UC Irvine
UC Irvine Electronic Theses and Dissertations

Title
The role of communicative efficiency in shaping linguistic choices in Arabic and beyond

Permalink
https://escholarship.org/uc/item/5w67f9tg

ISBN
9798173722171

Author
Kachakeche, Zeinab

Publication Date
2026-11-02

Copyright Information
This work is made available under the terms of a Creative Commons Attribution-NonCommercial-NoDerivatives License, available at https://creativecommons.org/licenses/by-nc-nd/4.0/

Peer reviewed|Thesis/dissertation

eScholarship.org Powered by the California Digital Library
University of California

UNIVERSITY OF CALIFORNIA,
IRVINE

The role of communicative efficiency in shaping linguistic choices in Arabic and beyond

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Language Science

by

Zeinab Kachakeche

Dissertation Committee:
Associate Professor Gregory Scontras, Chair
Associate Professor Richard Futrell, Co-advisor
Assistant Professor Connor Mayer

2026

© 2026 Zeinab Kachakeche

DEDICATION

To those who gave their life to the homeland, for it was dearer to them than themselves.
To the loved ones they left behind, in the hopes of giving them the future...

TABLE OF CONTENTS

Page

LIST OF FIGURES **v**

LIST OF TABLES **viii**

ACKNOWLEDGMENTS **ix**

VITA **x**

ABSTRACT OF THE DISSERTATION **xiii**

1 Introduction **1**
- 1.1 Efficiency as a balance between cost and reward . . . 3
- 1.2 The different costs of successful communication . . . 7
  - 1.2.1 Cost as word-length . . . 8
  - 1.2.2 Availability-based production . . . 9
  - 1.2.3 Cost of ambiguity . . . 10
  - 1.2.4 Cost as motor effort . . . 11
  - 1.2.5 Comprehension cost . . . 11
- 1.3 Rewards of successful communication . . . 12
  - 1.3.1 Reward as successful transmission . . . 12
  - 1.3.2 Reward as faster or earlier comprehension . . . 13
  - 1.3.3 Reward as message clarity . . . 14
  - 1.3.4 Reward as ultimate survival . . . 14
- 1.4 Looking ahead . . . 15

2 The efficiency of adjective modification **18**
- 2.1 Introduction . . . 18
- 2.2 Background . . . 18
  - 2.2.1 How speakers use adjectives to establish reference . . . 18
  - 2.2.2 Overmodification . . . 22
  - 2.2.3 What happens post-nominally? . . . 30
- 2.3 Methods . . . 36
  - 2.3.1 Wikipedia Corpus Analysis . . . 36
  - 2.3.2 Universal Dependencies analysis . . . 41

2.3.3 Color-size asymmetry in Arabic . . . 45
2.3.4 Discussion . . . 46
2.3.5 Expanding the corpus analysis . . . 52
2.3.6 Behavioral experiments . . . 55

3 Adjective ordering in Arabic **59**
3.1 Introduction . . . 59
3.2 Background . . . 60
3.2.1 Empirical observations of adjective ordering preferences . . . 60
3.2.2 Proposed explanations of ordering preferences . . . 62
3.3 Behavioral Experiments . . . 72
3.3.1 Experiment 1: Ordering preferences in native Arabic . . . 75
3.3.2 Experiment 2: Assessing subjectivity . . . 79
3.3.3 Experiment 3: Ordering preferences in heritage Arabic . . . 81
3.3.4 Discussion of Experimental Results . . . 85

4 Vowel Minimization in Arabizi **89**
4.1 Introduction . . . 89
4.2 Background . . . 90
4.2.1 Word length and predictability . . . 90
4.2.2 Arabic, Arabizi, and vowel variation . . . 95
4.2.3 Efficiency and Vowel dropping . . . 100
4.3 Methods . . . 102
4.3.1 Choosing short and long forms of the words . . . 102
4.3.2 Training an n-gram model . . . 103
4.3.3 Obtaining informativity measures . . . 106
4.3.4 Scaling up the Twitter analysis . . . 109
4.4 Discussion . . . 117

5 General Conclusion and Future Directions **121**
5.1 Connecting threads . . . 121
5.2 Chapter conclusions and future directions . . . 125

Bibliography **129**

LIST OF FIGURES

Page

1.1 Figure from Gibson et al. (2019). The information theoretic pathway of successful communication. This figure shows that every communication has a source and a sender that sends a signal, the signal is then impacted by noise before it arrives at the receiver which deciphers the signal before it reaches its final destination. 7
1.2 Plot from Levshina (2022a) illustrating the different costs involved in efficient communication. 8
1.3 Plot from Levshina (2022a) about the different benefits of language communication 15
1.4 Different cost and reward considerations of efficient communication. Speakers and listeners navigate a trade-off between minimizing cost and maximizing reward during conversations. Generated with the help of Claude. 16

2.1 An example of a display from Sedivy et al. (1999) showing contrast conditions, where a participant would narrow down the potential referent of tall objects (the tall glass and the pitcher) before hearing the noun. 21
2.2 Results from Degen et al. (2020) showing that speakers significantly use redundant color adjectives relative to size in their first experiment. The results are shown alongside estimates of their RSA model for redundant utterance probability as a function of scene variation. 29
2.3 Example of visual paradigm from Rubio-Fernández (2016). English and Spanish speakers were asked to refer to one of the items in this 4-item display. On the right is an example of a polychrome display where all the items differ in color while on the left is a monochrome display where all the items exhibit the same color. English speakers produced redundant color adjectives more than Spanish speakers in both displays, although the color adjective is not needed to identify the referent in either case. 31

2.4 Example of visual paradigm from Rubio-Fernández (2019) where the authors addressed the issue of color adjective being central to the noun property. In this paradigm, both English and Spanish speakers were asked to refer to one of the 4 items in the display following the same polychrome/monochrome condition. In both cases, the color adjective is not required for reference identification.Yet, English speakers still produced redundant color adjectives on roughly 46% of trials, against 7% for Spanish speakers. 34
2.5 Wikipedia analysis results: percentage of noun modification in the analysis is similar across the three languages: Arabic, English, and Spanish. 38
2.6 UD analysis results: proportion of nouns with any adjectival modification compared with the proportion of modified nouns with post-nominal adjectives ($r^2 = 0.18$; 95% CI [0.05, 0.32]). 43
2.7 UD dependency analysis results: rate of adjective modification compared with the log probability of post-noimnal adjectival modification (Spearman's $\rho = -0.16$, p = 0.18). 45
2.8 Tentative Arabic results on color-size asymmetry with scene variation. 47
2.9 Sample displays from the referential communication paradigm used in Rubio-Fernández et al. (2020). Here, the target is the blue triangle (top left corner in all displays). In the shape competitor condition (A), color is necessary for disambiguation. In the color competitor condition (B), shape is sufficient for disambiguation. In the no competitor conditions (C, D), both color and shape are redundant with difference in the size of competitor objects. 58

3.1 Summary of the different predictors of adjective ordering preferences discussed in this section. This figure was generated with the help of Claude. 71
3.2 An example trial from the ordering preferences experiment. On this trial, participants were presented with two phrases: *the brown small chair* and *the small brown chair*. They were asked to adjust the slider towards the preferred phrase. To the right is the Arabic version of the experiment, and to the left is the original experiment by Scontras et al. (2017). 78
3.3 Native Arabic speakers' mean preferred distance measures from the noun grouped by adjective semantic class. Values range from 0 (preferred closest to the noun) to 1 (preferred farthest from the noun). The dashed line indicates chance level, or the absence of stable preferences. Error bars represent bootstrapped 95% confidence intervals drawn from 10,000 samples of the data. 79
3.4 An example trial from the faultless disagreement task. 81
3.5 Ordering preferences obtained in Experiment 1 plotted against subjectivity scores for each of the 25 adjectives tested. Subjectivity accounts for 76% of the variance in the ordering preferences ($r^2 = 0.76$, 95% CI $[0.57, 0.88]$). . . . 82
3.6 An example trial from the ordering preference task adapted for heritage speakers. Instructions were presented bilingually in Arabic and English, while the adjective-noun descriptions were presented only in Arabic. 84
3.7 Naturalness ratings from Experiment 3 (heritage) grouped by adjective semantic class. Error bars represent bootstrapped 95% confidence intervals drawn from 10,000 samples of the data. 85

3.8 Heritage ordering preferences plotted against adjective subjectivity scores obtained from baseline speakers in Experiment 2. 86

4.1 Corpus results from Mahowald et al. (2013). The y-axis shows the difference in surprisal between long and short forms of 22 English word pairs, and the x-axis shows the combined count of both variants. Most differences fall above the y=0 line, indicating that short forms are more predictable given context. 95
4.2 Model architecture and parameter counts. The model has 18,758,181 trainable parameters, with an embedding dimension of 100 and a linear dimension of 128. 105
4.3 Training and development loss across epochs. The development loss reaches its minimum at epoch 11 and begins increasing thereafter, indicating overfitting. Loss is computed as negative log likelihood divided by the number of bigrams, placing training and development losses on the same scale. 105
4.4 Difference in average surprisal between long and short spelling variants of lexical items plotted against log frequency of the word pairs combined. Most of the word pairs lie above the $y = 0$, indicating that longer spelling variants carry higher average surprisal (more informative) than their shorter counterparts. 108
4.5 Difference in surprisal between long and short spelling variants for each word in the Twitter data. Here, frequency of the spelling variants is assumed to be similar. Nine of the thirteen pairs fall above the y=0 line, showing the expected pattern. Unlike the BOLT chat data, some pairs pattern differently here; for example, the word ⟨ana/anaa⟩which was predominantly used in its short form in the private chats, appears more often in its longer form on Twitter. 110
4.6 Difference in average surprisal between the long and short forms of the original word pairs, tested on held-out Twitter data using the Twitter-trained bi-gram model. Context size is n=1. 114
4.7 Difference in average surprisal between the long and short forms of the original word pairs, tested on held-out Twitter data using the Twitter-trained models. Context size on the left is n=2, and on the right n=3 115
4.8 Difference in average surprisal between the long and short forms of the expanded word pairs after excluding the outliers, tested on held-out Twitter data using models with increasing context sizes. 116
4.9 Difference in average surprisal between the long and short forms using a count-based n-gram model on held-out Twitter data. 117

LIST OF TABLES

Page

2.1 Results of Wikipedia analyses for English, Spanish, and Arabic. # Nouns gives the total number of nouns extracted. Columns AN, NA, and ANA give the percentage of nouns with adjacent adjectives of the specified type. Total mod. gives the percentage of modified nouns which are adjacent to any adjective. 39
2.2 Data sources compiled for the dialectal Arabic corpus from ten different existing datasets. 54
2.3 Dialect distribution of the compiled Arabic corpus. The dialects are grouped into the common five dialect groups, in addition to MSA (Modern Standard Arabic). 54

3.1 Dialect distribution of participants in Experiment 1 (N=24). 76
3.2 The seven adjective classes used in the experiment, which are translations of the adjectives used in the original experiment by Scontras et al. (2017). . . . 77
3.3 The set of ten nouns from two different classes I adopted from Scontras et al. (2017) with their translations. 78
3.4 Dialect distribution of participants in Experiment 2 (N=16). 80

4.1 Arabizi word pairs used in the analysis. The short variant reflects vowel dropping. Many pairs are function words such as pronouns and prepositions. 103
4.2 New Arabizi word pairs used in the scaled-up Twitter analysis. Unlike the original set (Table 4.1), these pairs include both function and content words. 113
4.3 Summary of mixed-effects logistic regression results across all analyses. ✓ = significant in the predicted direction; ✗ = not significant; ✗* = significant but in the opposite direction; **?** = inconclusive. Updated content words list excludes three word pairs: *walahy/wallahy*, *ayza/3ayza*, and *gam3a/gama3a*; the latter was driving the effect in the opposite direction. 117

ACKNOWLEDGMENTS

I am deeply grateful to my advisors, Gregory Scontras and Richard Futrell. Greg, you have shaped the way I think about language since my undergraduate years, and you have been patient, flexible, and endlessly generous ever since — answering my questions with better questions, navigating my drafts, and steering me away from more academic cliffs than I can count. Richard, your insight runs through every information-theoretic bone of this dissertation, and this work is immeasurably sharper for your guidance. And to Connor, thank you for engaging with my work with a care and precision most students only hope for. Your comments and questions made me a more careful thinker.

To the Language Science community at UCI: thank you for being an academic home. To Shiva, thank you for hearing out every half-formed idea before it was ready for anyone else. To Niels, thank you for the accountability, and the reminders that words eventually arrive if you sit with them long enough. To Yanting, thank you for your support and encouragement when I needed them most.

To my parents, thank you for believing in me long before there was any evidence this was possible. You raised me to love learning in a country that gave you every reason to stop dreaming. Every page of this dissertation was paved by your sacrifices. This degree is as much yours as mine.

To my husband, Ali Kassir: this degree asked a great deal of me, and somehow more of you. You absorbed the long nights and the chaos, held our family steady, and never made me feel that my dream was a burden. Your support wasn't behind this work — it was underneath all of it, holding everything up. I could not have done this without you, and I would not have wanted to.

And finally, to Hassan, Fatima, and Narjis: every hour this work took from you, I was trying to give back something bigger. I did this to become a better mother, a better example, and a person who leaves something impactful behind, but the truth is you were already teaching me all three. This is for you.

VITA

Zeinab Kachakeche

EDUCATION

Doctor of Philosophy in Language Science **2026**
University of California, Irvine *Irvine, California*

Masters of Arts in Language Science **2024**
University of California, Irvine *Irvine, California*

Bachelor of Science in Language Science, minor in Psychology **2020**
University of California, Irvine *Irvine, California*

Irvine Valley College **2016**

RESEARCH EXPERIENCE

Graduate Research **2020–2026**
University of California, Irvine *Irvine, California*

Undergraduate Research **2018–2020**
University of California, Irvine *Irvine, California*

TEACHING EXPERIENCE

Teaching Assistant **2020–2026**
University of California, Irvine *Irvine, California*

REFEREED CONFERENCE PUBLICATIONS

Kachakeche, Zeinab, Gregory Scontras, and Richard Futrell. 2022. The efficiency of dropping vowels in Romanised Arabic script. *Proceedings of the 44th Annual Meeting of the Cognitive Science Society.*

Kachakeche, Zeinab, Richard Futrell, and Gregory Scontras. 2021. Word order affects the frequency of adjective use across languages. *Proceedings of the 43rd Annual Meeting of the Cognitive Science Society.*

Scontras, Gregory, Galia Bar-Sever, **Zeinab Kachakeche**, Cesar Manuel Rosales Jr., and Suttera Samonte. 2020. Incremental semantic restriction and subjectivity-based adjective ordering. *Proceedings of Sinn und Bedeutung (SuB) 24, 253–270.*

Kachakeche, Zeinab, and Gregory Scontras. 2020. Adjective ordering in Arabic: Post-nominal structure and subjectivity-based preferences. *Proceedings of the Linguistic Society of America 5, 419–430.*

CONFERENCE PRESENTATIONS

Zeinab Kachakeche*, Greg Scontras, and Richard Futrell. 2023. Predictability given context affects word length in non-standard Romanised Arabic script. *5th California Meeting on Psycholinguistics conference*

Judith Degen, Brandon Waldon*, Adolfo Hermosillo, Anthony Velaquez, and **Zeinab Kachakeche**. 2022. Spanish exhibits a color/size asymmetry and scene variation in redundant referring expression production – just like English. *Xprag conference 2022*

Kachakeche, Zeinab*, Gregory Scontras, and Richard Futrell. 2022. The efficiency of dropping vowels in Romanised Arabic script. *The 44th Annual Meeting of the Cognitive Science Society.*

Kachakeche, Zeinab*, Richard Futrell, and Gregory Scontras. 2022. Adjective use across languages. The 96th Annual Meeting of the Linguistic Society of America.

Kachakeche, Zeinab*, Richard Futrell, and Gregory Scontras. 2021. Word order affects the frequency of adjective use across languages. Poster presented at the 43rd Annual Meeting of the Cognitive Science Society.

Kachakeche, Zeinab*, Richard Futrell, and Gregory Scontras. 2021. Adjective use across languages. California Annual Meeting on Psycholinguistics.

Scontras, Gregory*, **Zeinab Kachakeche**, Austin Nguyen, Cesar Rosales, Suttera Samonte, Einat Shetreet, Yuxin Shi, Elli Tourtouri, and Nitzan Trainin. 2021. Cross-linguistic evidence for subjectivity-based adjective ordering preferences. Theoretical and Experimental Approaches to Modification.

Kachakeche, Zeinab*, and Gregory Scontras. 2020. Adjective ordering in Arabic: Post-nominal structure and subjectivity-based preferences. The 94th Annual Meeting of the Linguistic Society of America.

Scontras, Gregory*, Galia Bar-Sever, **Zeinab Kachakeche**, Cesar Manuel Rosales Jr., and Suttera Samonte. 2019. Incremental semantic restriction and subjectivity-based adjective ordering. Sinn und Bedeutung (SuB) 24.

ABSTRACT OF THE DISSERTATION

The role of communicative efficiency in shaping linguistic choices in Arabic and beyond

By

Zeinab Kachakeche

Doctor of Philosophy in Language Science

University of California, Irvine, 2026

Associate Professor Gregory Scontras, Chair

My dissertation investigates the role of communicative efficiency in shaping linguistic form, bringing evidence from Arabic and Arabizi across different levels of linguistic structures. Chapter 1 reviews the essential background needed for the study of communicative efficiency. In Chapter 2, I examine whether the rate at which speakers modify nouns with adjectives is affected by adjective position relative to the noun by conducting a corpus analysis of English, Spanish, and Arabic Wikipedia, as well as a broad analysis of 74 languages in the Universal Dependencies treebanks. In Chapter 3, I test whether the adjective ordering preferences well established for languages with pre-nominal adjectives also hold in Arabic, a language with post-nominal modification through behavioral experiments for both baseline native Arabic speakers and heritage Arabic speakers. In Chapter 4, I test whether contextual predictability shapes vowel dropping in Arabizi, the Romanized Arabic script used in digital communication, comparing private chat data with public Twitter data. Together, the three studies show that the efficiency pressures documented in English and other well-studied languages are also at work in Arabic, while revealing that their expression is modulated by the communicative context.

Chapter 1

Introduction

Human language is a system of extraordinary complexity, one that is shaped by pressures operating simultaneously at multiple levels of linguistic structure, from the phonological and morphological units that compose individual words, to the syntactic structures that combine those words into phrases, to the typological patterns that distinguish one language from another. A growing body of research has proposed that an overarching pressure, the drive to communicate efficiently, cuts across all of these levels, shaping the form and function of language in systematic and predictable ways (Zipf, 1949; Hawkins, 2004; Piantadosi et al., 2011; Gibson et al., 2019; Levshina, 2022a). According to this view, language is not merely a formal system governed by arbitrary conventions; it is a tool for communication, and many of its properties reflect the demands of successful information transfer under cognitive, articulatory, and environmental constraints.

Efficiency, broadly construed, involves balancing the costs of producing an utterance against the rewards of communicating successfully. Yet the content of this claim depends heavily on what we take efficiency, cost, and reward to mean. That is, these terms are defined differently throughout the literature and are often left implicit. This thesis takes the efficiency

hypothesis as its starting point and asks: what kinds of cost and reward shape linguistic choices, whether the same notions apply equally across languages, and how a careful look at Arabic, alongside cross-linguistic evidence, can sharpen our understanding of how efficiency shapes language use.

Arabic is one of the world's most widely spoken languages, with millions of speakers across the Middle East and North Africa, and across large communities of heritage speakers and immigrants in the United States and Europe. Yet it remains markedly understudied in the psycholinguistics literature, and almost entirely absent from the efficiency-based conversation that this thesis takes up. Several factors have contributed to this gap: relatively limited linguistic resources and annotated data in Arabic, a tradition of Arabic-language scholarship that has not always been translated into English, and a long-standing focus on Modern Standard Arabic at the expense of the spoken dialects that speakers actually use in everyday communication. Importantly, the diglossic relationship between Standard Arabic and the colloquial varieties means that questions about communicative pressures cannot be settled by looking at written Standard Arabic alone. Arabic is also structurally distinct from the languages on which most efficiency-based claims have been built: it is post-nominal in its adjectival modification, a feature that two of the chapters in this thesis take up directly. Together, these properties make Arabic a useful test case for asking whether the cost and reward notions developed in the existing literature carry across typologically different systems whose structural profile differs sharply from English.

Before turning to Arabic specifically, it is worth unpacking what the efficiency hypothesis assumes. Discussions of communicative efficiency have long centered on the idea that linguistic form is shaped by trade-offs between competing pressures. Central among these pressures is the balance between the costs of encoding a message and the need to convey it successfully. Crucially, both sides of this balance are themselves multifaceted. Cost is distributed between speakers, who plan and articulate utterances, and listeners, who perceive and inter-

pret them. Communicative success, in turn, depends on how reliably the intended message can be recovered given the signal, the context, and the knowledge the interlocutors already share. Across languages, commonly used forms tend to be shorter and less complex (Zipf, 1949; Hawkins, 2004), suggesting that speakers favor structures that reduce effort while still supporting effective communication. At the same time, communicative success requires that expressions remain sufficiently informative for listeners to recover intended meanings, highlighting an inherent tension between economy and clarity. How this tension is negotiated, and more specifically how cost and reward are defined on each side of it, is the question the rest of this chapter takes up.

Not all approaches to language take this communicative premise for granted. A long-standing tradition in linguistics, most closely associated with the generative view, holds that the core properties of language arise from constraints internal to the grammatical system rather than from pressures exerted by communicative use (Chomsky, 2017). According to this view, language is optimized to be an efficient instrument of thought rather than an ideal instrument of communication. The efficiency-based view departs from this position: it proposes that many observable properties of language are best understood as responses to communicative pressures. From this perspective, efficiency is not reducible to a single factor, but emerges from the interaction of multiple constraints that jointly shape language use and structure. Understanding how speakers navigate these trade-offs offers insight into the principles governing linguistic choice and the cognitive mechanisms that support successful communication, and provides the conceptual foundation for the present work.

1.1 Efficiency as a balance between cost and reward

The hypothesis that language is shaped by efficiency is not new. Zipf (1949) famously proposed that speakers balance competing pressures, minimizing effort for both themselves and

their listeners, producing the well-known statistical regularities we observe in word frequency and length distributions. Zipf's central observation was that more frequently used words tend to be shorter: in English, frequent lexical items such as *I*, *pen*, *run*, *eat*, and *on* are reliably shorter than rarer items like *pulchritudinous*, *incomprehensible*, and *uncopyrightable*. Zipf interpreted this regularity as a trade-off between two factors that jointly shape word form: production cost on the speaker's side captured by word length, and ease of comprehension on the listener's side, since frequent words carry meanings that are often retrievable in context and thus can afford to be short. Zipf's principle thus provides a quantitative perspective on the trade-off between production cost and comprehension ease, such that more frequent messages are easier to understand and less costly to produce, putting two specific factors at the center of the efficiency picture: **information transfer** and **word length**. Cost can take other forms as well: the duration of an utterance in a speech stream, the mental energy that is required to plan it, or the cognitive resources demanded of the listener to comprehend it. Similarly, information transfer is not a single notion: it encompasses everything from establishing common ground to relying on a listener's ability to infer the intended message from context. Despite these factors, Zipf's principle of least effort serves as a pivotal starting point for the literature focusing on efficiency, illustrating how multiple factors can contribute to a trade-off that shapes our language use.

In fact, the observation of a relationship between word length and frequency predates Zipf by several centuries. According to Leslau (1969), this relationship was already articulated by Arab grammarians, most notably by Al-Zamakhsharī in his 12th century book *Kitāb al-Mufassal fī San'at al-I'rāb* (al Zamakhsharī, 1905). This phenomenon was known as 'al-takhfīf li kathrat al-isti'māl' "weakening for frequent usage". *Takhfīf* referred to the tendency to abbreviate, or simplify linguistic forms that are used with high frequency, on the grounds that frequent use creates pressure toward ease of articulation. The grammarians observed that the most common elements of Arabic (pronouns, prepositions, particles, negation markers) are systematically shorter and less morphologically complex than rarer, more semantically

specific words. This is not merely a descriptive observation; it was treated as a principled explanation: high-frequency use justifies reduced form because the reduced form remains recoverable given shared knowledge between speaker and listener. Although the Arab grammarians did not frame their observations in information-theoretic terms, they essentially articulated a version of the same cost-reward trade-off that Zipf would later formalize mathematically; indeed, Leslau (1969, 181) explicitly draws the connection between the medieval Arabic observations and Zipf's modern work on relative frequency as a determinant of sound changes. For example, the Arabic phrase *layta-nī ghaniyyun* "may I be wealthy" becomes *layt-ī ghaniyyun* by eliminating a vowel /a/ and a consonant /n/ from the word: the first person suffix *-nī* is reduced to *-ī* when attached to the high-frequency particle *layta*, because the frequent co-occurrence licenses the shorter form. The principle was recognized across centuries of the Arabic grammatical tradition. Al-Ashmūnī, a fifteenth-century grammarian commenting on Ibn Mālik's *Alfiyya*, similarly explained the reduction of *mundh* 'since' to *mudh* as a consequence of frequent usage (Leslau, 1969, 180).

While Zipf's principle captures a statistical regularity in language use, Grice (1975) approached the same trade-off from a different angle, proposing the maxim of quantity: speakers should contribute just enough information to the conversation as is necessary — no more, no less. Like Zipf's principle, this maxim implicitly reflects a balance between saying too much and saying too little. But the framing is fundamentally different: where Zipf's law is descriptive and statistical, capturing a pattern observed in language use, Grice's maxim describes what cooperative speakers *should* do rather than what they actually do. Speakers violate these maxims frequently and systematically, both by saying less than they should and, more often, by saying more than they need to, a pattern that has become a central source of evidence in recent efficiency-based accounts of language use, and one that Chapter 2 takes up in detail.

A different but complementary articulation of the same trade-off comes from Hawkins (2004),

who proposed that grammatical forms and structures reflect a balance between **communicative efficiency** and **cognitive complexity**. Within this framework, speakers prefer shorter and simpler forms over longer or more complex ones, a principle Hawkins calls *minimizing forms*, unless there is a specific communicative, semantic, or pragmatic reason to use the longer form. What makes Hawkins's contribution distinctive is its scope: where Zipf focused on statistical regularities in word form and Grice on conversational implicature, Hawkins extends the efficiency argument to grammar itself. He argues that the structural patterns distinguishing one language from another reflect, in part, the same cost-reward pressures operating at the syntactic and morphological level. One natural extension of this view, taken up in detail later in this chapter, is that the accessibility of frequent forms is itself a form of reduced cost: highly frequent words demand less cognitive processing for both speaker and listener. I will return to Hawkins's principle of minimizing forms in Chapter 4, where it serves as the organizing concept for my investigation of how speakers balance form and information across linguistic levels.

Gibson et al. (2019) give this trade-off its most explicit formulation. Grounding their definition in information theory, they picture communication as the transmission of a message from speaker to listener through a noisy channel. Efficient communication, according to this view, is achieved when speakers minimize the effort of production while maximizing the likelihood that the listener successfully recovers the intended message (Figure 1.1). What this framing makes clear is that efficiency is not a property of form alone, nor of meaning alone, but of the pairing between them under conditions of noise. Crucially, the cost and reward entering into the trade-off need not be defined in a single way: articulatory effort, processing load, and informativity can all, in principle, play a role in efficient communication. The next section takes up exactly this point, surveying how cost has been operationalized across the efficiency literature before turning, in Section 1.3, to the question of what counts as reward.

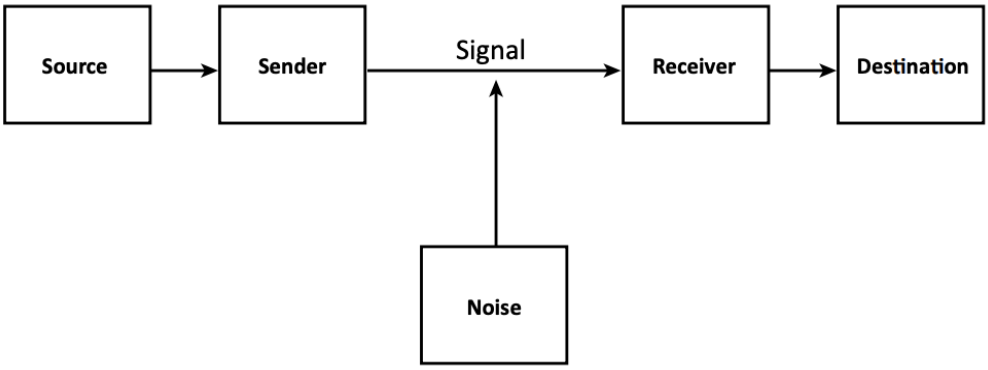


Figure 1.1: Figure from Gibson et al. (2019). The information theoretic pathway of successful communication. This figure shows that every communication has a source and a sender that sends a signal, the signal is then impacted by noise before it arrives at the receiver which deciphers the signal before it reaches its final destination.

1.2 The different costs of successful communication

The efficiency literature has not converged on a single definition of cost. Some researchers focus on production cost: the effort speakers expend in planning, articulating, or writing their utterances. Others emphasize comprehension cost: the cognitive effort listeners invest in decoding and interpreting what they hear. Cost has also been framed in terms of the time a message takes to transmit in a verbal channel or the space it occupies in a written channel. These different emphases can be organized into a small number of recurring categories: (i) the effort required to articulate or write a message, (ii) the processing load involved in producing or comprehending it, (iii) the time or space required to transmit the message, and (iv) a broader category covering social and other costs that fall outside the cognitive or articulatory picture. These categories are not sealed off from one another, and the same phenomenon can often be described in more than one way, but they correspond to recognizably different emphases in the literature. Word length, for instance, is simultaneously a matter of articulatory effort and processing time. Figure 1.2, from Levshina (2022a), summarizes these distinctions in a single diagram. The paragraphs that follow walk through these categories, beginning with the most extensively studied: word length and articulatory

effort.

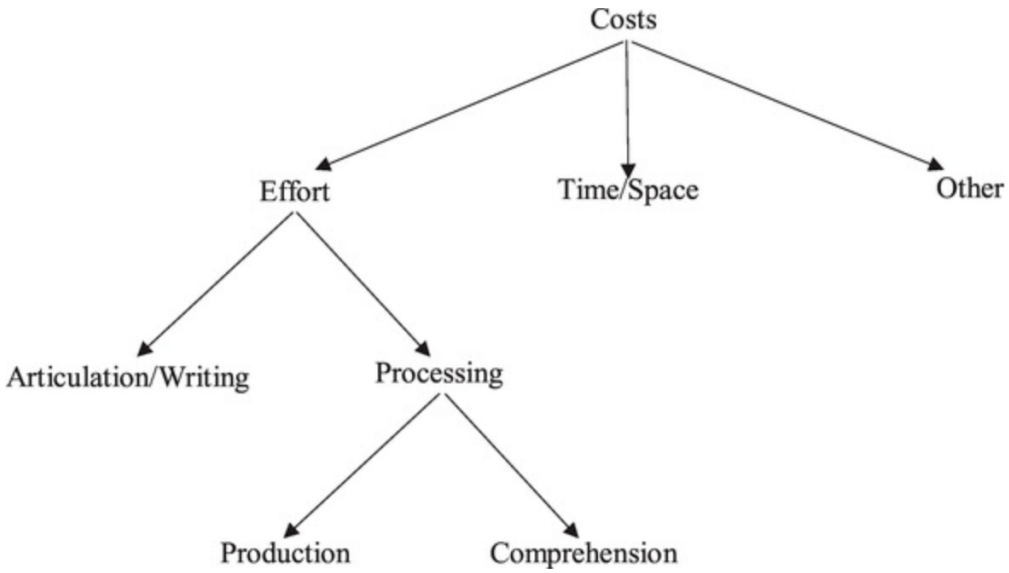


Figure 1.2: Plot from Levshina (2022a) illustrating the different costs involved in efficient communication.

1.2.1 Cost as word-length

Word length is perhaps the most straightforward operationalization of cost, and the one with the longest empirical tradition. As discussed in Section 1.1, Zipf (1949) established that more frequent words tend to be shorter, a regularity he attributed to the pressure to minimize production effort. But frequency is not the only factor that predicts word length. Piantadosi et al. (2011) demonstrate that word length is better predicted by a word's average surprisal than by its raw frequency alone. Surprisal is the information content of a word given its preceding context (see Equation 1.1).

$$S(w_i) = -\log P(w_i \mid w_1, w_2, \ldots, w_{i-1}). \tag{1.1}$$

Where $S(w_i)$ is the surprisal of word w_i and $P(w_i \mid w_1, w_2, \ldots, w_{i-1})$ is its conditional probability given the preceding context. Their work shifts the efficiency story from a static property of words in the lexicon to a contextual one, in which the cost of a word depends not only on how often it is used but on how predictable it is given what has come before. Mahowald et al. (2013) extend this line of reasoning, showing that the information content of a word, rather than its frequency, is a stronger predictor of its phonological length across languages. Together, these findings suggest that articulatory cost is not simply a matter of how long or short a word is in the abstract, but of how much information it carries relative to its context. This point connects directly to the information-theoretic framing introduced in Section 1.1 and is one that my own work in Chapter 4 develops further.

