertisement	eco	ecological	pic	picture
ag	agricultural	exam	examination	pics	pictures
ammo	ammunition	exec	executive	pol	politician
aqua	aquamarine	expat	expatriate	polio	poliomyelitis
berg	iceberg	expo	exposition	porn	pornography
bi	bisexual	fab	fabulous	prof	professor
bike	bicycle	fax	facsimile	prot	protestant
bio	biography	fem	female	pup	puppy
biotech	biotechnology	flu	influenza	rad	radical
biz	business	frat	fraternity	ref	referee
bod	body	fridge	refrigerator	regs	regulations
boob	booby	grad	graduate	rehab	rehabilitation
boobs	boobies	grid	gridiron	roach	cockroach
bot	robot	hubby	husband	sax	saxophone
bra	brassiere	hyper	hyperactive	servo	servomechanism
bro	brother	info	information	sig	signature
burger	hamburger	intro	introduction	sis	sister
carbs	carbohydrates	limo	limousine	stats	statistics
celeb	celebrity	lit	literature	synth	synthesizer
cello	violincello	logo	logotype	teen	teenager
cert	certainty	math	mathematics	telecom	telecommunications
champ	champion	med	medical	toon	cartoon
clit	clitoris	memo	memorandum	tv	television
combo	combination	metro	metropolitan	undies	underwear
comfy	comfortable	mil	million	uni	university
comm	communication	mis	miserable	vid	video
condo	condominium	nav	navigator	vom	vomit
cop	copper	neg	negative	za	pizza
decal	decalcomania	phone	telephone	zine	fanzine
disco	discotheque	photo	photograph		

Table B.2: List of English *-ic/-ical* pairs used in Experiment 6. Only the short form from each pair is included here for reasons of space.

acoustic	botanic	elliptic	lexicographic	phantasmagoric	stoic
aerodynamic	brahmanic	emblematic	linguistic	pharisaic	strategic
aeronautic	brahminic	emphatic	liturgic	pharmaceutic	stratigraphic
aesthetic	calendric	empiric	macroscopic	phenotypic	subtropic
agronomic	canonic	energetic	magnetic	philanthropic	symbolic
ahistoric	cartographic	enigmatic	majestic	philosophic	symmetric
alchemic	casuistic	epic	mathematic	phonetic	syndic
algebraic	categoric	epigraphic	metallurgic	photoelectric	synodic
allegoric	characteristic	episodic	metalogic	photographic	synoptic
alphabetic	chemic	eremitic	metaphoric	phylogenetic	syntactic
alphanumeric	chimeric	ethnic	methodic	physiognomic	synthetic
amic	chronic	ethnographic	metonymic	physiographic	systematic
anagogic	classic	ethnohistoric	microanalytic	phytogeographic	talmudic
analogic	cleric	eugenic	microscopic	piratic	taxonomic
analytic	climatic	evangelic	mineralogic	poetic	technic
anarchic	conic	exegetic	monarchic	politic	thaumaturgic
anatomic	cosmic	fanatic	mythic	polytechnic	theatric
angelic	cosmogonic	fantastic	neoclassic	postclassic	thematic
anthropometric	cosmographic	galenic	neuroanatomic	pragmatic	theoretic
antisymmetric	cubic	gastronomic	nonanalytic	preclassic	theosophic
antithetic	cyclic	genealogic	noncyclic	prehistoric	therapeutic
apocalyptic	cylindric	genetic	nonelectric	problematic	thermodynamic
apologetic	cytogenetic	genotypic	nonmetric	prophetic	thetic
apostolic	deistic	geographic	nonnumeric	prototypic	topographic
aristocratic	demagogic	geometric	numeric	psychiatric	tragic
arithmetic	democratic	grammatic	obstetric	psychoanalytic	trigonometric
artistic	despotic	hagiographic	oceanographic	puritanic	typic
ascetic	diabolic	hemispheric	oligarchic	pyrotechnic	typographic
aspheric	diagrammatic	hermetic	optic	rabbinic	tyrannic
astronomic	dialogic	heroic	organic	radiographic	uneconomic
asymmetric	diametric	hierarchic	orographic	rhapsodic	unhistoric
atheistic	didactic	historic	orthographic	rhythmic	unphilosophic
atmospheric	dogmatic	historiographic	palaeogeographic	sabbatic	unpoetic
autobiographic	domic	homiletic	palaeographic	satanic	unproblematic
automatic	dramatic	hydrodynamic	paleogeographic	satiric	unsymmetric
axisymmetric	dramaturgic	hydrographic	paleographic	satyric	zoogeographic
barometric	druidic	hyperbolic	panegyric	schismatic	
basilic	ecclesiastic	hypocritic	parabolic	scientific	
bathymetric	economic	hypothetic	paradoxic	semantic	
bibliographic	ecumenic	hysteric	parasitic	sophic	
biochemic	egoistic	iconographic	parenthetic	sophistic	
bioelectric	egotistic	identic	pathetic	specific	
biogeographic	electric	ironic	pedagogic	spectroscopic	
biographic	electrodynamic	juridic	periodic	spheric	
biometric	electrooptic	kinematic	petrographic	stereotypic	