1.2.2 Availability-based production

A second way cost has been operationalized is as processing load: the cognitive effort involved in producing or comprehending an utterance, rather than the physical effort of articulating it. According to this view, a word's cost is a function of how easily it can be retrieved from the lexicon and integrated into an unfolding structure, and a structure's cost is a function of how easily it can be planned and executed. Ferreira and Dell (2000) argue that production is most efficient when syntactic structures allow quickly selected lemmas to be mentioned as soon as they become available. A speaker who has retrieved the embedded subject of a sentence-complement structure like *The coach knew you missed practice* can begin articulating *you* immediately, whereas a speaker who has not yet retrieved the embedded subject may opt for *The coach knew that you missed practice*, using the optional complementizer *that* to buy planning time. Across six experiments, Ferreira and Dell show that speakers do in fact omit the optional *that* when subsequent material is readily available, but do not include it to spare listeners from temporary garden-path ambiguities. This finding is telling: when production pressure and comprehension pressure pull in opposite directions, production pressure wins.

What the speaker treats as costly, then, is not simply the articulation of extra words, but the moment-by-moment planning and retrieval effort of composing an utterance in real time. More broadly, this view suggests that cost is shaped by how accessible a linguistic unit is; a perspective that extends beyond sentence production to the structure of the lexicon itself.

1.2.3 Cost of ambiguity

A related implication of this accessibility-based view of cost is the functional role that ambiguity plays in language. Piantadosi et al. (2012) argue that ambiguity is not a design flaw but a feature of efficient communication. Grounding their argument in information theory, they show that any efficient communication system in which context is informative about meaning will necessarily be ambiguous. Allowing linguistic units to map to more than one meaning lets speakers reuse the shortest, most accessible forms across contexts, rather than inventing a distinct form for every possible message. The reasoning is the following: language tolerates, and even exploits, ambiguity where context is reliable enough to disambiguate; short and frequent forms can do double duty across many meanings. Consider polysemy, where a single word encompasses multiple meanings. The English verb *run* can denote physical motion (*John runs to the store*), smooth continuous movement (*she runs her fingers through her hair*), or process execution (*the scientist runs a language model on his laptop*), among many other senses. Out of context *run* is ambiguous; in context, speakers and listeners converge on the intended meaning with essentially no disambiguating effort. Piantadosi et al. show further cross-linguistic empirical support for their claim through the use of homophones. Homophones, words that share the same form but not the same meaning, are on average shorter and more frequent than non-homophonous words across English, German, and Dutch. This is exactly what an efficiency-based account would predict: the forms that speakers use most often, and that are therefore most accessible, are precisely the ones that the language allows to carry multiple meanings. Ambiguity, on this view, is

not incidental but is a direct consequence of the pressure to minimize production cost while relying on context to preserve communicative success.

1.2.4 Cost as motor effort

Cost can also take the form of motor-production effort. Fedzechkina and Jaeger (2020) investigate this through miniature artificial language learning experiments in which learners are exposed to languages with optional case marking. In their design, producing case markers requires additional effort in the form of extra keyboard clicks in the experimental interface. Learners systematically reduced their use of case marking when it required this additional effort, but they did not eliminate it entirely. Instead, they retained case markers in contexts where they were needed to disambiguate the intended message. This pattern suggests that speakers balance production effort against the need for robust message transmission: grammatical structure is simplified where doing so does not compromise communicative clarity, but preserved where it is necessary for the listener to recover the intended meaning.

1.2.5 Comprehension cost

Cost does not apply only to language production; language comprehension can be costly as well. Speakers may adjust their utterances to reduce the effort listeners spend in arriving at the intended message. Rubio-Fernández (2016) illustrate this using redundant color adjectives. In a referential context where a single blue cup is the only cup on the display, saying *the blue cup* is technically overinformative since *the cup* would suffice. Yet speakers routinely include the color adjective because color is salient and facilitates object identification. As a result, the listener can begin narrowing down potential referents earlier in the utterance, arriving at the intended referent faster than they would without the adjective. Rubio-Fernández and Jara-Ettinger (2020) develop this argument further, connecting

redundant modification to the incrementality of language processing. From an efficiency perspective, what looks like unnecessary redundancy on the speaker's side turns out to reduce comprehension cost on the listener's side.

The cost categories surveyed so far do not exhaust the ways cost has been discussed in the literature, but they cover the dimensions most directly relevant to the scope of the present work. The next section turns to the other side of the trade-off: what counts as reward in efficient communication.

1.3 Rewards of successful communication

Cost has received extensive attention in the efficiency literature, but the discussion of reward has been comparatively thin. Communication is typically deemed successful when the intended message reaches the listener, and the analytical work in the literature has gone into specifying what makes this transmission costly rather than what makes it rewarding. This is understandable, since reward is harder to measure: cost can often be represented by the signal itself, in the form of word length, articulation time, or processing load, whereas reward depends on what the listener does with the signal, which is less directly visible. In practice, then, reward is often inferred rather than measured directly: once cost has been pinned down on both ends of the channel, reward emerges as whatever remains once cost has been accounted for. This section surveys the reward notions that have been implicit in the efficiency literature.

1.3.1 Reward as successful transmission

The most direct consideration of reward comes from Gibson et al. (2019). On this view, reward is the successful reconstruction of the intended message and they make this explicit:

communicative success is "the event where the information source's intended message is reconstructed accurately at the destination". A code that delivers the message reliably despite the presence of noise is what Gibson et al. call robust. All of the other notions of reward mentioned in the remainder of this section hold the same intuition that the listener should end up with the intended message. Incrementality-based accounts, for example, sharpen the temporal grain of reward, asking not just whether the message is recovered but how quickly.

1.3.2 Reward as faster or earlier comprehension

Listeners process incoming speech in real time, integrating each word the moment it arrives rather than waiting for a complete utterance (Tanenhaus et al., 1995). A word that lets the listener rule out candidate meanings earlier therefore has communicative value beyond the bits it adds — it moves the listener toward the referent sooner. Rubio-Fernández and Jara-Ettinger (2020) make this reward implication explicit in what they call the **visual efficiency hypothesis**: the efficiency of a referring expression should be evaluated "not only on its informational value, but also on its discriminatory value". Their eye-tracking evidence bears this out directly: For example, participants were faster to locate *the blue star* than *the star* in a display containing only one star, even though the color adjective is informationally redundant. Whether the same anticipatory benefit arises when the color word follows the noun, as in a post-nominal language, is precisely the cross-linguistic question Chapter 2 takes up. Reward here is not merely arriving at the right message, because the listener would have arrived at it anyway without the extra modifier, but a matter of how quickly the arrival happens; a form that lets the listener commit earlier is more rewarding than one that does not, even when both ultimately deliver the same message.

1.3.3 Reward as message clarity

Gibson et al. (2019) locate reward at signal-to-message transmission, which is mainly concerned with whether the message arrived successfully. Piantadosi et al. (2012), by contrast, highlight form-to-meaning recovery: what matters is whether the listener arrived at the correct interpretation of the message. This shift matters because a signal can arrive perfectly intact and still be ambiguous. Conversely, a signal that is ambiguous out of context can still deliver its intended meaning reliably if context makes one interpretation much more probable than the others. Piantadosi et al.'s term for this notion of reward is **clarity**: "a clear communication system is one in which the intended meaning can be recovered from the signal with high probability" (Piantadosi et al., 2012, p. 281). Reward, on this view, is the probability that the listener arrives at the intended meaning, a probability that short, reusable, and potentially ambiguous forms can maintain just as well as longer unambiguous ones.

1.3.4 Reward as ultimate survival

Levshina situates communicative reward within a broader evolutionary picture: the ultimate reward of communication is survival, such that successful communication leads to cognitive benefits, and ultimately to behavior crucial for survival (illustrated in Figure 1.3). What this adds is not another reward alongside the three above, but an evolutionary rationale for why speakers and listeners optimize for them at all.

Taken together, these notions of reward sharpen rather than replace one another. Gibson et al.'s transmission criterion fixes the baseline: communication succeeds when the listener recovers the intended message, Rubio-Fernández and Jara-Ettinger's incrementality criterion refines when that recovery counts, pushing the evaluation earlier into the unfolding signal, and Piantadosi et al. relaxes exact recovery into high-probability recovery. What they share

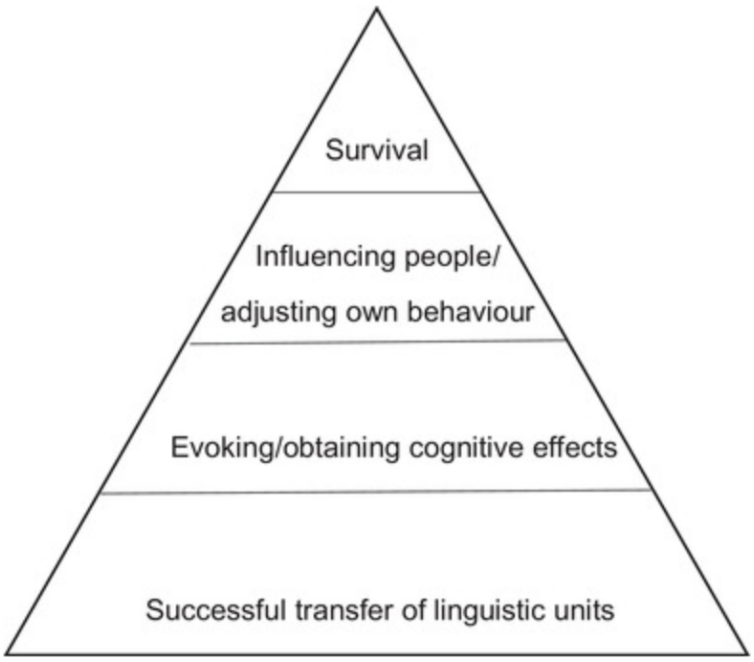


Figure 1.3: Plot from Levshina (2022a) about the different benefits of language communication

is a listener-anchored conception of reward.

1.4 Looking ahead

The picture this chapter has developed is one in which speakers and listeners navigate a trade-off of cost and reward. Cost is distributed across articulation, planning, and memory on the production side, and across processing and memory on the comprehension side, while reward is anchored on the listener's side in what is recovered, and how clearly or quickly that recovery happens. Research over the past two decades has applied this cost-reward framework to a range of linguistic phenomena. At the heart of this framework is the claim that linguistic form interacts with communicative function: the forms speakers produce and the structures languages settle into are the ones that lower cost and raise reward. Three lines of work stand out because they map onto the threads that this thesis will take up: how speakers adapt their utterances to their listeners' needs under structural linguistic constraints, how

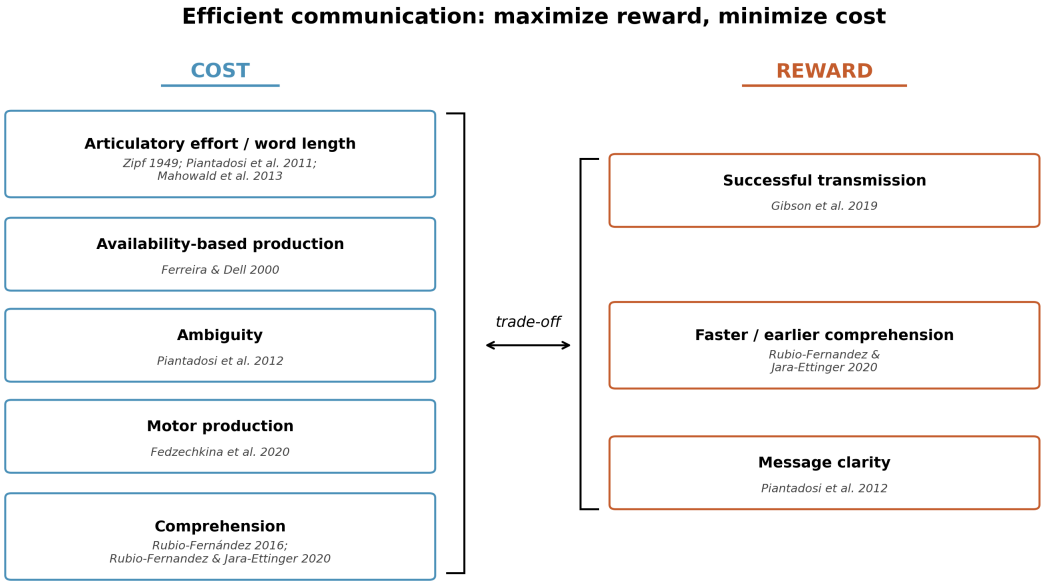


Figure 1.4: Different cost and reward considerations of efficient communication. Speakers and listeners navigate a trade-off between minimizing cost and maximizing reward during conversations. Generated with the help of Claude.

languages evolved differently in response to these communicative pressures, and how context and predictability shape linguistic form to maximize its function.

The first line of work concerns referential communication. Speakers' referring choices repeatedly show that utterances are tailored not to the speaker's own production cost alone, but to what the listener needs to identify the intended referent. Redundant adjective modification and the inclusion of perceptually salient features under pressure from incremental language processing both point to audience-aware production (Rubio-Fernández, 2016; Degen et al., 2020). A second line of work turns to cross-linguistic variation in word order, such as whether adjectives precede or follow the noun. Efficiency-based accounts argue that languages arrange linguistic material so as to keep processing cost low, a principle captured by dependency-length minimization (Futrell et al., 2015) and Hawkins' principle of maximizing online processing (Hawkins, 2004). A third line of work concerns the lexicon, where

word length and phonetic reduction have been tied to contextual predictability (Zipf, 1949; Piantadosi et al., 2011).

The efficiency literature has been built predominantly on English and a small set of typologically adjacent languages, with recent cross-linguistic extensions reaching Spanish, Mandarin, and a handful of others. Arabic, along with languages with structurally distinct profiles from English, has remained largely outside this empirical conversation and significantly understudied. In this thesis, I take Arabic as a principal test case for the three lines of work mentioned above. Chapter 2 takes up adjective modification, asking whether overmodification patterns established in pre-nominal languages hold when the adjective follows the noun. Chapter 3 turns to cross-linguistic variation in word order, using adjective subjectivity as a diagnostic to ask whether post-nominal Arabic exhibits the hierarchical ordering preferences documented for pre-nominal languages like English. Chapter 4 shifts to the lexicon, investigating vowel dropping in Arabizi, a Romanized colloquial Arabic script used online, and testing whether contextual predictability shapes whether a word is written in its long or short form. Together, these chapters investigate the role of communicative efficiency in shaping linguistic choices, extending the empirical conversation beyond the languages that have dominated the literature to date.

Chapter 2

The efficiency of adjective modification

2.1 Introduction

Successful communication depends on the ability of the speaker and the listener to coordinate on the intended meaning in real time. One mechanism through which this coordination is achieved is adjective modification. By adding adjectives to nouns, speakers constrain the set of possible referents, which reduces uncertainty and thus ease of reference resolution.

2.2 Background

2.2.1 How speakers use adjectives to establish reference

I begin my investigation of adjective modification by adopting a central assumption from the literature on referential processing: that adjectives contribute to reference resolution

by incrementally constraining the set of potential referents available to the listener as an utterance unfolds [1]. Adjectives are one of the primary linguistic tools that speakers use to support reference resolution, the process by which a listener identifies the entity a speaker intends to denote. The intended referent may be a perceptually present object, *"the green cup on the table"*, a previously encountered entity, *"I saw a green box in the store yesterday"*, a location, *"Are you near the green building yet?"*, or an object not currently in view, *"Can you grab the green folder from my office?"* Across these cases, successful communication depends on the listener's ability to identify the intended referent efficiently as the utterance unfolds in real time. Adjectives play a particularly useful role here because the descriptive content they contribute can be integrated the moment it is heard: on encountering *the green*, a listener can already begin restricting attention to the subset of green items in their visual field before the noun is even encountered.

The incremental nature of referential processing has been demonstrated in a series of influential eye-tracking studies. Tanenhaus et al. (1995) introduced the visual-world paradigm, showing that listeners direct their gaze toward relevant objects before the utterance is complete, which establishes that listeners integrate visual and linguistic information in real time during spoken language comprehension. Building on this, Sedivy et al. (1999) propose that incremental processing relies on listeners' ability to quickly identify contrast, whether against other objects in the visual context or against a typical category representation in memory. For example, upon hearing "Pick up the tall glass," listeners immediately fixated on the tall object, narrowing their referential search before the head noun was processed (see Figure 2.1 for an example of the experimental paradigm used). Crucially, this anticipatory behavior was modulated by the referential context; listeners were faster to restrict attention when a contrasting object (e.g., a short glass) was present, suggesting that they use adjectival information not merely to describe but to actively discriminate among potential referents during

[1] Reference resolution is one of several functions adjectives serve; I focus on it here because it is the locus of the efficiency-based literature this chapter engages with.

online processing (see also Sedivy (2003); Qing et al. (2018) for more nuanced interpretations of these eye-tracking patterns). When adjective information is provided early (particularly in languages with pre-nominal adjectives such as English), listeners can immediately narrow the set of candidate referents before the noun is encountered, enabling faster and more accurate reference resolution. This advantage becomes especially pronounced in contexts where multiple objects share overlapping properties.

These dynamics sit within a broader view of language use as shaped by efficiency: speakers produce utterances that balance the cost of production against the benefit of successful communication, selecting more accessible and readily retrievable forms when doing so aids the listener (Piantadosi et al., 2012; Ferreira and Dell, 2000). Hawkins (2004) formalizes one aspect of this pressure in the principle of *maximizing online processing*, according to which speakers arrange linguistic material so that listeners can build interpretations as quickly as possible. In other words, adjective use exemplifies this principle: by front-loading discriminating information, speakers reduce the listener's processing burden and distribute the work of reference resolution across the unfolding utterance.

This picture, however, rests on two simplifications. The first is empirical: it draws almost entirely on evidence from English and other languages where adjectives precede the noun, leaving open whether the same incremental dynamics hold when adjectives follow the noun. The second is pragmatic: it assumes that speakers include only as much modifying information as is strictly needed to single out the intended referent. In practice, neither simplification holds reliably. Speakers frequently produce referring expressions that contain more adjectival information than the minimum required for successful reference resolution, a phenomenon known as overmodification, and the cross-linguistic picture is considerably more varied than the English-based literature alone would suggest. I take up overmodification in the next subsection, and return to the question of post-nominal modification in the subsection following.

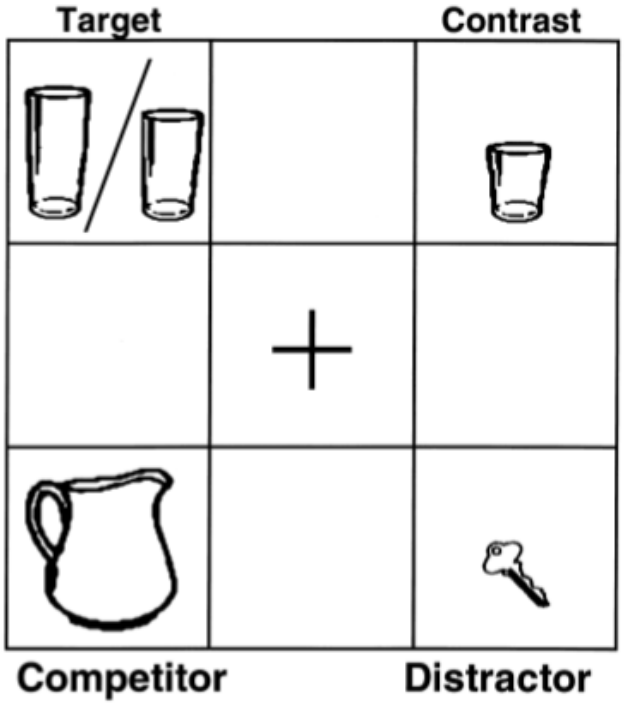


Figure 2.1: An example of a display from Sedivy et al. (1999) showing contrast conditions, where a participant would narrow down the potential referent of tall objects (the tall glass and the pitcher) before hearing the noun.

2.2.2 Overmodification

Speakers frequently include adjectives that are not strictly necessary for successful reference resolution (Pechmann, 1989; Sedivy, 2003; Koolen et al., 2011, among others). In referential communication tasks, where a speaker is asked to describe a target object so that a listener can pick it out from a visual display, speakers routinely produce expressions that exceed the minimal information required to uniquely identify the intended referent in context. Consider a simple contrast: in a display containing a wooden box and a metal box, the expression *the wooden box* is minimally informative, since the material adjective is needed to distinguish the target from its competitors. In a display containing only a single box, however, the same expression is informationally redundant: the unmodified noun *the box* would suffice. This pattern, known as overmodification or overspecification, has been documented repeatedly across a range of experimental paradigms and adjective types (Belke and Meyer, 2002; Sedivy, 2003; Koolen et al., 2013).

From a Gricean perspective, overmodification appears to be in tension with the Maxim of Quantity (Grice, 1975). The Maxim of Quantity has two sides: speakers should make their contribution as informative as is required for the conversation, but they should not make it more informative than is required. The redundant adjective in the example above, uttered in a context where only one box is present, is a clear violation of the second clause. Yet a striking and consistent cross-linguistic pattern emerges in the production literature: speakers reliably over-specify rather than under-specify (Pechmann, 1989; Rubio-Fernández, 2019). This asymmetry is itself telling: under-modification might be explained by the communicative cost it carries; by omitting words or modifiers from an utterance, a speaker risks reference failure, whereas including an extra one does not. Over-informativeness is therefore the safer choice. Crucially, the prevalence of overmodification is robust, which suggests that it is not random noise in production but a systematic feature of how speakers design referring expressions.

Several accounts have been proposed for why speakers reliably overspecify, and most of them share a common assumption: that referring expressions are not generated in isolation by a speaker minimizing their own production effort, but rather tailored to what the speaker believes the listener needs to know. Eye-tracking evidence supports this picture: speakers attend to the listener's visual context during production, suggesting that audience design is integrated into utterance planning in real time (Sedivy et al., 1999). Others have argued that speakers estimate the amount of information required for successful reference on the fly and adjust their utterances accordingly (Pechmann, 1989). Under this view, when speakers anticipate increased processing difficulty for the listener, they may take extra measures to ensure clarity — slowing their speech, simplifying lexical choices, or adding descriptive material. What unites these accounts is the idea that overmodification reflects a trade-off between production cost and the anticipated benefit to the listener, one that speakers appear to manage rapidly and with little apparent effort.

A more recent line of work pushes this trade-off framing further, arguing that overmodification is itself a form of efficient communication, because it can facilitate the listener's perceptual search for the intended referent (Rubio-Fernández, 2016, 2019; Rubio-Fernández et al., 2020; Wu and Gibson, 2020). On this view, a redundant adjective can carry real communicative value: it gives the listener a feature that can be matched against the visual scene quickly, lowering the overall cost of reference resolution even though it adds extra production effort. What motivates the speaker here is not minimal production but the listener's anticipated difficulty in identifying the referent. Koolen et al. (2013) demonstrated that overmodification rates increase with visual scene variation, showing that the complexity of the visual context directly influences the form of referring expressions. Rubio-Fernández (2019) extend this reasoning, arguing that speakers anticipate the listener's search difficulty rather than strictly adhering to a principle of minimal modification. This reframes the Gricean Maxim of Quantity in practice: speakers aim not at minimal informativeness, but at sufficient informativeness for the listener's search. Taken together, these findings suggest

that overmodification is not a violation of efficiency principles but a manifestation of them, provided that efficiency is measured at the level of the joint speaker-listener communication rather than the speaker's production alone. The broader takeaway is that efficiency does not necessarily mean reduction: overmodification reflects systematic pressures operating at the level of both production and comprehension.

Despite this empirical support, the existing evidence for the efficiency of overmodification is narrow in several important respects. The bulk of the relevant findings come from controlled referential communication experiments involving color adjectives in visual displays of physical objects, and the cross-linguistic comparisons that have been conducted typically span only two or three languages. These are valuable test cases, but they leave open whether the efficiency pressures observed in such settings extend to other adjective types, to communicative contexts that go beyond visual identification of physical objects, and to a typologically broader sample of languages. The two sub-subsections that follow take up the parts of this picture that have received the most empirical attention. I first examine the most extensively studied case, the overmodification of referring expressions with **color** adjectives, and then turn to the asymmetry between **color** and **size** adjectives in overmodification rates. Together, these two cases set up the broader cross-linguistic question that motivates the remainder of the chapter.

Color overmodification

Empirical observations suggest that redundancy is not uniform across different adjective types. Color adjectives, in particular, tend to overmodify nouns at strikingly high rates relative to other adjective types, even when color is not needed to establish reference (Pechmann, 1989). This disproportionate tendency to include color has been replicated across experimental paradigms and populations (Sedivy, 2003; Koolen et al., 2011; Rubio-Fernández, 2019), and cross-linguistic extensions of this work have established color as the central test case for

production accounts of redundant modification (Rubio-Fernández, 2016; Rubio-Fernández et al., 2020; Wu and Gibson, 2020). It has also been the primary empirical target of recent computational models of overinformative reference resolution (Waldon and Degen, 2021; Degen et al., 2020), making color adjectives the most extensively studied case of redundant adjective modification in the production literature.

Several factors have been shown to modulate the likelihood that a speaker will include a redundant color adjective in their utterance. The first is the complexity of the visual scene itself. The more distractor objects are present in the interlocutors' shared visual field, the more likely speakers are to produce a redundant color term (Koolen et al., 2013). This pattern is consistent with the intuition that speakers anticipate the listener's difficulty in locating the target and respond by providing an additional perceptual cue, even when the bare noun alone would suffice for unique identification. A second factor is color typicality. Speakers are less likely to mention stereotypical colors (e.g., *yellow* for a banana) but are more likely to specify color when it is atypical (e.g., *a green banana* or *a yellow strawberry*) (Sedivy, 2003; Westerbeek et al., 2015; Rubio-Fernández, 2016; Kreiss and Degen, 2020). This typicality effect suggests that speakers take into account more than just the immediate visual context when choosing to include a color adjective. They are also sensitive to how informative the color is relative to what listeners already know about the category. A third factor is noun type: objects such as clothing and toys elicit more redundant color modification than simple geometric shapes, suggesting that lexical-semantic properties of the head noun, and perhaps the extent to which the referent naturally varies along the color dimension, also shape speakers' decisions to include color (Rubio-Fernández, 2016).

Importantly, color adjectives do more than add descriptive content: they have measurable consequences for comprehension. In displays containing contrast objects, the early use of color adjectives facilitates faster identification of the intended referent, a phenomenon known as *contrastive inference* (Sedivy et al., 1999; Sedivy, 2003). For example, consider a display

that has a yellow banana and a green one. On hearing *the yellow*, listeners immediately begin to fixate on the yellow banana, often before the head noun has been processed at all, because they expect that a cooperative speaker would use the color adjective contrastively, i.e., to distinguish the target from a competitor that shares its category but not its color. Crucially, this expectation carries over to non-contrastive contexts: even when no competitor is present, listeners interpret a color adjective as if it were contrastive and begin narrowing down referents immediately. The facilitative effect of color adjectives in contrastive referential contexts indicates that redundant color modification may serve a communicative function beyond strict reference resolution. Even when color is not required for unique identification, it allows listeners to exploit their pragmatic expectations about why a speaker would include it, thereby accelerating reference resolution.

This view of color overmodification as communicatively beneficial has been formalized within the Rational Speech Act (RSA) framework, which models production and comprehension as the output of probabilistic reasoning between a speaker and a listener, each of whom models the other (Frank and Goodman, 2012). In particular, Degen et al. (2020) develop an RSA model of referring expressions with redundant color adjectives and show that a speaker who takes the listener's search costs into account prefers overmodified descriptions to unmodified ones when color provides a perceptually distinctive cue. Their model also predicts that redundant color modification should be sensitive to fine-grained properties of the visual scene and to how predictable the target is. The model captures a key intuition: the listener's processing constraints are part of what determines which referring expression counts as optimally informative, in addition to the information content of the utterance itself. This supports the view that speakers behave as rational communicators rather than mechanically minimizing or maximizing the amount of descriptive content in their utterances.

A separate but complementary line of empirical work has extended these findings into a wider range of communicative settings (Rubio-Fernández, 2016, 2019; Rubio-Fernández et al.,

2020). Across a series of studies, it has been shown that speakers reliably overmodify with color in contexts that vary in visual complexity, referent familiarity, and task demands. Listeners in turn use the resulting color adjectives to speed visual search in a way that outweighs the added production cost. In Rubio-Fernández (2019), for instance, speakers overspecified with color most heavily in displays where the target was embedded among perceptually similar distractors, precisely the conditions under which a listener would benefit most from an early-arriving perceptual cue. Taken together, these findings paint a consistent picture: color overmodification is not a pragmatic slip but a systematic response to the joint demands of reference and visual search. The result is an efficiency trade-off such that the benefit to the listener outweighs the extra effort on the speaker's side.

Color overmodification thus offers a strong empirical case for the efficiency story of redundant modification. That is, if efficiency is measured at the level of the collaborative speaker-listener communication rather than in terms of utterance length alone, then producing a redundant color adjective is an efficient choice whenever the benefit to the listener's visual search outweighs the small cost of articulating an extra word. This is exactly the trade-off that the work of Degen et al. (2020) converges on for the efficiency of color overmodification.

At the same time, the disproportionate focus on color adjectives in this literature raises an obvious question: do these findings generalize to other adjective types, or do they reflect properties specific to color as a visually salient feature? The literature on adjective overmodification has long noted that size adjectives behave differently in production. Speakers are markedly less likely to include redundant size adjectives than redundant color adjectives, even in visual displays where size is just as useful for discriminating the target. This asymmetry has attracted considerable attention in the overmodification literature, because it suggests that the efficiency account cannot rest on visual salience alone: something about the specific properties of color must be driving the difference in production.

Color-size asymmetry

A consistent asymmetry has been observed between color and size adjectives: speakers reliably overproduce color but rarely overspecify size (Pechmann, 1989; Belke and Meyer, 2002). For example, when a display contains only one pen, a speaker is likely to say *the red pen* rather than simply *the pen*, including color even though it is not needed to identify the referent. Size, by contrast, is rarely included when it is not distinguishing. The magnitude of this effect is striking. Pechmann (1989) found that 98% of all overmodification in his data involved color, while size was added redundantly in only a single utterance across the entire dataset. Belke and Meyer (2002) report the same pattern in a larger sample: when either color or size could serve as the minimal modifier, speakers chose redundant color modifiers significantly more often than redundant size modifiers ($p=.00016$, $N=18$). The asymmetry has since been replicated across multiple studies and extended to other dimensions such as material (Degen et al., 2020; Kursat and Degen, 2021). In particular, Degen et al. (2020) provide a direct production test of the color-size asymmetry under varying scene complexity. In a web-based interactive reference game, speakers were reliably more likely to include a redundant modifier when color was sufficient than when size was sufficient, and redundant modification increased with scene variation (see Figure 2.2). Crucially, speakers tend not to include redundant size or material adjectives when color alone suffices to establish reference, yet they frequently include redundant color adjectives even when size or material is already sufficient (Sedivy, 2003; Gatt et al., 2011; Rubio-Fernández, 2016, among others).

These explanations operate at three connected levels. At a perceptual level, color is visually available early and processed rapidly, making it more accessible during incremental planning (Pechmann, 1989); see also Rubio-Fernández and Jara-Ettinger (2020). Because speakers begin formulating a referring expression before they have fully scanned the scene, highly salient features are more likely to be retrieved in time to appear pre-nominally, and color is a strong candidate for this kind of early retrieval. Size, by contrast, typically requires a

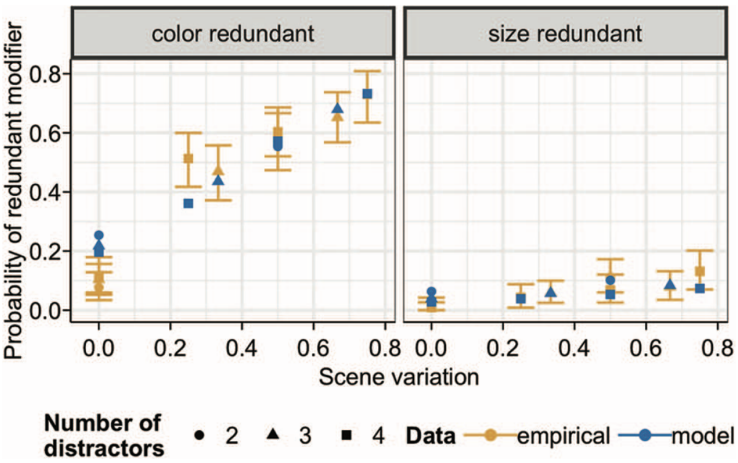


Figure 2.2: Results from Degen et al. (2020) showing that speakers significantly use redundant color adjectives relative to size in their first experiment. The results are shown alongside estimates of their RSA model for redundant utterance probability as a function of scene variation.

comparison across objects in the display, and this comparison takes time that the incremental production system may not have.

At the level of semantic structure, unlike size adjectives, which are context-dependent and relative (what counts as *big* depends on a comparison class), color adjectives denote relatively stable and absolute properties (Belke and Meyer, 2002). A red pen is red regardless of what sits next to it, but a large pen is only large relative to the other pens in view. This makes color cheaper to encode: the speaker does not need to compute a comparison class before committing to the word, and the listener does not need to resolve one before interpreting it. Redundant color is therefore low-cost to produce and low-cost to understand, while redundant size carries overhead on both ends.