Table B.3: List of French clipped pairs used in Experiment 7.

short form	long form	short form	long form
asso	association	impro	improvisation
collec	collection	info	information
compil	compilation	interro	interrogation
compo	composition	intox	intoxication
conso	consommation	intro	introduction
corpo	corporation	manif	manifestation
déco	décoration	promo	promotion
démo	démonstration	récré	récréation
dissert	dissertation	rédac	rédaction
distrib	distribution	réduc	réduction
éva	évaluation	relègue	relégation
expo	exposition	sono	sonorisation
fedé	fédération		

Bibliography

- Antieau, L. (2013). George Kingsley Zipf. In Strazny, P., editor, *Encyclopedia of Linguistics*, pages 1205–1206. Routledge.
- Antoine, F. (2000). *An English-French dictionary of clipped words*, volume 106. Peeters Publishers.
- ATILF - CNRS and Université de Lorraine (2016). Base textuelle FRANTEXT. <http://www.frantext.fr>.
- Aylett, M. and Turk, A. (2004). The smooth signal redundancy hypothesis: A functional explanation for relationships between redundancy, prosodic prominence, and duration in spontaneous speech. *Language and Speech*, 47(1):31–56.
- Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48.
- Bell, A., Brenier, J. M., Gregory, M., Girand, C., and Jurafsky, D. (2009). Predictability effects on durations of content and function words in conversational english. *Journal of Memory and Language*, 60(1):92–111.
- Bezerra, B. M., Souto, A. S., Radford, A. N., and Jones, G. (2011). Brevity is not always a virtue in primate communication. *Biology Letters*, 7(1):23–25.
- Brants, T. and Franz, A. (2006). Web 1T 5-gram Version 1 LDC2006T13.
- Bybee, J. (2006). *Frequency of use and the organization of language*. Oxford University Press.
- Bybee, J. (2010). *Language, usage and cognition*. Cambridge University Press.
- Chomsky, N. (1965). *Aspects of the theory of syntax*. MIT Press.
- Chomsky, N. (2011). Language and other cognitive systems. What is special about language? *Language Learning and Development*, 7(4):263–278.
- Christiansen, M. H. and Chater, N. (2008). Language as shaped by the brain. *Behavioral and Brain Sciences*, 31(5):489–509.

- Clark, H. H. and Wilkes-Gibbs, D. (1986). Referring as a collaborative process. *Cognition*, 22(1):1–39.
- Culbertson, J., Smolensky, P., and Legendre, G. (2012). Learning biases predict a word order universal. *Cognition*, 122(3):306–329.
- Dautriche, I., Mahowald, K., Gibson, E., Christophe, A., and Piantadosi, S. T. (2017). Words cluster phonetically beyond phonotactic regularities. *Cognition*, 163:128–145.
- Davies, M. (2008). The corpus of contemporary american english: 520 million words, 1990-present. Available online at <http://corpus.byu.edu/coca/>.
- Degen, J., Franke, M., and Jäger, G. (2013). Cost-based pragmatic inference about referential expressions. In *Proceedings of the 35th annual conference of the Cognitive Science Society*, pages 376–381.
- Ellis, S. R. and Hitchcock, R. J. (1986). The emergence of Zipf’s law: Spontaneous encoding optimization by users of a command language. *IEEE Transactions on Systems, Man and Cybernetics*, 16(3):423–427.
- Esper, E. A. (1925). A technique for the experiment investigation of associative interference in artificial linguistic material. *Language Monographs*.
- Esper, E. A. (1966). Social transmission of an artificial language. *Language*, 42(3):575–580.
- Esper, E. A. (1973). *Analogy and association in linguistics and psychology*. University of Georgia Press.
- Evans, N. and Levinson, S. C. (2009). The myth of language universals: Language diversity and its importance for cognitive science. *Behavioral and Brain Sciences*, 32(5):429–448.
- Fedzechkina, M., Jaeger, T. F., and Newport, E. L. (2012). Language learners restructure their input to facilitate efficient communication. *Proceedings of the National Academy of Sciences*, 109(44):17897–17902.
- Fehér, O., Wonnacott, E., and Smith, K. (2016). Structural priming in artificial languages and the regularization of unpredictable variation. *Journal of Memory and Language*, 91:158–180.
- Fenk, A. and Fenk, G. (1980). Konstanz im kurzzeitgedächtnis-konstanz im sprachlichen informationsfluß. *Zeitschrift für Experimentelle und Angewandte Psychologie*, 27:400–414.