On the listener's side, Degen et al. (2020) reframe the asymmetry in terms of listener-oriented efficiency. On their account, speakers include redundant modifiers when doing so helps the listener identify the referent more quickly, and the benefit of a redundant modifier depends on how reliably it narrows the referential search. Color is a near-perfect cue: it is easy to perceive, easy to verify against the scene, and rarely ambiguous. Size is a noisier cue, and a

redundant size adjective may even slow the listener down by inviting a contrastive inference that the scene does not support. Under this view, the asymmetry is not a quirk of production but a reflection of which features actually contribute to communicative success.

Taken together, these accounts suggest that overmodification is not governed by a uniform principle of minimal production, but instead reflects systematic interactions between perceptual accessibility, semantic representation, and communicative cost. The asymmetry between color and size therefore poses a challenge for accounts that treat all adjectival modification as functionally equivalent in referential communication, and it motivates a more fine-grained view in which different dimensions contribute differently to the speaker-listener trade-off.

So far, the evidence reviewed has treated overmodification as a question of which features speakers choose to include, with color emerging as the dimension most likely to be produced redundantly. But the story does not end with feature choice. A separate line of work has shown that where an adjective appears relative to the noun also shapes whether it is produced. The next subsection turns to the cross-linguistic dimension of this asymmetry: how adjective position relative to the noun shapes overmodification rates across languages.

2.2.3 What happens post-nominally?

Cross-linguistic asymmetry

So far I have only focused on languages with pre-nominal adjectives, and it's not obvious whether the efficiency account extends to post-nominal languages. Cross-linguistic evidence suggests that the tendency to overmodify with color is tightly bound to word order. Speakers of languages with pre-nominal adjectives behave differently from speakers of languages with post-nominal adjectives, and these differences map onto the incremental pressures that shape production in the first place. Rubio-Fernández (2016) found that Spanish, a language with

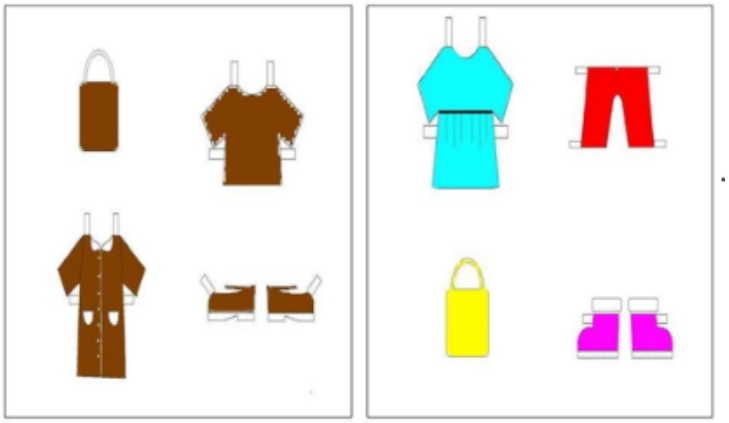


Figure 2.3: Example of visual paradigm from Rubio-Fernández (2016). English and Spanish speakers were asked to refer to one of the items in this 4-item display. On the right is an example of a polychrome display where all the items differ in color while on the left is a monochrome display where all the items exhibit the same color. English speakers produced redundant color adjectives more than Spanish speakers in both displays, although the color adjective is not needed to identify the referent in either case.

post-nominal adjective modification, exhibits a lower rate of color adjective modification compared to English, a language with pre-nominal adjective modification, particularly in referential tasks involving objects such as clothes and toys that can vary in color. This cross-linguistic difference is especially pronounced in contexts where color is redundant and not required to establish reference. In a Paper Dolls task, Rubio-Fernández (2016) asked English and Spanish speakers to refer to garments in 4-item displays of paper clothes and shoes (see figure 2.3). In polychrome displays, where the target was the only object of its color, English speakers produced redundant color adjectives on 95% of trials, while Spanish speakers did so on only 59% ($U = 42, Z = 4.26, p < 0.001$). The same asymmetry held in monochrome displays, where color carried no discriminating value: English speakers still produced redundant color on 37% of trials, against 5% for Spanish speakers.

The proposed explanation for this asymmetry rests on the incremental nature of language processing. In a pre-nominal language like English, a color adjective arrives before the noun and can begin to narrow the referential domain immediately: a listener hearing *the blue* can

already restrict their visual search to blue objects before the noun is even produced. In a post-nominal language like Spanish, the same color adjective arrives after the noun, by which point the listener has already committed to searching for the noun. The redundant color adjective therefore offers less anticipatory benefit, and the speaker has correspondingly less incentive to produce it. Rubio-Fernández (2016) illustrates this with the instruction *Give me the blue cup*: the English listener can use *blue* to guide the visual search before the noun is even produced, whereas a Spanish listener would only encounter the color adjective after the noun, at a point when the reference has already been resolved. Subsequent work replicated this asymmetry, further supporting the claim that speakers of post-nominal languages produce fewer redundant color adjectives than speakers of pre-nominal languages (Rubio-Fernández and Jara-Ettinger, 2020; Rubio-Fernández et al., 2020).

One concern in the original 2016 study is that the experiments used clothes and shoes, categories for which color is a central property, leaving open the question of whether the asymmetry would survive when color is less central to the noun category. Rubio-Fernández (2019) addressed this directly by running the same English-Spanish comparison on geometrical shapes. Crucially, the experiments were timed so that both language groups had equal planning time, ruling out the possibility that Spanish speakers overmodify less simply because post-nominal syntax gives them more time to determine whether color is needed before producing the adjective. English speakers still produced redundant color adjectives on roughly 46% of trials, against 7% for Spanish speakers.

Several open questions remain, however. First, the cross-linguistic comparisons so far have been conducted in face-to-face referential settings, raising the question of whether similar patterns would emerge in other communicative contexts, for example online experiments where the speaker does not necessarily see their listener. Second, the contrast is based on cases of overmodification, situations in which the adjective is not required for successful reference and is considered redundant. It is not clear whether this asymmetry can be gener-

alized across the language as a whole: do Spanish speakers generally use fewer adjectives to modify nouns? Finally, if this is true about Spanish, does it generalize to other languages with post-nominal modification, or is it particular to Spanish?

Two further pieces of evidence bear on the incrementality account, one supporting it and one complicating the picture. First, the logic of the account is not specific to color. Wu and Gibson (2020) tested English and Spanish speakers with numerals, which precede the noun in both languages, and found that the cross-linguistic asymmetry in redundant modification disappeared: speakers of both languages produced numerals at comparable rates, even when the numeral was not required to establish reference. This is consistent with the hypothesis that word order, rather than language per se, is what drives the asymmetry. Second, the picture from signed languages is more complicated.

Kursat et al. (2022) investigated redundant modifier production in Central Taurus Sign Language (CTSL), a young village sign language from Southern Turkey in which modification is overwhelmingly post-nominal. In an interactive reference game with 22 native signers, they found that signers nonetheless produced redundant modifiers at substantial rates and showed the same color/size asymmetry attested in pre-nominal languages. Notably, while English speakers in their comparison experiment never produced redundant size modifiers, CTSL signers did so frequently. The authors conclude that these patterns reflect the referential usefulness of the modifiers themselves rather than the anticipatory benefit of early (pre-nominal) placement. That is, redundancy driven by what the modifier contributes, not by when it arrives. This counterexample raises the possibility that post-nominal languages may adopt alternative strategies for redundant modification. Together, these two pieces of evidence leave the cross-linguistic empirical picture mixed: incrementality predicts the numeral pattern across English and Spanish but is challenged by emerging post-nominal sign languages, suggesting the story may need more qualification once a broader typological sample is considered.

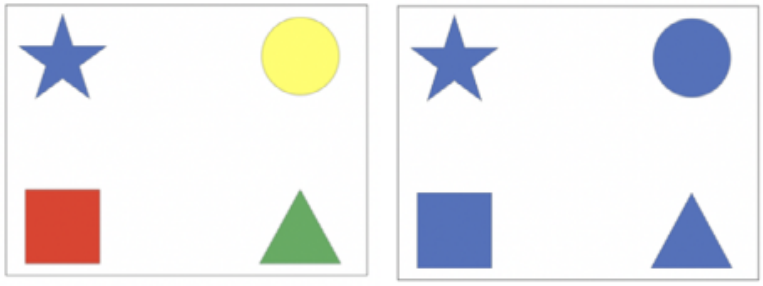


Figure 2.4: Example of visual paradigm from Rubio-Fernández (2019) where the authors addressed the issue of color adjective being central to the noun property. In this paradigm, both English and Spanish speakers were asked to refer to one of the 4 items in the display following the same polychrome/monochrome condition. In both cases, the color adjective is not required for reference identification.Yet, English speakers still produced redundant color adjectives on roughly 46% of trials, against 7% for Spanish speakers.

From experiments to corpora

The cross-linguistic evidence reviewed above connects to a broader conception of efficiency in language. Gibson et al. (2019) define efficiency from an information-theoretic perspective: languages allow humans to communicate successfully with minimal effort, subject to cognitive and environmental constraints. If a speaker's intended message matches the message received by the listener, the communication is successful; if this success is achieved with minimal effort, it is efficient. The incremental efficiency hypothesis (Rubio-Fernández, 2016; Rubio-Fernández and Jara-Ettinger, 2020) is one instantiation of this broader principle: speakers produce redundant color adjectives pre-nominally because doing so aids the listener's visual search at low cost, and they produce them less often post-nominally because the anticipatory benefit is absent.

Despite the empirical support for this hypothesis, three gaps remain. First, the cross-linguistic evidence rests almost entirely on the English-Spanish contrast, with only limited extensions to other post-nominal languages. Second, the existing computational accounts (Degen et al., 2020; Waldon and Degen, 2021) have been developed and tested in narrow experimental settings (typically, object-request tasks with small visual displays and short

referring expressions), and it remains unclear whether the efficiency pressures observed in these paradigms shape production in broader communicative contexts. Third, the evidence base is almost entirely experimental; we know little about whether the same pressures shape naturalistic written and spoken language use, where speakers are not constrained by a fixed referential task.

The present work addresses these gaps by turning to corpus data. Corpora of naturally-occurring written and spoken language offer a complementary window on adjective modification: they capture production at scale, across communicative contexts that are not constrained by a referential task, and across languages for which no experimental data currently exist. In particular, I extend the cross-linguistic comparison beyond English and Spanish to include Arabic, a language with post-nominal adjective modification that has not, to my knowledge, been tested in this paradigm. Whether Arabic speakers behave like Spanish, like English, or show a distinct profile altogether is an open question, and one this thesis takes up directly.

The hypothesis that adjectives are used at a higher rate in pre-nominal languages than in post-nominal languages follows directly from the incremental efficiency account: if redundant adjectives are used when they can anticipatorily narrow the referential search, then speakers of post-nominal languages should produce fewer of them overall. Assuming that incremental reference resolution is at least sometimes an objective in written and spoken production, it is expected that this efficiency-based pressure applies in naturally-occurring language as well, such that languages with pre-nominal adjectives use adjectival modification more frequently than languages with post-nominal adjectives.

2.3 Methods

2.3.1 Wikipedia Corpus Analysis

Building on Rubio-Fernández's hypothesis regarding pre-and post-nominal asymmetry of adjective modification, the supporting evidence from Wu and Gibson (2020), and the RSA model results of Waldon and Degen (2021), I test whether this prediction, originally formulated for redundant color modification in referential tasks, extends to overall rates of adjective use in naturalistic written text. I begin with the analysis of Wikipedia corpora for Spanish, English, and Arabic.

Tagging the Wikipedia data

I downloaded the Wikipedia corpus from the Wikimedia Dumps,[2] a large dataset of encyclopedia articles, open-sourced by Wikipedia. My work takes advantage of several tools for data collection and analysis. Using WikiExtractor (Attardi, 2012), I extracted plain text from the xml-formatted Wikipedia data. The English corpus consisted of 12 GB of data, while the Arabic and Spanish corpora were 1.2 and 3.7 GB, respectively. To identify nouns and the adjectives that modify them, I first need part-of-speech information. I used the StanfordCoreNLP Part-Of-Speech tagger (Manning et al., 2014) to tag the Spanish and English corpora. For Arabic, I used the Part-Of-Speech tagger from the Farasa Arabic NLP toolkit, which is made specifically for Arabic, and is more than an order of magnitude faster than comparable tools such as MADAMIRA and the Stanford POS tagger (Abdelali et al., 2016). An initial attempt using the Stanford POS tagger took much longer and led to lower accuracy. With the Farasa tagger, I noticed performance on a par with the Stanford tagger's performance on Spanish and English.

[2] `https://dumps.wikimedia.org/`

Estimating the rate of adjective modification

Once I had part-of-speech tags for each language, I identified all of the nouns (tagged as NOUN, NN, or NNS). I then collected all the adjectives (tagged as ADJ, JJ, or JJS) that were adjacent to the noun. The logic is as follows: I sequentially iterated over all the words in the corpus, and if I hit a noun, I check the two adjacent words (to the right and to the left). The noun will then fall into one of four buckets: (1) not modified, where there are no adjacent adjectives, such as the three-word segment *the box was*, (2) pre-nominally modified (AN), such as the three-word segment *blue box on*, (3) post-nominally modified (NA), such as *ra'aytu al-ṣundūq al-bunnī* (the equivalent of saw-I the-box the-brown) in Arabic, and (4) mixed modified (ANA), where the noun is modified both pre- and post-nominally, such as *gran caja marrón* (big box brown) in Spanish.

This method of identifying adjectives modifying nouns is approximate, as (1) some adjectives not directly adjacent to nouns are missed, and (2) some non-attributive adjectives that happen to be adjacent to a noun by chance may be unintentionally included (see Example (1)). However, this method has the advantage that it can be scaled up to large amounts of text and does not require parsing of structure.

Once I had the noun and adjective counts, I then calculated the rate at which adjectives are used to modify nouns in each of these languages. I calculated the rate by dividing the number of nouns with any adjectival modification (one of the three types mentioned above) by the total number of nouns.

Results

The results of the Wikipedia corpus analysis appear in Figure 2.5 and are summarized in Table 2.1. One striking pattern in the data is that the estimated modification rate is nearly

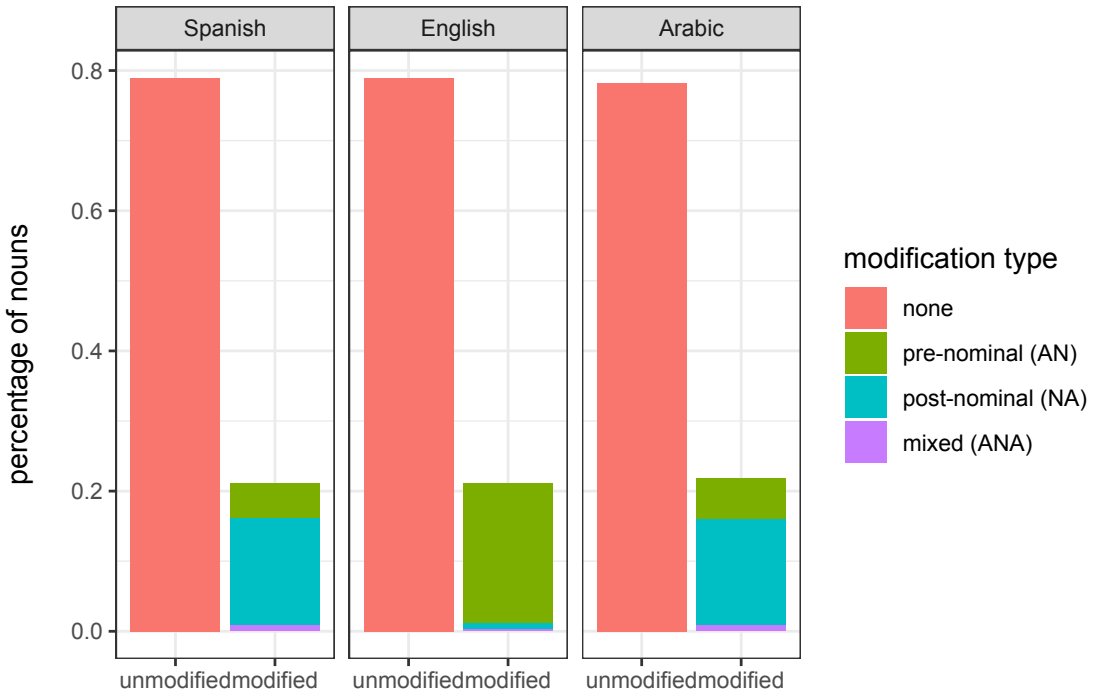


Figure 2.5: Wikipedia analysis results: percentage of noun modification in the analysis is similar across the three languages: Arabic, English, and Spanish.

identical across the three languages: roughly 21% of nouns are modified by an adjacent adjective in each, with English at 21.1%, Spanish at 21.1%, and Arabic at 21.8%. This convergence is notable given how structurally different the three languages are, particularly with respect to where attributive adjectives sit relative to the noun. If the incremental efficiency hypothesis predicts that pre-nominal languages should overmodify at higher rates than post-nominal ones, one might have expected English to use adjectives substantially more than Spanish or Arabic. Instead, all three languages settle on essentially the same total modification rate. A logistic regression of modification probability on language confirmed that the three rates are nearly identical. Spanish was indistinguishable from English (odds ratio ≈ 1.00, 95% CI $[1.00, 1.00]$), and Arabic was only marginally higher (odds ratio $= 1.04$, 95% CI $[1.04, 1.04]$), a difference of only 0.7 percentage points in probability. With hundreds of millions of nouns in the analysis, even minute differences register as statistically significant ($\chi^2(2) = 23{,}117$, $p < 0.001$), and the effect sizes here are negligible, so the three languages should be treated as having the same overall modification rate.

Table 2.1: Results of Wikipedia analyses for English, Spanish, and Arabic. # Nouns gives the total number of nouns extracted. Columns AN, NA, and ANA give the percentage of nouns with adjacent adjectives of the specified type. Total mod. gives the percentage of modified nouns which are adjacent to any adjective.

Language	# Nouns	%AN	%NA	%ANA	Total % Mod.
English	435,089,686	19.8	1.0	0.2	21.1
Spanish	128,946,337	4.9	15.4	0.8	21.1
Arabic	92,218,932	5.7	15.2	0.8	21.8

Looking beyond total rates, the internal composition of these rates lines up with each language's typological profile. English modification is overwhelmingly concentrated in the AN order (19.8%), with only minor contributions from NA (1.0%) and the mixed ANA configuration (0.2%). Spanish and Arabic show the mirror image: their modification is concentrated in the NA order (15.4% and 15.2% respectively), with smaller AN components (4.9% and 5.7%) and very low ANA rates (0.8% in both). Each language places its adjectives where its grammar would predict. The similarity in overall rate, however, sits in some tension with the literature, which predicts that pre-nominal languages should use adjectives more frequently than post-nominal ones.

While these results align with the typological profile of each language, the small but non-zero rates of AN modification in Arabic (5.7%) and NA modification in English (1.0%) deserve some unpacking. The Arabic rate is largely an artifact of construction rather than evidence against the post-nominal generalization. Arabic does allow adjective-first surface orders in certain constructions, but these follow the syntax of *idafa* rather than ordinary adjective-noun modification. In a phrase like (1), the adjective *jamila* takes the construct-state form and the modified noun appears in the genitive with the definite article; the surface order is adjective-noun, but the syntactic relation is distinct from the standard NA modification structure that dominates the rest of Arabic adjective use. A POS-tagger that labels by surface category will count these instances as AN.

(1) labis-tu jamīl-a al-thiyāb-i
wore-I beautiful.ACC the-clothes-GEN
'I wore the (most) beautiful (of) clothes.'

The 1% NA modification in English could be explained by instances like *attorney general*, or *the cup tall enough to hold the water*. These are licensed structures in English, but they sit outside the dominant pre-nominal pattern and contribute the small NA tail observed.

Finally, some portion of both residual rates likely reflects POS-tagger noise. At the scale of tens of millions of tokens, even a low per-token error rate will yield non-trivial counts of mislabeled adjective-noun sequences. The 5.7% AN rate in Arabic is large enough that genuine *idafa*-derived constructions are likely the main contributor, but at the 1.0% level seen in English NA, tagger noise becomes a more plausible candidate for a substantial fraction of the count.

Setting these residual patterns aside, the central finding from the Wikipedia analysis is that although the logistic regression detected a statistically significant difference across the three languages, the effect size is negligible, and the analysis therefore does not provide meaningful support for the efficiency hypothesis on this measure. However, two limitations should be noted. First, the Wikipedia text I investigate may commonly feature translations from one language's articles (often English) into another language. These translations could contribute to similarities across languages that would not arise in the absence of translated content, thereby artificially leveling differences in modification rates. Second, by only investigating three languages, the analysis is severely limited and may therefore miss important generalizations that only become apparent in the consideration of a larger set of languages. In the next section, I address both of these concerns through the analysis of UD corpora.

2.3.2 Universal Dependencies analysis

The broad corpus analysis is conducted on the Universal Dependencies (UD) Treebanks from the Linguistic Data Consortium (LDC) (Nivre et al., 2020). UD is a framework for consistent annotation of grammar (parts of speech, morphological features, and syntactic dependencies) across different languages, currently providing over 200 treebanks in more than 150 languages. Each treebank is constructed from a source corpus of naturally-occurring text in a particular genre or set of genres (spoken language, weblogs, newspapers, emails, fiction, religious texts, and others), and is annotated either manually by trained annotators or semi-automatically with human correction of automatic parses. Unlike the Wikipedia corpora described in the previous section, which I used a POS tagger to annotate, UD treebanks come manually tagged with part-of-speech information. Comparing the UD results with the targeted Wikipedia analysis allows for a more robust test of the efficiency hypothesis: do these trends extend beyond the handful of languages typically investigated in psycholinguistic experiments?

Part of Speech Analysis

Using the same logic as the Wikipedia query, I collected all the nouns and adjacent adjectives. Nouns were again classified into one of four categories: not-modified, pre-nominally, post-nominally, and mixed modified. I excluded languages with zero adjectival modification (Tagalog and Classical Chinese), as well as languages with fewer than 1000 nouns in the search (14 languages total); these criteria yielded data from 74 languages.[3] Using these counts, I then calculated the rate at which adjectives are used to modify nouns in each of the 74 languages by dividing the number of nouns with any adjectival modification by the

[3]It is likely that the reason no adjectival modification appears in the data for Tagalog and Classical Chinese is that these languages commonly feature a linking particle between adjectives and the nouns they modify, and this linking particle breaks up the adjacency of adjectives and nouns (Scontras and Nicolae, 2014) and (Shi and Scontras, 2020).

total number of nouns.

To test the efficiency hypothesis, I needed a way of characterizing a language as either pre- vs. post-nominal in its use of attributive adjectives. However, a quick look at the data revealed that languages rarely fit neatly into either category: although a language may favor pre-nominal modification, it is common for there to also be post-nominal uses, and vice-versa.

Rather than relying on a potentially-arbitrary binary classification of languages as pre- vs. post-nominal, I therefore characterized languages in terms of their propensity toward post-nominal adjectival modification: among the nouns with adjectival modification, what proportion of those nouns feature post-nominal adjectives? I calculated a language's post-nominal propensity by dividing the number of nouns with post-nominal adjectives (note that this number will include both nouns with strictly post-nominal adjectives, and mixed pre- and post-nominal adjectives, as in an ANA template common to languages like Spanish) by the number of nouns with any adjectives at all.[4]

Following the hypothesis of Rubio-Fernández (2016), and if the function of adjectives in writing is sufficiently similar to their function in visual-world-paradigm experiments, it is expected that languages with a higher post-nominal propensity feature fewer instances of adjectival modification. In Figure 2.6, I plot the results of my analysis, which shows that this prediction holds: as a language features more post-nominal modification, the proportion of modified nouns decreases; post-nominal propensity accounts for 18% of the variance in the proportion of noun modification ($r^2 = 0.18$, 95% CI [0.054, 0.325], Spearman's $\rho = -0.41$). Although the effect may be small, it is reliable. A mixed-effects linear regression predicting proportion of nouns with modification by proportion of nouns with post-nominal modification (i.e., post-nominal propensity), with random intercepts by language family confirmed this pattern ($\beta = -0.12$, $t = -3.30$, $p < 0.01$).

[4] I used an analogous 'pre-nominal propensity' measure and got nearly-identical quantitative results (i.e., a robust statistical correlation between pre-nominal propensity and modification rate).

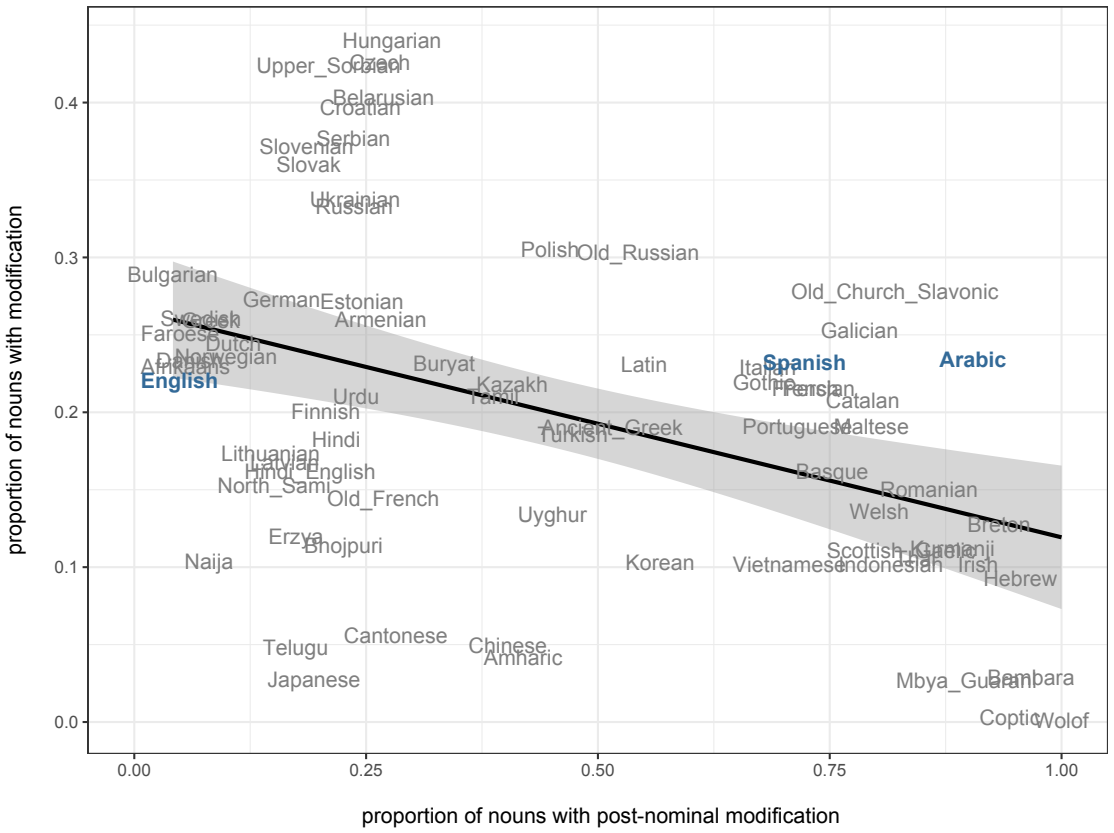


Figure 2.6: UD analysis results: proportion of nouns with any adjectival modification compared with the proportion of modified nouns with post-nominal adjectives ($r^2 = 0.18$; 95% CI [0.05, 0.32]).

One might worry that this trend is driven by the ten languages in the upper-left quadrant of Figure 2.6 with low post-nominal propensity and high modification rates; these languages (Hungarian, Czech, Upper Sorbian, Belarusian, Croatian, Serbian, Slovenian, Slovak, Ukrainian, and Russian) all have nominal modification rates above 0.33. Importantly, even if these ten potential outliers were excluded, the pattern persists, such that post-nominal propensity predicts nominal modification rates for the 64 languages that remain ($r^2 = 0.12$, 95% CI [0.008, 0.299], Spearman's $\rho = -0.40$). Therefore, I find reliable support for the efficiency hypothesis.

Dependency Analysis

The adjacency-based analysis reported above relies on POS tags alone, considering only nouns and adjectives that appear directly next to each other. By limiting the search to adjacent words, the analysis may miscount adjectives that are attributed to nouns, and it ignores a rich source of information available in these corpora: dependency relation annotations. To address this, I extended the part-of-speech analysis to take advantage of the dependency relation annotations in the UD treebanks. I collected all nouns and the adjectives related to them regardless of whether they are adjacent. Whenever the adjective occurred before the noun, it was classified as pre-nominal; whenever it occurred after the noun, it was classified as post-nominal.

The dependency analysis provides a more accurate count of adjectival modification than the POS-based analysis, because many adjectives that appear adjacent to nouns and were counted as attributive modifiers are now correctly reclassified. Consider the phrase *the box is brown*: in English, the linking verb *is* identifies *brown* as a predicative adjective rather than an attributive one. However, many languages lack a copula, so the equivalent phrase surfaces as something closer to *the box brown*. In these languages, the adjective *brown*, which is not an attributive modifier, would be counted as one in the POS-based analysis; the dependency annotations in the UD treebanks correctly distinguish these cases.

The dependency-based analysis does not show the same effect as the POS-based analysis: the correlation is weaker and not significant (Spearman's $\rho = -0.16$, $p = 0.18$). I plotted the correlation between post-nominal adjective use and overall adjective modification by taking the log probability of post-nominal adjective modification and plotting it against adjective modification overall (see Figure 2.7). As expected, many languages shifted downward and to the left in the dependency plot relative to the POS plot, indicating that a substantial number of adjectives had been miscounted in the adjacency-based analysis.

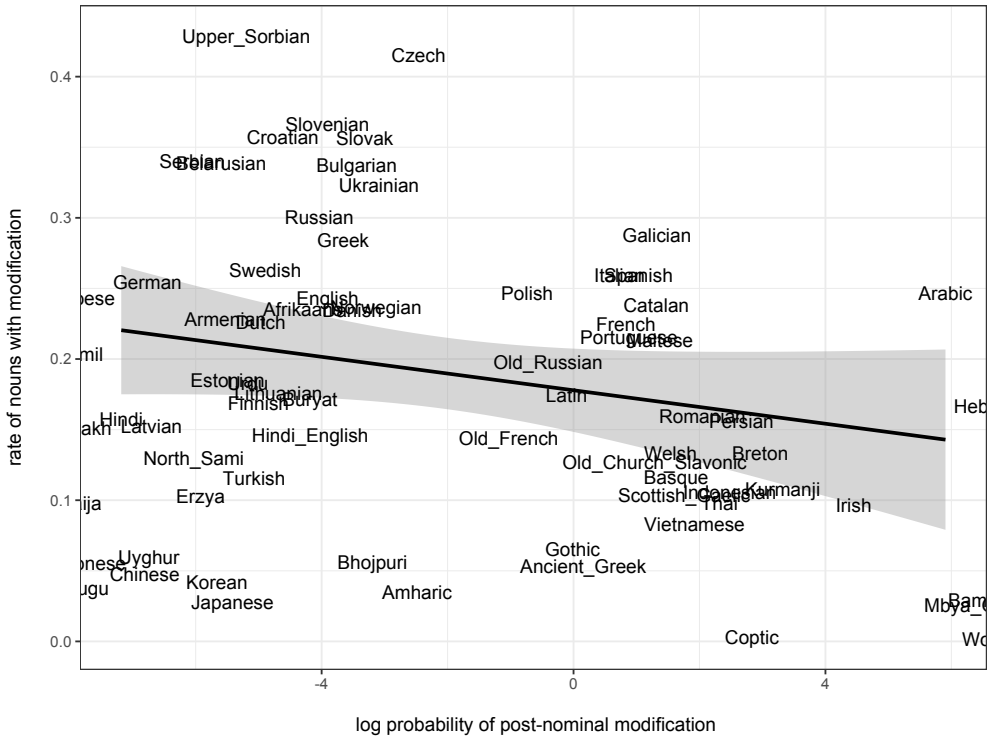


Figure 2.7: UD dependency analysis results: rate of adjective modification compared with the log probability of post-noimnal adjectival modification (Spearman's $\rho = -0.16$, p = 0.18).

2.3.3 Color-size asymmetry in Arabic

This subsection is based on joint work with Judith Degen and Brandon Waldon presented at X-Prag 2022, 9th Experimental Pragmatics Conference (Degen et al., 2022).

The corpus analyses reported above provide partial support for the prediction that post-nominal languages use adjective modification at lower rates: The UD corpus analysis found that languages with higher post-nominal propensity tend to use adjective modification at lower overall rates, though the Wikipedia analysis of English, Spanish, and Arabic alone showed no reliable difference. However, these results cannot tell us whether Arabic speakers show the specific color overmodification patterns reported in experimental settings. Similarly, they do not inform us about the color-size asymmetry observed experimentally in English and Spanish. We collected preliminary behavioral data on color and size overmod-

ification in Arabic, extending the interactive referential communication paradigm reported in Degen et al. (2020) whose cross-linguistic predictions were modeled in Waldon and Degen (2021). In the original study, English speaker pairs played an online communication game in which a director saw a highlighted target object in a display of two to five everyday objects varying in color and size, and typed a referring expression in a chat box to allow a guesser, who saw the same objects but did not know which was the target, to identify it. The critical manipulation was scene variation: the proportion of distractors that shared the redundant property (color or size) with the target. Waldon and Degen (2021) predicted that English speakers should produce more redundant color adjectives than Spanish speakers. We adapted the experiment for Arabic speakers, translating the interface and instructions into Arabic. We recruited 26 Arabic-speaking participants (13 pairs of participants since the study was interactive) through personal networks and community connections. In the Arabic pilot, participants showed a clear color-size asymmetry: they produced redundant color adjectives at higher rates than redundant size adjectives, replicating the asymmetry documented in English and Spanish (see Figure 2.8). Notably, Arabic speakers appear to exhibit less redundant modification overall than English speakers, consistent with the prediction of the efficiency hypothesis that speakers of post-nominal languages should overmodify at lower rates. However, the sample is small and further data collection is needed to confirm this pattern and to determine whether color overmodification increases with scene variation in Arabic as it does in English and Spanish.

2.3.4 Discussion

I performed an in-depth analysis on Spanish, English, and Arabic Wikipedia text and a large-scale corpus analysis on Universal Dependencies Treebanks, covering a wide range of languages. In the Wikipedia analysis, I failed to find a difference in the rates of adjective modification across the three languages, and thus no evidence to support the incremental

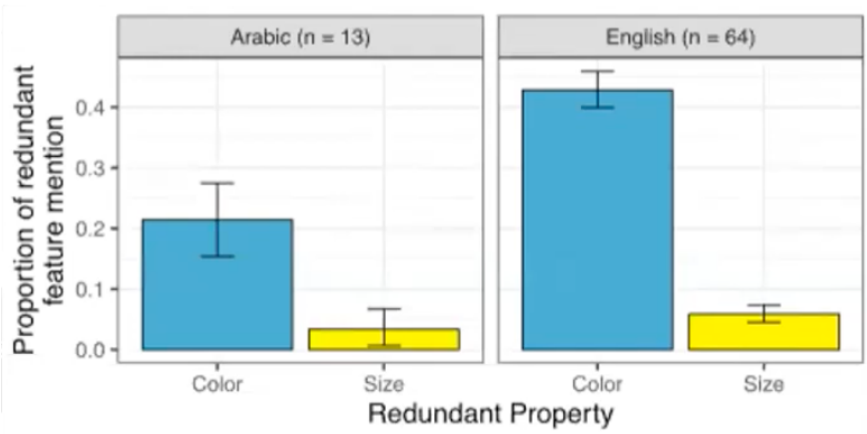


Figure 2.8: Tentative Arabic results on color-size asymmetry with scene variation.

efficiency hypothesis. When looking at the broader analysis of the UD treebanks, however, a robust trend consistent with the hypothesis is observed: the rates of post-nominal adjective modification predict the rates of noun modification overall.

The two POS analyses may look like they point in opposite directions: no obvious effect in Wikipedia, and a robust trend in UD. However, the apparent conflict dissolves on closer inspection. Spanish, English, and Arabic sit closer together in the UD data as well: the three languages the Wikipedia analysis happen to occupy only a narrow portion of the cross-linguistic range sampled in UD. Had I looked only at these three languages in the UD analysis, I would have failed to detect the relationship between post-nominal propensity and modification rate. By conducting a large-scale corpus analysis of dozens of languages, most of them under-studied in psycholinguistic research, I find that this relationship does in fact exist.

Taken together, the corpus analyses provide only qualified support for the efficiency hypothesis, and the two analyses should be read as complementary rather than as a primary result with a secondary check. The dependency-based analysis is the more accurate measure of attributive modification, since it distinguishes genuine attributive modifiers from adjacent predicative adjectives, and it is the analysis I weight most heavily. It shows the predicted

negative direction (languages with higher post-nominal propensity tend toward lower overall modification rates) but the relationship is weak and does not reach statistical significance. The POS-based analysis shows the same direction more strongly and reaches significance, but it rests on a coarser adjacency heuristic that overcounts attributive modification. Given all this, I draw a deliberately cautious conclusion: the corpus data are consistent with the predicted relationship in their direction across both analyses, but they do not establish its magnitude, and the gradient correlation on its own is too fragile to carry the efficiency claim. The categorical pattern is more robust than the gradient one: no language in the sample combines heavy post-nominal modification with a high overall modification rate. In the POS analysis the upper-right quadrant is empty outright; in the dependency analysis the region remains sparse. I therefore treat the distributional gap, rather than the correlation coefficient, as the principal corpus-level evidence, with the behavioral pilot in 2.2.3 providing convergent experimental support of the kind corpus data cannot supply on their own.

The interpretation these converging patterns point toward is the following. languages with pre-nominal adjectives are more likely to overmodify nouns given the contribution adjectives make to incremental reference resolution, and as a result, these languages have higher overall rates of adjective modification. This reasoning rests on an assumption: that pre- and post-nominal languages have comparable rates of adjectival modification in cases where adjectives are strictly necessary for successful communication. If this assumption holds, then the increases observed in pre-nominal languages arise from overmodification. Given that overmodification arises in referential contexts, my results suggest that the referential function of adjectives is not confined to controlled experiments but extends to spoken and written language more broadly.

One might worry that instances of referential language use (and overmodification specifically) are relatively rare in the corpora I analyzed, making it unlikely that the effect observed (greater adjective use in pre-nominal languages) is driven entirely by cases of overmodifica-

tion. It is possible, however, that the more frequent use of pre-nominal adjectives in genuine contexts of overmodification bleeds over to the use of adjectives overall, such that pre-nominal adjectives are used across the board at a higher rate. For example, if speakers of pre-nominal languages habitually retrieve and slot in adjectives before nouns during overmodification (because doing so consistently aids the listener's anticipatory referential search), they may come to produce adjectives more readily even when overmodification is not at stake. Moreover, small biases in production get amplified into robust categorical trends as language is passed from speakers to listeners and from one generation to another (Kirby, 2017). Over time, a small initial bias toward pre-nominal adjectives in overmodification contexts could grow into a broader pattern of higher adjective use across the language.

One might reasonably ask whether the incremental-efficiency prediction should apply to encyclopedic text at all, given that a Wikipedia author is not engaged in live referential communication with a reader. The prediction, however, is not that corpus writers perform online audience design; it is that the cumulative pressure of everyday referential communication leaves an imprint on a language's conventional production preferences. (Hawkins (2004) point that grammars conventionalize performance preferences). On this view, register differences are themselves a prediction rather than a nuisance: if the pressure is communicative, its signature should strengthen in interactive registers and weaken in encyclopedic ones. This is precisely the motivation for extending the analysis to conversational dialectal data, and it parallels the register modulation observed within a single language in Chapter 4, where the predictability effect emerges in private chat but not in public broadcast.

This efficiency reading aligns with a prediction that has been advanced in the literature on multiple grounds. The cross-linguistic claim that speakers of post-nominal languages should produce redundant color adjectives at lower rates than speakers of pre-nominal languages was first proposed empirically by Rubio-Fernández (2016) as the incremental efficiency hypothesis, and has been supported in subsequent behavioral and computational work (Wu

and Gibson, 2020; Rubio-Fernández et al., 2020; Cohn-Gordon et al., 2019; Rubio-Fernández et al., 2021; Waldon and Degen, 2021). Waldon and Degen (2021) make the most general version of the prediction: a lower overall rate of redundant modification in Spanish than in English, not just of redundant color adjectives as originally proposed by Rubio-Fernández (2016). This prediction is derived by aggregating across the size-sufficient and color-sufficient scene types of their experimental paradigm. What their model cannot establish is whether it generalizes beyond English and Spanish, beyond color, and beyond the visual-world task on which it was parameterized. The corpus analysis I conducted supplies that external test; my contribution is thus not to compete with their formal account but to extend the empirical base it needs to be taken as a cross-linguistic claim rather than a two-language fit. A complementary extension that remains incomplete is the color-size paradigm in Arabic, with preliminary results pointing in the predicted direction (Figure 2.8). Together with the corpus data reported here, this ongoing work begins to broaden the empirical base of the Rubio-Fernández and Waldon and Degen accounts beyond the English-Spanish pair on which it was built.

Beyond the gradient relationship between post-nominal propensity and modification rate, a second, more categorical pattern in the UD results warrants attention: the absence of languages that are both heavily post-nominal and use adjectives heavily (the empty upper right quadrant in the UD results in Figure 2.6) is itself a typological generalization that needs explaining: No language in the sample combines a high rate of post-nominal adjectival modification with a high overall rate of adjective use. That is, the distributional gap corresponds exactly to what the efficiency hypothesis would lead us to expect. On a soft reading, the empty quadrant reflects a probabilistic pressure against high rates of adjective modification when adjectives do not aid incremental reference resolution, consistent with the efficiency account. On a hard reading, it reflects a near-universal upper bound on post-nominal modification rates that may be imposed by processing or production constraints independent of the efficiency mechanism Rubio-Fernández (2016) proposes. Such constraints could include

working memory limitations on integrating multiple modifiers after the head noun has been processed. The more modifiers a listener has to keep track of, the harder working memory has to work, and at some point there is likely a cap on how many can be held at once. In a language whose modification is heavily post-nominal, this per-phrase cap also caps overall modification rates, directly producing the empty upper-right quadrant observed in the UD analysis.

Distinguishing these readings would require follow-up work. In particular, within-language studies on post-nominal languages are needed to test whether the efficiency pressure we see across languages also holds within a single language. Such studies could take several forms. The first is a behavioral overmodification paradigm in Arabic. The preliminary results shown earlier in Figure 2.8 represent a first step in this direction, with results pointing in the predicted direction. A second form is a comparison within the diglossic Arabic situation, testing whether modification rates differ between spoken dialects (e.g., Egyptian, Levantine, or Gulf Arabic) and Modern Standard Arabic. MSA, as the formal written register, carries less direct face-to-face communicative pressure than the spoken dialects. If communicative pressure shapes modification, the dialects might show higher rates than MSA. I flag the empty quadrant here as a potentially stronger and more distinctive finding than the gradient correlation itself, and as a potential target for future typological and experimental work.

Alongside the tests the empty quadrant calls for, a second extension concerns the specificity of my findings to color adjectives. The analyses I conducted treated all adjectives uniformly, without distinguishing among types. This kept the method scalable across dozens of languages, but it leaves an obvious question unasked: the experimental production literature is almost entirely about color (Pechmann, 1989; Sedivy, 2003; Rubio-Fernández, 2016; Degen et al., 2020), and the English-Spanish asymmetry that motivated this work is specifically an asymmetry in redundant color. If resources permit, isolating the color subclass in each of the 74 UD treebanks and asking whether the cross-linguistic pattern concentrates in color

or holds across adjective types would sharpen the claim that color carries distinctive communicative work.

More broadly, the findings reported in this chapter illustrate a connection between linguistic form and function. The rate at which speakers use adjectives is a communicative function that is governed by the position of adjectives relative to nouns in the language. The corpus results (especially the empty upper right quadrant result) show that these two are not independent: languages whose grammar places adjectives before the noun allows adjectives to serve an incremental referential function that post-nominal position may not.

2.3.5 Expanding the corpus analysis

One limitation of the Wikipedia and UD analyses is that the Arabic data come exclusively from Standard Arabic, a written register that lacks the direct communicative pressures of face-to-face interaction. To address this, I plan to expand the corpus analysis to include spoken Arabic dialects, which would provide a more representative picture of how Arabic speakers use adjective modification in everyday communication.

Compiling the Arabic dialects datasets

To build a dialectal Arabic corpus, I compiled several existing datasets that cover the major dialect families. Arabic dialects are generally grouped into five broad families based on geographic region: Maghrebi (covering Libya, Tunisia, Algeria, Morocco, and Western Sahara), Egyptian (spoken in Egypt and parts of Sudan), Levantine (spanning Palestine, Syria, Lebanon, and Jordan), Iraqi, and Gulf (including Qatar, Kuwait, Bahrain, the United Arab Emirates, Saudi Arabia, Oman, and Yemen) (Zbib et al., 2012). For Levantine dialect data, the SHAMI corpus (Kwaik et al., 2018) provides 66K lines from four countries (Lebanon,

Syria, Palestine, and Jordan). The MSDA dataset (Boujou et al., 2021) consists of 53K lines of user-generated social media content spanning five dialects from North Africa and the Middle East. The PADIC corpus (Meftouh et al., 2015) contributes 36K lines from a parallel corpus covering both Maghrebi dialects (two Algerian varieties and Tunisian) and Levantine dialects (Syrian and Palestinian). The Arabic Online Commentary (AOC) dataset (Zaidan and Callison-Burch, 2011) provides 20K lines drawn from reader commentary on an Arabic news website, with content in Gulf, Egyptian, and Levantine dialects. Two Twitter-based datasets round out the collection: NADI (Abdul-Mageed et al., 2021), which provides 26K tweets labeled by country of origin across the Arab world, and QADI (Abdelali et al., 2021), which contains 3.5K tweets from 18 Arab countries. The GUMAR corpus (Khalifa et al., 2016), drawn from online forum novels, contributes 15K lines of Gulf Arabic. For the Tunisian dialect, the TSAC (Tunisian Sentiment Analysis Corpus) dataset (Medhaffar et al., 2017) contains 12K lines of Tunisian Arabic user comments from social media. An additional Levantine dataset, the LET-MI corpus (Levantine Twitter dataset for Misogynistic language), contributes 5K lines of data. Finally, the DART corpus, drawn from the IADD (Zahir, 2022), contributes 1.8K lines of additional Gulf dialect data.

After standardizing the format and mapping dialect labels to the five dialect families described above, plus MSA as a reference category, the compiled corpus contains approximately 239,000 lines of text. The distribution of data by source is shown in Table 2.2, and the dialect distribution of the data is shown in Table 2.3

The remaining step is to run a POS tagger on the compiled data and apply the same adjacency-based noun modification analysis used in the Wikipedia and UD corpora. This would allow a direct comparison of adjective modification rates between Standard Arabic and spoken dialects. If the communicative pressures of social interaction shape modification rates, we might expect spoken dialects to differ from Standard Arabic in ways that align with the patterns observed in the UD analysis. I also plan to apply the same analysis to

Table 2.2: Data sources compiled for the dialectal Arabic corpus from ten different existing datasets.

Source	**Number of Lines**
SHAMI	66,247
MSDA	53,403
PADIC	36,065
NADI (train and dev sets)	26,000
AOC	19,713
GUMAR	15,225
TSAC	11,998
LET-MI	5,240
QADI	3,503
DART	1,781
Total	**239,175**

Table 2.3: Dialect distribution of the compiled Arabic corpus. The dialects are grouped into the common five dialect groups, in addition to MSA (Modern Standard Arabic).

Dialect	**Number of Lines**
Levantine	111,467
Maghrebi	72,273
Gulf	30,612
Egyptian	18,336
Iraqi	3,787
MSA	2,700
Total	**239,175**

comparable English and Spanish discussion forum data from the Linguistic Data Consortium, which contains more than 1190 discussion forums. This would extend the cross-linguistic comparison beyond the encyclopedia genre of Wikipedia and the genres in UD.

Tagging the data, analysis plan, and predictions

Prior to tagging, the compiled data will be pre-processed by removing special characters, punctuation, and non-Arabic tokens to ensure consistent input for the POS tagger. For Arabic, I plan to use CAMeL Tools, which provides dialect-specific BERT-based POS taggers

for Egyptian, Gulf, and Levantine Arabic, with an MSA model as fallback for Iraqi and Maghrebi (Obeid et al., 2020). For the English and Spanish comparison, I will use discussion forum data from the Linguistic Data Consortium and tag it with the StanfordCoreNLP POS tagger (Manning et al., 2014). Once tagged, the same method used in the Wikipedia and UD analyses will be applied to all three languages: nouns will be classified by whether they are modified pre-nominally, post-nominally, both, or not at all, and modification rates will be compared across dialects and languages using mixed-effects logistic regression. Because the compiled corpus is overwhelmingly dialectal with MSA representing only a small fraction of the data, the primary comparison will be across dialect families rather than between dialects and MSA. If the communicative pressures shape modification rates, we may observe variation across dialects that parallels the cross-linguistic observed in the UD analysis.

2.3.6 Behavioral experiments

The corpus analyses reported above test the efficiency hypothesis at the level of overall adjective modification rates, but they cannot disentangle the contextual factors that drive a speaker's decision to include or omit a particular adjective in a given referential expression. Corpus data can reveal that one language or dialect modifies nouns more frequently than another, but it cannot tell us whether a given adjective was produced because a competitor object was present, because the property was perceptually salient, or simply because the speaker defaulted to a fuller description. Teasing apart these factors requires controlled experimental settings in which the visual context and the properties of the referent can be systematically manipulated. Indeed, much of the work on adjective modification and overspecification has relied on exactly this approach (Rubio-Fernández, 2016, 2019; Rubio-Fernández et al., 2020; Wu and Gibson, 2020; Degen et al., 2020).

To complement the corpus work, I propose a set of referential communication experiments

that test whether Arabic speakers replicate the asymmetry in overmodification that has been documented in English and Spanish (Rubio-Fernández, 2016; Degen et al., 2020). Extending this paradigm to Arabic, a language that has not been tested in this literature, broadens the empirical base of the efficiency hypothesis.

The aim of this proposal is to collect behavioral data on the production of referring expressions in Arabic. Specifically, I ask whether the production of color adjectives in spoken Arabic dialects follows the pattern observed in previous behavioral work on post-nominal languages, or whether it follows the pattern observed in my corpus analysis of the general uses of adjectives in Standard Arabic (2.3.1 and 2.3.2). If Arabic speakers produce lower rates of redundant adjective modification than English speakers, this would provide direct evidence that adjective processing is incremental and efficient: speakers of a post-nominal language produce fewer redundant adjectives, consistent with the reduced incremental benefit of post-nominal modification. If, on the other hand, Arabic and English speakers produce similar rates of adjective modification, the question becomes what communicative purpose these adjectives serve when they follow the noun and therefore cannot aid the listener's incremental referential search. The pilot data reported in 2.3.3 offers preliminary evidence for the first possibility; Arabic speakers showed suggestively lower overmodification rates under scene variation, but the sample was too small to be conclusive; the experiments proposed here are designed to test that pattern with adequate power. Both Arabic and English speakers are expected to use adjectives efficiently, but when the adjective occurs after the noun in Arabic, it does not contribute to narrowing down the referent before the noun is processed (Rubio-Fernández et al., 2020). Producing a redundant adjective in that position may instead reflect the salience of the property to the speaker, for instance because the color of the object is particularly distinctive or contrastive in the visual scene. Arabic is thus a particularly informative test case for the efficiency hypothesis: if overmodification rates are lower in Arabic than in English, this would be consistent with the reduced communicative utility of post-nominal adjectives for incremental reference resolution, particularly if the asymmetry

is specific to post-nominal modifiers rather than reflecting a general reluctance to produce extra words.

To test this, I plan to extend the referential communication paradigm developed by Rubio-Fernández (2016) and used in subsequent work by Rubio-Fernández et al. (2020). In the study by Rubio-Fernández et al. (2020), twenty-two participants from each of English and Spanish were recruited in person. Participants viewed visual displays containing colored geometric shapes arranged in a grid, with one shape marked as the target. The participant sat behind the experimenter, and each visual display was presented on the screen in front of the experimenter. Each participant had a printed sheet that showed a star on the position of the target item on the grid, and their task was to ask the experimenter to click on the target item in each trial. The experimenter did not know the target object in each trial, and they were presented with 20 visual displays. The critical manipulation was the number of competitor objects in each display. In a shape competitor display, another object of the same shape but a different color appeared alongside the target (e.g., a red triangle alongside a blue triangle target), making the color adjective necessary for disambiguation. In a color competitor display, another object of the same color but a different shape appeared (e.g., a blue star alongside a blue triangle target), so that shape alone was sufficient. In the no-competitor displays, no object shared either the color or the shape of the target, making any adjective redundant (see Figure 2.9). In their experiment, English speakers used color adjectives significantly more than Spanish speakers, consistent with the prediction that speakers of a pre-nominal language overmodify more because their adjectives contribute to incremental reference resolution. I will adapt this paradigm to an online format, with participants' spoken responses recorded via their microphone. On each trial, the participant's task will be to produce a referring expression (in their dialect or MSA) that allows a simulated online partner to identify the target.

Together, the expanded corpus analysis and the behavioral experiments proposed in this

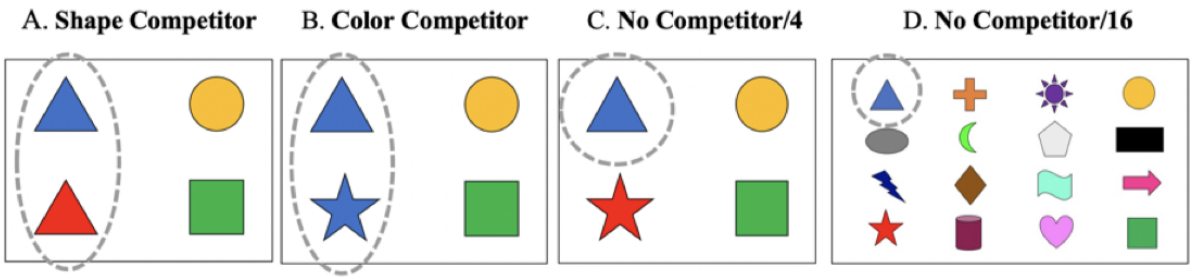


Figure 2.9: Sample displays from the referential communication paradigm used in Rubio-Fernández et al. (2020). Here, the target is the blue triangle (top left corner in all displays). In the shape competitor condition (A), color is necessary for disambiguation. In the color competitor condition (B), shape is sufficient for disambiguation. In the no competitor conditions (C, D), both color and shape are redundant with difference in the size of competitor objects.

chapter address complementary gaps in the efficiency literature. The corpus work extends the cross-linguistic coverage of the modification rate analysis to Arabic dialects, testing whether the relationship between adjective position and modification frequency that holds across the UD treebank languages also holds within the diglossic Arabic situation. The behavioral work provides the controlled experimental conditions needed to isolate the contextual factors that corpus data alone cannot disentangle, testing directly whether Arabic speakers produce fewer redundant adjectives than English speakers in referential communication. Beyond the question of how frequently adjectives are used, the structure of multi-adjective strings raises a separate efficiency-related puzzle: if speakers are sensitive to the informational contribution of adjectives, this sensitivity may also shape the order in which adjectives appear. Chapter 3 takes up this question, turns to the phenomenon of adjective ordering preferences, examining whether the subjectivity-based ordering preferences documented in English extend to a language with post-nominal modification.

Chapter 3

Adjective ordering in Arabic

3.1 Introduction

The use of adjectives in description and reference resolution sometimes extends to more than a single adjective per phrase (as we saw in the examples from Degen et al.'s use of size and color adjectives). How speakers order these adjectives relative to each other and to the modified noun is not arbitrary. Imagine, for example, that you come across a box that is small, brown, and made of wood. English-speaking adults prefer the order *small brown wooden box* over *wooden brown small box*; they are uncomfortable with the latter ordering, yet most are unable to articulate why (Sproat and Shih, 1991; Scontras et al., 2017). This observation is not unique to English — speakers of many languages converge on specific orderings of multi-adjective strings, a phenomenon known as **adjective ordering preferences**.

3.2 Background

3.2.1 Empirical observations of adjective ordering preferences

Adjective ordering preferences have received considerable attention from the linguistics community. The cross-linguistic robustness of these preferences evidences cognitive properties that shape language. Most of the research in this area arrives at the same core finding: in multi-adjective strings, some adjectives are systematically preferred closer to the modified noun than other adjectives, and this preference is not random.

Languages with pre-nominal modification

Adjective ordering preferences have been most extensively documented in languages with pre-nominal adjectives. Beyond English, similar preferences have been observed in other languages with pre-nominal adjectives like Hungarian (Uralic), Telugu (Dravidian), Mandarin Chinese, and Dutch, to name a few (Dixon, 1982; Hetzron, 1978; LaPolla and Huang, 2004; Sproat and Shih, 1991). Additionally, the tendency to order adjectives a certain way starts earlier in life. Evidence from English-speaking children as found in corpus data demonstrates adjective ordering preferences as young as four years old (Bar-Sever et al., 2018).

While these cross-linguistic observations establish the robustness of the phenomenon, the majority of experimental work on adjective ordering has focused on languages with pre-nominal adjective modification, and on English in particular. Martin (1969) performed some of the first behavioral studies on adjective ordering in English, documenting that speakers exhibit reliable preferences. Since then, a growing body of work has confirmed that English ordering preferences are robust and gradient. Although orders like *brown small box* are not ungrammatical, speakers show a reliable tendency toward considering them dispreferred (Scontras et al., 2017).

Explanations for where these preferences come from have taken a variety of forms. A null hypothesis posits that adjective ordering preferences merely reflect distributional learning: adults reproduce sequences such as *the small brown box* because they have previously encountered them in that form. However, this account fails to explain why speakers show consistent preferences for novel, previously unattested combinations, such as *the tiny green magical mouse-riding gnomes* (Bar-Sever et al., 2018). Moreover, if speakers simply reproduce the orderings they hear, the account offers no explanation for why those orderings are being produced in the first place. The productivity of adjective ordering preferences thus suggests that speakers draw on some underlying knowledge beyond the repetition of familiar sequences.

Languages with post-nominal modification

While most of the literature has focused on pre-nominal languages, a smaller but growing body of work has investigated adjective ordering in languages where adjectives follow the noun. Dixon (1982) notes that post-nominal adjectives surface in the mirror image of the pre-nominal order. The following English-Arabic pair illustrates the pattern:

(1) the small brown box

(2) al-ṣundūq al-bonnī al-ṣaghīr
the-box the-brown the-small
'the small brown box'

In these examples, the noun in Arabic comes first and the adjectives follow, displaying the reverse ordering of those in English. Indonesian was also reported to have mirror-image adjective ordering preferences in the post-nominal position (Martin, 1969). More

recent investigations of post-nominal languages have yielded mixed results. Rosales Jr. and Scontras (2019) investigated ordering preferences in Spanish, a language with post-nominal adjectives. Results revealed that Spanish speakers do not exhibit adjective ordering preferences. However, the multi-adjective strings tested featured conjunctions of adjectives similar to (3), which could be responsible for the lack of stable preferences, and indeed Scontras et al. (2020c) show that you get stable preferences in Spanish without conjunction.

(3) la caja pequeña **y** marrón
'the small **and** brown box'

3.2.2 Proposed explanations of ordering preferences

The observations I have discussed so far document the ordering patterns that speakers exhibit. I now turn to proposed explanations of where these preferences come from. These accounts have taken a variety of forms, ranging from syntactic accounts rooted in hierarchical structure to information-theoretic approaches grounded in efficiency. In surveying them, it is worth drawing a distinction that organizes the discussion to follow: between identifying a predictor of ordering preferences and explaining why that predictor should have the force it does. I survey these accounts in the next few subsections.

Syntactic and hierarchical explanations

Beyond distributional learning, several theoretical proposals have sought to identify the source of these preferences. A dominant line of research adopts a lexical-class approach, according to which adjectives are inherently organized into discrete semantic classes and patterned according to shared properties. On this view, adjectives denoting SIZE constitute one class, those denoting COLOR another, and so forth. Crucially, these classes are assumed

to stand in a hierarchical relation, such that certain semantic categories are systematically positioned closer to the modified noun than others. Dixon (1982) systematically classifies adjectives into semantic categories such as size, color, and quality, and proposes that these categories are ordered in a certain way with respect to the noun. Based on data from 17 languages, Dixon arrives at the hierarchy in (4), where classes farther to the left are preferred farther from the noun.

(4) $Quality > Size > Shape > Color > Nationality > Noun$

Cinque (2010) suggests a different but closely preferred order (5). Cinque argues that adjective classes are encoded in the syntax through a hierarchy of functional projections, each hosting a different class of adjectives. For example, a size adjective and a color adjective each occupy a designated position in the syntactic structure. Because these positions are fixed and hierarchically ordered, the result is a predictable linear order of adjectives across languages.

(5) $Value > Size > Shape > Color > Provenance > Noun$

Although this hierarchy captures a robust generalization across linguistic data, it remains fundamentally descriptive. It does not explain why these preferences exist: why should the hierarchy be ordered as in (4) or (5), and not, say, in a different order? The need for a principled explanation of such observations has motivated earlier and subsequent research to explore the mechanisms underlying these preferences. Some approaches focus on the linear order in which the adjectives and nouns appear in language, while others investigate structural or relational proximity independent of linear sequencing. This distinction becomes especially salient in languages other than English with different positions for adjectival modification.

If such a hierarchy is indeed fixed across languages, then the question arises of why this should be the case. Scott (2002) builds on Cinque's approach and argues that the consistency of ordering preferences across languages reflects Universal Grammar. Speakers of different languages converge on similar orderings because the underlying syntactic structure is innate. This syntactic approach describes the ordering but does not offer an independent explanation for why certain classes of adjectives are consistently preferred closer to the noun than others.

Meaning-based explanations

Other approaches suggest a relation between preferred distance and the meaning of adjectives. For example, Sweet (1898) suggests that the adjective most closely connected in meaning with the noun appears nearest to it. The idea is that some adjectives form a tighter conceptual unit with the noun than others. Consider the example *the small brown box*: a *brown box* picks out a recognizable kind of object (maybe a cardboard box), while *small* merely adds a property on top of that unit. On this view, *brown* appears closer to the noun than *small* because it is more closely connected to what the noun denotes. Whorf (1945) proposes a related distinction, splitting adjectives into two broad categories: those naming inherent properties of the noun and those naming non-inherent ones. Adjectives that name properties inherent to the noun generally appear closer to it. For example, a cardboard box is inherently brown while size is non-inherent: a box is not inherently small or large. Sproat and Shih (1991) propose a different but related notion, absoluteness: adjectives that hold absolute properties, properties that do not depend on a standard of comparison, occur closer to the noun. For example, a box is brown regardless of context, but whether it counts as small depends on what it is being compared to. On this account, absolute adjectives like color and shape adjectives appear closer to the noun, while relative adjectives like size and quality adjectives appear farther away. Although inherentness and absoluteness are different notions, what these proposals share is the intuition that the semantics of the adjective plays

a role in determining order.

Accessibility-based explanations

Accessibility-based approaches propose that the order in which adjectives appear is influenced by how easily they are retrieved during production. More accessible adjectives tend to appear earlier in the sequence because they come to mind more quickly. This is consistent with a broader observation in the production literature that easily accessible words are less costly to produce (Ferreira and Dell, 2000), and words that are highly frequent are considered to be readily accessible (Piantadosi et al., 2012). Indeed, Martin (1969) was among the first to show that more frequent adjectives tend to occur earlier in production. In languages with pre-nominal modification like English, if a speaker is producing the *big brown box*, the adjective that is easier to retrieve comes out first, which places it farther from the noun. Hetzron (1978) argues that adjective length also plays a role, with shorter adjectives tending to appear earlier; so more frequent or shorter adjectives end up farther from the noun in pre-nominal languages. The problem comes with post-nominal languages. In a language like Arabic or Spanish, the noun comes first and the adjectives follow. If accessibility is driving order, then the most accessible adjective should still come out first, but now "first" means closer to the noun, not farther from it. Accessibility would therefore predict the same linear order across languages with pre-nominal and post-nominal modification, rather than the mirror-image pattern that has been documented (Dixon, 1982). This limitation suggests that accessibility alone cannot account for the full range of ordering preferences observed cross-linguistically.

Information-theoretic explanations

Communicative efficiency pressures extend to shape the way we prefer to order adjectives across languages. Speakers who decide to use more than one adjective in their phrases are in a real-time incremental decision-making process to order adjectives a certain way or another. Yet, the ordering preferences we see among speakers from a single language and the patterns we observe across languages are far from random: there is a systematic pattern that reflects a general need to be communicatively efficient, optimizing the cost-reward trade-off of the utterance.

A growing body of work proposes that adjective ordering preferences fall out from more general efficiency pressures inspired by information theory (Shannon, 1948). Information theory is concerned with optimizing digital communication codes while minimizing entropy. Entropy can be thought of as a measure of uncertainty over a probability distribution. When the distribution is concentrated over few outcomes, entropy is low; when outcomes are equally probable, entropy is high. In the context of language, we can think of outcomes as possible words that follow a given word. Consider the adjective-noun pair *polar bear*: upon hearing *polar*, the distribution over possible next words is sharply concentrated, and so entropy is low. In contrast, a word like *nice* could be followed by a wide range of nouns (*nice day*, *nice person*, *nice idea*), making the distribution much flatter and entropy considerably higher. Entropy relates to information content as quantified by surprisal: a high-probability word drawn from a low-entropy context is less surprising, and therefore less informative, than one drawn from a high-entropy context.

One of the main information-theoretic measures that predict adjective ordering preferences is point-wise mutual information (PMI), a measure of how much two words predict each others' occurrences in a phrase (Futrell, 2017, 2019). While entropy captures uncertainty over an entire probability distribution, PMI zooms in on a particular word pair and asks how

much more likely they are to co-occur than we would expect by chance. Returning to the earlier example, *polar* and *bear* have high PMI because their joint probability far exceeds the product of their marginal probabilities. In other words, they co-occur far more frequently than their individual frequencies would predict. Under Futrell's account, comprehension processing cost decreases when words that are highly predictive of one another, as measured by PMI, appear closer to each other in linear order. Given that comprehension happens under memory constraints from the listeners, placing high PMI adjectives nearer to the noun thus maximizes prediction accuracy and reduces processing cost. This principle assumes that contextual information degrades over distance due to memory noise, an assumption later formalized in Futrell et al. (2020b). Under this model, the listener's internal representation of preceding context degrades as a function of linear distance, meaning that two highly associated elements may fail to facilitate each other's processing if they are too far apart. An efficient language should minimize the linear distance between highly associated words, so that their mutual predictability is preserved. Adjective ordering preferences constitute one manifestation of this broader pressure.