- Fenk-Oczlon, G. (2001). Familiarity, information flow, and linguistic form. In Bybee, J. and Hopper, P., editors, *Frequency and the Emergence of Linguistic Structure*, pages 431–448. John Benjamins Publishing.
- Ferrer-i-Cancho, R. and Hernández-Fernández, A. (2013). The failure of the law of brevity in two new world primates. Statistical caveats. *Glottology International Journal of Theoretical Linguistics*, 4(1):45–55.
- Ferrer-i-Cancho, R., Hernández-Fernández, A., Lusseau, D., Agoramoorthy, G., Hsu, M., and Semple, S. (2013). Compression as a universal principle of animal behavior. *Cognitive Science*, 37(8):1565–1578.
- Ferrer-i-Cancho, R. and Lusseau, D. (2009). Efficient coding in dolphin surface behavioral patterns. *Complexity*, 14(5):23–25.
- Ferrer-i-Cancho, R. and Moscoso del Prado, F. (2012). Information content versus word length in random typing. *JSTAT*.
- Fowler, C. A. (1988). Differential shortening of repeated content words produced in various communicative contexts. *Language and Speech*, 31(4):307–319.
- Franke, M. (2017). Game theory in pragmatics: Evolution, rationality & reasoning. In *Oxford Research Encyclopedia of Linguistics*.
- Futrell, R., Mahowald, K., and Gibson, E. (2015). Large-scale evidence of dependency length minimization in 37 languages. *Proceedings of the National Academy of Sciences*, 112(33):10336–10341.
- Gahl, S. and Garnsey, S. M. (2004). Knowledge of grammar, knowledge of usage: Syntactic probabilities affect pronunciation variation. *Language*, 80(4):748–775.
- Genzel, D. and Charniak, E. (2002). Entropy rate constancy in text. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 199–206.
- Granger, C. W. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: Journal of the Econometric Society*, 37(3):424–438.
- Hamming, R. W. (1950). Error detecting and error correcting codes. *Bell Labs Technical Journal*, 29(2):147–160.
- Hauser, M. D., Chomsky, N., and Fitch, W. T. (2002). The faculty of language: What is it, who has it, and how did it evolve? *Science*, 298(5598):1569–1579.

- Hauser, M. D., Yang, C., Berwick, R. C., Tattersall, I., Ryan, M. J., Watumull, J., Chomsky, N., and Lewontin, R. C. (2014). The mystery of language evolution. *Frontiers in Psychology*, 5:401.
- Hawkins, R. X., Frank, M., and Goodman, N. (2017). Convention-formation in iterated reference games. In *Proceedings of the 39th annual conference of the Cognitive Science Society*, pages 482–487.
- Hudson Kam, C. L. and Newport, E. L. (2005). Regularizing unpredictable variation: The roles of adult and child learners in language formation and change. *Language Learning and Development*, 1(2):151–195.
- Huffman, D. A. (1952). A method for the construction of minimum-redundancy codes. *Proceedings of the IRE*, 40(9):1098–1101.
- Hupet, M. and Chantraine, Y. (1992). Changes in repeated references: Collaboration or repetition effects? *Journal of Psycholinguistic Research*, 21(6):485–496.
- Hyndman, R. J. and Khandakar, Y. (2008). Automatic time series forecasting: the forecast package for R. *Journal of Statistical Software*, 26(3):1–22.
- Jaeger, T. F. (2010). Redundancy and reduction: Speakers manage syntactic information density. *Cognitive Psychology*, 61(1):23–62.
- Jamet, D. (2009). A morphophonological approach to clipping in English. Can the study of clipping be formalized? *Lexis: Journal in English Lexicology*, HS 1.
- Kanwal, J., Smith, K., Culbertson, J., and Kirby, S. (2017a). Language-users choose short words in predictive contexts in an artificial language task. In *Proceedings of the 39th annual conference of the Cognitive Science Society*, pages 643–648.
- Kanwal, J., Smith, K., Culbertson, J., and Kirby, S. (2017b). Zipf’s Law of Abbreviation and the Principle of Least Effort: Language users optimise a miniature lexicon for efficient communication. *Cognition*, 165:45–52.
- Kanwal, J. S., Matsumura, S., Ohlemiller, K., and Suga, N. (1994). Analysis of acoustic elements and syntax in communication sounds emitted by mustached bats. *The Journal of the Acoustical Society of America*, 96(3):1229–1254.
- Kemp, C. and Regier, T. (2012). Kinship categories across languages reflect general communicative principles. *Science*, 336(6084):1049–1054.
- Kirby, S. (2001). Spontaneous evolution of linguistic structure—an iterated learning model of the emergence of regularity and irregularity. *IEEE Transactions on Evolutionary Computation*, 5(2):102–110.