Another information-theoretic measure of adjective ordering preferences is information gain which captures how much a word reduces a listener's uncertainty about the speaker's intended meaning (Dyer et al., 2021). In other words, as each adjective is encountered, it narrows the set of possible referents by partitioning them into those that have the relevant feature and those that do not. In pre-nominal languages, this process begins before the noun is known, so early adjectives reduce uncertainty about what the referent could be. In post-nominal languages, the noun is already known, and each subsequent adjective narrows the set of referents that match that noun. In both cases, the information gained by this narrowing determines the preferred ordering: adjectives that yield the greatest reduction in uncertainty tend to be placed linearly earlier. Dyer et al. (2021) test this prediction using corpora from 32 typologically diverse languages, and across all three ordering templates: pre-nominal (AAN), post-nominal (NAA), and mixed (ANA), the latter order being particularly challenging for

distance-based measures like PMI, which rely on the linear proximity between adjective and noun. Information gain successfully predicts attested orderings in 29 of these languages, achieving above-chance accuracy across all three templates.

While measures like PMI and information gain effectively capture communicative efficiency in adjective ordering, a separate line of work has identified a single cognitive property that independently predicts ordering preferences: Subjectivity.

Subjectivity

In an attempt to explain why adjective preferences appear the way they do, Scontras et al. (2017) documented a robust empirical generalization concerning these preferences in English such that the order of adjectives in multi-adjective strings is predicted by subjectivity: less subjective adjectives occur closer to the modified noun. In *the big blue box*, *blue* is considered less subjective than *big*; two people are more likely to agree on whether the box is blue, but whether or not the box is big depends on each person's experience with box sizes. That is, one person might consider a box big, because he or she is used to small-sized boxes, while the other person disagrees and considers it small.

Scontras et al. (2017) constructed two experiments to test their hypothesis. In the first experiment, they measured adjective ordering preferences using native English speakers' judgments about multi-adjective strings; they asked participants to indicate their preferences on phrases like *the small brown box* vs. *the brown small box*. Then, in their second experiment, they established a behavioral measure to determine subjectivity. They presented participants with adjectives and asked them how "subjective" each was on a scale from "completely subjective" to "completely objective". The authors then validated their subjectivity measure with a faultless disagreement task (Kölbel, 2004; Kennedy, 2013; Barker, 2013; MacFarlane, 2014). In a faultless disagreement task, participants encounter a scenario in which two

people argue about whether an object has some property, for example whether a box is big or not. The participant has to determine if both people can be right or if one of them must be wrong. If the participant indicates that both people in the scenario can be right, this suggests a higher degree of subjectivity for that adjective.

The methodology of Scontras et al. (2017) has been extended to Tagalog (Samonte and Scontras, 2019), Spanish (Rosales Jr. and Scontras, 2019), Mandarin Chinese Shi and Scontras (2020), Greek, Hebrew, and Vietnamese Scontras et al. (2021), and more recently Italian (Scontras et al., 2025). Spanish, a language with post-nominal adjectives where speakers prefer to use conjunction to form multi-adjective strings, does not exhibit adjective ordering preferences — at least not in multi-adjective strings formed with conjunction (cf. Scontras et al., 2020b).

The importance of subjectivity lies in the success of communicative efficiency: speakers aim to maximize their listeners' understanding of their utterances, and thus they minimize the distance between the objective adjectives and the nouns these adjectives modify. In an attempt to formalize this notion, recent work by Scontras et al. (2019) proposes that adjectives that are linearly closer to the noun are usually structurally closer, meaning they compose with the noun first. When adjectives are used to establish reference, having less subjective adjectives closer to the nouns aids in making that referent clear to the listener. This is because adjectives narrow down the set of possible referents incrementally: when more than one adjective modifies a noun, the adjective closest to the noun restricts the set first. A listener is more likely to correctly identify which boxes are brown than which boxes are small. By placing less subjective adjectives closest to the noun, speakers ensure that the most reliable restrictions happen first, reducing the overall chance of miscommunication. Franke et al. (2019) formalize this reasoning using simulated reference games, demonstrating that ordering adjectives with respect to decreasing subjectivity increases the probability of successful reference resolution. The logic follows from the structure of restrictive modifica-

tion: each adjective restricts the set of potential referents identified by the noun, and less subjective adjectives, which restrict the set more reliably, should compose with the noun first in the hierarchical structure, operating over the largest set of potential referents and thereby minimizing opportunities for miscommunication.

Unlike the reference-resolution accounts discussed above, Hahn et al. (2018) assume that adjectives are used not to pick out a referent but to communicate the speaker's judgments and attitudes about objects. In their model, listeners process language incrementally, and material heard earlier is more likely to be forgotten due to memory limitations. Placing more subjective adjectives farther from the noun means that if any adjective is lost from memory, it is more likely to be the subjective one. This is the less costly loss, because objective adjectives carry information that generalizes across speakers, making them more useful for the listener to retain. Therefore, integrated with a memory limitation model, ordering adjectives by decreasing subjectivity helps preserve the speaker's intended meaning. In other words, it is a better strategy for sharing the speakers' mental state and aiding in successful communication.

In an attempt to compare the predictive power of subjectivity to other measures like PMI, integration cost, and information gain, Futrell et al. (2020a) found that no single predictor accounts for the full range of ordering preferences. Their results suggest that adjective ordering reflects at least two broad factors operating in parallel: a semantic factor, captured by subjectivity, and a wordform-based factor captured by PMI. A model combining these two factors outperforms any individual predictor. Dyer et al. (2023) extend this comparison cross-linguistically, evaluating eight cognitive predictors across 32 languages and all three ordering templates: pre-nominal (AAN) orderings, mixed (ANA) orderings, and post-nominal (NAA). They found that subjectivity measure performs best for AAN, adjective length for ANA, and PMI for NAA. While subjectivity may not be the strongest single predictor in post-nominal languages, it remains a significant factor across templates. However, the evidence from

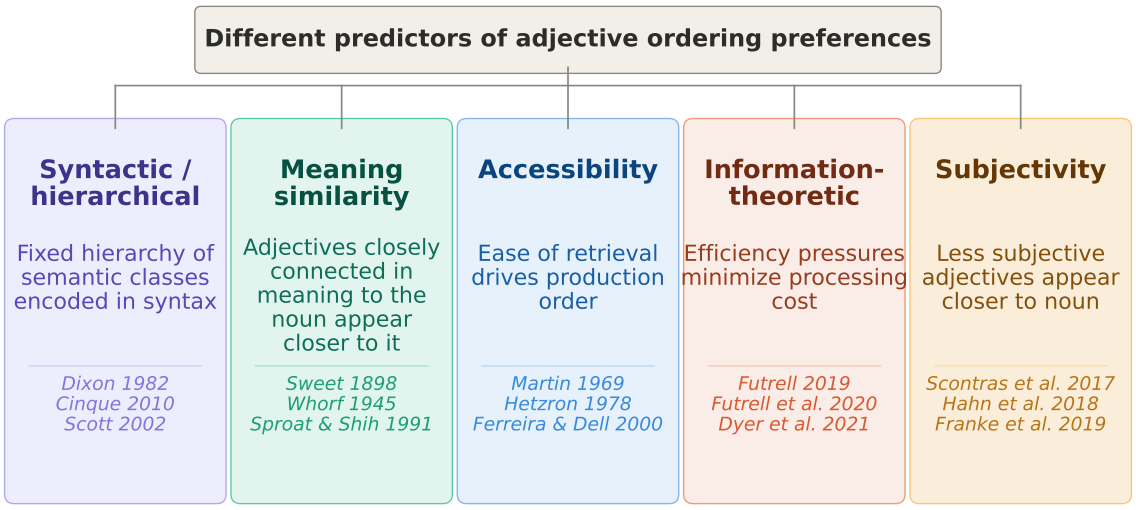


Figure 3.1: Summary of the different predictors of adjective ordering preferences discussed in this section. This figure was generated with the help of Claude.

post-nominal languages is based almost entirely on corpus data; behavioral investigations of ordering preferences in these languages remain scarce.

All of the predictors of adjective ordering preferences I discussed in this section are summarized in Figure 3.1. While much of the work surveyed above has focused on English and other pre-nominal languages, the question of how well these predictors generalize to post-nominal languages remains underexplored. Scontras et al. (2025) for instance, find that Italian ordering preferences are strongest in the ANA template and that subjectivity predicts ordering most weakly in the NAA order, though subjectivity still offers a consistent cross-linguistic predictor. In what follows, I test adjective ordering preferences in Arabic, a language with post-nominal adjectives, and examine how well subjectivity predicts these preferences. If adjective ordering in post-nominal languages mirrors the pre-nominal pattern, we would expect subjectivity to be a strong predictor; however, the findings reviewed above suggest this may not be the case. By collecting behavioral data from Arabic, I aim to contribute much-needed evidence from a post-nominal language to this line of work.

3.3 Behavioral Experiments

This section provides exactly this behavioral data, investigating whether subjectivity predicts adjective ordering preferences in Arabic, a post-nominal language without obligatory conjunction in multi-adjective strings. Arabic consists of two varieties: Spoken Arabic, which consists of the different dialects spoken across the Arab world, and Standard Arabic (also referred to as Modern Standard Arabic, MSA, classical Arabic, or literary Arabic). The two varieties are used in different situations. Spoken Arabic is used in everyday communication, while Standard Arabic is used in schools, media, politics, speeches, academic research, and formal documents. In the written form, Standard Arabic remains almost exclusively the language of literacy across the Arabic-speaking world. Of relevance to the present study, adjectives in both Standard and Spoken Arabic attach to the noun post-nominally as the default. While it is the case that pre-nominal adjectives are possible (see example (6); Fehri, 1999), these adjectives come with special syntactic properties and have their own interpretations in Arabic grammar. Post-nominal adjectives occur much more frequently (as discussed in Chapter 2), and they agree with the noun in definiteness, case, number, and gender. In this chapter, I only consider post-nominal adjectives.

(6) labis-tu jamīl-a al-thiyāb-i
wore-I beautiful.ACC the-clothes-GEN[1]
'I wore the (most) beautiful (of) clothes.'

(7) al-kursiyy-u al-'azraq-u al-saghīr-u
the-chair-NOM the-blue-NOM the-small-NOM
'the small blue chair'

Although Standard Arabic is not the variety used in daily conversation, there are good

[1]Abbreviations used: ACC = accusative, GEN = genitive, IND = indefinite, NOM = nominative.

reasons to test in Standard Arabic. First, all the spoken dialects of Arabic are tied to Standard Arabic. That is, dialects have a considerable amount of shared vocabulary with Standard Arabic. The amount of vocabulary sharing varies from one dialect to another, but Standard Arabic is more broadly intelligible across dialect groups than any single dialect. Consistent with this, recent work by Abou-Ghazaleh et al. (2018) investigating the neural basis of the diglossic situation in Arabic using fMRI showed that there is no difference in brain activity between naming objects in Standard and Spoken Arabic. Given this, and since the experiments present stimuli in written form, I use Standard Arabic.

In terms of adjective ordering, a substantial body of research has been conducted on the adjectival system and adjective constructions in Arabic (Al-Sharifi and Sadler, 2009; Al-Shurafa, 2006; Kremers, 2003), but very little research has studied the ordering of multiple adjectives within a single noun phrase. Sproat and Shih (1991) argue that Arabic adjectives are freely ordered when combined together, meaning that both possible orderings of a given adjective pair are equally acceptable. For example, they report that both (8) and (9) are acceptable in Arabic, and neither is preferred over the other, a claim that Shlonsky (2004) suggests is not confirmed in the Arabic literature.

(8) kalbu-n ahmaru-n kabiiru-n
dog-IND red-IND big-IND
'a big red dog'

(9) kalbu-n kabiiru-n ahmaru-n
dog-IND big-IND red-IND
'a big red dog'

Fehri (1999) proposes that post-nominal Arabic adjectives follow a mirror image order to languages with pre-nominal adjectives. For example, the Arabic equivalent of *the big red*

ball surfaces as (10) following the order: noun < size < quality. This mirror-image pattern has also been observed in Selepet, a language with post-nominal adjectives in which the pre-nominal English order is largely preserved in reverse (Dixon, 1982, p. 26).

(10) ʔ-alʕab-u bi-l-kurat-i l-ħamraʔ-i l-kabirat-i
I-play with-the-ball-GEN the-red-GEN the-big-GEN
'I play with the big red ball.'

Panayidou (2014) reaches a similar conclusion for Cypriot Maronite Arabic, a dialect spoken by the Maronite community of Cyprus and highly influenced by Greek, arguing that adjective ordering in this variety is inflexible.

The picture that emerges from this literature is thus mixed: some authors predict free ordering in Arabic, while others argue for fixed, mirror-image preferences, and the evidence base for Arabic remains largely observational. Crucially, none of the existing work on Arabic adjectives has employed experimental preference tasks, nor does it offer an explanation for ordering preferences grounded in cognition. Furthermore, if pre-nominal and post-nominal languages share the same hierarchical structure but differ in linear order, then speakers who have both systems present an especially informative test case. Heritage speakers of Arabic who also speak a pre-nominal language like English have two competing linear orders, yet the underlying hierarchy is thought to be similar across both languages. Studying this population can shed light on how robust ordering preferences are under reduced input and cross-linguistic competition.

In this chapter, I use experimental data to address these gaps. I show that Arabic exhibits adjective ordering preferences that mirror those found in pre-nominal languages, consistent with the efficiency-based explanations discussed above. Specifically, subjectivity predicts these preferences, with less subjective adjectives preferred closer to the noun, just as in pre-

nominal languages like English (Scontras et al., 2017). I then explore the robustness of these preferences by studying English-dominant heritage speakers of Arabic.

3.3.1 Experiment 1: Ordering preferences in native Arabic

Replicating the methodology of Scontras et al. (2017), I begin by measuring adjective ordering preferences in native speakers of Arabic.

Participants

135 participants were recruited via Amazon Mechanical Turk, a crowd sourcing service. Only 24 participants were identified as native speakers of Arabic (Table 3.1 shows the dialect distribution of these speakers). To assess language status, participants were asked a series of demographic questions. For example, participants were asked what their first language is, and what language they use at home. I identified native speakers as those participants who indicated their first language as Arabic and who continue to speak Arabic as their dominant language today, and who lived in an Arabic-speaking country for more than five years before and after the age of eight. To ensure that the participants were not answering randomly, I appended three simple Arabic catch questions which contained phrases that required diagnosing gender agreement, and asked the participants to mark the correct phrase (see the examples in (11)). I only considered participants who marked all three catch questions correctly.

(11) a. al-baħr al-azraq al-waseʕ
the-sea.M the-blue.M the-wide.M
'the wide blue sea'

b. *al-bahr al-zarqāʔ al-waseʕa
the-sea.M the-blue.FEM the-wide.FEM

Table 3.1: Dialect distribution of participants in Experiment 1 (N=24).

Dialect	**Count**
Moroccan	10
Levantine	6
Egyptian	5
Saudi	2
Other	1
Total	**24**

(intended: 'the wide blue sea')

Design

Participants were asked to indicate their preferences for multi-adjective strings. To do so, participants judged which order of a multi-adjective string was more natural. Multi-adjective strings were random combinations of 25 adjectives that came from 7 semantic classes (i.e., quality, age, texture, size, shape, color, and material adjectives; see Table **??**), with the constraint that no description contained adjectives from the same semantic class. The adjectives described nouns that were sampled randomly from a set of 10 nouns (see Table **??**). Participants completed 26 trials each. For example, participants were presented with two phrases: the Arabic equivalents of *the small brown chair* and *the brown small chair*, and they were asked which order they preferred (Figure 3.2). Where possible, adjectives and nouns were direct translations of the materials from Scontras et al. (2017). The only exception was the adjective *tiny*, whose closest Arabic equivalent, *ḍaʔīl*, is largely restricted to formal registers and is uncommon in everyday speech, making it unsuitable for a naturalness judgment task. Participants answered by adjusting a continuous slider with endpoints labeled with the competing multi-adjective strings. The distance from the slider to each endpoint indexes the degree of preference.

I used the slider position to compute the naturalness score for each adjective, which cor-

Table 3.2: The seven adjective classes used in the experiment, which are translations of the adjectives used in the original experiment by Scontras et al. (2017).

adjective	pronunciation	translation	class
أحمر	ʔaħmar	red	color
أصفر	ʔasfar	yellow	color
أخضر	ʔaxdar	green	color
أزرق	ʔazraq	blue	color
بني	bonnī	brown	color
بنفسجي	banafsajī	purple	color
كبير	kabīr	big	size
صغير	saġīr	small	size
ضخم	daxm	huge	size
قصير	qasīr	short	size
طويل	tawīl	tall	size
جيد	jayyed	good	quality
سيئ	sayyeʔ	bad	quality
فاسد	fāsed	rotten	age
طازج	tāzaj	fresh	age
جديد	jadīd	new	age
قديم	qadīm	old	age
مستدير	mustadīr	round	shape
مربع	murabbaʕ	square	shape
ناعم	nāʕem	soft	texture
قاسي	qāsī	hard	texture
أملس	ʔamlas	smooth	texture
خشبي	xašabī	wooden	material
بلاستيكي	bilastīkī	plastic	material
حديدي	ħadīdī	metal	material

responds to how far away from the noun the adjective is preferred. If a participant, for example, preferred *the brown small chair*, *small* would have a naturalness score close to 0 since it is preferred closer to the noun, and *brown* would have a naturalness score closer to 1 since it is preferred farther away from the noun.

Table 3.3: The set of ten nouns from two different classes I adopted from Scontras et al. (2017) with their translations.

noun	pronunciation	translation	class
تفاحة	toffāħa	apple	food
موزة	mawza	banana	food
جزرة	jazara	carrot	food
جبن	jobn	cheese	food
طماطم	tamātim	tomato	food
كرسي	korsī	chair	furniture
كنبة	kanaba	couch	furniture
مروحة	merwaħa	fan	furniture
تلفاز	telfāz	tv	furniture
مكتب	maktab	desk	furniture

Results

For each adjective, I computed a mean preferred distance measure, which ranges from 0 (preferred closest to the noun) to 1 (preferred farthest from the noun), with 0.5 indicating chance, or the absence of stable preferences. Error bars in Figure 3.3 represent bootstrapped 95% confidence intervals drawn from 10,000 samples of the data. The results show that Arabic does have stable ordering preferences, despite also having post-nominal adjectives. As shown in the figure, adjective classes pattern along a gradient: quality adjectives are

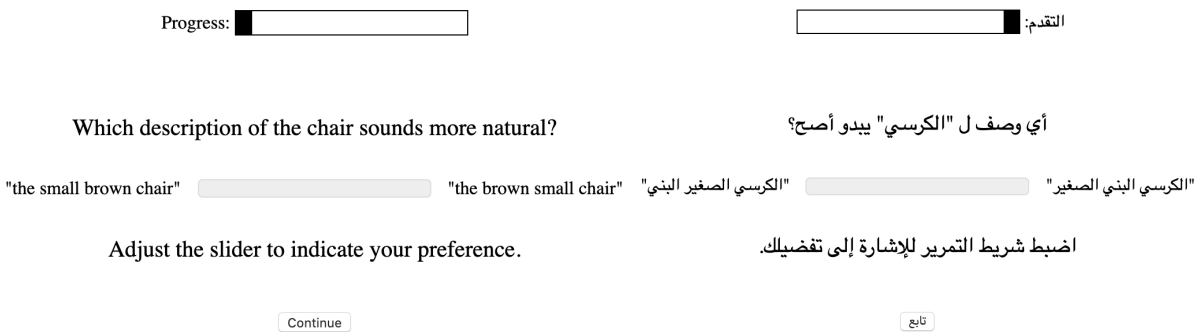


Figure 3.2: An example trial from the ordering preferences experiment. On this trial, participants were presented with two phrases: *the brown small chair* and *the small brown chair*. They were asked to adjust the slider towards the preferred phrase. To the right is the Arabic version of the experiment, and to the left is the original experiment by Scontras et al. (2017).

preferred farthest from the noun, followed by age, texture, and size, while shape, color, and material adjectives are preferred closer to the noun. The bootstrapped confidence intervals do not overlap with the chance line, indicating reliable preferences for these classes. This pattern is consistent with the class hierarchies proposed in the literature (Dixon, 1982; Cinque, 2010) and mirrors the qualitative pattern reported for English by Scontras et al. (2017).

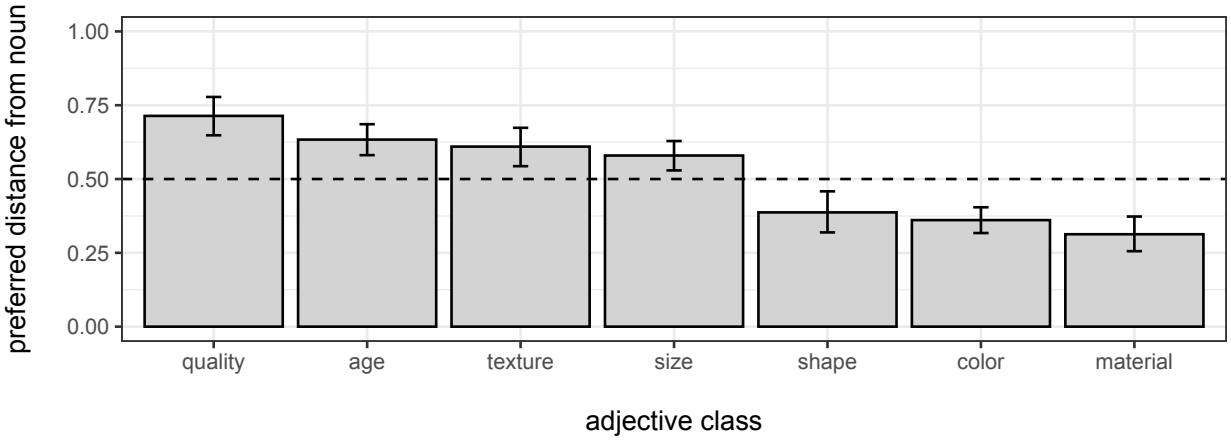


Figure 3.3: Native Arabic speakers' mean preferred distance measures from the noun grouped by adjective semantic class. Values range from 0 (preferred closest to the noun) to 1 (preferred farthest from the noun). The dashed line indicates chance level, or the absence of stable preferences. Error bars represent bootstrapped 95% confidence intervals drawn from 10,000 samples of the data.

3.3.2 Experiment 2: Assessing subjectivity

After obtaining measures for ordering preferences, I next measured whether these preferences are subjectivity-based. To do so, I measured adjective subjectivity using a faultless disagreement task.

Participants

135 participants who did not take part in the first experiment were recruited through Amazon's Mechanical Turk. All participants were compensated for their participation. Because

Arabic speakers represent a small proportion of the platform's workforce, the same language background questionnaire and catch questions described in Experiment 1 were administered to verify their native speaker status. Of the 135 participants, only 16 were identified as native Arabic speakers (Table 3.4 shows the dialect distribution of these participants); the remaining participants were excluded from analysis.

Table 3.4: Dialect distribution of participants in Experiment 2 (N=16).

Dialect	**Count**
Egyptian	6
Levantine	3
Other	2
Moroccan	2
Iraqi	1
Gulf	1
Saudi	1
Total	**16**

Design

Participants were given a scenario in which two speakers see the same object but they disagree about its description (see Figure 3.4). This task, known as a faultless disagreement task, has been shown to yield scores that are highly correlated with independent measures of subjectivity (Scontras et al., 2017): if two speakers can be right while they disagree on a description, then that adjective is subjective. For example, the extent to which two speakers can faultlessly disagree about whether a *chair* is *brown* indexes the subjectivity of the adjective *brown*. On each trial, participants adjusted a slider to indicate their judgment from 0 to 1: "No, somebody must be wrong" coded as 0, and "Yes, it's a matter of opinion" coded as 1. Each participant completed 25 trials, one for each adjective (I used the same 25 adjectives from Experiment 1), with nouns randomly selected for each trial. I averaged these scores across participants to obtain a single subjectivity measure for each adjective.

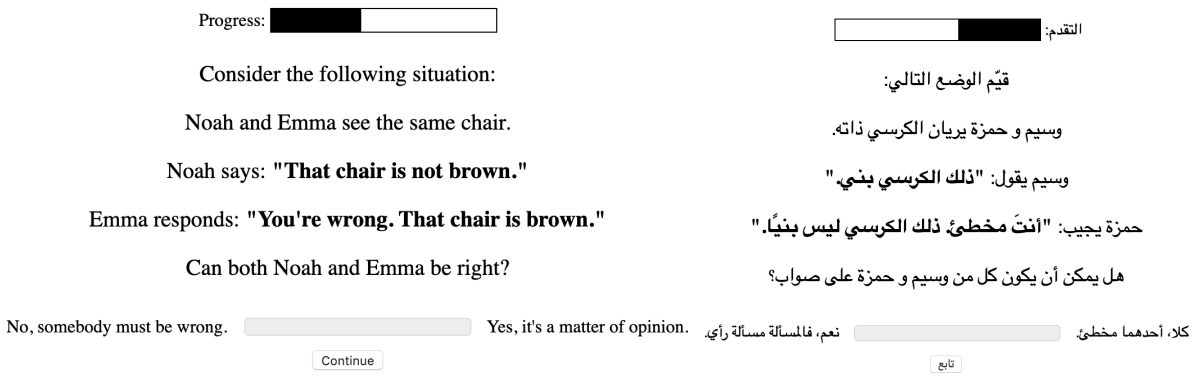


Figure 3.4: An example trial from the faultless disagreement task.

Results

Comparing subjectivity scores with the ordering preferences from Experiment 1, I find that subjectivity is a strong predictor of adjective ordering in Arabic ($r^2 = 0.76$, 95% CI $[0.57, 0.88]$; see Figure 3.5). That is, more subjective adjectives such as *jayyed* "good" and *sayyeʔ* "bad" are preferred farther from the noun than less subjective adjectives such as *xašabī* "wooden" or *ʔazraq* "blue". Crucially, these preferences mirror those found in pre-nominal languages, with less subjective adjectives closer to the noun, consistent with the cross-linguistic generalization documented for English by Scontras et al. (2017).

3.3.3 Experiment 3: Ordering preferences in heritage Arabic

After obtaining results from baseline native speakers, I investigated whether English-dominant heritage speakers of Arabic exhibit similar ordering preferences to the native baseline. Heritage speakers are unbalanced bilinguals who have experienced a shift in language use from their home language to the dominant language of their society. This shift often results in bilingualism unbalanced in favor of the dominant societal language, but research has shown that abilities of the heritage language persist (Scontras et al., 2015; Polinsky and Scontras, 2020). In some cases, knowledge from the dominant language transfers to the heritage lan-

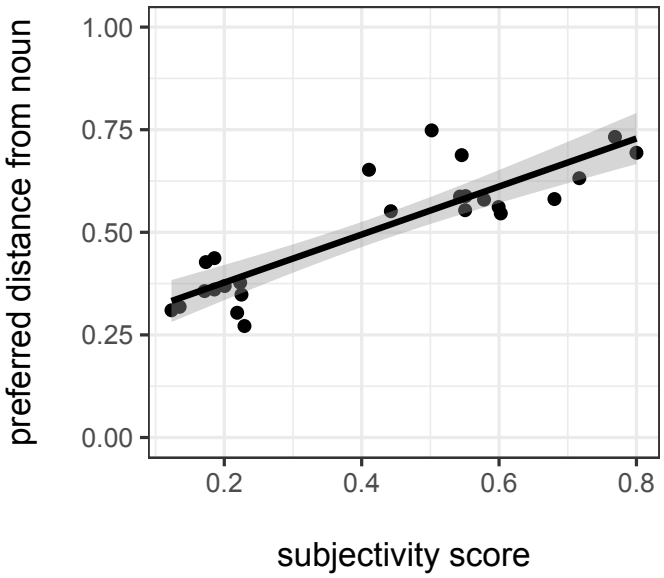


Figure 3.5: Ordering preferences obtained in Experiment 1 plotted against subjectivity scores for each of the 25 adjectives tested. Subjectivity accounts for 76% of the variance in the ordering preferences ($r^2 = 0.76$, 95% CI $[0.57, 0.88]$).

guage. In other cases, heritage speakers simply fail to acquire the relevant knowledge, owing to a lack of suitable input, or they acquire the knowledge and then lose it in the absence of proper maintenance (i.e., without sufficient use). By studying the behavior of heritage speakers, we can understand how language functions in the mind of a bilingual. More specifically, we can examine how robust ordering preferences are to reduced input and maintenance.

Importantly, with adjective ordering, we have seen that both English and the Arabic baseline order adjectives hierarchically with respect to decreasing subjectivity, so transfer from English would result in the same preferences. However, heritage speakers might transfer the linear order of adjectives from English to Arabic, rather than following hierarchical ordering preferences. If heritage speakers are attending to linear order and transferring that order from their dominant English, we should find heritage preferences that are the mirror image of baseline Arabic.

Participants

I began by recruiting English-dominant heritage speakers of Arabic through Mechanical Turk but was unable to obtain a sufficiently large sample. I therefore recruited participants from the Arab-American community in Southern California using email chain and social media to reach out to different organizations and associations. Many potential participants from the younger generation raised in the U.S. either did not speak Arabic or could not read it, the latter being a requirement for the written experiment. I recruited 34 participants. English-dominant heritage speakers were identified as those whose first language was Arabic and whose main language now is English; participants who lived in an Arabic-speaking country after the age of eight were excluded, as were participants who failed to answer all three catch questions correctly. Twelve participants were finally identified as heritage Arabic speakers, all of whom identified their dialect as Levantine.

Design

I measured heritage ordering preferences using a version of Experiment 1 adapted for heritage speakers with bilingual instructions to accommodate heritage speakers' dominant English proficiency (see Figure 3.6). The instructions and interface prompts were presented in both Arabic and English to ensure heritage speakers understood the task, but the actual stimuli (the adjective-noun descriptions they were choosing between) remained in Arabic only, to avoid priming English ordering preferences. Participants indicated their preferences for multi-adjective strings by completing the same 26 trials using the same adjectives and nouns from Experiment 1.



Figure 3.6: An example trial from the ordering preference task adapted for heritage speakers. Instructions were presented bilingually in Arabic and English, while the adjective-noun descriptions were presented only in Arabic.

Results

Figure 3.7 plots results from heritage speakers grouped by semantic class. Although there is more variance overall compared to the native baseline (likely due both to the smaller number of participants and the heterogeneity inherent to heritage populations) we continue to see stable preferences: some classes are preferred closer to the noun, and others are preferred farther away. Moreover, we see a similar qualitative pattern of results between the native baseline and the heritage results, suggesting that both exhibit subjectivity-based preferences. The one point of qualitative deviation from the baseline involves material adjectives, which baseline native speakers prefer maximally close to the noun.

Comparing heritage ordering preferences with adjective subjectivity

To evaluate the predictive power of subjectivity in explaining the heritage ordering preferences, I used the subjectivity scores obtained from native speakers in Experiment 2. Figure 3.8 plots the heritage ordering preferences against these subjectivity scores. While the heritage results are not nearly as robust as the native results, there is a positive correlation between the preferred distance of adjectives from the noun and the subjectivity of these

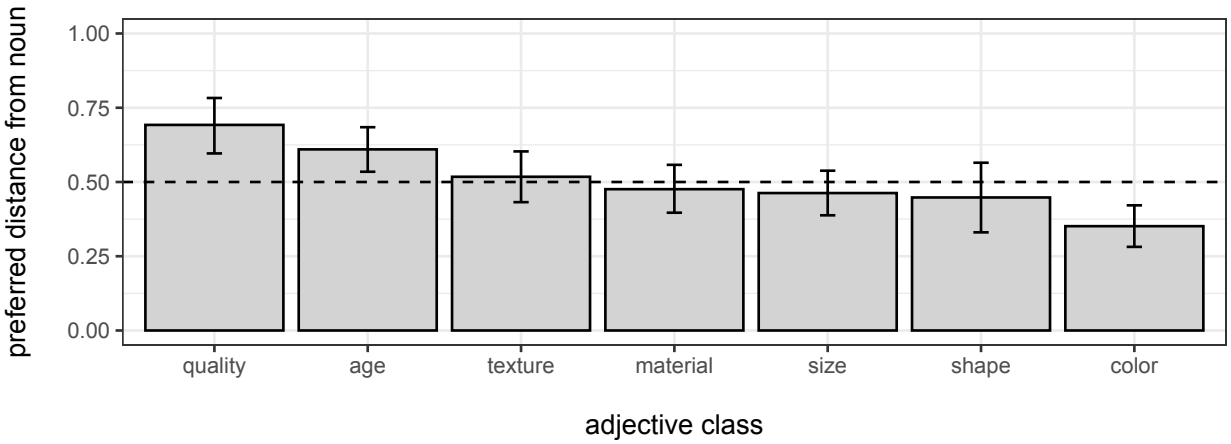


Figure 3.7: Naturalness ratings from Experiment 3 (heritage) grouped by adjective semantic class. Error bars represent bootstrapped 95% confidence intervals drawn from 10,000 samples of the data.

adjectives ($r^2 = 0.26$). In other words, despite the dominant influence of English, a pre-nominal language with the reverse linear order, heritage speakers do not simply transfer English linear preferences to Arabic. Instead, they maintain the same direction of ordering as native speakers, with less subjective adjectives preferred closer to the noun.