- Kirby, S., Cornish, H., and 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*, 105(31):10681–10686.
- Kirby, S., Tamariz, M., Cornish, H., and Smith, K. (2015). Compression and communication in the cultural evolution of linguistic structure. *Cognition*, 141:87–102.
- Krauss, R. M. and Weinheimer, S. (1964). Changes in reference phrases as a function of frequency of usage in social interaction: A preliminary study. *Psychonomic Science*, 1(1-12):113–114.
- Krauss, R. M. and Weinheimer, S. (1966). Concurrent feedback, confirmation, and the encoding of referents in verbal communication. *Journal of Personality and Social Psychology*, 4(3):343.
- Kuperman, V. and Bresnan, J. (2012). The effects of construction probability on word durations during spontaneous incremental sentence production. *Journal of Memory and Language*, 66(4):588–611.
- Levy, R. and Jaeger, T. F. (2007). Speakers optimize information density through syntactic reduction. In Schlokopf, B., Platt, J., and Hoffman, T., editors, *Advances in Neural Information Processing Systems (NIPS)*, volume 19, pages 848–856. MIT Press.
- Lieberman, P. (1963). Some effects of semantic and grammatical context on the production and perception of speech. *Language and Speech*, 6(3):172–187.
- Lindsay, M. and Aronoff, M. (2013). Natural selection in self-organizing morphological systems. *Morphology in Toulouse*, pages 133–153.
- Luo, B., Jiang, T., Liu, Y., Wang, J., Lin, A., Wei, X., and Feng, J. (2013). Brevity is prevalent in bat short-range communication. *Journal of Comparative Physiology A*, 199(4):325–333.
- Mahowald, K., Fedorenko, E., Piantadosi, S. T., and Gibson, E. (2013). Info/information theory: Speakers choose shorter words in predictive contexts. *Cognition*, 126(2):313–318.
- Mahowald, K., Piantadosi, S. T., Alper, M., and Gibson, E. (2015). Lexical items are privileged slots for meaning. Poster presented at CUNY 2015.
- Mandelbrot, B. (1954). Simple games of strategy occurring in communication through natural languages. *Transactions of the IRE Professional Group on Information Theory*, 3(3):124–137.

- Manin, D. (2006). Experiments on predictability of word in context and information rate in natural language. *arXiv preprint cs/0612136*.
- Meylan, S. C. and Griffiths, T. L. (2017). Word forms—not just their lengths—are optimized for efficient communication. *arXiv:1703.01694*.
- Michel, J.-B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., The Google Books Team, Pickett, J. P., Hoiberg, D., Clancy, D., Norvig, P., Orwant, J., Pinker, S., Nowak, M. A., and Aiden, E. L. (2011). Quantitative analysis of culture using millions of digitized books. *Science*, 331(6014):176–182.
- Miller, G. A. (1951). *Language and communication*. McGraw-Hill.
- Miller, G. A. (1957). Some effects of intermittent silence. *The American Journal of Psychology*, 70(2):311–314.
- Moscoso del Prado, F. (2013). The missing baselines in arguments for the optimal efficiency of languages. In *Proceedings of the 35th annual conference of the Cognitive Science Society*, pages 1032–1037.
- Moscoso del Prado, F. (2014). Grammatical change begins within the word: Causal modeling of the co-evolution of Icelandic morphology and syntax. In *Proceedings of the 36th annual conference of the Cognitive Science Society*.
- Pate, J. K. (2017). Optimization of American English, Spanish, and Mandarin Chinese over time for efficient communication. In *Proceedings of the 39th annual conference of the Cognitive Science Society*.
- Piantadosi, S. T., Tily, H., and Gibson, E. (2011). Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences*, 108(9):3526–3529.
- Piantadosi, S. T., Tily, H., and Gibson, E. (2012). The communicative function of ambiguity in language. *Cognition*, 122(3):280–291.
- Piantadosi, S. T., Tily, H., and Gibson, E. (2013). Information content versus word length in natural language: A reply to Ferrer-i-Cancho and Moscoso del Prado Martin [arxiv:1209.1751]. *arXiv:1307.6726*.
- Pinker, S. (1995). *The language instinct: The new science of language and mind*. Penguin UK.
- Prün, C. (2002). Biographical notes on GK Zipf. *Glottometrics*, 3:1–10.
- Pustet, R. (2004). Zipf and his heirs. *Language Sciences*, 26(1):1–25.

- R Core Team (2015). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. <https://www.R-project.org/>.
- Real, F. and Griffiths, T. L. (2009). The evolution of frequency distributions: Relating regularization to inductive biases through iterated learning. *Cognition*, 111(3):317–328.
- Regier, T., Carstensen, A., and Kemp, C. (2016). Languages support efficient communication about the environment: Words for snow revisited. *PloS one*, 11(4):e0151138.
- Regier, T., Kemp, C., and Kay, P. (2015). Word meanings across languages support efficient communication. *The handbook of language emergence*, pages 237–263.
- Semple, S., Hsu, M. J., and Agoramoorthy, G. (2010). Efficiency of coding in macaque vocal communication. *Biology Letters*, 6(4):469–471.
- Seyfarth, S. (2014). Word informativity influences acoustic duration: Effects of contextual predictability on lexical representation. *Cognition*, 133(1):140–155.
- Shannon, C. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423.
- Sigurd, B., Eeg-Olofsson, M., and Van Weijer, J. (2004). Word length, sentence length and frequency—Zipf revisited. *Studia Linguistica*, 58(1):37–52.
- Smith, K. and Wonnacott, E. (2010). Eliminating unpredictable variation through iterated learning. *Cognition*, 116(3):444–449.
- Strauss, U., Grzybek, P., and Altmann, G. (2007). Word length and word frequency. In *Contributions to the science of text and language*, pages 277–294. Springer.
- Tanimura, S., Kuroiwa, C., Mizota, T., et al. (2006). Proportional symbol mapping in R. *Journal of Statistical Software*, 15(i05).
- Teahan, W. J., Wen, Y., McNab, R., and Witten, I. H. (2000). A compression-based algorithm for Chinese word segmentation. *Computational Linguistics*, 26(3):375–393.
- Tily, H., Gahl, S., Arnon, I., Snider, N., Kothari, A., and Bresnan, J. (2009). Syntactic probabilities affect pronunciation variation in spontaneous speech. *Language and Cognition*, 1(2):147–165.
- Vouloumanos, A. (2008). Fine-grained sensitivity to statistical information in adult word learning. *Cognition*, 107(2):729–742.

- Ward, G. (2002). Moby Words II.
<http://onlinebooks.library.upenn.edu/webbin/gutbook/lookup?num=3201>.
- Whalen, D. H. (1991). Infrequent words are longer in duration than frequent words. *The Journal of the Acoustical Society of America*, 90(4):2311–2311.
- Wilkes-Gibbs, D. and Clark, H. H. (1992). Coordinating beliefs in conversation. *Journal of Memory and Language*, 31(2):183–194.
- Winters, J., Kirby, S., and Smith, K. (2015). Languages adapt to their contextual niche. *Language and Cognition*, 7(3):415—449.
- Wright, C. E. (1979). Duration differences between rare and common words and their implications for the interpretation of word frequency effects. *Memory & Cognition*, 7(6):411–419.
- Zeileis, A. and Hothorn, T. (2002). Diagnostic checking in regression relationships. *R News*, 2(3):7–10.
- Zipf, G. K. (1929). Relative frequency as a determinant of phonetic change. *Harvard Studies in Classical Philology*, 40:1–95.
- Zipf, G. K. (1935). *The Psycho-biology of Language*, volume ix. Houghton Mifflin, Oxford, England.
- Zipf, G. K. (1949). *Human Behavior and the Principle of Least Effort*. Addison-Wesley Press.