3.3.4 Discussion of Experimental Results

The results from Experiment 1 demonstrate that baseline native Arabic speakers do have stable adjective ordering preferences. Moreover, the results of Experiment 2 show that subjectivity is a reliable predictor of these preferences in Arabic: more subjective adjectives are preferred farther away from the noun. In Experiment 3, results from twelve English-dominant heritage speakers of Arabic show that heritage speakers have adjective ordering preferences that, as in the baseline, are predicted by adjective subjectivity. Not only does Arabic have subjectivity-based ordering preferences, but those preferences are robust to the reduced input and language maintenance characteristic of heritage speakers.

While a qualitative pattern of preferences similar to the native baseline is observed in her-

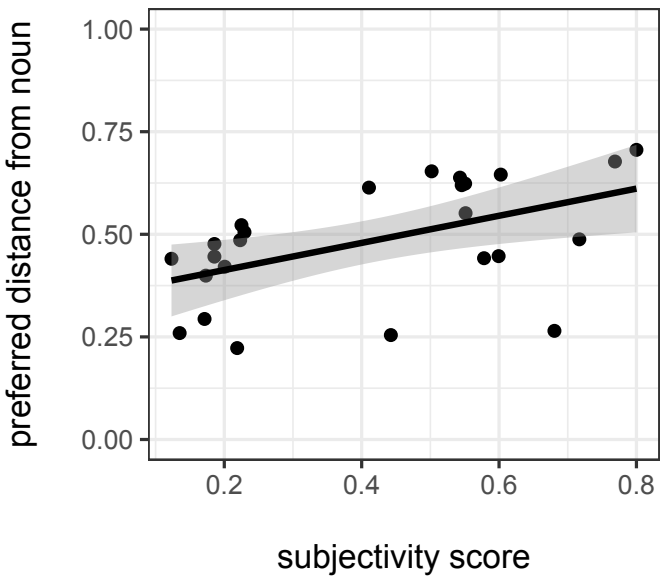


Figure 3.8: Heritage ordering preferences plotted against adjective subjectivity scores obtained from baseline speakers in Experiment 2.

itage speakers, one group of adjectives stands out in the heritage data: material adjectives. Material adjectives are preferred closest to the noun in the native baseline, but fail to exhibit stable preferences in heritage speakers. Investigating the special status of material adjectives in the heritage grammar is needed to explain this observation, along with looking at frequency, length, and age-of-acquisition, together with possible effects from the dominant English grammar. Also, note that the number of participants in the heritage experiment is half the number of participants in Experiment 1. Too few participants could lead to increased noise in the data, thereby obscuring the empirical picture.

Regarding heritage subjectivity, there is a clear correlation between the preferred distance from the noun and the subjectivity of the adjectives. However, this correlation is not very strong. The strength of the correlation suffers in part because of two sets of outliers. First are the adjectives 'hard' and 'long', which are preferred closer to the noun than we would expect given their subjectivity. The second group of outliers is material adjectives: 'metal', 'wooden', and 'plastic', which are preferred farther from the noun than we would expect given their subjectivity. However, despite these outliers in the heritage results, native speakers

subjectivity does correlate with heritage preferences. It remains to be seen whether heritage-specific subjectivity estimates would reduce these deviations. Investigating this question is a natural next step, but will require recruiting a larger sample of heritage speakers — both to obtain reliable subjectivity estimates and to achieve greater statistical power in detecting ordering preferences.

These results situate Arabic adjective ordering within the broader communicative efficiency framework developed in Chapter 1. As discussed there, linguistic form is shaped by a trade-off between the cost of producing and processing utterances and the reward of successful communication (Gibson et al., 2019). The finding that subjectivity predicts adjective ordering preferences in Arabic provides evidence that this trade-off extends beyond the pre-nominal languages where ordering preferences have been most extensively studied. According to Scontras et al. (2020a), subjectivity-based orderings maximize communicative success by ensuring that less subjective adjectives compose with the noun earlier.

Moreover, Hahn et al. (2018) suggest that placing more subjective adjectives farther from the noun preserves the speaker's intended meaning against the pressure of incremental forgetting. Both accounts predict that ordering adjectives based on subjectivity is more efficient and applies cross-linguistically regardless of the position of the adjective relative to the noun. The Arabic results confirm this prediction: speakers order adjectives hierarchically with respect to subjectivity, with the linear order reversed to reflect the post-nominal position of adjectives. The fact that subjectivity accounts for Arabic ordering preferences suggests that the same communicative pressures are at work across typologically distinct languages. Accessibility-based explanations, on the other hand, tie ordering to ease of retrieval during production and predict the same linear order across pre-nominal and post-nominal languages. The latter prediction is not borne out by the mirror-image pattern we observe.

The consistency of subjectivity-based ordering preferences across both baseline native and heritage speakers suggests that these preferences reflect deep cognitive pressures rather than

a product of full language proficiency. Despite their unbalanced use of Arabic, heritage speakers exhibit qualitatively similar ordering patterns to native speakers, pointing to a robust underlying mechanism that persists even under conditions of reduced input and maintenance. What the source of these preferences is remains an open question: they may reflect online communicative reasoning, learned distributional regularities, or some combination of both. Moreover, although subjectivity performs well as a predictor of ordering preferences in Arabic, recent work on other post-nominal languages suggests that distributional measures such as PMI may be at least as strong a predictor in post-nominal configurations (Hahn et al., 2018; Dyer et al., 2023). What is clear from the present results is that heritage speakers are not simply transferring the linear order of their dominant language, English, onto Arabic. Instead, they maintain ordering preferences that are qualitatively consistent with those of native speakers, suggesting that whatever drives these preferences is resilient to the reduced input conditions characteristic of heritage language development.

I turn in the next chapter to a different dimension of efficiency: the minimization of linguistic form in Romanized Arabic script.

Chapter 4

Vowel Minimization in Arabizi

4.1 Introduction

In the previous two chapters, I explored the phenomenon of redundant adjective modification and adjective ordering preferences, and showed that both are shaped by pressures toward successful and efficient communication. In this chapter, I examine how speakers reduce linguistic form and show that this reduction, too, is shaped by communicative efficiency pressures. Specifically, I investigate vowel dropping in Arabizi, an emerging Romanized writing system for Arabic, as a case of word length minimization driven by efficiency.

4.2 Background

4.2.1 Word length and predictability

Frequency and word length

A central prediction of the efficiency framework outlined in Chapter 1 is that speakers should prefer shorter linguistic forms when doing so does not compromise successful communication. One of the most influential formulations of this idea came from Zipf (1949), who observed that more frequently used words tend to be shorter, a regularity he attributed to the pressure to minimize production effort while maintaining comprehension ease for the listener, reflecting a balance between reducing cost and preserving information transfer. As discussed in Chapter 1, Arab grammarians arrived at a similar insight centuries earlier under the label *al-takhfīf li kathrat al-istimal* (reduction for frequent usage), noting that the most common elements of Arabic are systematically shorter and less morphologically complex (al Zamakhsharī, 1905). Hawkins (2004) generalizes this observation into the principle of *minimizing forms*, which states that linguistic forms are shaped to ease cognitive processing for both speaker and listener. This idea emerges from the concept that shorter, simpler forms are processed more quickly, and therefore reduce the cognitive effort required during communication. For example, English speakers routinely say *lab* instead of *laboratory* or *phone* instead of *telephone*; the shorter forms are easier to produce and, because they are so frequent, are immediately understood by the listener. Even when a shortened form is ambiguous, context typically disambiguates the intended meaning without additional effort from the listener. The word *phone*, for instance, is a clipped form of *telephone*, but it is also a technical term in linguistics referring to an individual speech sound. In everyday conversation, a sentence like *I left my phone at home* is immediately understood as referring to the device; in a phonetics lecture, *the phone was aspirated* is just as unambiguous. The poly-

semy of the short form does not impede communication, because the conversational context constrains interpretation, and the production advantages of the shorter form are preserved. Language, then, naturally gravitates toward shorter forms as a way of optimizing the balance between production cost and communicative success.

Hawkins's principle could apply to duration of an utterance (Aylett and Turk, 2004; Seyfarth, 2014), or the length of a word in written form (Piantadosi et al., 2011, 2012; Mahowald et al., 2013, 2018; Levshina, 2022b), which is where most research in this area has focused. Speakers naturally develop shorter descriptions for entities that are more accessible or more frequently used in language. This tendency is most commonly explained through frequency: words that occur more often face greater pressure toward brevity. The intuition is that speakers optimize their lexicon so that the forms they produce most often are the least costly to articulate and process. From an information-theoretic perspective, frequent words carry less surprisal, meaning they can afford to be shorter without risking miscommunication (Piantadosi et al., 2011). For example, the most frequent words in English are *the, be, to, and, of*, which are short, unambiguous, and easy to produce, serving as functional words in comprehension. Pronouns reflect a similar pressure: once a referent is established in discourse, speakers replace the full noun phrase with a shorter form, reducing production cost without sacrificing comprehension. Another example is when speakers create new verbs such as *googled* instead of *searched on Google*, where both production and comprehension effort are reduced while the meaning remains intact and easy to extract. However, while frequency has been the most prominent account of why common words tend to be shorter, more recent work suggests that it does not capture the full picture.

Predictability and information content

In information theory, the information content of a word reflects how surprising it is in context: a word that is highly expected carries little information, while an unexpected

word carries more. While Zipf's account links word length to frequency, Piantadosi et al. (2011) build their framework on information theory and propose that the average information content is a better predictor of word length. Words with high average information content tend to be longer and less frequent, while words with low information content tend to be shorter and more frequent. To estimate information content, the authors use an n-gram model to calculate the surprisal of each word given its preceding context. This highlights the importance of predictability in determining the length of the next word: if the reader can predict the next word easily from the word before, then that word has low information content. However, if the word is not easily predicted, and hence is more surprising, then that word has higher information content.

However, whether informativity consistently outperforms frequency as a predictor of word length is not settled. Indeed, even before Piantadosi et al.'s proposal, Jurafsky et al. (2001) had argued along similar lines in the domain of speech production. Their Probabilistic Reduction Hypothesis states that word forms are reduced when they have a higher probability, where probability is conditioned on multiple aspects of context, including neighboring words, syntactic structure, and discourse factors. Studying conversational English, the authors found that more predictable words are more likely to undergo vowel reduction and word duration shortening, suggesting that the pressure toward brevity operates not only on word length in the lexicon but also on the phonetic realization of words in context. Building on this work, Bell et al. (2009) demonstrated that frequency, contextual predictability, and repetition each make independent contributions to word duration in spontaneous English speech. Their analysis of the Switchboard corpus shows that even after controlling for word frequency, words in more predictable contexts are produced with shorter durations.

These findings, however, are largely based on English. Levshina (2022b) tested the relationship between word length, frequency, and informativity across seven typologically diverse languages, including Arabic, and found that the results vary by language. In Arabic, infor-

mativity given the previous word showed a negative correlation with word length, meaning that words that are highly predictable from the preceding context tend to be longer, not shorter. Levshina attributes this to the structure of Arabic definite noun phrases, in which post-nominal modifiers carry the definite article and derivational morphology, inflating their length despite their high predictability. Crucially, however, this analysis is based on standard Arabic orthography. Whether this pattern holds beyond standard Arabic orthography remains an open question, and one that I will address in this chapter.

Word form choice in context

Building on previous results from Piantadosi et al. (2011), Mahowald et al. (2013) investigated whether word length is affected by context, but zoomed in on a specific phenomenon: word length reduction in words that have short and long variants. For example, English speakers alternate between forms like *math/mathematics, exam/examinations*, and *chimps/chimpanzees*. Mahowald et al. tested whether more predictable contexts favor the use of one form over the other in these word pairs: specifically, whether factors pertaining to efficiency shape this production choice, or whether alternating between variants is entirely arbitrary. The authors developed a behavioral experiment with a paradigm such that each pair had two contexts: a supportive context and a non-supportive (neutral) context. Participants had to choose between short and long variants of words such as *math* and *mathematics* in examples (1) and (2). In supportive contexts like (1), participants were expected to use the shorter variants of words.

(1) Susan was very bad at algebra, so she hated...

1. Math 2. Mathematics

In neutral contexts, however, no clues about the target word were provided. In information-

theoretic terms, when the target word is less predictable given its context, it carries higher surprisal, and speakers compensate by using the longer form. Participants in these contexts were therefore expected to choose the longer variant of the word, as in (2).

(2) Susan introduced herself to me as someone who loved...

1. Math 2. Mathematics

The authors indeed found that shorter forms are more likely to be used in more predictable contexts. Moreover, Mahowald et al. computed surprisal values for each word using the n-gram model from Piantadosi et al. (2011), and found that the mean surprisal for long forms is higher than that of the short forms. As Figure 4.1 shows, most of the surprisal differences between short and long forms fall above zero, indicating that the long form of a word carries more information on average than its shorter counterpart.

While these studies establish a clear link between predictability and word length reduction, the evidence is drawn almost entirely from English and other Indo-European languages. Although recent work has provided strong empirical support for efficiency's role across multiple levels of linguistic structure (Gibson et al., 2019; Jaeger and Tily, 2011; Pimentel et al., 2020; Piantadosi et al., 2012), experimental data from a wider set of languages remain limited, leaving open the question of how universal these patterns are. For instance, in a morphologically rich language like Arabic, where information is packed into words through multiple morphemes and a root pattern morphology system, Levshina's findings already suggest that Arabic departs from the patterns observed in English. However, that analysis was limited to standard Arabic orthography, leaving open the question of whether these patterns hold in other written forms of Arabic. One such form is Arabizi, a Romanized writing system for Arabic used in digital communication, where speakers routinely alternate between longer and shorter spellings of the same word through vowel dropping.

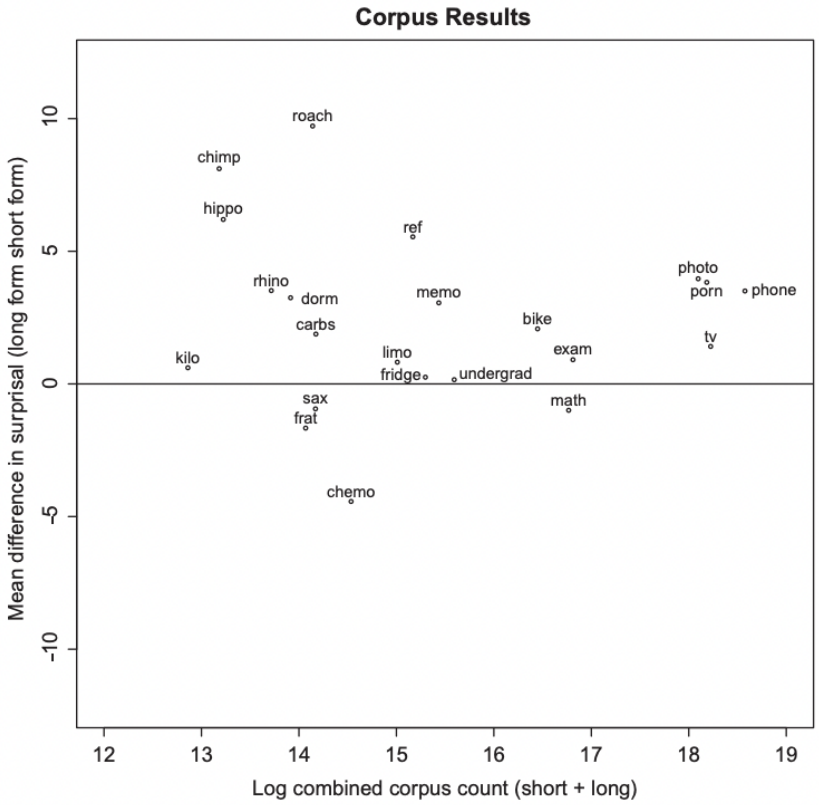


Figure 4.1: Corpus results from Mahowald et al. (2013). The y-axis shows the difference in surprisal between long and short forms of 22 English word pairs, and the x-axis shows the combined count of both variants. Most differences fall above the y=0 line, indicating that short forms are more predictable given context.

4.2.2 Arabic, Arabizi, and vowel variation

Vowel Length in Arabic

Arabic has three vowels: /a/, /i/, and /u/, each of which may be long or short. The distinction between long and short vowels is phonemic, and changing vowel length can change the meaning of a word (Gordon, 2011). For example, the pair *ḥasaba* 'he calculated' and *ḥaasaba* 'he held accountable' differ only in the length of the initial vowel /a/, and *naama* 'he slept' and *namaa* 'he grew' differ in the position of the long and short vowels. These examples illustrate that vowel length is not merely a phonetic detail but a meaningful distinction that listeners rely on to differentiate words.

In Arabic script, long vowels are represented as full letters: *alif* (ا) for /aa/, *waw* (و) for /uu/, and *yaa* (ي) for /ii/. Short vowels, by contrast, are represented as small diacritical markers written above or below consonants: *fatḥa* (بَ) for /a/, *ḍamma* (بُ) for /u/, and *kasra* (بِ) for /i/. To illustrate the difference, consider the word kitaab 'book'.

(3) *kitaab* 'book'

a. With diacritics: كِتَابُ

b. Without diacritics: كتاب

In (3a) the *kasra* beneath the first letter marks the short /i/, and the *ḍamma* above the final letter represents the nominative case vowel /u/. In (3b), these diacritics are omitted entirely, as is standard in everyday writing. The long vowel /aa/, represented by the letter (ا), remains visible in both forms.

These omissions are not exceptional; they are the norm when writing in everyday Arabic. Short vowel diacritics are consistently used in religious texts, children's educational materials, and texts where ambiguity must be avoided. In all other contexts, including newspapers, books, academic writing, and digital communication, diacritics are dropped and readers rely on context and morphological knowledge to recover the intended word. This means that Arab readers are already accustomed to reconstructing vowel information from incomplete orthographic input, a skill that makes vowel dropping in Arabizi a natural extension of existing reading practices rather than an entirely novel phenomenon.

Arabizi as a writing system

The rise of technological tools and digital communication in the early 1990s had a profound impact on language use worldwide. These technologies facilitated the global reach of English, but they also introduced a practical problem: early operating systems, mobile phones, and

messaging platforms did not support non-Roman scripts, leaving speakers of languages like Arabic without a way to write in their native script in digital spaces. In response, Arabic speakers developed a new writing system for informal digital communication: Arabizi. The name originates from a blend of "Arabic" and "Englizi" (the Arabic word for English), and the system represents spoken Arabic, primarily the various Arabic dialects, using Roman letters. Where Arabic has phonemes with no English equivalent, speakers substitute numerals that visually resemble the original Arabic letters. For example, the Arabic pharyngeal consonant *ayn* (ع) is replaced with the number 3, and the voiceless pharyngeal fricative *ḥaa* (ح) is replaced with the number 7.

This emerging system for writing Arabic is mainly used on social media platforms and SMS messaging and is far from standardized. This lack of standardization reflects several factors. One major factor is that Arabizi is a written form of spoken dialects, and dialects differ widely across the Arab world. While many dialects are mutually intelligible, variation in pronunciation across and within dialects affects how speakers spell what they say. Another factor is which Roman-script language the speaker knows. For example, an individual who knows French as a second language will write phonemes differently from another individual who knows English as a second language. The sound /ʃ/ would be written as ⟨sh⟩ by an English-influenced writer but as ⟨ch⟩ by a French-influenced one. Similarly, the long vowel /uː/ might be written as oo under English influence but as ou under French influence. Consider the following example from a Lebanese Arabic speaker, which illustrates several of these features:

(4) Kteer na3sene! Mesh adra rakez abadan lyoum
'Very sleepy! I can't focus at all today'

This single message illustrates several characteristic features of Arabizi: the numeral 3 stands in for the Arabic letter *ayn* in *na3sene* 'sleepy', and the spelling ou in *lyoum* 'today' reflects

French orthographic influence. An Egyptian speaker, by contrast, might say *naharda* instead, using an entirely different dialectal form. These factors mean there is no one-to-one mapping between Arabic and Roman letters. This variability, however, provides a natural testing ground for studying how efficiency shapes written forms.

Vowel representation in Arabizi

In Arabizi, the representation of vowels is considerably more variable than in Arabic script. Since there is no standard orthography, speakers represent the same vowel in multiple ways. For example, the long /ii/ vowel, (*Yaa*, ي), could be written as ⟨ee⟩, ⟨ii⟩, ⟨ea⟩, ⟨y⟩ or shortened to ⟨i⟩, or ⟨e⟩, among other representations. Similarly, the short /i/ vowel, *al-kasra*, can be written as ⟨e⟩, ⟨i⟩, or dropped altogether. Consider the word *kataba* 'to write' in Standard Arabic, which has three short /a/ vowels. In Arabizi, it can be written as ⟨katab⟩, ⟨ktb⟩, ⟨katb⟩, or ⟨ktab⟩. By contrast, the word *kitaab* 'a book', which has a long /aa/ vowel in the second syllable, can be written in Arabizi as ⟨ktab⟩ or ⟨kteb⟩ depending on the dialect. Notice that ⟨ktab⟩ could correspond to either *kataba* or *kitaab*, meaning that readers must rely on context to determine the intended word, much as they do when reading unvowelized Arabic script. This variability extends to all three vowel pairs.

The fact that a single consonant-only spelling can correspond to multiple Arabic words might seem problematic, but from an information-theoretic perspective, this ambiguity is expected. Piantadosi et al. (2012) argue that ambiguity is a functional property of any efficient communication system. Because context is informative about meaning, an efficient system will not redundantly encode information that the context already supplies. As a result, linguistic forms that are ambiguous in isolation can be used without compromising comprehension, as long as context reliably disambiguates them. Moreover, ambiguity allows speakers to re-use forms that are short and easy to produce across multiple meanings, reducing overall production effort. In Arabizi, consonant-only spellings serve precisely this function: they reduce

the number of keystrokes while relying on context to resolve which word is intended. The range of possible spellings for a single word highlights a key property of Arabizi: because the system is still emerging and has not yet been conventionalized, speakers have a choice about whether to include or drop vowels, making Arabizi a productive domain for studying how efficiency shaping written forms.

Vowel dropping in Arabizi

Much of the existing research on Arabizi has approached it from a computational perspective, building machine learning models to recognize Arabizi text (Tobaili et al., 2019; Shazal et al., 2020; Baert et al., 2020), or to create automatic translation systems between Arabizi and standard Arabic (Darwish, 2014; Bies et al., 2014) or between Arabizi and English (May et al., 2014). These studies have focused primarily on optimizing system performance, as Arabizi remains under-studied in the field of Natural Language Processing. While this computational work has advanced the tools available for processing Arabizi, it has not addressed the linguistic question of why Arabizi takes the forms it does.

A smaller body of work has approached Arabizi from a sociolinguistic perspective, examining the linguistic features of this emerging writing system. Several of these studies have investigated vowel dropping, the tendency for speakers to omit vowels when writing in Arabizi. Gordon (2011) explores the orthography of Arabizi in the written Levantine dialect and reports that short vowels are frequently dropped. Akbar (2019) investigates vowel and consonant deletion in Kuwaiti Arabizi using data collected from group e-conversations and private WhatsApp messages of 35 students, and finds that consonants are rarely deleted while short vowels are deleted more often. The authors also report that long vowels are slightly shortened. For example, the Kuwaiti Arabic word ⟨7aalıch⟩'your situation' has two vowels: a long /aa/ in the first syllable and a short /ɪ/ in the second. Speakers deleted the short vowel in the second syllable and shortened the long vowel in the first, resulting in

the spelling ⟨7alch⟩. Sullivan (2017) explores orthographic variation in Lebanese Arabizi on Twitter and documents the range of spellings used for each Arabic letter in the data, finding that short vowels exhibited the most variability. Across dialects and platforms, these studies converge on a consistent finding: short vowels are the most variable and most frequently omitted element in Arabizi spelling.

Some of these studies offer explanations for the phenomenon of vowel dropping. One proposal is that Arabic speakers are carrying over the conventions of Arabic orthography when writing in Roman script. That is, since short vowels are represented as optional diacritics in Arabic script, speakers tend to omit them in Arabizi as well. However, this explanation does not account for the instances when speakers do write the short vowels in Arabizi. If vowel dropping were simply an orthographic transfer, we would expect short vowels to be dropped uniformly, yet the data show considerable variation. Crucially, an orthographic transfer account predicts when vowels should be dropped (always, for short vowels) but not why speakers sometimes retain them. An efficiency-based account, by contrast, can explain this variation.

4.2.3 Efficiency and Vowel dropping

None of the studies discussed above examine vowel dropping from an information-theoretic perspective or explain this phenomenon through the lens of communicative efficiency. Yet the pattern is well suited to such an analysis. As established in Section 4.2.1, Mahowald et al. (2013) showed that English speakers are more likely to use shorter word forms in more predictable contexts — contexts where the word carries less information and is therefore less costly to reduce. Vowel dropping in Arabizi presents a natural parallel: when a word is predictable given its context, the vowels contribute less information, and speakers can afford to drop them without compromising the reader's ability to recover the intended meaning.

Evidence from spoken language suggests that reduction driven by predictability is not limited to online processing but can become part of how words are represented. Seyfarth (2014) found that English words with low average informativity have shorter acoustic durations even when they occur in unpredictable contexts, which indicates that habitual reduction in predictable contexts leaves a trace in lexical representation. In other words, words that are frequently reduced come to be produced in their reduced form as a default. A similar process may be at work in Arabizi. If certain words are routinely spelled without vowels because they tend to appear in predictable contexts, the vowel-dropped spelling may become the conventional form for those words, used even when the context does not strongly predict them. This would mean that what begins as an efficiency-driven choice (such as dropping vowels when the context makes the meaning recoverable) could over time become a fixed spelling convention, independent of any particular context.

This rest of this chapter investigates whether the same efficiency pressures that shape word length reduction in English extend to vowel dropping in Arabizi. Specifically, I ask: does contextual predictability affect the length of the word form chosen? If efficiency is at play, we expect shorter spellings to occur in more predictable contexts, where the word carries less information and the reader can still recover the intended meaning from context alone. To test this, I follow Mahowald et al.'s methodology, calculating the surprisal of short and long spellings of Arabizi words as a measure of predictability. I also test the impact of frequency on the length of the words. That is, if a word is more frequent, are Arabic speakers more likely to drop the vowels when writing Arabizi?

4.3 Methods

4.3.1 Choosing short and long forms of the words

Following the Mahowald et al. (2013) methodology, I compare the use of short and long pairs of Arabizi words with respect to predictability of the word in context. The data I used came from the Egyptian Arabic text chat conversations in the Arabic Treebank of the BOLT (Broad Operational Language Translation) project, obtained through the Linguistic Data Consortium (LDC). The data set consists of 157,569 lines and 694,910 tokens, of which 120,129 are unique. The data are Egyptian Arabic written in both Arabic script and Arabizi, and was not annotated. Since the data was collected from SMS and chat messages, it reflects informal naturalistic digital communication — precisely the register in which Arabizi is most commonly used. I first pre-process the data into a usable format, then classified it into Arabic script and Arabizi. My classification technique was simple: I identified text written in Roman characters as Arabizi using a Unicode function in Python (`encode().isalnum()` function). This was possible because, based on the LDC documentation, Arabizi is the only Romanised script in the data. While the data may contain occasional English or French words embedded in otherwise Arabizi messages, the BOLT corpus was collected from Egyptian Arabic chat conversations, and participants were instructed to communicate in Arabic. The Roman-character text therefore overwhelmingly represents Arabic written in Arabizi rather than a different language. The resulting Arabizi subset consists of 103,141 lines and 460,079 tokens, of which 81,857 are unique.

From the obtained subset, I generated a list of the most frequent words and manually chose 13 pairs of words with short and long spellings (see Table 4.1). The short spelling in each pair reflects vowel dropping. For example, the first-person pronoun /ana:/ ('I' or 'me' in English) is written as <anaa> or might be alternatively written as <ana> in its reduced form, where the long vowel has been shortened. Because the most frequent words in the data

Table 4.1: Arabizi word pairs used in the analysis. The short variant reflects vowel dropping. Many pairs are function words such as pronouns and prepositions.

short variant	long variant	translation
ana	anaa	I, me
msh	mesh	not, will not
bs	bas	just, only, stop
kda	keda	like this
mn	min	from
tyb	tayeb	okay
3shan	3ashan	because
y3ni	ya3ni	meaning that, that means
knt	kont	I was
m3	ma3	with
hwa	howa	he
tmam	tamam	perfect
kman	kaman	also

set tend to be function words, many of the selected pairs are pronouns and prepositions.[1] Some of these words have additional spelling variants beyond the two considered here. For example, *anaa* could also be written as ⟨ane⟩, ⟨ani⟩, or other variants depending on the dialect. However, for the current study, I restrict the comparison to a single short-long pair per word in order to maintain a clear contrast between reduced and unreduced forms. To determine the longer spelling of some of these words, I consulted the MADAR lexicon (Bouamor et al., 2018), which provides dialectal variants of words as they are pronounced in different Arabic-speaking cities. I used the Cairo city pronunciations to match the Egyptian Arabic data in the BOLT corpus. For words whose MADAR spellings did not appear in the BOLT data set, I modified the spelling to the closest attested form in the data.

4.3.2 Training an n-gram model

After selecting the word pairs, I trained a neural language model on the BOLT Arabizi data set, using a single preceding word as context (equivalent to a bigram model). I first split

[1]The list of pairs will be revisited in future extensions of this work

the data into a training set (90%) and development set (10%). The split was performed at the conversation level, ensuring that messages from the same conversation did not appear in both sets. I trained the n-gram model using PyTorch, experimenting with different model architectures, learning rates, hyperparameters, and optimizers. The best-performing model was a bigram model, which estimates the probability of each word given the single word immediately preceding it, optimized with Adam with an initial learning rate of 1e-5. While larger context windows (such as a trigram model) might in principle capture richer contextual information, they require more data to estimate probabilities reliably, and in our case did not outperform the bigram model. During each epoch, the model trains on the entire training set, and attempts to minimize the loss. I evaluated the improvements of the model by calculating the loss on the DEV set. The loss function I used to measure model performance is negative log likelihood. Optimally, the DEV set loss should be as close to zero as possible, indicating that the model has learned to assign high probabilities to the words that actually occur in the data.

The model had a total of 18,758,181 trainable parameters (see Figure 4.2). I trained the model on a GPU for faster running time. Each epoch took an average of 38 seconds. I used the full vocabulary of the Arabizi data set (81,857 unique tokens) without applying any frequency cutoff or replacing rare words with an unknown token. I did not use any smoothing techniques for the model because it is implicitly smoothed. That is, the neural network model is unlikely to produce zero probabilities, but instead will produce extremely small probabilities.

One common issue with training models is overfitting: the model achieves a small loss on the training set, but does not perform as well on the DEV set. When this happens, it is an indication that the model has memorized patterns specific to the training data rather than learning generalizable patterns, and therefore does not generalize well to unseen data. To prevent overfitting, I monitored the development set loss at the end of each epoch.

```
+------------------+------------+
|     Modules      | Parameters |
+------------------+------------+
| embedding.weight |  8185700   |
|  linear1.weight  |   12800    |
|   linear1.bias   |    128     |
|  linear2.weight  |  10477696  |
|   linear2.bias   |   81857    |
+------------------+------------+
Total Trainable Params: 18758181
```

Figure 4.2: Model architecture and parameter counts. The model has 18,758,181 trainable parameters, with an embedding dimension of 100 and a linear dimension of 128.

During experimentation, I found that the model started overfitting after the 11th epoch (the development loss began increasing even as the training loss continued to decrease), and therefore selected 11 as the optimal number of epochs (see Figure 4.3). I also trained a trigram model, which took more time to learn (16 epochs) but did not reach as low a loss as the bigram model.

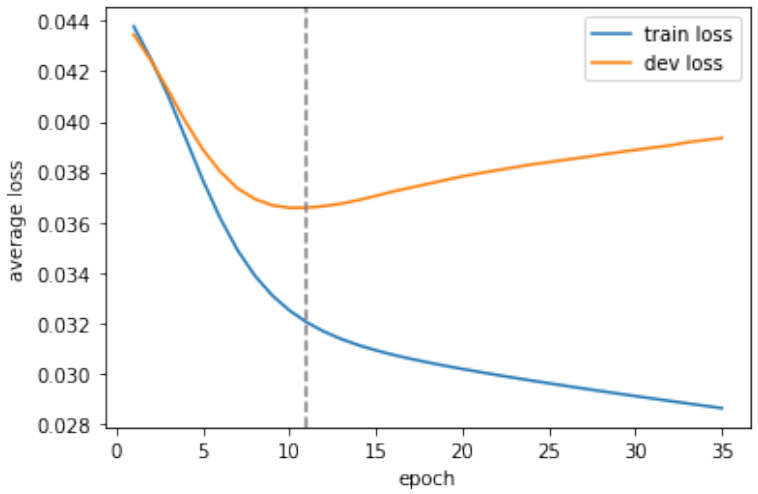


Figure 4.3: Training and development loss across epochs. The development loss reaches its minimum at epoch 11 and begins increasing thereafter, indicating overfitting. Loss is computed as negative log likelihood divided by the number of bigrams, placing training and development losses on the same scale.

4.3.3 Obtaining informativity measures

Using the bi-gram model, I calculated the average surprisal of each lexical item given its context. Since surprisal serves as a measure of informativity, a higher surprisal value for a given word, regardless of which spelling variant it appears in, indicates that it carries more information in that context. Consequently, words with higher surprisal values are less predictable from context alone. My hypothesis is that writers are more likely to use shorter spelling variants in more predictable contexts, where the additional orthographic material in the longer variant is less necessary for successful communication. To test this hypothesis, I measured the surprisal (and hence predictability) of the meaning of a word by combining the probability of that word occurring in either spelling in a given context. I calculated the average surprisal using the following equation:

$$-1/N \sum_{i=1}^{N} \log P(W = w|C = c_i) \tag{4.1}$$

where N is the total frequency of a lexical item w across all contexts, and c_i represents each context in which w actually occurs. Since the bi-gram model proved to be the most reliable (as discussed in the previous section), the context c_i is the single word immediately preceding the target word w. In other words, this equation computes the average of the log probabilities of the word across all of its attested contexts: the higher this value, the more surprising (and thus more informative) the word is on average.

Crucially, because I am interested in the informativity of the meaning rather than of a specific spelling, I combine the probabilities of the two spelling variants before computing surprisal. That is, for each context c_i, I sum the probability of the short form and the probability of the long form appearing in that context, and then take the log of this combined probability.

This ensures that the surprisal measure reflects how predictable the **meaning** is in a given context, regardless of which spelling variant was used. I then label each context according to which spelling variant actually appeared: if the long spelling occurred, the context is labeled as "long" and if the short spelling occurred, it is labeled as "short". For each word pair, I computed the average surprisal, represented by the log-probabilities, over the long and the short spelling variants of the words. I then subtracted the average surprisal of the short spellings from that of the long spellings. Figure 4.4 shows these differences plotted against the combined log frequency of each word pair. The results show that for most word pairs, the surprisal difference lies above the *y=0* line, which means that the long spelling variants tend to carry higher average surprisal than their short counterparts. If statistically significant, this pattern would suggest that the average predictability of a meaning across contexts influences which spelling of the word is used.

To test whether this pattern is statistically reliable, I ran a mixed-effects logistic regression predicting word form (short vs. long) by surprisal of the word meaning, with random intercepts for word meaning. The average surprisal for the long spelling variants was 4.92, compared to 4.51 for the short variants. The mixed-effects model confirmed this difference ($\beta = -0.111, z = -2.76, p < 0.01$). This confirms that longer spelling variants tend to occur in less predictable contexts, consistent with the efficiency-based hypothesis that writers reserve additional orthographic material for contexts where the word meaning is harder to anticipate. A t-test comparing the average surprisal rates of all short versus all long spellings was also significant ($t = 21.262, p < 0.001$). These results further support the overall pattern that the longer spellings of the words have higher information content than their shorter counterparts. I also performed t-tests for each word pair separately to see whether the surprisal difference between short and long spellings was significant for each pair. None of these individual comparisons reached significance, likely because the data for any single word pair was too sparse to detect the effect reliably. Finally, Figure 4.4 is a visual assessment of the role of word frequency: the correlation between combined log frequency and the surprisal

difference was not significant ($r^2 = 0.29$, Spearman's $\rho = -0.483$, p= 0.097) suggesting that the overall frequency of a word pair does not reliably predict the surprisal differences between its long and short forms.

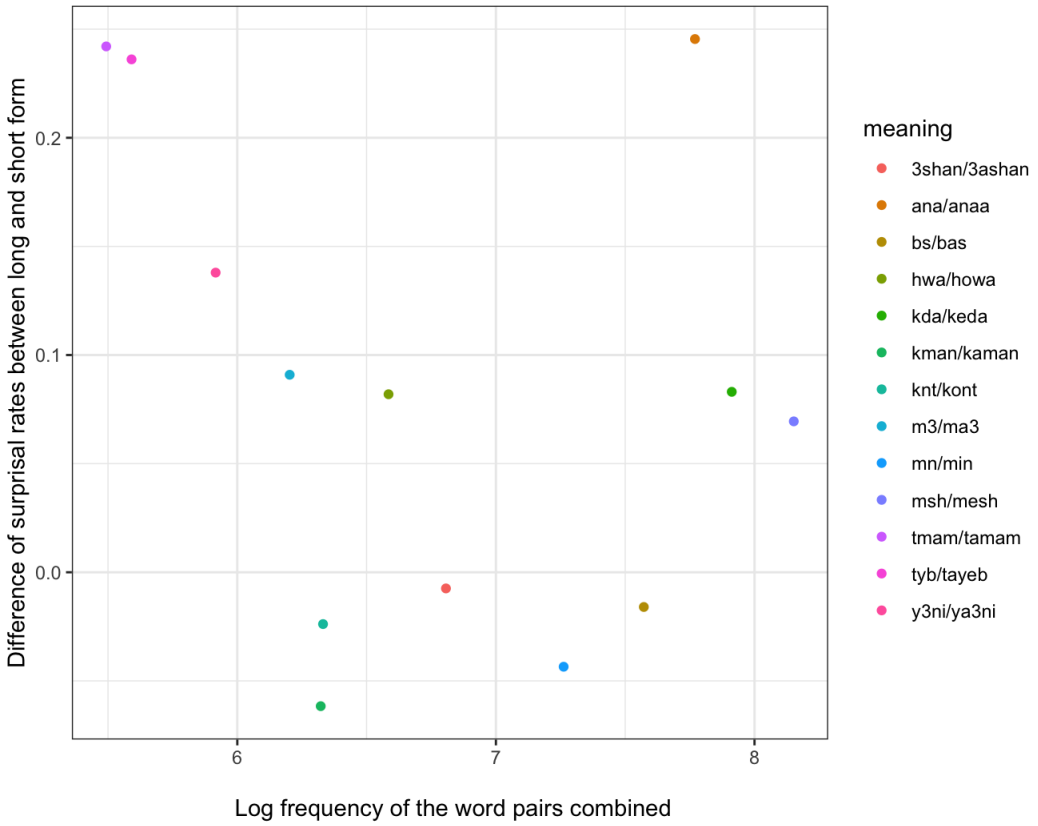


Figure 4.4: Difference in average surprisal between long and short spelling variants of lexical items plotted against log frequency of the word pairs combined. Most of the word pairs lie above the $y = 0$, indicating that longer spelling variants carry higher average surprisal (more informative) than their shorter counterparts.

After obtaining the informativity measures from the BOLT data, I ran the model on a separate set of data from Twitter[2] that the model had not seen during training. I used the Twitter API to extract tweets containing the 13 pairs of spelling variants for which I had obtained informativity scores. To match the Egyptian Arabic dialect of the BOLT data, I restricted the search to tweets geo-located to Cairo, Egypt. The initial data consisted

[2]The platform was renamed X in 2023. I retain the name Twitter throughout, since the data were collected under that name and via the Twitter API, and the platform's affordances at the time of collection are what bear on the analysis.

of approximately 6400 tweets. I filtered out bigrams in which the preceding word did not appear in the training vocabulary. Because the model had not encountered these context words during training, it could not provide reliable probability estimates for the target word. After filtering, 4,719 bi-grams remained. I then computed the difference in surprisal between the long and short spelling variants using the same procedure described above. The results are shown in Figure 4.5. I did not plot surprisal against frequency for this data set because the search query was designed to pull tweets containing each spelling variant, so the frequencies of the pairs are roughly comparable by construction. An alternative approach would be to use frequency estimates from an independent corpus such as BOLT, which I leave for future work. A t-test aggregating over all lexical items showed the expected effect: the long spelling variants carry higher surprisal than the short ones ($t = 11.801, p < 0.001$). However, the effect is not significant in a mixed-effects logistic regression with random intercepts for word meaning ($\beta = -0.03628, z = -0.586, p = 0.558$). Overall, 9 out of 13 lexical items showed the expected pattern, with the long spelling variant carrying higher average surprisal than the short variant. While the aggregate t-test confirms this overall trend, the mixed-effects regression suggests that the effect is not consistent enough across individual lexical items to reach significance.

4.3.4 Scaling up the Twitter analysis

The results from Section 4.5 showed that while the expected efficiency effect held in the private LDC chat data, it did not emerge clearly in the public Twitter data. One possible explanation is that the Twitter test set used in the previous analysis was relatively small (approximately 6,400 tweets yielding 4,719 usable bigrams), and the tweets were pulled using a simple API search targeting only the 13 word pairs. Additionally, the model used to obtain the informativity measures was trained on the private LDC data and tested on public Twitter data, introducing a potential domain mismatch between training and test sets. To address

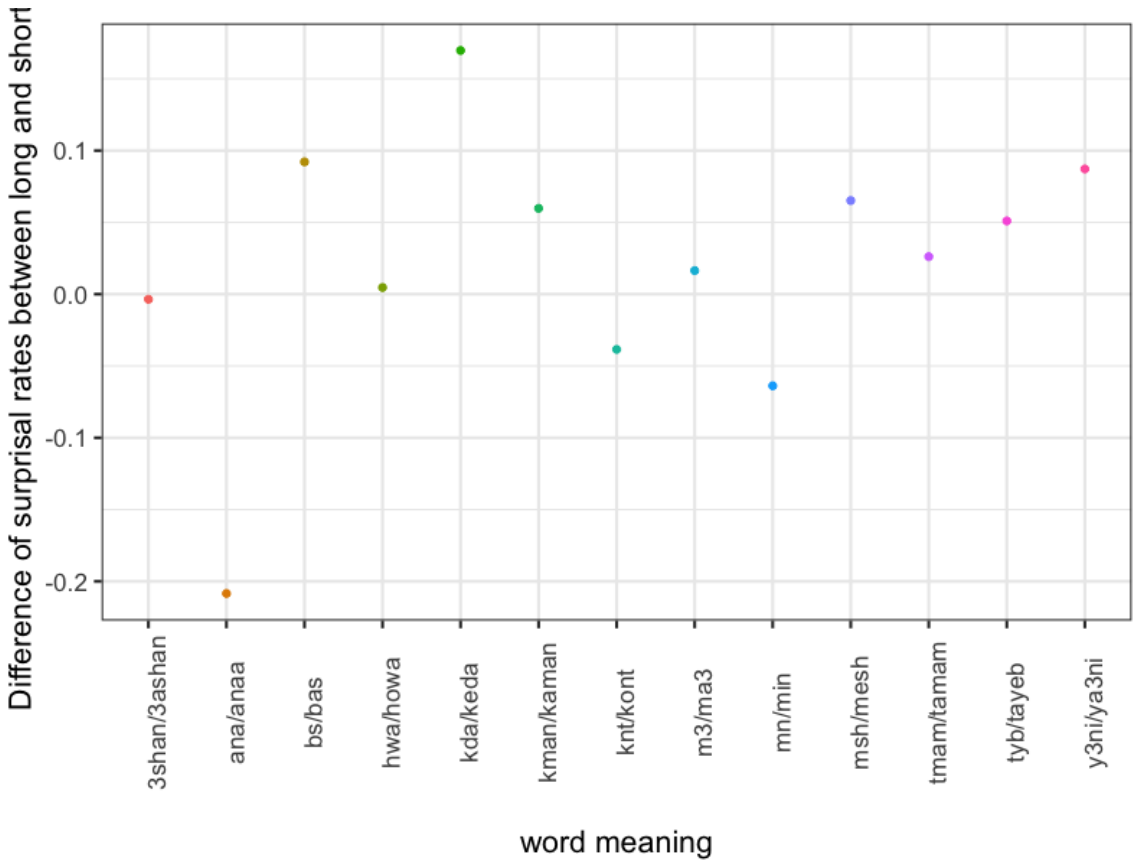


Figure 4.5: Difference in surprisal between long and short spelling variants for each word in the Twitter data. Here, frequency of the spelling variants is assumed to be similar. Nine of the thirteen pairs fall above the y=0 line, showing the expected pattern. Unlike the BOLT chat data, some pairs pattern differently here; for example, the word ⟨ana/anaa⟩which was predominantly used in its short form in the private chats, appears more often in its longer form on Twitter.

these limitations, I conducted a more thorough investigation of the Twitter data, scaling up the data collection, refining the word pair selection, and training a new model directly on Twitter data.

Collecting a larger Twitter data set

Using the Twitter API Academic access, I collected over 10 million tweets written in Arabizi. Unlike the original data collection, which relied on a simple API search for the 13 word pairs, this larger-scale collection required a more involved pre-processing pipeline. The initial data

could not be easily separated into Arabizi and non-Arabizi text, as there was no built-in function to distinguish Arabizi from Arabic script, English, or French text. I therefore applied several preprocessing steps: I excluded tweets automatically detected as English, Arabic-script, or French, removed mentions, links, and hashtags, and filtered the remaining data using a distinctive feature of Arabizi writing, namely, the use of numerals within words (e.g., 3 for (ع) and 7 for ح). Specifically, I compiled a list of approximately 35 words containing these distinctive numeral-letter substitutions and used them as seed terms to identify Arabizi-writing users. This yielded approximately 60,000 authors. I then collected all available tweets from these authors over the course of six months. After preprocessing, the data set consisted of approximately 3.5 million tweets.

Building an Egyptian Arabic classifier

Since Arabizi is used across a wide range of Arabic dialects, and the analysis targets Egyptian Arabic specifically (to remain consistent with the LDC data), I needed a method to filter out non-Egyptian tweets. I trained a binary classifier to identify Egyptian tweets. To build training data for the classifier, I manually labeled a subset of 1,000 lines as Egyptian (EG) or non-Egyptian, relying on my knowledge of Arabic dialects to identify dialect-specific lexical and phonological markers. I then trained a CatBoost classifier [3], a gradient boosting model, to distinguish Egyptian Arabizi from other dialects. The classification relied on several dialect-specific features, including lexical items unique to Egyptian Arabic (e.g., ⟨keda⟩ for 'like this', ⟨ezayak⟩ for 'how are you'), characteristic phonological patterns reflected in spelling (e.g., the use of ⟨g⟩ for the Egyptian pronunciation of ج, where other dialects would write ⟨j⟩), and other common Egyptian discourse markers. While the training set of 1,000 lines is relatively small, it was sufficient to capture the most salient dialect distinctions, and the resulting classifier achieved 86% accuracy on a held-out test set. I then applied the

[3]CatBoost (Prokhorenkova et al., 2018) is a gradient boosting algorithm designed for categorical data. I used it here because the features used for classification were primarily categorical.

classifier to the full 3.5 million tweet data set, to obtain only Egyptian Twitter messages; this was data unseen to the classifier model. The resulting data set was approximately 850,000 tweets, roughly eight times larger than the BOLT corpus used to train the original model and more than 100 times larger than the Twitter data set used in the previous analysis.

Training a new n-gram model on Twitter data

In the previous analysis, the n-gram model was trained on the private LDC chat data and tested on public Twitter data, introducing a potential confound: the model may not generalize well across these two registers. To eliminate this domain mismatch, I trained a new neural n-gram model directly on the Twitter data. Following the same procedure described in Section 4.3.2, I split the Twitter data into training and development sets (90% and 10%, respectively) and trained bigram, trigram, and unigram models using Pytorch. The best-performing model, as measured by perplexity on the development set, was again a bigram model. I used the same surprisal equation (Equation 4.1) to obtain informativity measures, summing the log probabilities of the short and long forms of each word pair across all contexts in which they occurred.

Choosing a new set of word pairs

One limitation noted in the previous analysis was that the 13 word pairs used were predominantly function words such as pronouns and prepositions (see Table 4.1). Function words tend to behave differently from content words in terms of their distribution and predictability, and many of them frequently occur as the first word in a sentence or as standalone messages, limiting the available preceding context for the bigram model. To address this, I constructed a new set of 15 word pairs that includes both function words and content words. The new pairs are shown in Table 4.2. This new set was designed to include words whose

Table 4.2: New Arabizi word pairs used in the scaled-up Twitter analysis. Unlike the original set (Table 4.1), these pairs include both function and content words.

short variant	long variant	translation
nas	naas	people
yom	youm	day
ayza	3ayza	I want feminine
msh	mesh	not
bgd	begad	seriously
aktr	aktar	more
gam3a	gama3a	group of people
anam	anaam	I sleep
nfsy	nefsy	I wish
wahd	wahed	one
y3ne	ya3ne	means
kda	keda	like this
hwa	howa	he
3ayz	3ayez	I want masculine
walahy	wallahy	I swear (used to mean seriously)

short and long variants differ specifically in whether a vowel is present or dropped, and to cover a broader range of word types than the original function-word-heavy list.

Results

I first tested the new Twitter-trained model using the original set of 13 function word pairs on held-out Twitter data. This analysis was based on 74,788 token instances across the 13 word pairs, compared to $\sim$ 5,000 instances in the LDC data. The results showed no clear correlation between predictability and word length. Unlike the LDC results, where most of the average surprisal differences were above the y=0 line, suggesting that shorter forms tend to be more predictable, the Twitter results showed the differences clustering around zero with no consistent direction (see Figure 4.6). A mixed-effects logistic regression predicting form choice (long vs. short) from log probability, with random intercepts for word pair, found no significant effect of predictability ($\beta = 0.003$, $z = 0.22$, $p = .83$). The Spearman correlation between log frequency and the average surprisal difference across the 13 word

pairs was also not significant ($\rho = 0.13$, $p = .67$).

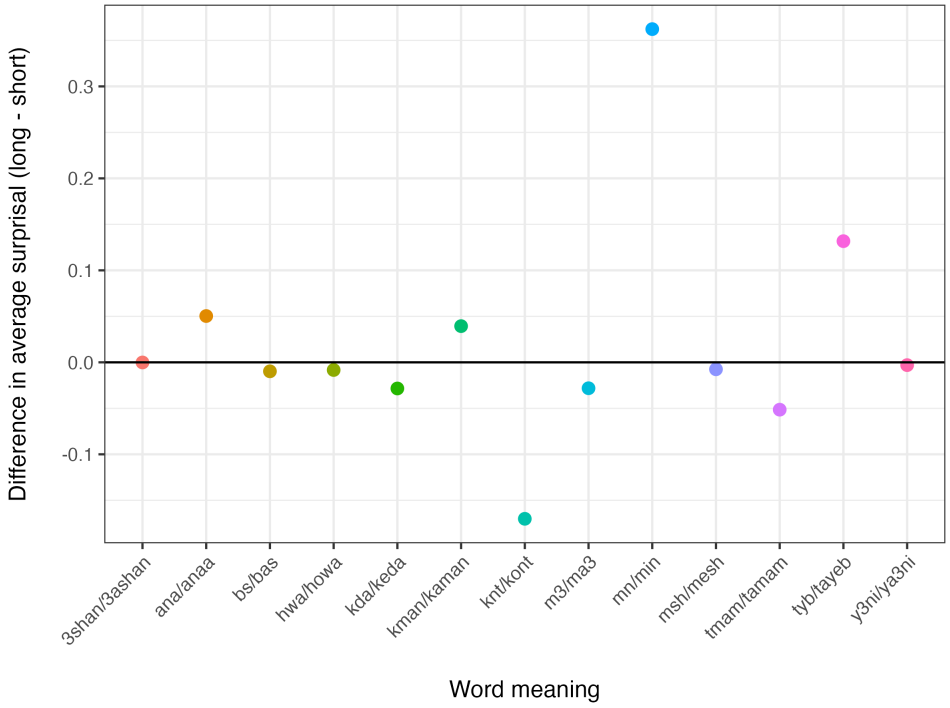


Figure 4.6: Difference in average surprisal between the long and short forms of the original word pairs, tested on held-out Twitter data using the Twitter-trained bi-gram model. Context size is n=1.

One possible explanation for the null result is that a unigram context window is too narrow to capture meaningful predictability differences on Twitter data. To test this, I also trained bigram and trigram models and repeated the analysis. As shown in Figure 4.7, increasing the context size did not change the pattern of results. Mixed-effects logistic regressions confirmed that predictability did not significantly predict form choice at either context size (β = -0.02, z = -1.14, p = .25 for n=2; β = -0.001, z = 0.07, p = .94 for n=3). Increasing the context window naturally reduces the number of usable instances, since more surrounding words must appear in the model's vocabulary (from 74,788 at n=1 to 60,715 at n=2 and 50,044 at n=3). Despite this reduction, the sample sizes remain large, and the null pattern persists.

I then tested all three models on the expanded set of 15 word pairs, which included both

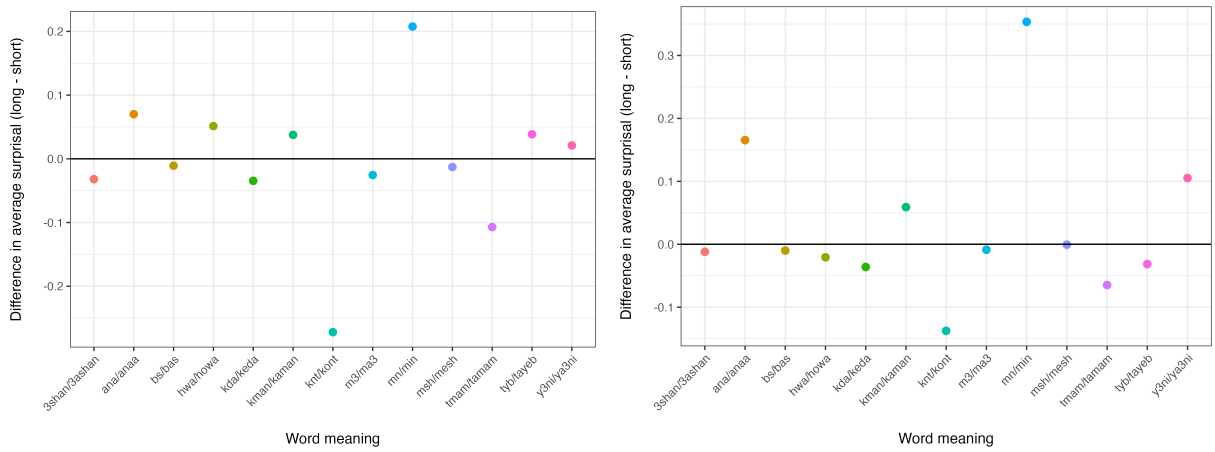
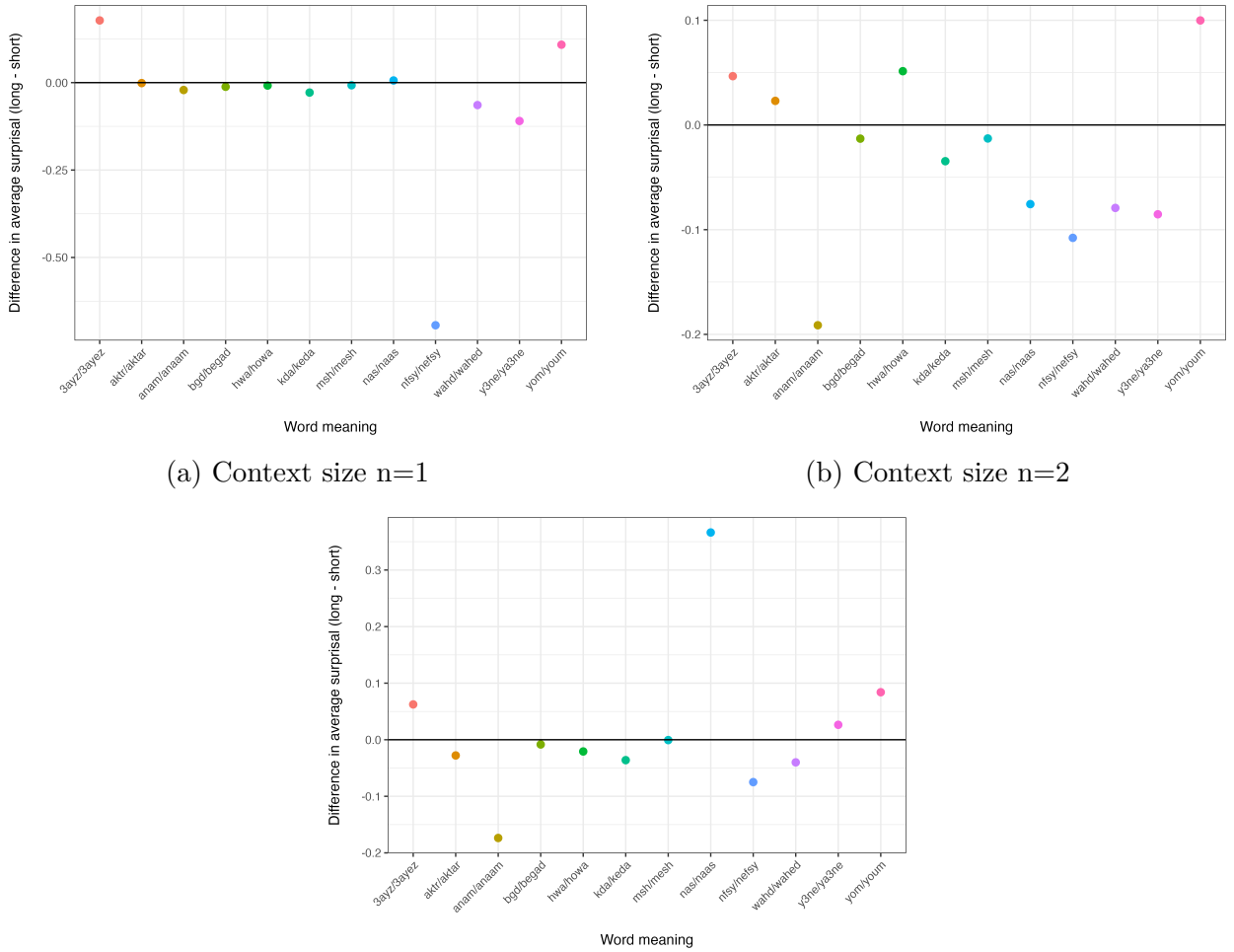


Figure 4.7: Difference in average surprisal between the long and short forms of the original word pairs, tested on held-out Twitter data using the Twitter-trained models. Context size on the left is n=2, and on the right n=3

function and content words. A mixed-effects logistic regression on the full 15 pairs revealed a significant effect of predictability on form choice, but in the opposite direction from what the efficiency hypothesis predicts: more predictable words were more likely to appear in their longer form (β = -0.11, z = -12.52, $p < .001$). This pattern was consistent across all three context sizes (β = -0.11, z = -10.79, $p < .001$ for n=2; β = -0.10, z = -7.95, $p < .001$ for n=3). However, this set includes two pairs whose short and long variants differ in consonant structure rather than vowel presence: *walahy/wallahy*, which involves consonant gemination, and *ayza/3ayza*, which involves deletion of the pharyngeal consonant *ayn* represented by the number 3. Excluding these two pairs did not change the pattern. However, the effect is largely driven by a single word pair, *gam3a/gama3a* ('group of people'). Removing this pair reduces the effect to marginal significance (β = -0.02, z = -1.81, p = 0.07 for n=1) and eliminates it entirely at larger context sizes (p = 0.52 for n=2; p = 0.55 for n=3).

To rule out the possibility that the null result was an artifact of the neural n-gram architecture, I also trained a traditional count-based n-gram model on the same Twitter data. Mixed-effects logistic regressions on the count-based surprisal values yielded nominally significant effects for both the original function word pairs (β = 0.002, z = 2.20, p = .03) and



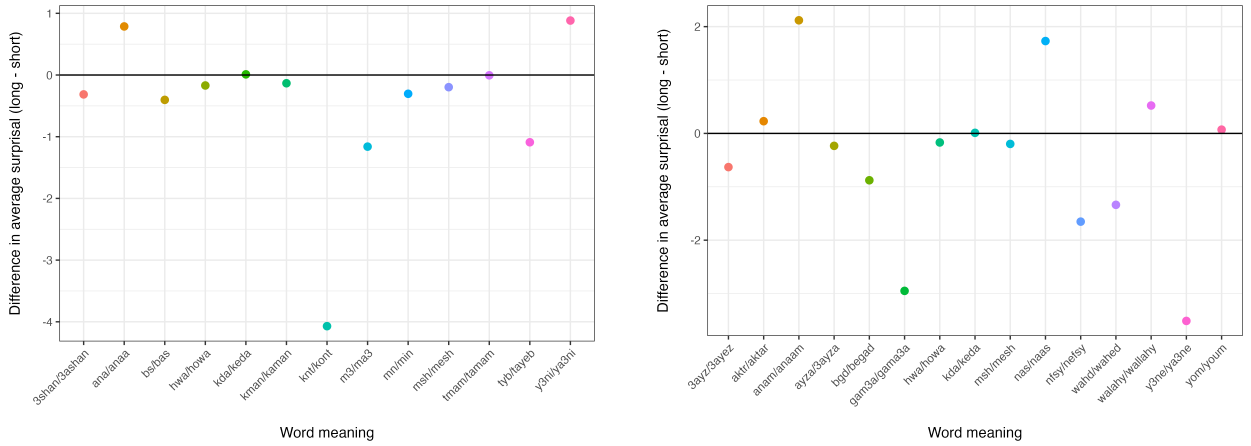
(a) Context size n=1

(b) Context size n=2

(c) Context size n=3

Figure 4.8: Difference in average surprisal between the long and short forms of the expanded word pairs after excluding the outliers, tested on held-out Twitter data using models with increasing context sizes.

the updated list consisting of mostly content words ($\beta = 0.002$, $z = 2.35$, $p = .02$) (see Figure 4.9). However, the effect sizes are extremely small, so these results are best interpreted with caution. In summary, none of the analyses on the Twitter data robustly replicated the predictability-length correlation observed in the LDC chat data, regardless of context size, word pair selection, or model architecture. Table 4.3 summarizes the results across all analyses. The only analysis that yielded a significant effect in the predicted direction was the BOLT private chat data. On Twitter, the effect was either absent, driven by a single outlier word pair, or too small to interpret with confidence.



(a) Function words

(b) Content words

Figure 4.9: Difference in average surprisal between the long and short forms using a count-based n-gram model on held-out Twitter data.

Table 4.3: Summary of mixed-effects logistic regression results across all analyses. ✓ = significant in the predicted direction; ✗ = not significant; ✗* = significant but in the opposite direction; ? = inconclusive. Updated content words list excludes three word pairs: *walahy/wallahy*, *ayza/3ayza*, and *gam3a/gama3a*; the latter was driving the effect in the opposite direction.

Context size	n=1	n=2	n=3
Neural BOLT function words	✓	—	—
Neural Twitter function words	✗	✗	✗
Neural Twitter content words (15 pairs)	✗*	✗*	✗*
Neural Twitter updated content words (12 pairs)	?	✗	✗
Count-based function words	?	—	—
Count-based updated content words	?	—	—

4.4 Discussion

The central question of this chapter was whether contextual predictability shapes vowel dropping in Arabizi, paralleling the efficiency-driven word length reduction documented in English by Mahowald et al. (2013) and the probabilistic reduction patterns reported by other literature (Jurafsky, 1996; Bell et al., 2009). The results I obtained present a mixed picture. In the BOLT private chat data, the efficiency prediction was confirmed: a mixed-effects logistic regression showed that longer spelling variants carry higher surprisal, and writers

are more likely to use shorter forms in more predictable contexts. In other words, when the surrounding context is informative enough to disambiguate a shorter form, Arabizi users forgo the less informative vowels for the sake of efficiency. This pattern holds even though I considered only one type of context: the single word immediately preceding the target, as captured by the bigram model. However, this effect did not replicate in the Twitter analysis. Across the original 13 function word pairs, predictability did not significantly predict form choice at any context size, regardless of whether the model was a neural n-gram or a traditional count-based model. This failure to replicate is robust: it holds across three context window sizes, two model architectures, and two sets of word pairs.

The divergence between the BOLT and Twitter results may appear contradictory, but on closer inspection, the two data sets differ drastically in communication setting. The BOLT data come from private chat conversations and SMS messages, where interlocutors may share background knowledge and can rely on that shared context to resolve ambiguity introduced by vowel dropping. Twitter, by contrast, is a public-facing platform where the intended audience is wider and less predictable. Writers posting publicly may tend toward fuller spellings to ensure comprehension of unknown readers. The broader view of efficiency is that it is sensitive to the demands of successful communication: when the cost of miscommunication is low (as in private chat), speakers reduce more freely; when it is higher (as in public Twitter posts), they tend to be more explicit. This could explain why the predictability-length correlation that emerged in the BOLT data did not replicate in Twitter. In other words, the efficiency pressure toward shorter forms may be weakened in a public setting where writers cannot rely on shared context. This account makes a further, directly testable prediction: independent of contextual predictability, the overall rate of short-form (vowel-dropped) spellings should be lower on Twitter than in the private BOLT data, since writers hedging against an unknown audience should favor fuller forms across the board. A direct comparison of aggregate short-form rates across the two corpora is a natural next step.

The scaled-up Twitter analysis of content word pairs introduced a finding that warrants separate discussion. After excluding two pairs whose alternation involves consonant structure rather than vowel reduction (*walahy/wallahy* and *ayza/3ayza*), the remaining 13 content word pairs showed a significant effect of predictability on form choice, but in the opposite direction from what the efficiency hypothesis predicts: more predictable words were more likely to appear in their longer form. This reversed effect, however, is largely driven by a single word pair, *gam3a/gama3a* ('group of people'). Removing this pair reduces the effect to marginal significance at n=1 ($p = 0.07$) and eliminates it at larger context sizes. This pair was subsequently excluded from the analysis for two reasons. First, the second vowel in the word *gama3a* is a long vowel and the other two vowels are short; this makes it unlikely that a writer would reduce this vowel while retaining the two short vowels, unlike the other pairs in the set where the alternation targets short vowels specifically. Second, the short spelling *gam3a* may correspond to two other meanings ('gathering' and 'university'), potentially introducing ambiguity if used as a short spelling of *gama3a*. After excluding this pair, the effect of predictability on form choice was no longer significant at any context size. The reversed direction of the content word effect, even if driven by a single pair, is nonetheless worth situating in the broader literature. Levshina (2022b) found that in standard Arabic, informativity given the previous word shows a negative correlation with word length. That is, more predictable words tend to be longer, not shorter. So it could be that this feature bleeds over from the use of Standard Arabic to writing in Arabizi.

Taken together, the results from this chapter suggest that the efficiency pressures documented in English do extend to Arabizi as seen in the BOLT analysis, but their expression is modulated by the communicative setting (BOLT vs Twitter) and by properties of the words themselves. Social and communicative pressures toward explicitness in the public Twitter are particularly noteworthy given that neither the short nor the long spellings in Arabizi are standardized. Spelling choices can vary significantly from one writer to another and from one dialect to another. Despite this variability, we find at least partial evidence

for this efficiency-driven pattern in the Bolt data, and this emerging pattern in a writing system without fixed orthographic conventions speaks to the robustness of the underlying communicative pressure.

A couple of limitations of the current study should be noted. First, the word pairs analyzed here were restricted to two variants per lexical item, though Arabizi words often have multiple possible spellings. A more comprehensive analysis that considers the full range of attested variants for each word could reveal finer-grained patterns in how efficiency shapes spelling choices. Second, the most frequent words in the data set were mostly function words, including prepositions and pronouns such as ⟨ana/anaa⟩ 'I' or ⟨tmam/tamam⟩ 'perfect'. This limited the choice of short/long spelling pairs, since many of these function words do not always appear with a preceding context in the data. For example, pronouns like ⟨ana/anaa⟩ can occur as the first word in a sentence, and a word like ⟨tmam/tamam⟩ might stand alone as a complete chat message. While I tried to address this point in the reconstructed new set of pairs, number of instances of some words still posed a real problem. The new set of pairs still included words that could be used as function or content words such as *bgd/begad* (really) and *kda/keda* (like this). This could be addressed in future work by considering a different set of short/long pairs of strictly content words drawn from a larger data set which will provide more insight into how vowel dropping is influenced by efficiency beyond function words.

Finally, the finding that Arabizi users drop vowels in predictable contexts raises a question about whether this reduction is driven by online sensitivity to context or by convention. Seyfarth (2014) found that English words with low average informativity have shorter acoustic durations even in unpredictable contexts, suggesting that habitual reduction in predictable contexts leaves a trace in the lexical representation itself. If a similar process operates in written language, words that are routinely spelled without vowels in Arabizi may come to be represented in their vowel-less form regardless of context.

Chapter 5

General Conclusion and Future Directions

5.1 Connecting threads

This dissertation has examined the hypothesis that communicative efficiency shapes linguistic form and function, bringing evidence from Arabic and Arabizi. Across three chapters, I examined efficiency at different levels of linguistic structure: the rate at which speakers modify nouns with adjectives (Chapter 2), the ordering of multiple adjectives within a noun phrase (Chapter 3), and the reduction of word length through vowel dropping in Arabizi (Chapter 4). Each chapter operationalized cost and reward differently, but all three converge on a shared conclusion: the pressures toward successful and efficient communication that have been documented in English and other well-studied languages are also at work in Arabic. At the same time, studying a typologically distinct language sharpens our understanding of when and how these pressures operate.

The clearest convergence across the chapters is on the role of cost in shaping linguistic form.

In Chapter 2, cost is realized as the production of additional adjectives that do not contribute to incremental reference resolution; in Chapter 4, cost is the additional keystrokes required to spell out vowels that the context already makes recoverable. Chapter 3's relationship to the efficiency framework is less direct: rather than testing a cost-reward trade-off, it establishes that Arabic adjective ordering follows the same subjectivity-based hierarchy documented in English, a pattern that has been argued to reflect pressures toward communicative success, whether through incremental restriction of the nominal denotation (Franke et al., 2019; Scontras et al., 2019, 2020a) or through preserving more informative content closer to the noun to reduce the risk of memory loss (Hahn et al., 2018).

In each chapter, the structural properties of Arabic offered a more direct test of the efficiency predictions that the literature suggests. In Chapter 2, the post-nominal position of adjectives in Arabic created a direct test of the incremental efficiency hypothesis: if pre-nominal adjectives contribute to reference resolution earlier in the utterance, then post-nominal languages should show lower rates of redundant modification. A comparison of English, Spanish, and Arabic alone did not support this prediction: all three languages modified nouns at approximately the same rate, and although the differences reached statistical significance given the large sample sizes, the effect sizes were negligible. The pattern only becomes visible across a larger sample: when I extended the analysis to 74 languages, the predicted relationship between post-nominal propensity and modification rate emerged.

The most striking pattern in the cross-linguistic data is the empty upper-right quadrant: no language in the sample combined a high proportion of post-nominal adjective use with a high overall modification rate. This distributional gap is consistent with the efficiency account. Using post-nominal adjectives may be more costly since it delays the incremental narrowing of referent candidates, and this bleeds over to the use of post-nominal modification overall. In other words, this gap could reflect a soft pressure that accumulates over generations of language transmission through iterated learning, or a constraint of language processing. Slight

preferences against post-nominal modification, if consistently present, would be expected to leave a cumulative imprint on the grammar of a language over time. The dependency-based reanalysis adds an important qualification. When modification is identified through syntactic dependency relations rather than adjacency of part-of-speech tags, the negative correlation weakens and is no longer statistically significant. This is likely because the POS-based analysis overcounts attributive adjectives. The dependency approach offers a more precise model of the modification relationship, and future work with more consistently annotated treebanks will help clarify the true magnitude of the effect. Yet, when we zoom in on a specific adjective type, as in the pilot behavioral experiment, the results converge with the corpus findings: Arabic speakers replicated the color-size asymmetry documented in English, with suggestive evidence that they overmodify at lower overall rates, consistent with the prediction in the literature.

In Chapter 3, testing adjective ordering in Arabic was important because behavioral evidence from post-nominal languages remains scarce (Dyer et al., 2023). The Arabic results confirmed that subjectivity predicts ordering preferences in a post-nominal language, and the mirror-image pattern with respect to English confirms that these preferences reflect hierarchical structure such that less subjective adjectives compose first with the noun. This mirror-image pattern observation is important because if adjective ordering was driven by the ease with which different properties come to mind during production, the most accessible adjective should be produced first in both English and Arabic. Instead, Arabic speakers reverse the English order, placing the same adjective classes closest to the noun. The heritage speaker results add a further dimension. Heritage Arabic speakers who grew up in English-dominant environments showed ordering preferences in the same direction as native speakers, suggesting that they acquire enough input to establish the mirror-image pattern, strengthening the hierarchical structure story.

In Chapter 4, my initial intuition was that the absence of short vowel diacritics in stan-

dard Arabic orthography means that Arabic speakers are already accustomed to reading incomplete orthographic input, making Arabizi vowel dropping a natural extension of existing literacy practices. However, the picture turned out to be more complicated. If vowel dropping were simply a transfer of Arabic orthographic conventions, we would expect short vowels to be dropped uniformly, yet the BOLT results showed that speakers are selective about when they drop vowels, and that their choice to include or drop a vowel is partly shaped by the predictability of the word in context. Part of the difficulty is that Arabizi is still an emerging system without fixed orthographic conventions. Unlike established writing systems, where spelling is largely settled, Arabizi writers make active choices about how to represent each word, and those choices vary considerably across individuals. A speaker's dialect, their familiarity with French or English orthography, and the social context of the interaction all shape how they spell. Isolating the role of predictability against this background of individual and social variation proved more difficult than I initially expected.

The results from Chapters 2 and 3, along with the BOLT analysis in Chapter 4, align with the general efficiency framework outlined in Chapter 1 (Zipf, 1949; Hawkins, 2004; Gibson et al., 2019). But the failure of the Twitter analysis to replicate the BOLT pattern suggests that efficiency is not a single pressure that applies equally in all contexts. If it were, the predictability-length correlation observed in private chat should have replicated on Twitter. The observation that it did not suggests that the balance between cost and reward shifts with the communicative context. In private chat, shared background increases the likelihood of successful communication even when forms are maximally reduced. On Twitter, where the audience is diffuse and unpredictable, the cost of being misunderstood may outweigh the production savings of vowel dropping, leading writers to adopt fuller spellings as a hedge against ambiguity. This interpretation aligns with Gibson et al.'s formulation of efficiency as the transmission of information through a noisy channel: as the noise increases, the speaker might invest more in redundancy to preserve the signal.

5.2 Chapter conclusions and future directions

Chapter 2 took up the question of whether overmodification patterns established in pre-nominal languages like English hold when the adjective follows the noun. I performed an in-depth analysis of English, Spanish, and Arabic Wikipedia text and a large-scale corpus analysis on Universal Dependencies Treebanks covering 74 languages. In the Wikipedia analysis, I failed to find a difference in the rates of adjective modification across the three languages. However, the broader UD analysis told a different story: as a language features more post-nominal modification, the proportion of modified nouns decreases. I also collected preliminary behavioral data confirming that Arabic speakers show the color-size asymmetry documented in English, with suggestive evidence of lower overmodification rates overall.

The dialectal Arabic data I collected opens a direction for examining how modification patterns may vary across different Arab countries and would provide a strong empirical evidence for this literature. Unlike Wikipedia, which is encyclopedic and relatively formal, dialectal data reflect the kind of spontaneous, conversational language in which efficiency pressures are most likely to shape production. Comparing modification rates across dialects would test whether the patterns observed in the UD treebanks hold. More broadly, the corpus I assembled (spanning five major dialect families and over 239,000 lines of naturalistic text) represents a resource that could extend the modification analysis in several directions. First, it opens the possibility of bridging the gap between the large-scale typological analysis and the behavioral experiments: corpus-derived modification rates from dialectal Arabic could serve as a baseline against which experimental overmodification rates are compared, grounding the behavioral predictions in naturalistic usage data. Second, it could also provide insight into modification rates within a single language family where dialects vary in their degree of contact with other languages, literacy traditions, and communicative norms.

Chapter 3 turned to cross-linguistic variation in word order, asking whether post-nominal

Arabic exhibits the adjective ordering preferences documented for pre-nominal languages like English. I found that native Arabic speakers show reliable ordering preferences, and that these preferences are well predicted by adjective subjectivity. I also tested heritage Arabic speakers and found that they show weaker but still significant preferences.

Several directions could extend this work. One direction is testing adjective ordering in specific Arabic syntactic constructions. In the idafa (construct state), the head noun enters a tight syntactic relationship with a following noun, and adjectives modifying the head noun must appear after the entire construct. This displacement creates a configuration in which the adjective is structurally further from its head noun than in simple post-nominal modification, and it would be informative to test whether subjectivity-based ordering preferences persist, weaken, or shift in this environment. Another extension is through a production study in which speakers describe objects in contexts that elicit multi-adjective descriptions naturally, rather than choosing between pre-set orders. Moreover, the corpora already assembled for the adjective modification analysis in Chapter 2 (both the Wikipedia data and the dialectal Arabic corpus) could be repurposed to examine how adjectives are ordered in naturally occurring text. This would complement the experimental findings with corpus-based evidence, testing whether the subjectivity-based ordering preferences observed in the experiment are reflected in spontaneous speech and in written production across different registers. A corpus approach would also make it possible to estimate noun-adjective co-occurrence statistics and obtain PMI-based measures of ordering preferences in Arabic, complementing the subjectivity-based measures from the experimental data.

Chapter 4 shifted to the lexicon, investigating vowel dropping in Arabizi and testing whether contextual predictability shapes whether a word is written in its long or short form. In the BOLT private chat data, I found support for the efficiency prediction: writers are more likely to use shorter, vowel-dropped spellings in more predictable contexts. However, this pattern did not replicate when I scaled up the analysis to a substantially larger Twitter data set,

trained a new model directly on Twitter data, and tested across multiple context window sizes and word pair selections. The divergence between the two data sets points to the role of communicative setting: in private chat, where interlocutors share background knowledge, speakers can afford to reduce more freely, while on a public platform like Twitter, where the audience is more widespread, the cost of being misunderstood may outweigh the production savings of dropping vowels.

Several extensions of this work could sharpen the picture. First, a character-level analysis could test whether the informativity of individual phonemes predicts vowel dropping: training an n-gram model at the character level and obtaining the surprisal of each phoneme given the preceding phonemes would reveal whether vowels are more likely to be retained when they are less predictable given the surrounding consonant sequence (Bell et al., 2009; Seyfarth, 2014). Second, the analyses in this dissertation relied entirely on forward predictability — the predictability of a word given its preceding context. However, work on spoken production has found that predictability from the future context can be a stronger predictor of reduction than past context predictability (Bell et al., 2009). Training a backward language model on Arabizi text would test whether the same asymmetry holds for written vowel dropping. Third, the current analysis paired each word with a single short form and a single long form, but Arabizi words often have many attested spellings. A model that clusters all orthographic variants of the same underlying word (grouping them by relative length rather than selecting two representative forms) would capture a fuller range of orthographic variation and provide a more sensitive test of whether writers draw from the shorter end of a word's spelling distribution in more predictable contexts.

Finally, the three studies in this dissertation show that Arabic, with its post-nominal adjectives, rich morphology, and diglossic variation, follow the efficiency pressures documented in English and other well-studied languages. The results suggest that speakers and writers are sensitive to production cost across different levels of linguistic structure, but that this sensi-

tivity is modulated by the communicative context in which language is used. Understanding when efficiency pressures prevail and when they are overridden by other factors remains an open and productive line of inquiry.

Bibliography

Abdelali, A., Darwish, K., Durrani, N., and Mubarak, H. (2016). Farasa: A fast and furious segmenter for arabic. In *Proceedings of the 2016 conference of the North American chapter of the Association for Computational Linguistics on Human Language Technology: Demonstrations*, pages 11–16.

Abdelali, A., Mubarak, H., Samih, Y., Hassan, S., and Darwish, K. (2021). Qadi: Arabic dialect identification in the wild. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 1–10.

Abdul-Mageed, M., Zhang, C., Elmadany, A., Bouamor, H., and Habash, N. (2021). Nadi 2021: The second nuanced arabic dialect identification shared task. *arXiv preprint arXiv:2103.08466*.

Abou-Ghazaleh, A., Khateb, A., and Nevat, M. (2018). Lexical competition between spoken and literary Arabic: A new look into the neural basis of diglossia using fMRI. *Neuroscience*, 393:83–96.

Akbar, R. (2019). Arabizi among kuwaiti youths: Reshaping the standard arabic orthography. *International Journal of English Linguistics*, 9(1):301–323.

Al-Sharifi, B. and Sadler, L. (2009). The adjectival construct in Arabic. In Butt, M. and King, T. H., editors, *Proceedings of the LFG09 Conference*, pages 26–43. CSLI Publications.

Al-Shurafa, N. (2006). Syntactic ordering and semantic aspects of adjectives and adjectival phrases in Arabic. *University of Sharja Journal (A Refereed Scientific Periodical)*, 3(1):1–21.

al Zamakhsharī, M. i. (1905). *Kitāb al-Mufaṣṣal fī Ṣanᶜat al-Iᶜrāb*. Maṭba at al-Taqaddum, Cairo. Edited by Muḥammad Badr al-Dīn al-Na sānī al-Ḥalabī.

Attardi, G. (2012). Wikiextractor. `https://https://github.com/attardi/wikiextractor`.

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.

Baert, G., Gahbiche, S., Gadek, G., and Pauchet, A. (2020). Arabizi language models for sentiment analysis. In *Proceedings of the 28th international conference on computational linguistics*, pages 592–603.

Bar-Sever, G., Lee, R., Scontras, G., and Pearl, L. (2018). Quantitatively assessing the development of adjective ordering preferences using child-directed and child-produced speech corpora. In *Proceedings of the Society for Computation in Linguistics*, volume 1, pages 203–204.

Barker, C. (2013). Negotiating taste. *Inquiry*, 56(2-3):240–257.

Belke, E. and Meyer, A. S. (2002). Tracking the time course of multidimensional stimulus discrimination: Analyses of viewing patterns and processing times during "same"-"different "decisions. *European Journal of Cognitive Psychology*, 14(2):237–266.

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.

Bies, A., Song, Z., Maamouri, M., Grimes, S., Lee, H., Wright, J., Strassel, S., Habash, N., Eskander, R., and Rambow, O. (2014). Transliteration of arabizi into arabic orthography: Developing a parallel annotated arabizi-arabic script sms/chat corpus. In *Proceedings of the EMNLP 2014 workshop on Arabic natural language processing (ANLP)*, pages 93–103.

Bouamor, H., Habash, N., Salameh, M., Zaghouani, W., Rambow, O., Abdulrahim, D., Obeid, O., Khalifa, S., Eryani, F., Erdmann, A., et al. (2018). The madar arabic dialect corpus and lexicon. In *Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018)*.

Boujou, E., Chataoui, H., Mekki, A. E., Benjelloun, S., Chairi, I., and Berrada, I. (2021). An open access nlp dataset for arabic dialects: Data collection, labeling, and model construction. *arXiv preprint arXiv:2102.11000*.

Chomsky, N. (2017). Language architecture and its import for evolution. *Neuroscience Biobehavioral Reviews*, 81:295–300. The Biology of Language.

Cinque, G. (2010). *The syntax of adjectives: A comparative study*, volume 57. MIT press.

Cohn-Gordon, R., Goodman, N., and Potts, C. (2019). An incremental iterated response model of pragmatics. In *Proceedings of the Society for Computation in Linguistics (SCiL) 2019*, pages 81–90.

Darwish, K. (2014). Arabizi detection and conversion to arabic. In *Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP)*, pages 217–224.

Degen, J., Hawkins, R., Graf, C., Kreiss, E., and Goodman, N. D. (2020). When redundancy is rational: A bayesian approach to 'overinformative' referring expressions. *Psychological Review*.

Degen, J., Waldon, B., Hormosillo, A., Velasquez, A., and Kachakeche, Z. (2022). Spanish exhibits color/size asymmetry and scene variation effects in redundant referring expression production — just like English. Talk presented by Brandon Waldon at X-Prag, Pavia, Italy.

Dixon, R. (1982). *Where have all the adjectives gone? And other essays in semantics and syntax*. Mouton, Berlin.

Dyer, W., Futrell, R., Liu, Z., and Scontras, G. (2021). Predicting cross-linguistic adjective order with information gain. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 957–967.

Dyer, W., Torres, C., Scontras, G., and Futrell, R. (2023). Evaluating a century of progress on the cognitive science of adjective ordering. *Transactions of the Association for Computational Linguistics*, 11:1185–1200.

Fedzechkina, M. and Jaeger, T. F. (2020). Production efficiency can cause grammatical change: Learners deviate from the input to better balance efficiency against robust message transmission. *Cognition*, 196:104115.

Fehri, A. F. (1999). Arabic modifying adjectives and DP structures. *Studia Linguistica*, 53:105–154.

Ferreira, V. and Dell, G. (2000). Effect of Ambiguity and Lexical Availability on Syntactic and Lexical Production* 1. *Cognitive Psychology*, 40(4):296–340.

Frank, M. and Goodman, N. (2012). Quantifying pragmatic inference in language games. *Science*, 336.

Franke, M., Scontras, G., and Simonic, M. (2019). Subjectivity-based adjective ordering maximizes communicative success. In *CogSci*, pages 344–350.

Futrell, R. (2017). *Memory and locality in natural language*. PhD thesis, Massachusetts Institute of Technology, Cambridge, MA.

Futrell, R. (2019). Information-theoretic locality properties of natural language. In *Proceedings of the First Workshop on Quantitative Syntax (Quasy, SyntaxFest 2019)*, pages 2–15, Paris, France. Association for Computational Linguistics.

Futrell, R., Dyer, W., and Scontras, G. (2020a). What determines the order of adjectives in english? comparing efficiency-based theories using dependency treebanks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2003–2012, Online. Association for Computational Linguistics.

Futrell, R., Gibson, E., and Levy, R. P. (2020b). Lossy-context surprisal: An information-theoretic model of memory effects in sentence processing. *Cognitive Science*, 44:e12814.

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.

Gatt, A., van Gompel, M., Krahmer, E., and van Deemter, K. (2011). Non-deterministic attribute selection in reference production. In *Proceedings of the Workshop on the Production of Referring Expressions (PRE-CogSci 2011)*, pages 1–7. Unknown Publisher.

Gibson, E., Futrell, R., Piantadosi, S. T., Dautriche, I., Mahowald, K., Bergen, L., and Levy, R. (2019). How efficiency shapes human language. *Trends in Cognitive Sciences*.

Gordon, C. (2011). From speech to screen: The orthography of colloquial arabic in electronically-mediated communication.

Grice, H. (1975). Logic and conversation. In Cole, P. and Morgan, J. L., editors, *Syntax and Semantics, Vol. 3, Speech Acts*, pages 41–58. Academic Press, New York.

Hahn, M., Degen, J., Goodman, N. D., Jurafsky, D., and Futrell, R. (2018). An information-theoretic explanation of adjective ordering preferences. In *Proceedings of the 40th annual conference of the Cognitive Science Society*, London. Cognitive Science Society.

Hawkins, J. A. (2004). *Efficiency and complexity in grammars*. Oxford University Press, Oxford.

Hetzron, R. (1978). On the relative order of adjectives. In Seiler, H., editor, *Language Universals*, pages 165–184. Narr, Tübingen, Germany.

Jaeger, T. F. and Tily, H. J. (2011). On language 'utility': Processing complexity and communicative efficiency. *Wiley Interdisciplinary Reviews: Cognitive Science*, 2(3):323–335.

Jurafsky, D. (1996). A probabilistic model of lexical and syntactic access and disambiguation. *Cognitive Science*, 20(2):137–194.

Jurafsky, D., Bell, A., Gregory, M., and Raymond, W. (2001). Probabilistic relations between words. *Frequency and the emergence of linguistic structure*, pages 229–254.

Kennedy, C. (2013). Two sources of subjectivity: Qualitative assessment and dimensional uncertainty. *Inquiry*, 56(2-3):258–277.

Khalifa, S., Habash, N., Abdulrahim, D., and Hassan, S. (2016). A large scale corpus of gulf arabic. *arXiv preprint arXiv:1609.02960*.

Kirby, S. (2017). Culture and biology in the origins of linguistic structure. *Psychonomic Bulletin & Review*, 24(1):118–137.

Kölbel, M. (2004). Faultless disagreement. In *Proceedings of the Aristotelian Society*, volume 104, pages 53–73.

Koolen, R., Gatt, A., Goudbeek, M., and Krahmer, E. (2011). Factors causing overspecification in definite descriptions. *Journal of Pragmatics"*, 43(13):3231 – 3250.

Koolen, R., Goudbeek, M., and Krahmer, E. (2013). The effect of scene variation on the redundant use of color in definite reference. *Cognitive Science*, 37(2):395–411.

Kreiss, E. and Degen, J. (2020). Production expectations modulate contrastive inference. In *CogSci*.

Kremers, J. (2003). *The Arabic noun phrase*. PhD thesis, Katholieke Universiteit Nijmegen.

Kursat, L. and Degen, J. (2021). Perceptual difficulty differences predict asymmetry in redundant modification with color and material adjectives. *Proceedings of the Linguistic Society of America*, 6(1):676–688.

Kursat, L., Waldon, B., Ergin, R., and Degen, J. (2022). Evaluating models of referring expression production on an emerging sign language. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 44.

Kwaik, K. A., Saad, M., Chatzikyriakidis, S., and Dobnik, S. (2018). Shami: A corpus of levantine arabic dialects. In *Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018)*.

LaPolla, R. J. and Huang, C. (2004). Adjectives in Qiang. In Dixon, R. M. and Aikhenvald, A. Y., editors, *Adjective Classes: A Cross-Linguistic Typology*, pages 306–322. Oxford University Press Oxford, Oxford.

Leslau, W. (1969). Frequency as determinant of linguistic changes in the ethiopian languages. *Word*, 25(1-3):180–189.

Levshina, N. (2022a). *Different Types of Efficiency in Language*, page 1–104. Cambridge University Press.

Levshina, N. (2022b). Frequency, informativity and word length: Insights from typologically diverse corpora. *Entropy*, 24(2):280.

MacFarlane, J. (2014). *Assessment sensitivity: Relative truth and its applications*. Clarendon Press, Oxford.

Mahowald, K., Dautriche, I., Gibson, E., and Piantadosi, S. T. (2018). Word forms are structured for efficient use. *Cognitive science*, 42(8):3116–3134.

Mahowald, K., Fedorenko, E., Piantadosi, S. T., and Gibson, E. (2013). Info/information theory: speakers choose shorter words in predictive contexts. *Cognition*, 126:313–318.

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., and McClosky, D. (2014). The stanford corenlp natural language processing toolkit. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60.

Martin, J. E. (1969). Semantic determinants of preferred adjective order. *Journal of Verbal Learning and Verbal Behavior*, 8(6):697–704.

May, J., Benjira, Y., and Echihabi, A. (2014). An arabizi-english social media statistical machine translation system. In *Proceedings of the 11th Conference of the Association for Machine Translation in the Americas: MT Researchers Track*, pages 329–341.

Medhaffar, S., Bougares, F., Esteve, Y., and Belguith, L. H. (2017). Sentiment analysis of tunisian dialects: Linguistic ressources and experiments. In *Proceedings of the third Arabic natural language processing workshop*, pages 55–61.

Meftouh, K., Harrat, S., Jamoussi, S., Abbas, M., and Smaili, K. (2015). Machine translation experiments on padic: A parallel arabic dialect corpus. In *Proceedings of the 29th Pacific Asia conference on language, information and computation*, pages 26–34.

Nivre, J., de Marneffe, M.-C., Ginter, F., Hajič, J., Manning, C. D., Pyysalo, S., Schuster, S., Tyers, F., and Zeman, D. (2020). Universal dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4034–4043.

Obeid, O., Zalmout, N., Khalifa, S., Taji, D., Oudah, M., Alhafni, B., Inoue, G., Eryani, F., Erdmann, A., and Habash, N. (2020). CAMeL tools: An open source python toolkit for Arabic natural language processing. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 7022–7032, Marseille, France. European Language Resources Association.

Panayidou, F. (2014). *(In) flexibility in Adjective Ordering*. PhD thesis, Queen Mary University of London.

Pechmann, T. (1989). Incremental speech production and referential overspecification. *Linguistics*, 27(1):89–110.

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.

Pimentel, T., Maudslay, R. H., Blasi, D., and Cotterell, R. (2020). Speakers fill lexical semantic gaps with context. *arXiv preprint arXiv:2010.02172*.

Polinsky, M. and Scontras, G. (2020). Understanding heritage languages. *Bilingualism: Language and Cognition*, 23(1):4–20.

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., and Gulin, A. (2018). Catboost: unbiased boosting with categorical features. *Advances in neural information processing systems*, 31.

Qing, C., Lassiter, D., and Degen, J. (2018). What do eye movements in the visual world reflect? a case study from adjectives. In *Proceedings of 40th annual conference of the Cognitive Science Society*, pages 2297–2302, London. Cognitive Science Society.

Rosales Jr., C. M. and Scontras, G. (2019). On the role of conjunction in adjective ordering preferences. In *Proceedings of the Linguistic Society of America*, volume 4, pages 1–12.

Rubio-Fernández, P. (2016). How redundant are redundant color adjectives? An efficiency-based analysis of color overspecification. *Frontiers in Psychology*, 7:153.

Rubio-Fernández, P. (2019). Overinformative speakers are cooperative: Revisiting the gricean maxim of quantity. *Cognitive science*, 43(11):e12797.

Rubio-Fernández, P. and Jara-Ettinger, J. (2020). Incrementality and efficiency shape pragmatics across languages. *Proceedings of the National Academy of Sciences*, 117(24):13399–13404.

Rubio-Fernández, P., Mollica, F., and Jara-Ettinger, J. (2020). Speakers and listeners exploit word order for communicative efficiency: A cross-linguistic investigation. *Journal of Experimental Psychology: General*.

Rubio-Fernández, P., Mollica, F., and Jara-Ettinger, J. (2021). Speakers and listeners exploit word order for communicative efficiency: A cross-linguistic investigation. *Journal of Experimental Psychology: General*, 150(3):583.

Samonte, S. and Scontras, G. (2019). Adjective ordering in Tagalog: A cross-linguistic comparison of subjectivity-based preferences. In *Proceedings of the Linguistic Society of America*, volume 4, pages 1–13.

Scontras, G., Bar-Sever, G., Kachakeche, Z., Rosales Jr, C. M., and Samonte, S. (2020a). Incremental semantic restriction and subjectivity-based adjective ordering. In *Proceedings of Sinn und Bedeutung*, volume 24, pages 253–270.

Scontras, G., Bar-Sever, G., Kachakeche, Z., Rosales Jr., C. M., and Samonte, S. (2020b). Incremental semantic restriction and subjectivity–based adjective ordering. In *Proceedings of the Sinn und Bedeutung 24*.

Scontras, G., Degen, J., and Goodman, N. D. (2017). Subjectivity predicts adjective ordering preferences. *Open Mind: Discoveries in Cognitive Science*, 1(1):53–65.

Scontras, G., Degen, J., and Goodman, N. D. (2019). On the grammatical source of adjective ordering preferences. *Semantics and Pragmatics*.

Scontras, G., Fuchs, Z., and Polinsky, M. (2015). Heritage language and linguistic theory. *Frontiers in Psychology*, 6(1545).

Scontras, G., Kachakeche, Z., Nguyen, A., Rosales, C., Samonte, S., et al. (2021). Cross-linguistic evidence for subjectivity-based adjective ordering preferences.

Scontras, G., Kachakeche, Z., Nguyen, A., Rosales, C., Samonte, S., Shetreet, E., Shi, Y., Tourtouri, E. N., and Trainin, N. (2020c). Cross-linguistic evidence for subjectivity-based adjective ordering preferences. In *Theoretical and Experimental Approaches to Modification (TExMod2020)*.

Scontras, G., Manzini, M. R., and Mazzaggio, G. (2025). Italian adjective order is not predicted by syntactic hierarchies and movement. *Linguistic Inquiry*, pages 1–19.

Scontras, G. and Nicolae, A. C. (2014). Saturating syntax: Linkers and modification in Tagalog. *Lingua*, 149:17–33.

Scott, G.-J. (2002). Stacked adjectival modification. *Functional Structure in DP and IP: The Cartography of Syntactic Structures, vol*, 1:91–122.

Sedivy, J. C. (2003). Pragmatic versus form-based accounts of referential contrast: Evidence for effects of informativity expectations. *Journal of psycholinguistic research*, 32(1):3–23.

Sedivy, J. C., Tanenhaus, M., Chambers, C., and Carlson, G. (1999). Achieving incremental semantic interpretation through contextual representation. *Cognition*, 71(2):109–147.

Seyfarth, S. (2014). Word informativity influences acoustic duration: Effects of contextual predictability on lexical representation. *Cognition*, 133(1):140–155.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, 27:623–656.

Shazal, A., Usman, A., and Habash, N. (2020). A unified model for arabizi detection and transliteration using sequence-to-sequence models. In *Proceedings of the fifth arabic natural language processing workshop*, pages 167–177.

Shi, Y. and Scontras, G. (2020). Mandarin has subjectivity-based adjective ordering preferences in the presence of de. *Proceedings of the Linguistic Society of America*, 5(1):410–418.

Shlonsky, U. (2004). The form of Semitic noun phrases. *Lingua*, 114(12):1465–1526.

Sproat, R. and Shih, C. (1991). The cross-linguistic distribution of adjective ordering restrictions. In Georgopoulos, C. and Ishihara, R., editors, *Interdisciplinary Approaches to Language: Essays in Honor of S.-Y. Kuroda*, pages 565–593. Kluwer Academic Publishers.

Sullivan, N. (2017). *Writing Arabizi: Orthographic Variation in Romanized Lebanese Arabic on Twitter*. PhD thesis.

Sweet, H. (1898). *A New English Grammar, Logical and Historical*. Clarendon Press, London.

Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., and Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*, 268(5217):1632–1634.

Tobaili, T., Fernandez, M., Alani, H., Sharafeddine, S., Hajj, H., and Glavaš, G. (2019). Senzi: A sentiment analysis lexicon for the latinised arabic (arabizi). In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 1203–1211.

Waldon, B. and Degen, J. (2021). Modeling cross-linguistic production of referring expressions. *Proceedings of the Society for Computation in Linguistics*, 4(1):206–215.

Westerbeek, H., Koolen, R., and Maes, A. (2015). Stored object knowledge and the production of referring expressions: The case of color typicality. *Frontiers in psychology*, 6:935.

Whorf, B. L. (1945). Grammatical categories. *Language*, 21(1):1–11.

Wu, S. A. and Gibson, E. (2020). Word order predicts cross-linguistic differences in the production of redundant color and number modifiers.

Zahir, J. (2022). Iadd: An integrated arabic dialect identification dataset. *Data in Brief*, 40:107777.

Zaidan, O. and Callison-Burch, C. (2011). The arabic online commentary dataset: an annotated dataset of informal arabic with high dialectal content. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 37–41.

Zbib, R., Malchiodi, E., Devlin, J., Stallard, D., Matsoukas, S., Schwartz, R., Makhoul, J., Zaidan, O., and Callison-Burch, C. (2012). Machine translation of arabic dialects. In *Proceedings of the 2012 conference of the north american chapter of the association for computational linguistics: Human language technologies*, pages 49–59.

Zipf, G. K. (1949). *Human behavior and the principle of least effort*. Addison-Wesley Press, Oxford, UK.